An Investigation into the Knowledge Discovery and Data Mining (KDDM) process to generate course taking pattern characterised by contextual factors of students in Higher Education Institution (HEI)

Thesis submitted for the degree of Doctor of Philosophy

by

Subhashini Sailesh Bhaskaran

Brunel Business School

Brunel University, London

June 2017

LIST OF ACRONYMS

DM	**Data Mining**
KDD	**Knowledge Discovery in Databases**
KDDM	**Knowledge Discovery and Data Mining**
CRISP-DM	**Cross Industry Standard Process for Data Mining**
GPA	**Graded Point Average**
CGPA	**Cumulative Grade Point Average**
EDM	**Educational Data Mining**
WEKA	**Waikato Environment for Knowledge Analysis**

ABSTRACT

The Knowledge Discovery and Data Mining (KDDM), a growing field of study argued to be very useful in discovering knowledge hidden in large datasets are slowly finding application in Higher Educational Institutions (HEIs). While literature shows that KDDM processes enable discovery of knowledge useful to improve performance of organisations, limitations surrounding them contradict this argument. While extending the usefulness of KDDM processes to support HEIs, challenges were encountered like the discovery of course taking patterns in educational datasets associated with contextual information. While literature argued that existing KDDM processes suffer from the limitations arising out of their inability to generate patterns associated with contextual information, this research tested this claim and developed an artefact that overcame the limitation.

Design Science methodology was used to test and evaluate the KDDM artefact. The research used the CRISP-DM process model to test the educational dataset using attributes namely course taking pattern, course difficulty level, optimum CGPA and time-to-degree by applying clustering, association rule and classification techniques. The results showed that both clustering and association rules did not produce course taking patterns. Classification produced course taking patterns that were partially linked to CGPA and time-to-degree. But optimum CGPA and time-to-degree could not be linked with contextual information. Hence the CRISP-DM process was modified to include three new stages namely contextual data understanding, contextual data preparation and additional data preparation (merging) stage to see whether contextual dataset could be separately mined and associated with course taking pattern. The CRISP-DM model and the modified CRISP-DM model were tested as per the guidelines of Chapman et al. (2000). Process theory was used as basis for the modification of CRISP-DM process. Results showed that course taking pattern contextualised by course difficulty level pattern predicts optimum CGPA and time-to-degree. This research has contributed to knowledge by developing a new artefact (contextual factor mining in the CRISP-DM process) to predict optimum CGPA and optimum time-to-degree using course taking pattern and course difficulty level pattern. Contribution to theory was in extension of the application of a few theories to explain the development, testing and evaluation of the KDDM artefact. Enhancement of genetic algorithm (GA) to mine course difficulty level pattern along with course taking pattern is a contribution and a pseudocode to verify the presence of course difficulty level pattern. Contribution to practise was by demonstrating the usefulness of the modified CRISP-DM process for prediction and simulation of the course taking pattern to predict the optimum CGPA and time-to-degree thereby demonstrating that the artefact can be deployed in practise.

Acknowledgements

My journey to the PhD was a challenging one. Throughout my journey many persons have supported me without which it would not have been possible for me to have completed this thesis.

Foremost, I would like to thank and express my sincere thankfulness and gratitude to Prof. Abdulla Al Hawaj, founding President and Managing Director, Ahlia University who was kind enough to encourage me to register in the Phd Programme and fully supported me as a Sponsored Student.

At the same time, it is pertinent to acknowledge the immense support provided by Dr. Kevin Lu, my principal supervisor at Brunel University, London whose committed and continuous guidance enabled me to complete this thesis.

Equally important was the support, encouragement and motivation provided by Prof. Mansoor Alaali, President Ahlia University, my supervisor at Ahlia University who continuously motivated me to reach this stage of submission.

Apart from the above I would like to record my appreciation to Dr. Tillal Eldabi who constantly followed up with my progress leading to submission of the thesis. In addition I would like to thank my colleagues Mr. Gowrishankar Srinivasan and Ms. Hessa Al Dhaen who provided continuous help in regards to many of the requirements that I had to fulfill before submitting the thesis.

I would like to acknowledge the tremendous support, encouragement and patience of my beloved husband who stood by me throughout the journey and motivated me during difficult situations.

Finally, I would like to thank my daughter, parents, brother and in-laws who motivated me and provided moral support which led to the culmination of this thesis.

Publications

Sailesh, S.B., Lu, K.J. and Al Aali, M.(2016).Context driven data mining to classify students of higher educational institutions. *In proceedings of Inventive Computation Technologies (ICICT), International Conference on* (Vol.2, pp.1-7). IEEE

Sailesh, S.B., Lu, K.J. and Al Aali, M., (2016). Profiling students on their course taking patterns in higher educational institutions (HEIs). *In Information Science (ICIS), International Conference* on (pp. 160-167).

Bhaskaran, S., Lu, K. and Al Aali, M.,(2015). Student Performance and Time-To-Degree Analysis Using J48 Decision Tree Algorithm. In *Managing Intellectual Capital and Innovation for Sustainable and Inclusive Society: Managing Intellectual Capital and Innovation; Proceedings of the Make Learn and TIIM Joint International Conference 2015*.

Bhaskaran, S., Lu, K. and Al Aali, M., 2017**,** Student Performance and Time-to-degree Analysis By the study of course taking patterns using J48 Decision Tree Algorithm, Inderscience, International Journal of Modelling in Operations Management.

To be published

Bhaskaran, S., Lu, K. and Al Aali, M., 2017,Facilitate Decision making in Higher Educational Institutions by linking course taking patterns and time-to-degree, Inderscience, International Journal of Society Systems Science (IJSSS) in Operations Management .

Table of Contents

List of Tables

List of Figures

Chapter 1 : Introduction

1.0 Overview of Student Performance Factors in HEIs

Higher education is growing at an accelerated pace. More and more students are enrolling in diverse programmes. Higher education institutions (HEIs) are forced to meet this growing demand of students who want to register in diverse programmes. In addition, job market requirements force HEIs to continuously enhance their programme quality and course offerings and make informed decisions about introducing new programmes to satisfy the needs of the job market. The burgeoning student numbers and the wide options available for them to choose a programme of their interest and enrol in an institution of their choice has thrown open both opportunities and challenges to the HEIs. Opportunities include higher rate of institutional growth, more revenue, wider market, improved brand image and international presence (Hua, 2011; Douglass, Edelstein and Haoreau, 2013). Challenges include ensuring better student performance, improving teaching and learning, improving student learning experience, finance issues, student demand, pressure arising out of quality assurance agencies, lack of funding, student diversity, lack of capacity, lack of availability of adequately qualified and experienced instructors, need for continuous augmenting of resources, inadequacy in infrastructural facilities and competition(Sheikh, 2017;Larkin, 2012). Amongst these challenges this research focusses on improving student performance and student learning experience.

While there are many factors that affect student performance and student learning experience, an area that has serious implication to the students as well as HEIs is found to the time taken by students to graduate called the time-to-degree (Stock, 2011). Although much less attention has been paid to examine time-to-degree as a factor that has bearing on student performance and learning experience the impact of time-to-degree on both student and HEIs appear to be very significant (Bowen et al. 2009). For instance, faster time-to-degree could enable students to complete their degrees and take up jobs earlier than the normal time prescribed by the HEIs. This could be highly beneficial to students. On the other hand HEIs could have a faster turnout of students which could facilitate higher number of student recruitment. Thus, enabling HEIs to efficiently use their resources and grow. However there are inherent risks in supporting these arguments as faster time-to-degree could lead students to score less than optimum grades due to possible overloading of courses in a particular semester or semesters. Similarly HEIs could fall prey to the fact that teaching quality could suffer due to compressing too many courses to be taught in a shorter duration of time. These examples of advantage and risks associated with the phenomenon called time-to-degree point towards the need to understand how time-to-degree affect both the students and HEIs. In addition, it is necessary to study the phenomenon along with other factors that can be related to it. For instance education system (e.g. credit hour system followed in U.S.A), course taking pattern of students in each semester, cumulative grade point average (CGPA), semester grade point average (SGPA), types of courses, number of courses offered as part of the curriculum plan, course difficulty, student potential, number of courses registered in a semester, number of semesters per year, course complexity and past history of student prior to enrolment in HEIs. Leaving behind these factors and studying the phenomenon of time-to-degree would not provide a comprehensive picture of time-to-degree as well as will leave an incomplete understanding of the phenomenon because the influence of each one of the factors could lead to different results during investigation.

While literature shows that time-to-degree has a major influence on the student performance and learning experience it is not clear how this phenomenon could be tackled by HEIs to improve both student performance and learning experience (Cullinane, 2014; Xu et al, 2016). Thus this study embarks on an investigation of the phenomenon of time-to-degree taking into account select related factors called contextual factors namely education system, course taking pattern of students in each semester, number of semesters, course difficulty, CGPA,SGPA, total number of courses required to be completed by the student successfully to graduate and type of courses(while classification of contextual factors in the field of HEI includes many different factors (e.g Appendix 25), addressing all the factors in one research effort was found to be beyond the scope of the research due to time constraints). Thus only limited factors discussed further have been addressed. The reason for choosing certain factors related to time-to-degree for investigation lies in the argument provided in the literature which indicates that at the minimum these factors are needed to be included in the study while researching on the phenomenon time-to-degree. In addition it is not clear how and which of those factors affect time-to-degree a gap found in the literature. These aspects are covered in the critical review of the literature provided in Chapter 2.

Furthermore, in order to investigate the phenomenon of time-to-degree this research has chosen data mining as a technique which according to literature provides a way to link time-to-degree to student related factors mentioned above although the usefulness of data mining to study related factors is grossly under researched (Herzog,2006; Letkiewicz, Lim and Heckman(2014)). For instance, researchers argue that time-to-degree can be related to gender, race, continuous enrolment using data mining, an argument that is not conclusive and not generalizable and not applicable to different contexts. Although extant literature shows that use of data mining techniques to study time-to-degree and related factors is under researched yet data mining has been found to have the potential to link the outcomes of investigations using datamining to business goals of HEIs which have direct bearing on student performance, student learning experience and decision making. Additionally, in order to employ data mining as a method this research utilised the core concepts of big data, KDDM and Educational Data Mining (EDM) an argument supported by literature(Kumar,2011; Alazmi, 2012). The necessity to involve these concepts in the current research stems from the argument found in the literature which say that hidden knowledge about time-to-degree and related factors in large datasets cannot be done using ordinary data mining techniques because they do not support decision making and achieving business goals. For instance, Picciano, 2012; Lane, 2014 argues that student retention and decisions related to student retention could be investigated using extracted knowledge hidden in student datasets of HEIs employing ordinary data mining techniques. But the outcomes of these researches are not found useful in supporting decision making concerned with business goals of HEIs. This and similar arguments found in the literature provide the basis for utilising the concepts of KDDM, EDM and big data in this research. Thus this research focusses on specific datasets of HEIs considered as big data, for investigating the phenomenon of time-to-degree, using the concepts of EDM and taking a particular KDDM process into account chosen through a critical review of the literature (See chapter 2). This statement gains significance as literature shows that no single KDDM process currently used can be accepted as a universal process for extracting hidden knowledge from big data which is a serious limitation in the literature. Lack of sufficient knowledge on how to choose a particular process to extract hidden knowledge from big data including student data is a major gap in the literature. Thus the study of KDDM processes and their utility to understand time-to-degree related factors gains currency.

Three things emerge from the foregoing discussions they are:

1. Time-to-degree is a phenomenon that has bearing on the student performance and learning experience an aspect that is under researched and it is not clear which data mining technique is most suitable for the purpose and there are contextual factors that affect time-to-degree knowledge about which is limited.

2. Educational Data mining as a technique could be used in extracting knowledge about the concept of time-to-degree and contextual factors hidden in student datasets although it is not clear which data mining technique is most useful to extract this knowledge.

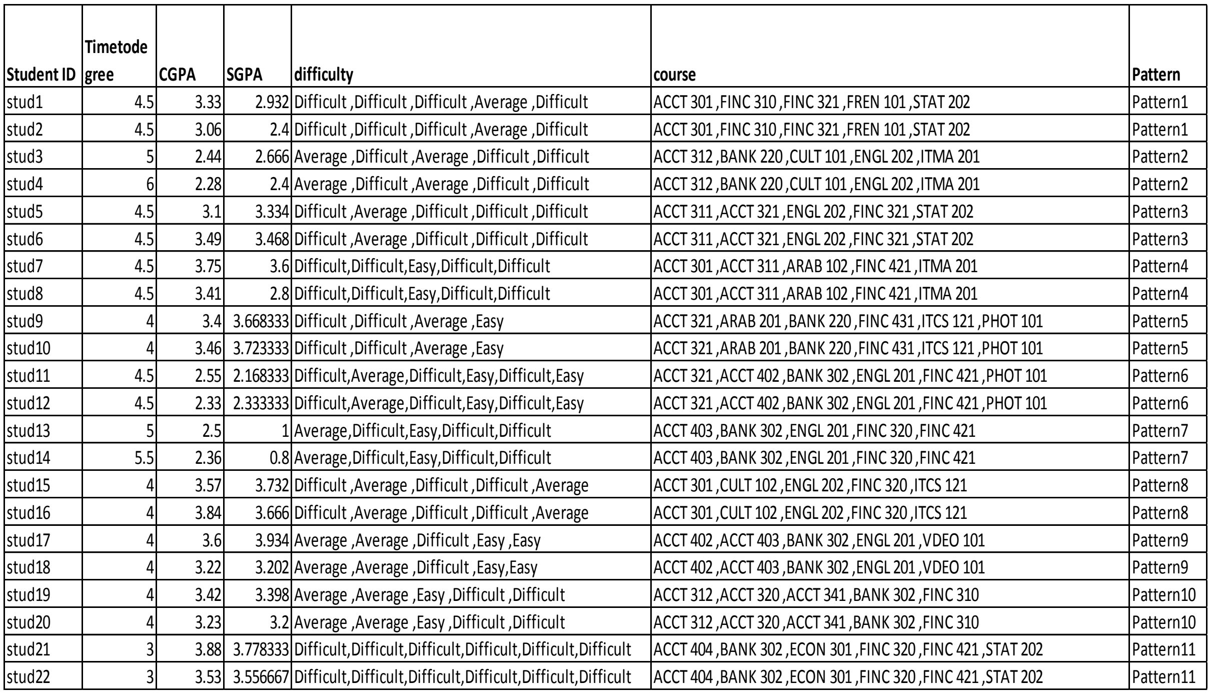
3. The need to utilise KDDM process to extract knowledge about time-to-degree and related contextual factors although none of the widely used KDDM processes have the ability to deal with contextual factors.

These emerging situations pose challenges to HEIs and addressing those challenges through research could provide HEIs with ways to tackle the challenges. Thus the following sections analyse the problems that arise due to those challenges mentioned above and how those problems could be investigated.

This chapter provides an overview of this research and outlines the scope of this thesis. Section 1.1 presents the research background and statement of problem, Section 1.2 presents the research gap. Section 1.3 defines the research questions. Section 1.4 presents the research aim and objectives, Section 1.5 presents the research method and Section 1.6 presents research significance. Finally, Section 1.7 describes the overall structure of the thesis.

1.1 Research background and statement of the problem

Foremost it can be seen that in an education system like the one in USA students have flexibility to register in the courses of their choice and the number of courses in a semester. This leads to a situation where certain students who register in less number of courses in a semester could in reality take longer time-to-degree. This flexibility creates further problems like possibility of students either scoring lower CGPA due to random choice of courses in a semester without understanding how difficult a course or a set of courses could be or overloading themselves with more number of courses than they could study leading to possible failure. In addition it is not always true that less number of courses registered in a semester by a student could enable a student to score the highest CGPA as the course difficulty of those courses could be a problem for the student to handle leading to lower CGPA. Therefore there is a set of courses in which a student could register which may have the course difficulty optimized leading to the student scoring a higher CGPA knowledge about which is hidden in the student dataset (Oladokun et al, 2008;Nandheshwar et al,2011). This situation can be visualized as a pattern courses in which a student can register alongside a pattern of course difficulty identified for each of the courses in that particular pattern of courses in which the student has registered. This can be explained using Table 1.1.



Student ID	Timetodegree	CGPA	SGPA	difficulty	course	Pattern
stud1	4.5	3.33	2.932	Difficult ,Difficult ,Difficult ,Average ,Difficult	ACCT 301 ,FINC 310 ,FINC 321 ,FREN 101 ,STAT 202	Pattern1
stud2	4.5	3.06	2.4	Difficult ,Difficult ,Difficult ,Average ,Difficult	ACCT 301 ,FINC 310 ,FINC 321 ,FREN 101 ,STAT 202	Pattern1
stud3	5	2.44	2.666	Average ,Difficult ,Average ,Difficult ,Difficult	ACCT 312 ,BANK 220 ,CULT 101 ,ENGL 202 ,ITMA 201	Pattern2
stud4	6	2.28	2.4	Average ,Difficult ,Average ,Difficult ,Difficult	ACCT 312 ,BANK 220 ,CULT 101 ,ENGL 202 ,ITMA 201	Pattern2
stud5	4.5	3.1	3.334	Difficult ,Average ,Difficult ,Difficult ,Difficult	ACCT 311 ,ACCT 321 ,ENGL 202 ,FINC 321 ,STAT 202	Pattern3
stud6	4.5	3.49	3.468	Difficult ,Average ,Difficult ,Difficult ,Difficult	ACCT 311 ,ACCT 321 ,ENGL 202 ,FINC 321 ,STAT 202	Pattern3
stud7	4.5	3.75	3.6	Difficult,Difficult,Easy,Difficult,Difficult	ACCT 301 ,ACCT 311 ,ARAB 102 ,FINC 421 ,ITMA 201	Pattern4
stud8	4.5	3.41	2.8	Difficult,Difficult,Easy,Difficult,Difficult	ACCT 301 ,ACCT 311 ,ARAB 102 ,FINC 421 ,ITMA 201	Pattern4
stud9	4	3.4	3.668333	Difficult ,Difficult ,Average ,Easy	ACCT 321 ,ARAB 201 ,BANK 220 ,FINC 431 ,ITCS 121 ,PHOT 101	Pattern5
stud10	4	3.46	3.723333	Difficult ,Difficult ,Average ,Easy	ACCT 321 ,ARAB 201 ,BANK 220 ,FINC 431 ,ITCS 121 ,PHOT 101	Pattern5
stud11	4.5	2.55	2.168333	Difficult,Average,Difficult,Easy,Difficult,Easy	ACCT 321 ,ACCT 402 ,BANK 302 ,ENGL 201 ,FINC 421 ,PHOT 101	Pattern6
stud12	4.5	2.33	2.333333	Difficult,Average,Difficult,Easy,Difficult,Easy	ACCT 321 ,ACCT 402 ,BANK 302 ,ENGL 201 ,FINC 421 ,PHOT 101	Pattern6
stud13	5	2.5	1	Average,Difficult,Easy,Difficult,Difficult	ACCT 403 ,BANK 302 ,ENGL 201 ,FINC 320 ,FINC 421	Pattern7
stud14	5.5	2.36	0.8	Average,Difficult,Easy,Difficult,Difficult	ACCT 403 ,BANK 302 ,ENGL 201 ,FINC 320 ,FINC 421	Pattern7
stud15	4	3.57	3.732	Difficult ,Average ,Difficult ,Difficult ,Average	ACCT 301 ,CULT 102 ,ENGL 202 ,FINC 320 ,ITCS 121	Pattern8
stud16	4	3.84	3.666	Difficult ,Average ,Difficult ,Difficult ,Average	ACCT 301 ,CULT 102 ,ENGL 202 ,FINC 320 ,ITCS 121	Pattern8
stud17	4	3.6	3.934	Average ,Average ,Difficult ,Easy ,Easy	ACCT 402 ,ACCT 403 ,BANK 302 ,ENGL 201 ,VDEO 101	Pattern9
stud18	4	3.22	3.202	Average ,Average ,Difficult ,Easy,Easy	ACCT 402 ,ACCT 403 ,BANK 302 ,ENGL 201 ,VDEO 101	Pattern9
stud19	4	3.42	3.398	Average ,Average ,Easy ,Difficult ,Difficult	ACCT 312 ,ACCT 320 ,ACCT 341 ,BANK 302 ,FINC 310	Pattern10
stud20	4	3.23	3.2	Average ,Average ,Easy ,Difficult ,Difficult	ACCT 312 ,ACCT 320 ,ACCT 341 ,BANK 302 ,FINC 310	Pattern10
stud21	3	3.88	3.778333	Difficult,Difficult,Difficult,Difficult,Difficult,Difficult	ACCT 404 ,BANK 302 ,ECON 301 ,FINC 320 ,FINC 421 ,STAT 202	Pattern11
stud22	3	3.53	3.556667	Difficult,Difficult,Difficult,Difficult,Difficult,Difficult	ACCT 404 ,BANK 302 ,ECON 301 ,FINC 320 ,FINC 421 ,STAT 202	Pattern11

Table 1.1, Course taking pattern and course difficulty pattern

Table 1.1 was tabulated using the data of students extracted from the student dataset of an anonymous university registered in a particular programme. The table shows the courses in which 22 students have registered in a semester, the course difficulty for each course expressed over a 5-point scale (very difficult, difficult, average, easy and very easy), the SGPA and CGPA scored by each student and time-to-degree taken by students. From the table the following inferences could be arrived at.

The courses registered in by the students can be visualized as course taking patterns. However for the same set of students registered in the same semester and programme the time-to-degree achieved and the pattern of courses registered in by the students vary. Additionally there are students who have achieved the same time-to-degree and have registered in the same set of courses in the same semester but have scored different SGPA and CGPA (e.g. compare Stud21 and Stud 22). Furthermore, some students who have registered in lesser number of courses (e.g. Stud17 who has registered in 5 courses) have scored a higher SPGA but lower CGPA when compared to some other students (e.g. Stud22 who has scored lower SPGA but higher CGPA) achieving different time-to-degree (e.g. Stud17 has taken four years whereas Stud22 has taken 3 years to graduate). This indicates that course taking pattern can affect the SGPA and CGPA alongside time-to-degree. This means that certain courses in which a student registers in a single semester as a set can impact the SGPA, CGPA and time-to-degree. However there is a paradoxical situation. That is Stud22 has achieved the highest CGPA with the shortest time-to-degree amongst the 22 students under study albeit registering the maximum number of courses in a semester. That is to say that even though Stud22 has registered in 6 courses in a semester, yet the student could achieve the highest CGPA with the shortest time-to-degree of 3 years indicating that the student did not under-perform despite being loaded with the maximum number of courses in a semester. This could possibly be attributed to either the ability of the student or the course difficulty (contextual factor) of the set of courses in which the student has registered in each semester or the number of courses per semester in which the student has registered or the order of semesters chosen by the student to register in a particular set of courses as a pattern. It is not easy to decipher how this situation happens or explain the situation or repeat the situation or predict the pattern of courses that could be the best possible choice of a student to register in a semester leading to the student achieving the optimum CGPA, SGPA and time-to-degree. Literature is also silent on this aspect. Here it must be seen that there emerges a clear and direct relationship between unobservable attributes namely the courses in which a student registers as a set in a semester and the course difficulty that could be associated with each one of the courses in the set, an argument not well articulated by the literature (Vialardi et al, 2011).

The questions that arise are

- Whether there exists a pattern of courses of unobservable attributes in which a student can register each semester whose course difficulty pattern as an unobservable attribute could be defined and enable a student to achieve the best possible SGPA, CGPA and time-to-degree.
- If such an unobservable pattern exists, can it be discovered for each student with the implied association between the course taking pattern and course difficulty pattern?

- How to discover this pattern of courses.

Answers to the above questions are not easy to find and require detailed investigation. However it can be seen that answers to the abovementioned questions can enable HEIs to make decisions to provide adequate support to students leading to improvement in the students’ performance and learning experience. Among the three questions raised above it is seen that if the third question is addressed first then the first two questions could be addressed easily. For instance it can be seen from Table 1.1 that if the number of students involved in an institution runs into its thousands, then it is not possible to determine or discover the course taking pattern of students in many semesters and the course difficulty pattern of courses for those many students manually. Use of an efficient method is inevitable. That is identifying a suitable method becomes important which is a challenge in itself. However, literature shows that discovery of patterns is possible from student data using datamining techniques although there are a number of types of datamining techniques that could be used to discover patterns and determining the most suitable datamining technique seems to be another challenge. Further, it appears that it may not be possible to achieve business goals using ordinary datamining techniques for instance deciding on the best method to achieve optimum student performance, but require datamining processes such as KDDM processes. Again whether currently used and widely recommended KDDM processes are suitable to determine the course taking and course difficulty patterns of students and link them to time-to-degree, SGPA and CGPA is another challenge that needs to be addressed. Overall it can be seen that the above problems appear to emerge when one analyses how student performance could be improved using time-to-degree, SGPA, CGPA and contextual factor concepts.

1.2 Research Gap

From the foregoing discussions it can be seen that literature (Thakar et al. 2015; Zeidenberg, 2011; Vialardi, 2009) is silent on the question of whether course taking pattern of students as an independent event hidden in educational data is related to time-to-degree and CGPA as dependent event with regard to HEIs. If this relationship were to be established, it may be possible for HEIs to make decisions that could enable students to optimize their time-to-degree and achieve best possible performance. However literature points out that finding knowledge about course taking patterns of large number of students involves huge data bases and requires use of special methodologies like Educational Data Mining(EDM) to produce knowledge about those patterns and demonstrate their relationship to time-to-degree and CGPA. But hardly any research outcome has been found in the extant literature that has dealt with this problem of discovering course taking patterns of students (unobservable attribute) and demonstrating their relationship to time-to-degree and CGPA using data mining methodology in the context of HEIs which is a major gap. Such a gap has left the HEIs and students with lack of knowledge about the pattern of courses they could choose and decide to register in a semester leading to optimum time-to-degree and best possible CGPA if such knowledge is derived on scientific or empirical evidence such as knowledge discovered through data mining.

Further the relationship between course taking pattern of students as an independent variable on the one hand and time-to-degree and CGPA as dependent on the other, is argued to be characterized by contextual attributes (unobservable) see section 2.3.3, although there has been no useful knowledge produced in the literature about this relationship as concept. KDDM are promising tools to investigate course taking patterns (Vialardi, 2009; Zeidenberg, 2011; Luan, 2002). However existing KDDM processes have been found to have limitations as they lack the ability to clearly unearth contextual knowledge from a given data set because of which the discovered data through the KDDM process is seen to be bereft of contextual knowledge. This knowledge that lacks contextual attributes is not considered to be of adequate quality for decision making.

Thus there is a need to develop a KDDM process model with a stage integrated into it which links the contextual knowledge domain with the main process, a need that is not addressed in the literature which is another gap in the literature. Such integration could enable the selection or rejection of a dataset at the outset of the KDDM process depending on whether the data set has hidden information about the context or not thereby qualifying the dataset for further mining to discover knowledge about the course taking pattern of students and its relationship to time-to-degree and GPA.

In order to address the research the following research questions have been formulated.

1.3 Research Questions

RQ1: Is there an unobservable relationship between course taking pattern, optimum CGPA and optimum time-to-degree hidden in the educational dataset of undergraduate students of a higher education institution? If there exists a relationship between the three factors then is there an *unobservable contextual factor course difficulty level pattern hidden in the educational dataset of undergraduate students of a higher education institution that affect the relationship between course taking pattern, optimum CGPA and optimum time-to-degree? Using KDDM process is it possible to establish the relationship between course taking pattern, course difficulty level pattern, optimum CGPA and optimum time-to-degree by discovering the unobservable relationship mentioned above?*

RQ2: Is it possible to predict the optimum CGPA and optimum time-to-degree of undergraduate students in terms of the course taking pattern and course difficulty level pattern extracted from the educational dataset of a higher education institution using a KDDM process?

In order to answer the research questions it was necessary to define the aim that could lead to those answers. This is defined next. The following aim and objectives have been set to address the research questions.

1.4 Aim and Objectives

The main aim of this research is to enhance a KDDM process model by adding contextual knowledge processing stage for implementation in HEIs in the decision making process to improve student performance (CGPA and time-to-degree) by mining student dataset and discovering course taking pattern of students characterized by course difficulty pattern (contextual factor).

In order to achieve this aim, the following objectives were identified

1.4.1 Identify factors that affect student performance in HEIs and examine in particular three factors namely course taking pattern of students (unobservable attributes), time to degree and CGPA characterized by contextual knowledge namely course difficulty pattern (unobservable attribute) through literature review.

1.4.2 Investigate the importance of course-taking patterns of students and their influence on time to degree and CGPA characterized by course difficulty pattern (contextual factors) in the decision making process regarding student performance.

1.4.3 Study the usefulness of KDDM process modelling to discover course taking pattern characterized by contextual knowledge and determine the relationship between course-taking patterns of students on the one hand and time to degree and CGPA for making decisions in HEIs regarding student performance.

1.4.4 Examine a KDDM process model and enhance the standard model to include contextual knowledge process stage for mining student dataset and discovering course taking pattern of students characterized by course difficulty pattern (contextual information) in HEIs.

1.4.5 Validate the enhanced KDDM process model using simulation method.

In order to achieve the above aim and objectives in this research the following research methodology has been adapted.

1.5 Research method

This research method has 3 steps: choosing appropriate research paradigm, development of hypotheses and research context. An understanding of these 3 will provide the way forward and achieve the aim and objectives.

1.5.1 Choosing an Appropriate Research Paradigm

There are two research paradigms namely Behavioural Science and Design Science. While the former has its origins in natural science research method the latter has its origin in engineering and the sciences of the artificial (Hevner et al, 2004; Simon, 1996). The behavioural science paradigm deals with development and justification of theories for e.g. principals and laws, that either describe or determine organisational and human happenings encompassing the examination design, execution and use of information systems. Literature shows that those theories eventually appraise researchers and practionisers of the inter relationship and interactions amongst people technology and organisations which need to be managed. If an information system is to realise the purpose it’s meant for, for instance, improving the success and efficiency an organisation could achieve (Hevner et al, 2004). It must be noted that those principles or laws affect or get affected by design decisions taken with respect to the system development methodology employed and the functional aspects, information substance and human interfaces executed within the information system. On the other hand, design science philosophy is basically concerned with problem solving. Its goal is to create innovations that outline the ideas, ways of doing things, technical expertise and products using which the testing design accomplishment and utility of information systems can efficiently and effectively achieved.

Even though most researches follow behavioural science, there is a potential scientific approach inherently involved namely the design science. The main goal of design science is to develop artefacts that solves particular problem (constructs, methods and models) (Cole, Purao, Rossi, & Sein, 2005; Hevner et al., 2004; March & Smith, 1995; Simon, 1996). Additionally, there exists a simple design that could be compared with design science which needs to be distinguished. Design science is concerned with artefacts that add novelty to current scientific knowledge base through improvements carried out on technical, social or informational resources. This implies novelty is not expected from a simple design. Furthermore, while it is argued that design science and behavioural science are related and one method can support the steps of the other (e.g. steps of design science research can be supported by behavioural science method) (Cole et al., 2005; Hevner et al., 2004; March & Smith, 1995), it must be pointed out that these two sciences are distinctly different. For instance, design science uses prescription to develop artefacts (prescriptive), whereas the other science deals with description and explanation (March & Smith, 1995, 254). Again, the behavioural science concerns with expansion and validation of theories or principals or laws which provide answers to problems related to organisational and human phenomena or predict happenings related to those phenomena.

From the above discussion on behavioural and design science it can be seen that a proper research paradigm could be chosen for this research to respond to the research questions formulated for this research. For instance, the research questions point towards the need to develop a KDDM model with a goal to extract or discover course taking patterns and course difficulty patterns which when used with other factors (time-to-degree and CGPA) can help in making better decisions in HEIs. That is to say a model of KDDM when developed, will assume a shape of an information system artefact (with a stage(s) or step(s) that extracts and processes contextual factors in the form of contextual KDDM process model) that addresses a solved problem (implementation of the KDDM process) but in more effective and efficient ways. This artefact needs to be developed through analysis, design, implementation, management and use of information systems (Hevner et al., 2004).

Here the concept of design science could be useful as it deals with the development of artefacts. On the other hand, behavioural science method is used then such a paradigm will not directly lead to the design, development or building of artefacts. In this situation use of design science to design an artefact appears to be more relevant, important and specific. And hence the choice of design science research could be justified for this research. Design science further can be used to support the fulfilment of the identified business needs (for e.g. improving student performance in terms of time-to-degree) by building appropriate artefact (March and Smith 1995; Järvinen 2000; Hevner, March et al. 2004). In addition, innovation is needed in this research to discover unobserved attributes to (e.g. course taking patterns) solve existing student performance problems more effectively. It is aimed that by using design science paradigm it is possible to implement a knowledge discovery and data mining artefact to solve the problem of HEIs related to student performance. The foregoing discussions provide the justification for use of design science paradigm for the research questions that have emerged. Following the justification of choice of research paradigm the next section deals with the assumptions to be made for answering research questions by developing hypotheses.

1.5.2 Development of Hypotheses

The hypotheses developed for this research aims to answer the research questions set. Thus taking into consideration the table 1.1 and discussions thereof it can be seen that there is scope to relate the factors CGPA, course taking patterns, time-to-degree, course difficulty and semester number (contextual factors). It is possible that if changes are made to course taking pattern of students in different semesters then the table shows that CGPA and time-to-degree could vary then a function can be generated which can be expressed as hypotheses provided below. The hypotheses developed thus are provided below.

- *H1: CGPA can be expressed as a function of student course registration data by semester, time-to-degree and course difficulty.*
- *H2: Time-to-degree can be expressed as a function of student course registration data by semester, CGPA and course difficulty.*

Hypotheses H1 and H2 covers every aspect of RQ1. As far as RQ2 is concerned it is argued that a reference KDDM model has to be chosen to develop a new artefact that will enable the testing of the hypotheses H1 and H2 using design science. In this context a review of different KDDM processes recommended was conducted through literature review (see section 2.6). A total of 6 KDDM processes namely generic, Fayyad, Cios et al, CRISP-DM, Anand and Buchner, Cabena et al were reviewed. The comparison is presented in Appendix 23. It was found that none of the KDDM processes were found capable of dealing with course difficulty (contextual factor). CRISP-DM model was chosen (see section 3.4) and used for developing a new artefact using design science research paradigm.

1.5.3 Research context

The research was conducted in Bahrain in an anonymous HEI. The anonymous HEI where this research was conducted offers both undergraduate and graduate programmes, for instance Bachelor's degree in Accounting and Finance (BSAF), Bachelor's degree in Management and Marketing (BSMM) and Master's degree in Business Administration (MBA). The HEI follows American system of awarding credits to students who successfully complete courses. Each course is typically of 3 credits. In American Education System, the students of Bachelor's degree programme graduate after completing the required number of credits which is equivalent to 44 courses with a minimum CGPA of 2.0 per course and usually complete in about 4 years of study up to a maximum of 8 years.

The focus of this research is restricted undergraduate levels only where the concept time-to-degree has serious implications to the student, the HEI and other stakeholders. At the undergraduate level if a student has to graduate, the student must have successfully completed 132 credits usually which is earned by completing 44 courses typically. The HEI offers semester system. Full time students are those who are registered in a minimum of 4 courses per semester but not exceeding 6. The maximum CGPA that could be scored is 4. The focus of this study is on the choice based credit system that facilitates the students to take courses of their choice, learn at their own pace, undergo additional courses and acquire more than the required credits, and adopt an interdisciplinary approach to learning. As the choice of courses is left to the discretion of the students, the graduation rates of the HEIs are affected, thereby affecting their reputation. The importance of grading is elevated to the extent where it can decide the future of a student or HEI. For Students, the grades are used as a means of selection within and between HEIs and as eligibility for employment. At institutional level, they are used for assessing the success of HEIs. Grades are awarded as mentioned in the Table 1.2.

Table 1.1, Grade and Grade points

Grade	Grade points
A	4
A-	3.67
B	3
B-	2.67
B+	3.33
C	2
C-	1.67
C+	2.33
D	1
D+	1.33

1.6 Research significance

Addressing concerns related to decision making in HEIs to improve student performance and associated aspects, has provided a new way to extract course taking patterns of students per semester and predict the time-to-degree and the grades of students using data mining method, a method that is part of the design science. In addition, a new KDDM process has been developed to introduce the concept of contextual factors (refer Appendix 25 for contextual factors) in extracting student course taking patterns to predict time-to-degree and grades, an aspect not addressed in the literature. This process is expected to provide a method by which hidden knowledge in student data could be extracted to take decisions by HEIs to enhance student performance and student experience, particularly related to time-to-degree and grades. Student profiling, evidence based student advising, offering special assistance to students to improve performance using course taking pattern and its relationship to time-to-degree, grades and contextual factors (refer Appendix 25 for contextual factors), identifying teaching needs of students at risk and determining the courses to be offered semester after semester are some of the areas that could be addressed using the outcome of this research. Researchers could investigate the outcomes of this research to identify newer contextual factors (refer Appendix 25 for contextual factors) and their impact on the course taking pattern of students.

1.7 Thesis layout

The second chapter discusses about current knowledge in the literature related to the research context, about data, importance of data in data driven decision making in HEIs, attributes of data, significance of contextual data in decision making in HEIs, limitations in the current literature related to critical factors that affect the decision process in HEIs and theoretical aspects concerning big data. It also dwells on data and data analysis, educational data mining and its application to HEIs, discovery of course taking pattern and time-to-degree.

Chapter 3 provides the development of basic functional relationships between course taking pattern, course difficulty, optimum CGPA and time-to-degree, KDDM process and synthesis of theories explaining KDDM process. The chapter also analyses functional relationships between course taking pattern, course difficulty, optimum CGPA and time-to-degree, develops the basic model using functional equations required for testing the KDDM process, KDDM models, their limitations and rationale behind the choice of CRISP-DM process and application of design science methodology.

Chapter 4 provides the guidelines to test the CRISP-DM process and results of testing of the basic model developed in Chapter 3 using clustering technique.

Chapter 5 details out results of testing of the basic model developed in Chapter 2 using association rule mining and classification techniques. In addition this chapter provides the basis for modification of the CRISP-DM model to integrate the course difficulty level as contextual factor.

Chapter 6 articulates the details of the modification carried out on the CRISP-DM process and the integration of the EDM to mine general as well as contextual datasets. The chapter describes the predictive ability of the modified CRISP-DM process and demonstrates how the model could be deployed using simulation.

Chapter 7 provides the discussion on whether the research questions have been addressed or not and the conclusions which include contributions to knowledge, theory, method and practice alongside the limitations of the current research and recommendations for future research.

Chapter 2 : Related Literature

2.1 Introduction

In the HEI sector large datasets could be analysed to enhance the performance and skill level of students, improve the personalized learning experiences of students that lead them to achieve their specific learning pathways, reduce drop-out rates and improve graduation rates (Daniel, 2015). In order to realize how large datasets in HEIs could be used to gain knowledge that could be helpful in enhancing student performance, it was necessary to discuss factors that have real importance and linkage to student performance. The factors investigated in this research are time-to-degree, course taking patterns of students, contextual factors namely course difficulty and cumulative grade point average (CGPA) of students. In fact that there could be a relationship between or association amongst those factors is hitherto not known and discovered and is a major challenge and gap. This aspect is discussed in Section 2.3. In order to discover the association amongst those factors one of the methods that has been identified in this research was data mining of large datasets. Despite its purported use to the HEI sector, research related to large data in the context of HEIs is limited (Daniel, 2015). This aspect is reviewed in Section 2.4 to gain an understanding of what is known about large data, its application to the HEI sector, theoretical and practical aspects, and what is not known. Furthermore, this chapter has critically reviewed the different datamining concepts and KDDM processes which are provided in Sections 2.5 and 2.6. Lastly there was a necessity to identify a KDDM process that could be used to address the research questions. CRISP-DM model was chosen as a reference model. This is reviewed in Sections 2.8 and 3.4.

2.2 Review of factors that could enhance Institutional performance

Many establishments including higher educational institutions (HEIs) have been struggling to improve their performance for instance making accurate and effective decisions (BIS, 2014). In general, researchers (Pheng & Arain, 2006) argue that decision making processes could be made effective if knowledge acquired from various sources is used in decision making. This argument is applicable to HEIs as well. In particular this research focuses on the decisions HEIs could take with regard to enhancement of student performance which is found to be a major area of concern. Decision making is one component that affects the performance of an institution and hence taken as an example for discussion here. Certain factors including course taking patterns, time-to-degree, CGPA and contextual knowledge are found to be affecting decision making processes in HEIs by researchers (Lotkowski, et al. 2004). However current level of understanding on ways these factors could be utilised to make useful decisions to improve student performance is not very clear and needs to be investigated further (Huebner, 2013; Osmanbegović & Suljić, 2012). It must be mentioned here that the focus of this research is not decision making process but how decision making can be aided by analysing certain factors and raw data pertaining to those factors, using latest analysing techniques. In this research CGPA of students is used as one of the reference factors to indicate performance of students and decision making although it is not the only reference factor. Therefore this factor will be commonly used throughout thesis to indicate the student performance quantitatively.

Knowledge that needs to be used for decision making is stored as data in data warehouses in almost all HEIs. For instance knowledge including demographic characteristics of students enrolled in various programmes, time-to-degree of students and GPA of students in HEIs is stored in the form of student data and warehoused using computer systems. This knowledge is used by institutions regularly to make decisions by HEIs in many areas including strategies related to student support, student assessment, student performance, student administration and teaching issues. Although the knowledge utilised by HEIs in decision making is obtained from the data that resides in student information systems of HEIs. However such knowledge seem to lack depth and might have omitted essential knowledge that is unseen in the data that could not be extracted by those processes mentioned above. The outcome is utilisation of incomplete knowledge in decision making processes in HEIs.

In order to overcome this issue researchers have established sophisticated processes for knowledge extraction, foremost amongst is the data mining process, also known as knowledge discovery data mining process (KDDM process). Although substantial advances are made in developing data mining processes, literature points out that in HEIs there is very little evidence of use of data mining processes in decision making (Goyal & Vohra, 2012). It could be inferred that one of the reasons for this situation could be the lack of consensus amongst the researching community with regard to identifying a single standardized process that could be used in all contexts. While data mining has been a very useful process for knowledge discovery, the availability of a varied variety of data mining processes having significant differences among them can cause difficulty for the users in determining which one of them is the best. This is a major challenge and finding the most suitable mining method to deal with a business task by itself requires in-depth study about different data mining techniques and their performance aspects. Besides, each process has its own advantages, disadvantages and limitations. In addition, there is no one process fit all situations. These arguments clearly make it difficult to choose and implement a particular data mining process in HEIs because of lack of sufficient guidance from research outcomes. Current research efforts related to data mining focus on improving mining methods and modify certain algorithms concerning those data mining methods leading to the development optimal data mining methods that make the mining process simpler and easier to implement and extract hidden knowledge. However there a number of areas in regards to HEIs where data mining as a method is yet to be accepted as an established method that could aid in improving the decision making or enhancing student performance by extracting complete knowledge hidden in the datasets as well as lead the investigation to create a relationship amongst student performance factors. What is significant is that it is still not clear how far the different data mining methods developed so far and tested in the industrial sector could support the HEIs in extracting complete and accurate knowledge (Hollands & Escueta, 2017) as well as enable discovery association amongst the factors. This is a major gap and requires in-depth study. For instance a single type of data could be analysed using multiple data mining methods making it difficult to know which of those methods are most suitable to be applied to HEIs and how to substantiate the claim. In addition the current level of knowledge about the various data mining methods and their application in the context of HEIs is considered to be insufficient to support HEIs and modifications may need to be carried out in terms of improving the algorithms or data mining processes so that complete knowledge that is accurate is extracted that could help in creating association amongst the factors is discovered. Carrying out such modifications require the application of data mining theories, models and elaborate testing before concluding that the modification is valid and acceptable.

Thus there is a need to study some of the widely used data mining processes and find out how HEIs could be guided in choosing the most appropriate processes to support their effort to implement data mining processes in discovering knowledge from datasets and improve their performance. Any knowledge contributed in enhancing a particular data mining process and eliminating the limitation through scientific research could significantly help in better understanding of the process, interpretation of the outcomes, extracting complete hidden knowledge and discovering associations amongst performance factors not known until now. Such an understanding could lead to knowledge that could be used in enhancing institutional performance. However prior to understanding about the data mining process or knowledge discovery process it is important to review what factors related to students and decision making in HEIs have been widely studied in the data mining literature and which factors have not been studied yet. It is also necessary to know whether data mining methods could be used to address those factors useful to enhance institutional performance not studied yet. Thus the following section discusses some of the important factors that have been studied widely and those that need to be studied further.

2.3 Critical factors that affect the decision processes in HEIs

HEIs are facing challenges that have surfaced due to a sharp rise in the demand for higher education and the corresponding massive growth of higher education sector. In this context, universities are forced to adopt planning and research methods that would help to detect and address the requirements of a larger, more diverse student body (Menon et al. 2014). In such a situation the risk of making wrong decisions could cost the HEIs dearly. Hence HEIs are forced to make evidence based decisions. Decision making theory highlights the importance of systematic research prior to the actual decision making and implementation stages, especially in cases of strategic planning (Milkman et al. 2009). While an extensive amount of research has been conducted regarding how postsecondary institutions are organized (e.g. Bess & Dee, 2008), less attention has been paid to the specific role that data and data-related systems play in college and university operations. However the situation is gradually changing wherein HEIs are forced to move towards data driven decisions. At the same time, given the fact that extensive quantum of data pertaining to students is available in most colleges and universities (e.g. graduation rates and annual tuition revenue), many observers argue that analytic techniques from the world of big data could be applied to higher education (Picciano, 2012; Lane, 2014).

Factors or variables that have been studied with regard to extracting hidden knowledge using data mining methods in HEIs include prior learning, course attendance, financial aid, gender and age. Factors that have not been studied in-depth with regard to extracting hidden knowledge using data mining methods in HEIs include course taking patterns and student performance (time-to-degree and CGPA). Study of factors hitherto not investigated with regard to extracting hidden knowledge using data mining pose a problem in regard to determining the most suitable data mining method using comparable performance factors, improving existing data mining methods to perform better in discovering and extracting hidden knowledge, developing ways to make the data mining process optimum, provide better ways to interpret mined results for more accurate prediction. The following section provides some idea about critical factors which need to be studied to explain about the ways to choose the most suitable data mining method and modify the data mining method to discover more complete knowledge hidden in the datasets pertaining to those factors that could be used to enhance the performance of the HEIs with a focus on decision making.

Some factors that are identified to be affecting the decision processes of HEIs in the literature are given in Table 2.1.

Table 2.1, Factors affecting decision processes in HEIs

Factors	Authors
Student dropouts	Astin, 1971
Student retention	Tinto, 1975; Daempfle, 2003
Student performance	Minaei-Bigdoli, et al. 2003
Student satisfaction	Athiyaman, 1997; Elliott & Healy, 2001; DeShields, et al. 2005; Helgesen & Nesset, 2007
Time-to-degree	Knight, 1994; Adelman, 1999
Course taking patterns	Ronco, 1996; Bahr, 2010; Kovacic, 2010; Vialardi, et al. 2011

Though these factors are found to be affecting decision making in HEIs, literature is silent on the impact of course taking pattern as a factor on student performance (GPA) in terms of time-to-degree. Further, literature has emphasized that data pertaining to these three factors namely student performance (GPA), course taking patterns, and time-to-degree, can have hidden knowledge which can be used in decision making in HEIs. This infers that data mining could be used by HEIs as a support mechanism to help in their decision making process. However research results produced in this area seem to be deficient in suggesting conclusive ways using which course taking pattern can be used as a factor to predict student performance and the time-to-degree using data mining as part of the decision making process in HEIs. This indicates that there is a need to examine the impact of course taking pattern on student performance and their time-to-degree and to find the appropriate one from the many data mining processes developed by researchers which can be used to bring out the knowledge hidden in student data characterized by attributes which include these factors. This research addresses this issue by examines the enhancement of student performance in terms of time-to-degree and grade point average (GPA) as a business case. Based on the above arguments, the researcher could hypothesize that there may be a pattern of courses enrolled by students in semesters that acts as the independent factor, knowledge about which if extracted from student data, can enable the students and HEIs to determine the most optimum time-to-degree and the best possible GPA, which then become two dependent factors. It is found from the literature that this aspect has not been well understood in general. In addition knowledge to apply data mining method to study this aspect is not found in the literature.

While this research focuses on time-to-degree, course taking pattern and CGPA, it is pertinent to note that educational datasets are characterized by contextual factors (refer Appendix 25) including student potential, course complexity and course difficulty which have a role to play in extracting more complete knowledge hidden in datasets alongside the three identified factors. The extraction of these factors from the dataset using data mining methods and discovering course taking patterns that are characterized by contextual factors (refer Appendix 25) is an area not investigated in the literature. Discovering this knowledge is expected to provide an understanding of how the factors could be related and hence better decisions can be made to improve student performance. This knowledge gap in the literature has serious implications to the ability of current data mining methods to discover knowledge or pattern that are characterized by contextual aspects (refer Appendix 25) from educational datasets. The significance of the need to study time-to-degree, course taking pattern and CGPA as well as contextual factors for this research is brought out in the discussions that follow.

2.3.1 Time-to-degree of students

Time-to-degree is the entire number of semesters excluding summer sessions taken to reach graduation. The time-to-degree measure includes only full-time students not including transfer students. Some examples of students and the time taken to achieve their degree are provided in Table 2.2.

Table 2.2, Example of time-to-degree and CGPA

Student ID	CGPA	Time-to-degree
Stud1	3.78	3.5
Stud2	3.84	3.5
Stud3	2.18	4
Stud4	2.62	4
Stud5	2.43	4
Stud6	2.24	8.5
Stud7	3.6	4
Stud8	3.85	4
Stud9	2.77	4
Stud10	3.36	3.5

The first column indicates the student ID and has been coded to maintain anonymity. This is an elementary factor in decision making. The Table 2.2 shows that different students have taken different times-to-degree and have scored different CGPA. What factors contribute to this phenomenon is a major question. For instance three students with ID Stud1, Stud2 and Stud10 have taken only 3.5 years to graduate and have scored CGPA equal to 3.78, 3.84 and 3.36 respectively whereas majority of the other students who have taken 4 years and above scored lower CGPA. This is an unusual situation where students who have taken shorter time-to-degree have scored higher CGPA than those who have taken longer time-to-degree. Moreover the 10 students who have been referred here were chosen randomly from the student dataset consisting of hundreds of thousands of students implying that the variation in time-to-degree could be even wider if the entire dataset is analysed with CGPA also showing wide variation. In this situation is it possible to conclude that students taking shorter time-to-degree will score higher CGPA when compared to those taking longer time-to-degree? What factors can contribute to this? To answer this question further study needs to be conducted. Thus the factor time-to-degree has been critically reviewed next.

To begin with some of the general characteristics of the concept of time-to-degree were explained. Time-to-degree not only affects students but also institutions in the name of graduation rates. The graduation rate was defined as the percentage of full-time, first-time, degree-seeking enrolled students who graduate after 150 percent of the normal time for completion; defined as six years for four-year colleges (8 semesters or 12 quarters excluding summer terms) and three years for two-year colleges (4 semesters or 6 quarters excluding summer terms). Graduation rates, regardless of how they are calculated, are a source of contention. Low rates are associated with poor performing institutions while high rates are associated with superior institutions (Gillmore & Hoffman, 1997; Underwood & Rieck, 1999; and Astin, 2005). Attention is frequently drawn to institutions' graduation rates when large numbers of students enrol but don't graduate, for fear that public funds are being wasted on these institutions (Fields, 2005) where at the end of the fiscal year, their losses are kept to a minimum and their gains are maximized to the fullest extent possible.

HEIs with bad graduation rates end up with bad admission rates and bad reputation. The statistics of (IPEDS, 2013) shows that only 50 of the 580 public four-year institutions in America have on-time graduation rates as shown in Table 2.3.

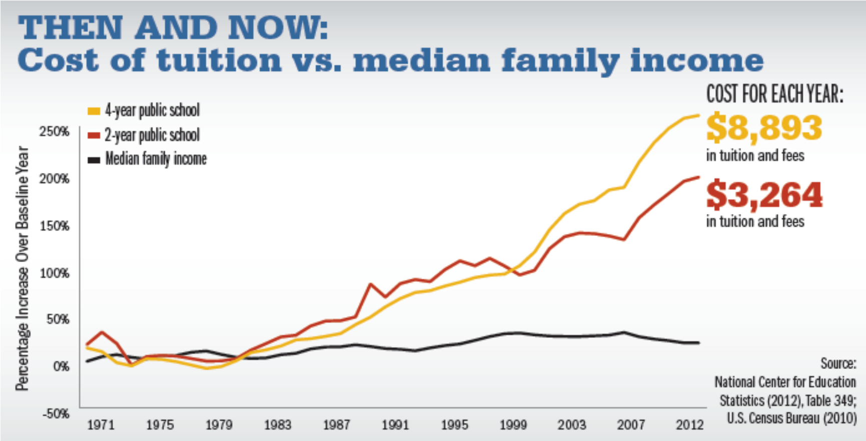
Table 2.3: On time graduation rates of U.S. Institutions (Source: Complete College America, 2014)

On-time Graduation rates for their full-time students **(only 50 of 500 public four-year institutions in America have on-time graduation rates)**

Michigan State University	**University of Iowa**	**North Carolina State University**	**Auburn University**	**University of Arizona**
48%	44%	41%	36%	34%

Time is money. Parents and students lose their money in loans/debts and missed opportunities. Every extra year of tuition and fees adds up, and borrowers who do not graduate on time take on far more debt in years 5 and 6. On average, an additional year now costs more than $3,000 extra at a two-year institution and nearly $9,000 extra in tuition at a four-year institution. Most colleges and universities raise tuition and fees each year, while financial aid stays nearly constant. As scholarships and savings run out, students and their families are left to borrow more of the costs of attending school. The above arguments are supported by researchers (Schneider, 2010; Skinner, 2011; Wellman, 2008, 2010a, 2010b, 2011).The statistics given in Appendix 20 clearly indicate this.

The cost of higher education has drastically outpaced increase in median family income. As a result, obtaining the education necessary for success has become far more difficult and costly, and students have been forced to pile on even more debt in the process.



Source:
National Center for Education Statistics (2012), Table 349;
U.S. Census Bureau (2010)

Figure 2.1, Cost of Tuition (*Source*: Complete College America, 2014)

The reasons for students not able to complete on time are provided in Appendix 19.

Guided Pathways (GPS) is the only direct route to graduation. There can be new policies and strategies that tackle head-on the institutional practices that are the great drag on student progress: credits lost in transfer, unavailable critical courses, uninformed choices of majors, low credit accumulation each semester, broken remediation sequences, and excessive credit requirements.

Addressing all this is possible through the implementation of a comprehensive, integrated restructuring of higher education delivery called Guided Pathways to Success (GPS).

Every major should be organized into a prescribed pathway of sequenced courses that lead to an on-time arrival on graduation day. And all students should be scheduled to maintain steady progress on their chosen path. Random acts of enrollment should be replaced with deliberate and directed advancement toward degrees. After highlighting the importance of time-to-degree to students and institutions, the below paragraphs details the contemporary literature on time-to-degree, its outcomes, current limitations in the research outcomes and the methods used.

The research on time-to-degree started in 1994 when Knight (1994) investigated time-to-degree using multiple regression. Knight (1994) used variables like graduates' gender, race, high school grade point average, date of birth, major at the time of graduation, admission status, SAT verbal and mathematics scores, number of quarters attended, number of courses dropped, and graduation date. Graduates' grade point averages (final and at the end of the freshman year), age at matriculation, total credit hours completed, residence status (whether or not they lived in a residence hall during their freshman year), financial aid status and whether or not graduates enrolled in an orientation course. The outcome of this research showed that gender, enrollment behaviours, academic ability, preparation and credits per semester were found to be the best forecasters of time-to-degree. The main limitations as pointed by Knight was that important predictor variables like interaction with faculty and students, and employment data, were not used. Further the effects between colleges could not be assessed as the study was limited to graduated students from only one university. Student background, college environment, and student involvement effects were not studied.

Volkwein and Lorang (1996) investigated time-to-degree using first phase transcript analysis, special student outcome survey and second phase multivariate regression and used background information namely student plans, levels of student satisfaction, a range of cognitive and non-cognitive knowledge and results including classroom experiences, course taking patterns, instructor contact, graduation plans, anticipated loan indebtedness, GPA and self-reported growth. The research outcome showed that getting grants, lower class loads per term and higher GPA was associated with longer time-to-degree. Also increased time-to-degree by students enabled protection of their grade and students to have more free time.

Lam (1999) used multiple regression with demographic variables like residency status, gender and ethnicity; academic variables like admission test scores and CGPA during graduation. 12 attributes were used in the study. The enrollment variables comprised of hours transferred from other institutions, number of semesters enrolled as part-time students, number of changes in major and count of summer sessions enrolled. The financial variables comprised of total family contribution, total financial aid dollars received by aid type, and employment. The study examined the relationship between financial aid type received by students and the time taken to attain a Bachelors of Sceince degree. It was found that students with loans completed the degree on time. The limitation of the study was having the assumption that loans have to be repaid; the role of loans in funding higher education may act as encouragement for students to complete the programme on time.

Further Knight (2000) used structural equation modelling approach to study the effect of student background features, summer freshman program participation and remedial course, financial aid data, enrollment behaviours, academic results, student experiences and perceptions on time-to-degree. It was found that summer term enrollment, per term average student credit hour load, credit hours transferred and count of courses failed were acting as strongest predictors of total enrolled terms and completed total terms before graduation. The study used financial aid data for the previous 3 years of enrollment. Complete financial aid data set could have produced different results .Regular data on non-campus-based student employment was not available.

Herzog, 2006 used decision trees (C&RT, CHAID-based, and C5.0) and three backpropagation neural networks (simple topology, multi topology, and three hidden-layer pruned) to analyse time-to-degree. Around fifty attributes including financial aid, demographics, campus and academic experience were used to predict student retention. The purpose of the study was Retention prediction and time-to-degree analysis. The research indicates that decision tree and neural networks worked better for larger data sets.

Letkiewicz et al. (2014) applied logistic regression models to study the effect of sociological and economic factors on time-to-degree. It was found that college atmosphere and personal monetary features are significant factors in defining time-to-degree. Students who overspend, have a car loan, credit cards, or high debt, and those who feel stress from their funds are more likely to take longer than 4 years. Students were expected to finish in 4 years or less if they live or work on campus, have a high GPA, or have met with a financial counsellor or advisor. However this study did not take into account the impact of courses or course taking pattern on the GPA which is a serious limitation because some of the factors like living or working on campus can easily be achieved with less problems when compared to course taking pattern or course difficulty pattern. Course taking pattern is a difficult variable that cannot be easily applied on every student without taking into account the student ability and together course taking pattern and course difficulty pattern provide a solid ground for investigation.

When viewed from another angle it was seen that increased time-to-degree can consume resources (e.g. space and money) of institutions (Bowen et al. 2009). Institutions can save money and space to admit new students if the students can graduate in lesser semesters thereby increasing the graduation rates although it was not a conclusive finding. For instance if a student wants to take a longer time-to-degree when compared to another student who wants to take a shorter time to degree, but with the intention to score a grade of A and CGPA of 4.0, then the question of space and money will matter in which case the previous argument is not valid. Astin's model was updated by Knight (1994 & 2002) so as to forecast and describe time-to-degree. In the early 1990s Knight noted that, there were concerns of monetary restriction, responsibility and anxiety over the results of the undergraduate experience (1994, pg. 7). In 1980, the financial downturn placed pressure on colleges and universities to exhibit their usefulness and efficiency. Less research was conducted on time-to-degree in 1994, but it was developing as a vital issue to be researched. Therefore, Knight developed a model by adding to his model in 2002 to measure time-to-degree within an individual institutional setting. Time-to-degree had resurfaced as an important outcome for students in the current downturn and financial restraints placed on colleges (Bowen et al. 2009). Many factors were known to be linked with an increased time-to-degree. Using Knight's Input-Environment-Output framework, inputs like race (OCSA, 1996), gender (Knight, 1994; Adelman, 1999), preparation (Zhu, 2003; Ishitani, 2003; Knight, 1994) and socioeconomic status (Astin, 1993; Campbell, 2003) have been known to have influence on time-to-degree either directly or indirectly. College surrounding variables, like changes in major (Klopfenstein, 2000; Adelman, 2006; Ma, 2010; Knight & Arnold, 2000), credits per semester (Knight, 1994; DesJardins et al. 2003; Ishitani & Snider, 2003; Volkwein & Lorang, 1996), continuous enrollment (Ishitani, 2005; 2006; Belcheir, 2000) ,summer enrollment (Volkwein & Lorang, 1996) and first year GPA (Belcheir, 2000; Volkwein & Lorang, 1996; DesJardins et al. 2002) also have an influence on time-to-degree. Lastly, college variables related to completion such as total credits earned (Pitter et al. 1996) and missed credits (Florida Board of

Governors, 2004) have been found to affect time-to-degree. Although there is a lot of research done on time-to-degree, less is known on how time-to-degree varies with course taking patterns.

From the above discussion it can be seen that wide ranging factors are contributing to time-to degree although there is no consensus among researchers on the factors affecting time-to-degree. Further other factors which affect the time-to-degree like the course taking patterns appear to have escaped the attention of researchers. Even though the importance of increased time-to-degree is highlighted in the contemporary literature, most of the researches have used simple descriptive methods. Robust methods like data mining which have the capability of unearthing patterns and predictions have not been used in most of the researches. Even though guided pathways have been suggested in the literature for ensuring on timely completion, course taking patterns have not been completely studied. Further course taking patterns of students is seen to vary based on the contextual attributes of the course and students (Vermunt, 2005). Hence a study concerning the courses enrolled by the students along with the contextual attributes (like course difficulty) (refer Appendix 25) can bring out interesting results in terms of student performance (time-to-degree and CGPA). Further the knowledge of a course if enrolled in combination with which other course(s) leads to shorter or on time degree completion may also help the students and advisors.

2.3.2 Course taking Patterns of Students

The discussions in the previous sections showed that shorter or longer time-to-degree has major implications to students and other stakeholders and pointed towards some factors that could contribute to shorter or longer time-to-degree. For instance one factor that has been identified as affecting time-to-degree but not well investigated is the course taking pattern of students (Adelman, 2006). Course taking pattern as a factor has been reviewed in this section with regard to its relationship to time-to-degree and CGPA.

Table 2.4, Course taking pattern of students

Student ID	GPA	Time-to-degree	Course Code	Semester
Stud1	3.78	3.5	ACCT 101,ARAB 101,ECON 101,ENGL 101, ENGL 102	1
Stud2	3.84	3.5	ARAB 101,ECON 101,ENGL 101,ITCS 101,MATH 103	1
Stud3	2.18	4	ARAB 101,ENGL 101,ITCS 101	1
Stud4	2.62	4	ARAB 101,ECON 101,ENGL 101,FREN 101,ITCS 101	1
Stud5	2.43	4	ACCT 101,ECON 101,ENGL 101,ITCS 101	1
Stud6	2.24	8.5	ARAB 101,ECON 101,ENGL 101,ITCS 101,MATH 103	1
Stud7	3.6	4	ARAB 101,ECON 101,ENGL 101,ITCS 101,MATH 103	1

Stud8	3.85	4	ACCT 101,ECON 101,ENGL 101,ITCS 101,MATH 103	1
Stud9	2.77	4	ACCT 101,ECON 101,ENGL 101,ITCS 101,MATH 103	1
Stud10	3.36	3.5	ARAB 101,ECON 101,ENGL 101,ITCS 101,MATH 103	1

Course taking pattern of students refers to that pattern of courses the students enrol in each semester (Table 2.4). In Table 2.4 it can be seen that course taking pattern indicates the set of courses the various students have registered in Semester 1 in a private University in Bahrain. In column four it can be seen that student with code Stud1 has registered in 5 courses whereas student with code Stud3 has registered in 3 courses. However the time-to-degree in both the cases differs by six months with Stud3 taking a slightly longer period of time. This example shows that although in Semester 1 both Stud1 and Stud3 have registered in differing number of courses with Stud3 registering in less number of courses when compared to Stud1, difference in the time-to-degree between the two is restricted to only six months. This indicates that Stud3 must have enhanced his or her performance by registering in more number of courses in subsequent semesters to complete the degree in 4 years. If Stud3 had continued to take 3 courses in each semester then it was likely that the student would have taken much longer period of time to graduate which is apparently not the case. Thus some factor or factors related to the courses or the student or both must have played a role in Stud3 graduating faster than expected. This could be due to course factor namely course difficulty or student factor namely student potential or both. But it is not clear from the table that either of the two factors is reported in a transcript. This indicates that these factors are hidden in the dataset and have relevance to time-to-degree and course taking pattern of students. Thus there is a need to know how it happens. Another notable feature is that there is a big difference in the CGPA scored by Stud1 and Stud3. Whether this aspect could be improved by course taking pattern is another area not understood implying that even if a student registers in less number of courses or more number of courses in any semester, is it possible to determine the course type that contributes to the performance of a student in terms of the CGPA. Answer to this question could enable the student to identify the set of courses he or she could register in each semester in a particular order to score the most optimum CGPA. How to determine this is a major challenge.

Further literature shows that ‘course taking’ had the highest effect on academic achievement in mathematics subject among the academic subjects examined by various authors (e.g. Schmidt, 1983; Jones et al. 1986). Researchers found that, even when students’ social background and previous academic achievement were controlled, course taking was the single best predictor of the student’s achievements in the subject (Lee et al. 1998).

The study done by Prineas and Cini (2011) investigates how the emerging techniques, like learning analytics and data mining, allow the behavioural data and performance to develop the student learning not only for the future iterations but also for the current time. Online education is considered as a platform for facilitating instructor-student communication and interaction for delivering the content in education over a computer network. Learning outcomes assessment (LOA) efforts were carried out in the face-to-face or traditional classroom with the online courses in which it assessed to determine the learning outcomes of the students' performance. They empower the students to take decision of their own choice of course and their learning behaviours. Universities and colleges offer online courses through CMS (course management system), which offers virtual classroom for the students and the faculty members to interact over the course for the whole semester. With a single online course section, the faculty member can access qualitative as well as quantitative data about the student, and requires the feedback from the student to know about the outcome of the respective students' performance on the course. In an online course program, the students can make adjustments in their courses like changing the course sequences, curricula, resource allocation, academic requirements and so on. Online technologies offer the program-level feedback loop which the traditional classrooms simply cannot. In a study conducted by Mlambo (2011), learning preferences were found to be independent of both the age and gender of students. These two variables have often been studied in most literature on course taking patterns and the student's performances. However, such studies did not go far enough to determine specific course taking pattern of each student in each semester which appears to be a major gap in the literature. Studying individual student's course taking pattern in each semester and testing the overall performance of the student taking into account course taking pattern of that student in all the semesters is a complex process not attempted in the literature yet.

Similarly, many researchers have used curriculum analysis or students course transcripts (one method to analyse the course taking pattern) to understand student performance for instance Ratcliff et al. (1993) and Ronco (1996). In course transcript analysis, Ratcliff et al. (1993) linked student coursework with general learning assessments whereas Ronco (1996) used student high school transcript analysis to investigate the association between student high school course work and student college success. While curriculum analysis, especially for general education, Ewell and Paulson (2000) suggested that analysis of student course taking patterns could be used to examine prerequisites and placement policies and in exploring how course sequence works and how students act out the curriculum. However the study suffers from the fact that the course taking pattern did not take into account any student attributes such as student potential or social profile in order to determine student response to curriculum as there is a distinct possibility to link student performance (which is linked to student potential and social profile in the literature) to curriculum and hence course taking pattern.

In another study an interaction of parents' pressure for their children's taking advanced level courses and schools' flexibility in terms of rules and regulations for course placements was found to be a crucial deciding factor. Another distinct finding in a study by Ozturk (2001), relates with the assertion that minority students were less informed about course placement practices in schools and, consequently, missed out on more opportunities to take advanced level courses than the members of the majority group. Insufficient counselling services in schools, especially in the large and highly bureaucratized ones, were reported to be a major reason for this drawback. Here again it can be seen that study of course taking patterns in every semester in-depth has not been attempted by the respective researchers. A number of such studies could be found in the extant literature which have attempted to understand some kind of a course taking pattern of students and relate to them to student performance in an anecdotal manner. But all those research efforts have not envisaged any possible relationship between the course taking pattern of students in all semesters and their performance in terms of time to degree, a major limitation of the literature related to data mining. Some of them have been reviewed below as examples.

Relationship between time to degree and course taking pattern

There are many researches that have studied the significance of finishing certain gateway courses earlier in the study period in order to successfully complete the degree on time which is an example of a study that has investigated time-to degree. Adelman (2006) argued that more than 70 percent of students who successfully completed their bachelor's degree had completed the math courses in the first two years of enrolment. The outcome of the study can only be concluded as inconclusive as the research relied upon the findings that were based on data collected on the first two years of the study only and has no evidence of the findings derived having been applied to other contexts and also does not explain about the 30% of the students which is a large percentage. While on the one side the research provides support to show that particular courses could impact the performance of the student, on the other side the research does not go far to include all the factors that are part of the problem.

In another instance Cabrera et al. (2005) found that the chance of degree completion increased by around 42 percent with the completion of three math courses which is another example of time-to-degree. In both the studies there was evidence that time-to-degree is dependent on the course taking pattern of a single course that is mathematics although the studies did not go beyond a single course. Similar that too the findings were very similar in both the studies as the studies tackled mathematics only which is a serious limitation.

Vialardi et al. (2009) developed a recommendation system to recommend courses to be enrolled by students based on the number of courses enrolled currently, courses, schedules, sections, classrooms, professors and GPA of students with similar academic yield while starting the term and ending the term, grade got by students. In order to find out the courses in which the student has to be enrolled Vialardi et al. (2009) analysed data of student who had already taken the course. However this study has not studied the pattern of courses that could yield knowledge on the set of courses a student could register in, semester by semester to achieve the optimum-time to degree. The above studies while showing that some relationship could exist between time to degree and course taking pattern did not clearly explain or establish such a relationship or test it practically in any specific HEI setting or predict time-to-degree in terms of course taking pattern.

Other factors that could affect course taking pattern and student performance

There are a number of other factors that have been found in the literature to be useful in understanding about the course taking pattern of students and their performance. Such factors include student's failure patterns, constructive recommendation, curriculum structure and modification, student's academic performance, failure rate, student potential, students' social profile, course difficulty, course complexity, demographics, the completion and transfer rates of students, under reporting successes that do not result in "completion" (meaning transfer, an associate's degree or certificate), equity gaps in students pursuing completion outcomes, and the high volume of units attempted by students pursuing a completion goal. In this list of factors there exists a sub-set of factors called contextual factors (e.g. course difficulty and course complexity) (refer Appendix 25 for contextual factors) that are considered to be important for an understanding of the performance of students, knowledge about which is argued to be hidden in the student data set. While many of the factors have not been investigated by researchers yet (e.g. contextual factors (refer Appendix 25)), it is clear that investigating into a large set of factors in one PhD is neither feasible nor practical at one point in time due to the limited time available, the tremendous complexity that could arise in analysing the large set of factors. Thus in this research only an example of the contextual factors is investigated details about which are discussed next.

2.3.3 Contextual factors

Literature shows that there is lack of consensus and complete understanding of the concept of context (see the taxonomy in Appendix 16). In the extant literature it can be seen that there are different definitions or explanations given for the context for instance Kumpost (2008) who defines context information as a type of information that is directly or indirectly linked to an individual or arises from the individual’s activity. However Sokol and Chan (2013) have defined context as the cumulative history that is derived from data observations about entities (people, places, and things) and is a critical component of analytic decision process. Further, Vajirkar et al. (2003) have explained that context could consist of any circumstantial factors of the user and domain that may affect the data mining process. The various definitions and theories of context that are found in the extant literature have been tabulated in Appendix 16 which provide a strong base ground and understand the concept of context.

Many researchers have highlighted the importance of context to HEIs as well as data mining methods. For instance Baker (2008) has emphasised that issues of time, sequence, and context show significant roles in the study of educational data. It has been observed in the literature that existence of huge data of students from similar learning experiences but in very different contexts gives leverage for studying the influence of contextual factors on learning and learners (Knoblauch & Hoy, 2008). Many researchers have argued that without context, business decisions might be flawed (Brinckmann et al. 2010). By using context, organizations can derive trends, patterns, and relationships which can help an organization to make fact-based decisions. Creating data within the appropriate context delivers higher quality models which can lead to better decisions and outcomes (Brinckmann et al. 2010). Even though there is a high emphasis on context in the literature, very little has been done in the field of data mining to explain how the knowledge discovered could be characterized by contextual factors (refer Appendix 25) especially in the field of HEIs. The above mentioned arguments clearly highlight the need to investigate the importance and effect of context on data mining in HEIs. Taking the above into account this research reviews one contextual factor namely course difficulty which is discussed next.

Difficulty The course difficulty is the weighted average of the grades of every student who has taken that course or its backward equivalences. It is represented by:

$$Difficulty_c = \frac{\sum_{t \in BE_c} \sum_{j=1}^{m_t} G_{j,t} * W_t}{\sum_{t \in BE_c} W_t * m_t}$$

→ **2.1**

where *c*, current course; *t*, course equivalent to the current one; *BEc*, Set of equivalence courses for course *c*; *mt* , Total number of students in course t; *G j,t* Grade of the *jth* student in course *t*; *Wt* Number of credits of course *t*.

For instance in the following example a course ACCT311 has been arbitrarily chosen as the current course whose backward equivalence has been assumed to be ACCT301 and ACCT201 offered before. The number of students who have registered in the above courses and the grades scored by them and the maximum credit they can achieve are given in Table 2.5.

Table 2.5, Students Registered in courses with grades and credits

Term	Course name	Student name	Grade	Credits (depends on curriculum)
2014	ACCT201	Student1	3.67	3
2015	ACCT301	Student2	3	3
2015	ACCT301	Student3	4	3
2016	ACCT311	Student4	3.67	3
2016	ACCT311	Student5	4	3

Then from equation 2.1 the following can be written:

$m_{ACCT201} = 1$

$m_{ACCT301} = 2$

$m_{ACCT311} = 2$

$W_{ACCT201} = 3$

$W_{ACCT301} = 3$

$W_{ACCT311} = 3$

$G_{ACCT201, Student1} = 3.67$

$G_{ACCT301,\ Student2} = 3$

$G_{ACCT301,\ Student3} = 4$

$G_{ACCT311,\ Student4} = 3.67$

$G_{ACCT311,\ Student5} = 4$

Course difficulty of ACCT311 = $\{(G_{ACCT201,\ Student1} \text{ x } W_{ACCT201}) + (G_{ACCT301,\ Student2} \text{ x } W_{ACCT301}) + (G_{ACCT301,\ Student3} \text{ x } W_{ACCT301}) + (G_{ACCT311,\ Student4} \text{ x } W_{ACCT311}) + (G_{ACCT311,\ Student5} \text{ x } W_{ACCT311})\} / \{(W_{ACCT201} \text{ x } m_{ACCT201}) + (W_{ACCT301} \text{ x } m_{ACCT301}) + (W_{ACCT311} \text{ x } m_{ACCT311})\}$

Course difficulty of ACCT311 = {(3.67 x 3) + (3 x 3) + (4 x 3) + (3.67 x 3) + (4 x 3)} / {(1 x 3) + (3 x 2) + (3 x 2)}

= {11.01 + 9 + 12 + 11.01 + 12} / {3 +6 + 6}

= {55.02} / {15} = 3.668

Using this example course difficulty as a contextual factor has been calculated for each student for each course.

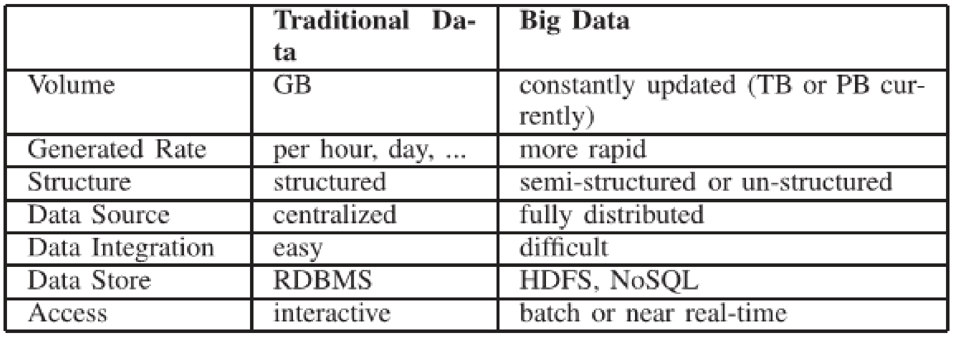
From the discussions given above it can be seen that time-to-degree, course taking patterns and contextual factors play an important role in the performance of the students in a university and hence the performance of an institution, for example in decision making process. In addition, two of the contextual factors (refer Appendix 25) that are the focus of this research have been explained. The gaps in the literature have been highlighted. It is argued that data mining as a concept could find application in understanding the association amongst the performance factors and enable the testing of more than one data mining method to determine which of the methods could be the most suitable to extract patterns and hidden knowledge and discover knowledge related to the association amongst the factors. There is a major gap in the literature which indicates that current data mining methods may not be adequate to mine educational data to discover patterns and association amongst the factors mentioned above (Baker, 2010). Thus this research embarks next to understand about data, data mining and knowledge discovery process. The discussions in this section focus only on large data whereas discussions on data mining and knowledge discovery of data mining (KDDM) process are discussed in the following sections.

2.4 Big data

Crawford et al. (2014) point out that the concept of large data or big data is not new and the challenges associated with big data continue to persist. Crawford et al. (2014) argue that only recently the concept of big data has found resonance amongst researchers although since many decades there have been discussions about big data. However contradicting this argument BSA (2015) claims that data need not be bigger always to be better as size may be one of the least important factors for many problems and small size data when analysed with the right tools can lead to keen insights. According to BSA (2015, p. 22) "*What matters most is creating robust data, securely storing the data, having access to the data, and being able to process the data — whatever its size — so that it can be utilized when and where it is needed to solve problems*". It is argued that quality of data is more important that its size and the belief larger the data size more objective it will be may not be true. These contradictory arguments pose questions regarding the very classification of data as big data. Despite such contradictions classification of datasets as traditional and large data can help as seen in the discussion in the following paragraph.

It is important to note here that in the discussions throughout this thesis the term big data has been used to represent large volume datasets and datasets are considered as big datasets when compared to traditional datasets based on the extant literature. For instance Table 2.6 shows how traditional and big datasets have been described and compared.

Table 2.6, Description and comparison of traditional and big data



	Traditional Data	**Big Data**
Volume	GB	constantly updated (TB or PB currently)
Generated Rate	per hour, day, ...	more rapid
Structure	structured	semi-structured or un-structured
Data Source	centralized	fully distributed
Data Integration	easy	difficult
Data Store	RDBMS	HDFS, NoSQL
Access	interactive	batch or near real-time

The comparison between the datasets has thrown up essential differences including the fact that traditional data is structured whereas big data is unstructured and traditional data is centralised whereas big data is fully distributed. The comparison provides the basis to treat big data differently and being the focus of this research, the discussions in this research further concentrate only on big data.

The foregoing discussions provided in Appendix 17 have provided an idea about big data, sectors where big data could find application, its uses, its advantages and the challenges that one could face while using big data. The discussion has provided a sound base on whether big data as a concept is needed in HEIs and if so how to use big data to derive benefits in terms of achieving organisational goals by successfully overcoming challenges. Turning to HEIs it can be seen that challenges faced by HEIs are formidable. For instance, Jones (2012) claims that institutions face ethical issues with regard to data collection including quality of data, privacy, security and ownership. Similarly some (e.g. Jones, 2012) argue that institutions are vested with the added responsibility of initiating steps based on the information available to them which are challenging sometimes. The problems faced by HEIs are aggravated further by challenges that concern data analysing processes required for big data analysis as implementing big data analysing processes are complex (Daniel & Butson, 2013). For example data quality is a major problem in big data analysing processes (Helfert & Ge, 2016). Warden (2011) argues that the data provided by the data sources can be messy many times and could require longer time to turn is usable than the rest part of the data analysis process combined. These arguments clearly demonstrate how data quality, one of the contextual aspects of big datasets, is significant and can be challenging. More importantly literature shows that big data analysing processes need to be dealt with in greater detail as a separate topic of research as they are complex, not easy to implement and challenging (Marcus, 2017). Recognising the importance of big data analysing processes to this research, this topic has been dealt with in greater detail in the following sections. It emerges therefore that while big data could be useful to HEIs it is also obvious that big data utilisation raises more issues in regard to its implementation, quality of data, privacy, security and ownership. These challenges unless addressed can have significant bearing on the HEIs to exploit the benefits offered by big data analysis.

In addition to the above it is necessary to ground the concept of data on theories to know that applications that are being spoken of in the literature are practically possible. Some of these aspects are discussed in Appendix 18.These aspects related to theory were required to gain a deeper knowledge about the outcomes of the analysis of big data are discussed in detail in the following section.

The foregoing discussions have presented the current level of understanding of big data and its utility to HEIs. The discussion has brought out in detail the challenges faced in adopting and using big data. The arguments show that still it is not known whether the big data analysing processes currently developed can comprehensively be applied to different contexts, particularly the HEI context. In fact the review of the literature given above shows that big data as a concept is yet to find its use in educational sector as the sector is criticized to be ill-equipped to handle the big data (Sin & Muthu, 2015) and it is not known how big data could be used in the context of HEIs characterized by such contextual aspects as data quality, privacy, security and ethical issues. Hardly any theoretical model or algorithm that could be applied exclusively to support the HEIs in analysing the student datasets and extracting knowledge hidden in those datasets to achieve better performance has been developed (Daniel, 2015). Particularly analysis of student dataset pertaining to specific student performance factors including time-to-degree, CGPA and course taking pattern of students have not been studied applying the concepts of big data and data mining that is characterized by contextual factors (refer Appendix 25) including course difficulty indicating the limitations existing in the current literature. While the currently available data analysing processes and theories are seen to be plagued by limitations, such limitations need to be overcome as those limitations deter users of big data and researchers from exploiting the advantages offered by the analysis of big data. Some of the key word definitions that are used in the different system of study and that will be referenced to in this research (See Appendix 15).

While data types and some examples of data analysis in terms of algorithms have been discussed in Appendix 21, literature shows that methods used to analyse data are many and there is no clarity on which one of the methods is the most suitable for a particular purpose. Contradictory arguments show that users are left with very few options to decide on the most suitable method for data analysis for a particular purpose (Kumar et al. 2014). These challenges force the researchers to use more than one type of data analysis method to analyse any dataset to decide on which one of the methods is most suitable for an application and what parameters of the data analysis must be considered analysing a dataset. This is an important gap in the literature (Han & Kamber, 2010; Morales et al.2016). It appears each time a research outcome is produced reporting comparison of the performance of two or more data analysing methods, useful knowledge is contributed to the growing body of data analysis (Bandaru et al.2017). This gap is being addressed in this research.

While this challenge remains and researchers are still talking about the need for continuous development in the areas related to data analysis one of the major shifts that has taken place in the analysis of data is the introduction of data processing from data analysis. One of the reasons for this is that the volume of data is growing exponentially and such growth is argued to be difficult to be analysed or processed or treated by humans or manual applications (García et al. 2016). With advancing technology providing facilities to store large data and connectivity to those data, other challenges arose namely the difficulty to obtain organized knowledge and information due to the huge increase in volume of data and difficulties in understanding and extracting knowledge and information hidden in the data. These aspects led to the advancement of a new branch of science called data science, alternatively called data mining (Aggarwal, 2015) .While data mining as a science is growing rapidly (see Appendix 22) this branch of science is yet to be exploited to its full potential (Kirchdoerfer & Ortiz, 2017; Baker, 2010). Particularly in the HEIs, its potential is being realized only now. In order to know what data mining is, how it is important to HEIs, what can be achieved using data mining and what are unknown areas that require study the following section reviews the literature about data mining. This is another gap in the literature (Liebowitz, 2017) and is being addressed in this research.

The above discussions show that early efforts to analyse data necessitated continuous development to enhance data analysis methods as the types of data, data storage facilities, access to data repositories, technologies used to analyse data and knowledge related to data analysis continuously changed over a period of time and in rapid succession. One important aspect that was unveiled during this evolutionary process was the discovery of hidden knowledge in large datasets that could not be extracted or understood using normal data analysing processes described above. A data science called data mining was needed as the hidden knowledge was not in the normal observable form. For instance, within large datasets there hide patterns of data that convey information and knowledge about phenomena. These patterns could be uncovered by the new data science called data mining. Thus the next section discusses about the science of data mining, what is known about it and what gaps exist in the literature that need to be addressed.

2.5 Data mining and Educational Data Mining (EDM)

2.5.1 Data Mining

That data mining could be used to discover course taking patterns is supported by some evidence in the literature. For instance Zhai et al. (2001) used Cluster Analyses to discover student course taking patterns followed by a Discriminant Function Analysis to test the validity of the cluster grouping. Chi-Square tests of Independence and correlation were used to detect if there were any associations between student course taking patterns and retention or between student course taking patterns and their majors. Similarly Bahr (2010) conducted a cluster analysis for the California Community Colleges Chancellor's Office and examined the course taking behaviour of first time students over an eight year period. The research revealed a number of interesting issues, including under reporting successes that do not result in "completion" (meaning transfer, an associate's degree or certificate), equity gaps in students pursuing completion outcomes, and the high volume of units attempted by students pursuing a completion goal.

Furthermore Kovacic (2010) presented a case study on educational data mining to identify the extent to which enrollment data can be used to predict student's success. The algorithms CHAID and CART were applied on student enrollment data of Open Polytechnic of New Zealand to get two decision trees classifying successful and unsuccessful students. The accuracy obtained with CHAID and CART was 59.4 and 60.5 respectively. In the same vein it can be seen that Vialardi et al. (2011) presented the rationale behind the design of a recommender system to support the enrollment process using the students' academic performance record. In this study two attributes were introduced. The first attribute estimates the inherent difficulty of a given course. The second attribute, named potential, is a measure of the competence of a student for a given course based on the grades obtained in related courses. Data was mined using C4.5, KNN (K-nearest neighbor), Naïve Bayes, Bagging and Boosting, and a set of experiments was developed in order to determine the best algorithm for this application domain. Results indicate that Bagging is the best method regarding predictive accuracy. The system was tested during the enrollment process.

Furthermore, Zeidenberg (2011) used clustering techniques to understand the course taking pattern of community college students in order to determine the programs of study. The research also examined the demographics and the completion and transfer rates of the students within each cluster, in order to get an idea of what types of students were in each program of study and how successful they seemed to be in college. The research found substantial variation on these dimensions as well as on the extent to which students' programs were either concentrated in a single subject or spread across several subjects. In another instance Oladipupo and Oyelade (2010) used association rule data mining technique to identify student's failure patterns by analysing a total of 30 courses pertaining to 100 and 200 levels. This study focused on constructive recommendation, curriculum structure and modification in order to improve student's academic performance and trim down failure rate.

The above discussion shows two aspects. One aspect is that data mining as a tool can be used to study different attributes of the students as well as their performance. The second one is that there are a number of factors that have not yet been fully studied by mining the student dataset for instance student potential, students' social profile, course difficulty or course complexity. Lack of knowledge about the possibility of using data mining techniques to unearth useful knowledge hidden in the student data set and extract course taking patterns of students to determine their performance is a major gap in the literature. Such a gap could hinder HEIs from effectively using data mining techniques to extract more important details about student performance that could support HEIs in decision making hither to no known. Thus while literature shows that data mining as a technique has not been widely used by HEIs to improve decision making and student performance, it is also clear lack of usage of hidden knowledge may make the decision making process incomplete in HEIs. Thus the next section deals with components of data mining.

Components of data mining

The major components of the architecture of a typical data mining system could be visualised as shown in Figure 2.2.

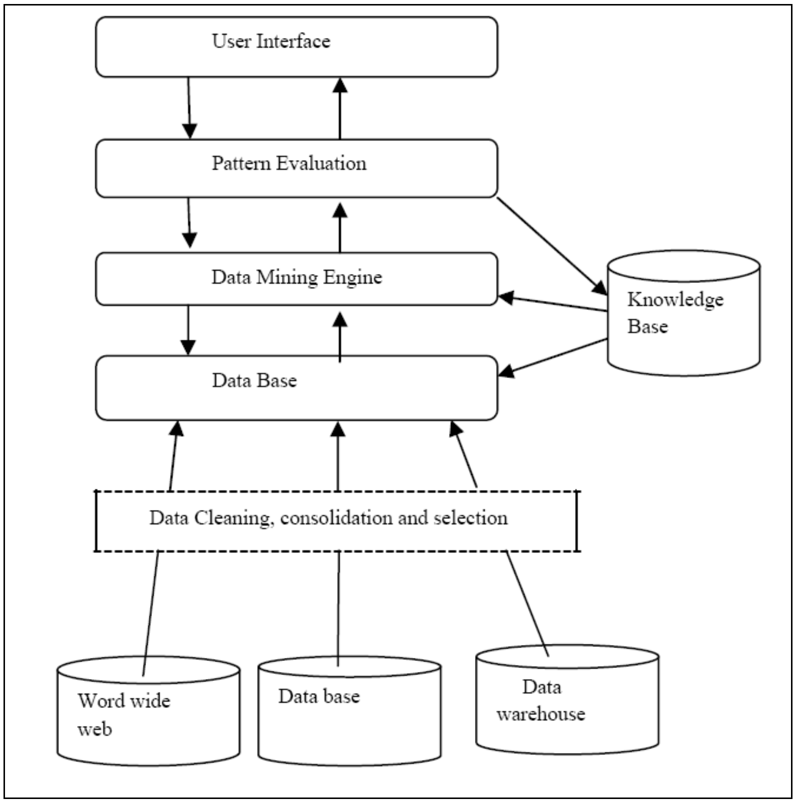


Figure 2.2 , Data mining architecture (*Source:* Han and Kamber, 2011)

The various blocks represented in Figure 2.2 are described below to gain an understanding of what a data mining process would be. This description is important to know the current state of the knowledge related to data mining, the advantages that are expected to accrue when data mining process is used, and the limitations of using data mining process and gaps that could exist in knowledge related to data mining.

1. Database, data warehouse or other information repository: The data present in databases or data warehouses or websites or spreadsheets, or other kinds of information repositories. The data has to be processed using data cleaning and data integration techniques.

2. Database: The database is responsible for fetching the appropriate data, based on the user's data mining request.

3. Knowledge base: This is the domain knowledge that is used to guide the search, or evaluate the interestingness of resulting patterns. Such knowledge can include concept

hierarchies, used to organize attributes or attribute values into different levels of abstraction. Knowledge such as user beliefs, which can be used to assess a pattern.

4. Data mining engine: This consists of a set of functional modules for data mining tasks like classification, association, characterization, cluster analysis, and evolution and deviation analysis.

5. Pattern evaluation module: This module uses interestingness measures and communicates with the data mining modules in order to focus the search towards interesting patterns.

6. User interface: This module interacts between users and the data mining system, permitting the user to communicate with the system by giving a data mining query or task, and produce patterns or data mining results. In addition, this allows the user to glance the database and data warehouse schemas or data structures, evaluate mined patterns, and visualize the patterns in different forms.

Data mining refers to the process of discovering knowledge from large volume of data which cannot be usually discovered by other methods including use of manual process,query processing,OLAP (Online Analytical Processing) and machine language. The significance of data mining lies in finding interesting, formerly unknown, significant and potentially useful pattern or knowledge from voluminous data. Such pattern or knowledge is found to help in improving the decision making process in businesses and the efficiency and productivity of a variety of activities(Fayyad,1996).

Over the last few years, spurred by the generation of large volume of data called as “Big Data”, there is a growing tendency to use Data Mining across the world. Such data is generated by many activities including transactions occuring in businesses, use of social media by a large population, logging of data by users of computer systems, and multimedia files created and used by various segments of users. The term Big Data refers to not only the volume of data but the velocity with which the data gets accumulated and the variety that characterises the data. Although data gets accumulated and data mining provides an oppurtunity to exploit the immense use of such data, still researchers (Khan,2013; Solanki, 2013) argue that there is no unified method of data mining that is applicable to all types of data or context or purpose. For instance the data mining techniques that could be used to discover knowledge in a dataset pertaining to medicine is different from the data mining technique that could be used to discover knowledge in a dataset pertaining to education as the hidden knowledge in these two datasets are expected to be different.

Further, datamining has been applied to different fields like health care, medical science, banking and finance, retail industry and education. However this research is concerned with data mining in Education called as Educational Data Mining(EDM). EDM has attracted many researchers as education data appears to have hidden knowledge hitherto not discovered making it necessary to conduct research to discover such knowledge. For example some researchers have used EDM for discovering knowledge and investigate how student performance, student success, student retention and the like could be enhanced although such investigations are not generalisable to all contexts. Considering the various characteristics associated with datasets, resesarchers have called for a deeper investigation into EDM inorder to gain useful insight into the hidden knowledge which has not been uncovered yet. Thus there is a need to identify the areas in the field of education that have not yet been investigated by researchers using EDM and a review of the relevant literature becomes imminent. One such area that has been not been fully investigated by researchers is the hidden knowledge in the dataset pertaining to students in Higher Education Institutions (HEIs) corresponding to credit hour system like the ones followed in HEIs in USA and other countries. However before reviewing the relevant literature on EDM, a wider discussion on the concept of DM will be useful as EDM is largely a derivative of DM and a deeper understanding of DM is expected to provide a strong basis to examine EDM which follows next. To begin with the taxonomy of DM is provided to gain knowledge on the various data mining applications that have been developed and the diverse areas in which those applications have been applied. This knowledge is critical to understand whether a particular application fits all situations or there are challenges that need to be addressed.

Taxonomy on Data Mining Applications

Due to its diverse application, data mining provides solutions to questions that a decision maker has earlier not thought of asking. Data mining is a multidisciplinary area in which several computing paradigms meet for instance decision tree, rule induction, artificial neural networks, instance-based learning, Bayesian learning, logic programming, statistical algorithms used. Further, some of the most useful data mining tasks and methods identified in the literature are statistics, visualization, clustering, classification, association rule mining, sequential pattern mining, text mining. In addition, datamining has been applied in many sectors like health care,medical science, banking and finance, retail industry and education. Literature shows that there has been consistent increase in the interest shown towards application and research on data mining over the past few years, for instance Wu (2010) who argued that there is considerable increase in interest in data mining applications/research. In order to understand the current interest on the topic of data mining the publications produced between 1995 and 2010 is provided in Figure 2.3 followed by Tables 2.7 and 2.8 which indicate the widely used methods in data mining as well as data mining types, application areas and other relevant details.

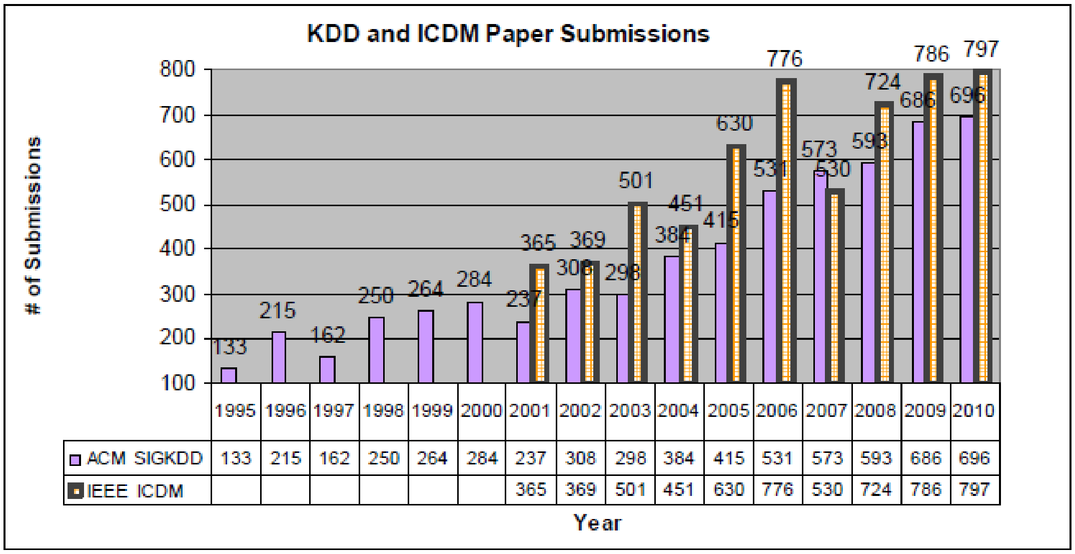


Figure 2.3: KDD and ICDM Paper Submissions

Table 2.7 : Widely used DataMining algorithms

	Authors	**Challenges**
Classification	– #1. **C4.5**: Quinlan, J. R. 1993. C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers Inc. – #2. **CART**: L. Breiman, J. Friedman, R. Olshen, and C. Stone. Classification and Regression Trees. Wadsworth, Belmont, CA, 1984. – #3. ***K* Nearest Neighbours (*k*NN):** Hastie, T. and Tibshirani, R. 1996. Discriminant Adaptive Nearest Neighbor Classification. IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI). 18, 6 (Jun. 1996), 607-616. – #4. **Naive Bayes**: Hand, D.J., Yu, K., 2001. Idiot's Bayes: Not So Stupid After All? Internat. Statist.	1.Developing a Unifying Theory of Data Mining 2.Scaling Up for High Dimensional Data/High Speed Streams 3.Mining Sequence Data and Time Series Data 4.Mining Complex Knowledge from Complex Data 5.Data Mining in a Graph Structured Data 6.Distributed Data Mining and Mining Multi-agent Data 7.Data Mining for Biological and Environmental Problems 8.Data-Mining-Process Related Problems 9.Security, Privacy and Data Integrity 10.Dealing with Non-static,

	Rev. 69, 385-398.	Unbalanced and Cost-sensitive Data
Statistical Learning	– #5. **SVM:** Vapnik, V. N. 1995. The Nature of Statistical Learning Theory. Springer-Verlag New York, Inc. – #6. **EM**: McLachlan, G. and Peel, D. (2000). Finite Mixture Models. J. Wiley, New York.	
Association Analysis	– #7. **Apriori**: Rakesh Agrawal and Ramakrishnan Srikant. Fast Algorithms for Mining Association Rules. In VLDB '94. – #8. **FP-Tree**: Han, J., Pei, J., and Yin, Y. 2000. Mining frequent patterns without candidate generation. In SIGMOD '00.	
Link Mining	– #9. **PageRank**: Brin, S. and Page, L. 1998. The anatomy of a large-scale hypertextual Web search engine. In WWW-7, 1998. – #10. **HITS**: Kleinberg, J. M. 1998. Authoritative sources in a hyperlinked environment. In Proceedings of the Ninth Annual ACM-SIAM Symposium on Discrete Algorithms, 1998.	
Clustering	– #11. ***K*-Means**: MacQueen, J. B., Some methods for classification and analysis of multivariate observations, in Proc. 5th Berkeley Symp. Mathematical Statistics and Probability, 1967. – #12. **BIRCH**: Zhang, T., Ramakrishnan, R., and Livny, M. 1996. BIRCH: an efficient data	

	clustering method for very large databases. In SIGMOD '96.	
Bagging and Boosting	– #13. **AdaBoost**: Freund, Y. and Schapire, R. E. 1997. A decision-theoretic generalization of on-line learning and an application to boosting. J. Comput. Syst. Sci. 55, 1 (Aug. 1997), 119-139.	
Sequential Patterns	– #14. **GSP**: Srikant, R. and Agrawal, R. 1996. Mining Sequential Patterns: Generalizations and Performance Improvements. In Proceedings of the 5th International Conference on Extending Database Technology, 1996. – #15. **PrefixSpan**: J. Pei, J. Han, B. Mortazavi-Asl, H. Pinto, Q. Chen, U. Dayal and M-C. Hsu. PrefixSpan: Mining Sequential Patterns Efficiently by Prefix-Projected Pattern Growth. In ICDE '01.	
Integrated Mining	#16. **CBA**: Liu, B., Hsu, W. and Ma, Y. M. Integrating classification and association rule mining. KDD- 98.	
Rough Sets	– #17. **Finding reduct**: Zdzislaw Pawlak, Rough Sets: Theoretical Aspects of Reasoning about Data, Kluwer Academic Publishers, Norwell, MA, 1992.	
Graph Mining	#18. **gSpan**: Yan, X. and Han, J. 2002. gSpan: Graph-Based Substructure Pattern Mining. In ICDM '02.	

Table 2.8: DataMining Types and application areas

Data Mining Type	Application Areas	Data Formats	Data Mining Techniques/Algorithms
Hypermedia data mining	Internet and Intranet	Hyper Text Data	Classification and Clustering Techniques
Ubiquitous data mining	Mobile phones, PDA, Digital cam etc	Ubiquitous data	Traditional data mining techniques drawn from statistics and machine learning
Multimedia data mining	Audio/video	Multimedia data	Rule based decision tree classification algorithms
Spatial data mining	Network, remote sensing and SIS	Spatial data	Spatial clustering techniques, Spatial OLAP
Time series data mining	Business and financial applications	Time series data	Rule induction algorithms

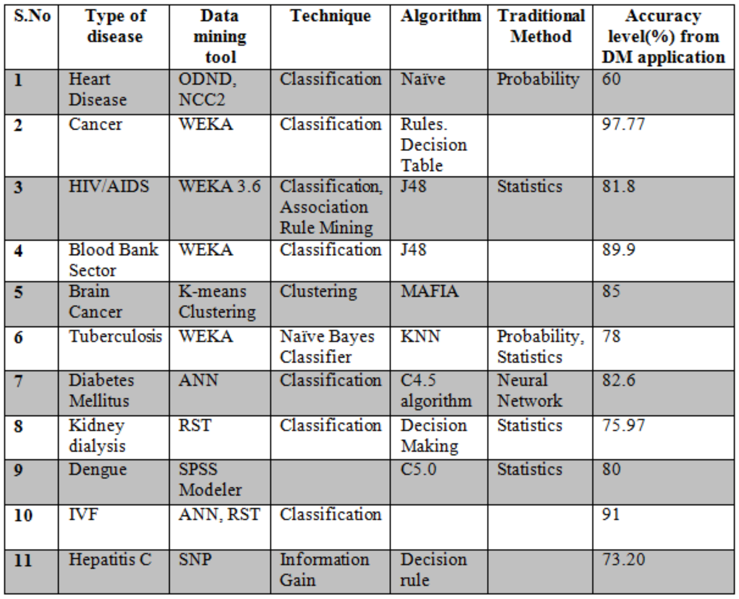
From Figure 2.3 it can be seen that the number of publications have been steadily increasing indicating that data mining as a technology has become popular and useful in the modern business environment. Tables 2.7 and 2.8 show how the data mining field is rapidly advancing with many authors investigating into the different methods of data mining using different types of data and algorithms.

The table below provides a brief taxonomy on data mining applications to various industries listing the significance, challenges and future directions. Although the information given in table is not exhaustive still the information provided could be useful to understand the current level of research taking place in datamining. The researcher acknowledges that the information gathered is limited and there is scope for adding more details.

Table 2.9: Taxonomy on data mining applications to various industries listing the significance, challenges and future directions

Field	**Researchers**	**Significance**	**Issues/Challenges**	**Future Directions**
Health care Industry	Hnin Wint Khaing(2011) Khemphila & Boonjing(2011)	1.Evidence based diagnosis 2.Early detection or prediction or prevention of diseases. 3. Non invasive diagnosis 4.Detect Insurance fraud 5.Predict Side effects of medicines	1.Privacy and ethical issues. 2. Lack of confidence on data mining	1. Text Mining(prescriptions) 2.Image Mining(Scans)
Retail Industry		1.Identify customer buying behaviors 2.Discover customer shopping patterns and trends 3.Improve the quality of customer service 4.Achieve better customer retention and satisfaction 5.Enhance goods consumption ratios 6.Design more effective goods transportation and distribution policies		

<table>
<tr>
<td>Finance</td>
<td>Fanning and Cogger 1998; Feroz et al. 2000; Green and Choi 1997;
<table>
<tr><th>Data</th><th>Researchers</th></tr>
<tr><td>Financial statement disclosures</td><td>Fanning and Cogger 1998; Feroz et al. 2000; Green and Choi 1997</td></tr>
<tr><td>Core financial statement data points and ratios</td><td>Ravisankar et al. (2011)</td></tr>
<tr><td>Auditing-process mining</td><td>Jans et al. 2013; Van der Aalst 2011</td></tr>
<tr><td>Compliance and risk management</td><td>Caron et al. 2013</td></tr>
<tr><td>Journal entries</td><td>Debreceny and Gray (2010)</td></tr>
</table>
</td>
<td>1.Fraud detection
2.Auditing</td>
<td>Complicated unstructured data</td>
<td>1.Text Mining</td>
</tr>
<tr>
<td>Telecom</td>
<td>Ravisankar et al. (2011)</td>
<td>1.Identify telecommunication patterns
2.Catch fraudulent activities
3.Make better use of resources
4.Improve the quality of service</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Medical Science</td>
<td>Jans et al. 2013; Van der Aalst 2011</td>
<td>1.Analysis of DNA sequences
2.Patient Profiling
3.History Generation</td>
<td>Complex medical data</td>
<td>Image Mining</td>
</tr>
</table>



S.No	Type of disease	Data mining tool	Technique	Algorithm	Traditional Method	Accuracy level(%) from DM application
1	Heart Disease	ODND, NCC2	Classification	Naïve	Probability	60
2	Cancer	WEKA	Classification	Rules. Decision Table		97.77
3	HIV/AIDS	WEKA 3.6	Classification, Association Rule Mining	J48	Statistics	81.8
4	Blood Bank Sector	WEKA	Classification	J48		89.9
5	Brain Cancer	K-means Clustering	Clustering	MAFIA		85
6	Tuberculosis	WEKA	Naïve Bayes Classifier	KNN	Probability, Statistics	78
7	Diabetes Mellitus	ANN	Classification	C4.5 algorithm	Neural Network	82.6
8	Kidney dialysis	RST	Classification	Decision Making	Statistics	75.97
9	Dengue	SPSS Modeler		C5.0	Statistics	80
10	IVF	ANN, RST	Classification			91
11	Hepatitis C	SNP	Information Gain	Decision rule		73.20

The advancements taking place in the field of data mining while has been useful to many industries (see Table 2.9), literature shows that as a technology, data mining is still evolving with many arguing that more needs to be done if the technology and its potential is to be exploited. For instance, it was argued that the process of using data mining varies according to the context and data size. This creates a situation of confusion as some of the algorithms used in data mining need to still changed or modified or new ones developed to takle various types of business requirements. That is to say the data mining method used in the field of medicine may not be suitable for the one used in education.

This in turn imposes a burden on the community using the data mining technique to explore how the data mining technique could be used fruitfully in specific fields an argument that could be extended to the field of education. Unless specific data mining techniques are developed it may not be possible to use the technology straightaway in specific contexts. Thus the use of data mining technique in the field of education termed as education data mining (EDM) gains importance. Prior to discussing EDM it is necessary to critically review some of the widely used data mining methods to gain knowledge on which of those methods either in isolation or in combination will be useful in dicovering knowledge about course taking patterns that are under investigation.

2.5.2 Review of Data Mining Techniques

2.5.2.1 Clustering

Clustering is an unsupervised data mining technique used for finding out groups and hidden patterns and distributions in the data. In this method class labels are not defined before the mining process (Richard, 2017). It partitions a given dataset into groups (clusters) so that the cluster data points are comparable to each other and unrelated to data points outside the cluster. The aim of clustering is to discover groups with data points having similar characteristics (Richard, 2017). For instance in a group of university students it is possible to use clustering to find groups of students with similar attributes as age, CGPA, gender, prior study and others (family background, family income). Of the various clustering algorithms that have been developed, widely used ones include k-means clustering, EM clustering, Fuzzy c-means clustering, hierarchical clustering Gaussian (EM) clustering algorithm and quality threshold clustering algorithm (Verma et al. 2016). While each type of algorithm has specific application area, still many times datasets are mined using more than one clustering algorithm to compare and find the most suitable for a particular application. This is tedious a process and requires the interaction between different entities involved in decision making. Additionally there is no unified method or generalised clustering technique that has been developed to overcome this problem in the literature. Usually algorithms are tested for their accuracy, speed, optimization and improvement in better prediction. Literature shows that in order to determine which algorithm suits the best in dealing with a particular task, it may be necessary to make modifications to the algorithm itself. This is a greater challenge. Furthermore, it must be noted that each particular data mining task may have to be mined with more than one algorithm to find the most useful one in terms of its performance which is usually a challenge and a gap in the literature (Kavakiotis, 2017; L'Heureux, 2017).

While review of the entire set of algorithms is beyond the scope of this research, from amongst the available algorithms the researcher has chosen to focus on k-means and EM clustering algorithms as examples a review of which will give an idea about the challenge that lies ahead when those algorithms are used in data mining tasks. The reason for choosing k-means clustering algorithm is that it has been widely used in the literature (Du et al. 2016) and hence chosen for study in this research. Similarly EM clustering algorithm was chosen as a second algorithm as it differs with k-means clustering algorithm in its functional aspects. Comparing the performance of the two provides knowledge about which one of the two can be chosen for implementing EDM based on performance. The process of comparison appears to be the same when multiple algorithms are used to mine data. However there is no specific rule found in the literature that informs which algorithm should be used in a particular project of data mining which is a serious gap that allows researchers to randomly pick an algorithm to execute data mining tasks leading to avoidable problems like complications in using the algorithm or modification of the algorithm itself. This aspect requires investigation. Especially when one looks at the problem of discovering course taking pattern to find out the optimum time to degree, application of these two algorithms could give a clue on how to identify the direction in which the researcher should proceed, meaning how to choose the right algorithm. Thus the next two section review critically the two algorithms namely K-means algorithm and EM clustering algorithm to see whether they are fit to be used in determining the time to degree using course taking pattern of student and what limitations could exist.

(A) K-means algorithm

This algorithm is useful in allocating instances to clusters. In the context of HEIs, instances refer to student records. Clusters are students who have similar features. Further k-means algorithm can deal with numeric values only (Du et al. 2016) (example CGPA measured in numbers) and not with nominal values (example gender = male/female). Every cluster is defined by two statistical measures namely centroid and Euclidean distance. Centroid is defined as center of the cluster. Euclidean distance is the distance between centroids of any two clusters (Neagu et al. 2017). Cases are fitted by the algorithm to different clusters based on features of students. Each cluster is called a profile. The limitation of k-means clustering algorithm lies in its requirement for the data miner to prescribe the k- value which determines the number of clusters (Hancer & Karaboga, 2017). Such a requirement necessitates that the data miner is very knowledgeable in choosing the k value and could lead to a situation where either higher or lower number of clusters than required are formed by the algorithm (Hancer & Karaboga, 2017). Such a situation could lead to lack of optimization in clustering thereby affecting the computational aspects of data mining process, for instance burdening the computation process or delaying the data mining process. Finding the optimum cluster is a major challenge.

(B) EM Clustering

EM clustering algorithm can be used to generate clusters that explain the probability of including a case in a cluster (Singh et al. 2016). EM algorithm includes the analysis of numeric and nominal values assigned to the features (Singh et al. 2016). The main limitation of EM algorithm is that "it can converge to or get trapped for many iterations at a local maximum, leading to failure to reach the global maximum and resulting in an inferior clustering solution" (O'Hagan & White, 2016; p. 2). It is also sensitive to initial parameter selection (Aggarwal & Reddy, 2013).

2.5.2.2 Association rule learning

It is a popular and well researched method for discovering hidden pattern between attributes in any dataset. It is intended to identify strong rules discovered in databases using different measures of interestingness. For instance with regard to student drop out literature shows that some factors that are attributes of students could be associated to the student drop out using association rules (Kumar et al. 2017). Association rule generation is usually split into two separate steps namely minimum support applied to find all frequent item sets in a database and use of those frequent item sets and the minimum confidence constraint to form rules (Agrawal et al.1993).

The original definition of Association rule mining given by Agrawal et al. (1993, pp. 207-216) is reproduced to explain the basic algorithm discussed in this research:

Let I = {i_1, $i_{2,\ldots,}i_n$ } be a set of n binary attributes called *items*.

Let D= {t$D = \{t_1, t_2, \ldots, t_m\}$ be a set of transactions called the *database*.

Each *transaction* in D has a unique transaction ID and contains a subset of the items in I.

A *rule* is defined as an implication of the form:

$X \Rightarrow Y$

Where $X, Y \subseteq I$ and $X \cap Y = \emptyset$.

Every rule is composed by two different set of items, also known as *item sets*, X and Y, where X is called *antecedent* or left-hand-side (LHS) and Y *consequent* or right-hand-side (RHS).

Let X an item-set, $X \Rightarrow Y$ an association rule and T a set of transactions of a given database.

Support

The support value of X with respect to T is defined as the proportion of transactions in the database which contains the item-set X.

Confidence

The *confidence* value of a rule, $X \Rightarrow Y$, with respect to a set of transactions T, is the proportion the transactions that contains X which also contains Y.

Confidence is defined as:

$$\mathrm{conf}(X \Rightarrow Y) = \mathrm{supp}(X \cup Y)/\mathrm{supp}(X).$$

Thus confidence can be interpreted as an estimate of the conditional probability $P(E_Y|E_X)$, the probability of finding the RHS of the rule in transactions under the condition that these transactions also contain the LHS.

Lift

The *lift* of a rule is defined as:

$$\mathrm{lift}(X \Rightarrow Y) = \frac{\mathrm{supp}(X \cup Y)}{\mathrm{supp}(X) \times \mathrm{supp}(Y)}$$

or the ratio of the observed support to that expected if X and Y were independent.

Conviction

The *conviction* of a rule is defined as $\mathrm{conv}(X \Rightarrow Y) = \frac{1 - \mathrm{supp}(Y)}{1 - \mathrm{conf}(X \Rightarrow Y)}$.

and can be interpreted as the ratio of the expected frequency that X occurs without Y (that is to say, the frequency that the rule makes an incorrect prediction) if X and Y were independent divided by the observed frequency of incorrect predictions. It must be noted that this basic association rule algorithm developed by Agrawal et al. (1993) has been further expanded to include derivatives for instance *a priori* and sequential patterns. Amongst these *apriori* is considered to be the best-known algorithm to mine association rules (Agrawal & Srikanth, 1994; Khobragade et al. 2015).Rules with 100% confidence were considered as interesting. It uses a breadth-first search strategy to count the support of item sets and a candidate generation function which exploits the downward closure property of support. Association rules need to be carefully used as there are inherent limitations which include obtaining non-interesting rules, high number of discovered rules leading to ambiguity in identifying rules that are useful to the task under mining and low performance of algorithm for instance efficiency. An example of association rule learning or mining is demonstrated using *a priori* algorithm.

***A priori* algorithm pseudo code**

```
procedure Apriori (T, min support) { //T is the database and min support is the minimum support
L1= {frequent items};
for (k= 2; Lk-1 !=∅; k++) {
 Ck= candidates generated from Lk-1
//that iscartesian product Lk-1 x Lk-1 and eliminating any k-1 size item set that is not
//frequent
 for each transaction t in database do{
 #increment the count of all candidates in Ck that are contained in t
 Lk = candidates in Ck with min support
 }//end for each
 }//end for
return ∪ ;
}
```

2.5.2.3 Linear Regression

Regression analysis is a statistical tool for the investigation of relationships between variables. It is a predictive data mining method (Sowan & Qattous, 2017). Usually, the investigator seeks to ascertain the cause and effect relationship between the dependent and independent variables. For instance with regard to universities it is possible to relate student satisfaction with quality of delivery of education (Khoo et al. 2017). The investigator gathers data about the variables under study and employs regression to estimate the effect on the dependent variable caused by the independent variables. The investigator also typically assesses the statistical significance of the estimated relationships, that is, the degree of confidence that the true relationship is close to the estimated relationship. Data is a numerical variable. The main drawbacks of regression include that the assumption of linear relationship between variables and lack of cause and effect relationship between variables (Darlington & Hayes, 2016).

2.5.2.4 Classification

Classification is a task in which the main objective is to assign a predefined label or class to a record using a set of known attributes (Hamsa et al. 2016). It is based on machine learning. Classification is often referred to as supervised learning because the classes are determined before examining the data. Each data instance consists of two parts, a set of predictor attribute values and a goal attribute value. Example of classification in the context of higher education could be the prediction of future student dropouts (Kumar et al. 2017). There are a number of classification algorithms that are used in different data mining projects by researchers, although the utility of those algorithms is highly restricted to specific domains. That is to say all the algorithms that can be used to classify in general cannot be uniformly used to mine data in different fields implying that finding the most appropriate algorithm for a particular data mining purpose is challenging (Kavakiotis, 2017;Cheng,2017). Some of the widely used classification algorithms in data mining include Naive Bayes Classifier, Nearest Neighbor Classifier, Decision Tree classifiers, Support Vector Machines and Genetic Algorithm (Allahyari, et al. 2017; Hamsa et al. 2016). While literature is replete with discussions on those algorithms, in this research genetic algorithm is the only algorithm that has been reviewed as an example of an algorithms that could be applied to this research. Discussing all the algorithms in one research project will be beyond the scope of that research due to constraints such as time, resource and complexity. Hence the next section reviews the genetic algorithm critically.

2.5.2.5 Genetic Algorithm

Genetic Algorithm (GA) is grounded on the Darwin's theory of natural evolution. It is based on the concept of survival of the fittest. The schema theorem of Holland, 1975 was the first explanation on how GA works.GA denotes the results in the form of chromosomes and evaluates the fitness of the chromosomes. Using cross over operator the more fit solutions are selected for reproduction. To maintain the diversity of the population the mutation operator is used. The less fit chromosomes are replaced with more fit chromosomes and the same is continued till optimal solution is reached based on a pre-set criteria.

For a set S, the member is called individual. In GA an individual is recognized with chromosome. The Information fixed in chromosome is called genotype. The values of source task variables corresponding to genotype are called Phenotype. Phenotype is nothing but decoded genotype. Chromosomes are binary string of finite length in simple genetic algorithm. Gene is a bit of this string. Allele is value of gene, 0 or 1. Population is finite set of individuals.

Objective function of optimization problem is called fitness function. Fitness of individual is value of fitness function on phenotype corresponding individual. Fitness of population is aggregative characteristic of fitness of individuals. Fitness of best individual or average fitness of individuals is commonly used as population fitness in genetic algorithms. In the process of evolution one population is replaced by another and so on, thus we select individuals with best fitness. So in the mean each next generation (population) is fitter than it predecessors. Genetic algorithm produces maximal fitness population, so it solve maximization problem. Minimization problem obviously reduced to maximization problem.

In simple genetic algorithm size of population n and binary string length m is fixed and don't changes in process of evolution. We can write basic structure of simple genetic algorithm in the following way:

```
Compute initial population;
WHILE stopping condition not fulfilled DO BEGIN
select individuals for reproduction;
create off springs by crossing individuals;
eventually mutate some individuals;
compute new generation;
END
```

As is evident from discussions given above about GA, the transition from one generation to the next is seen to consist of three basic components. They are:
Selection: Mechanism for selecting individuals for reproduction according to their fitness.

Crossover: Method of merging the genetic information of two individuals. In many respects the effectiveness of crossover is dependent on coding.

Mutation: In real evolution, the genetic material can by changed randomly by erroneous reproduction or other deformations of genes, e.g. by gamma radiation. In GA, mutation is realized as a random deformation of binary strings with a certain probability.

These components are called genetic operators which are described below.
Compared to conventional continuous optimization methods, such as gradient descent methods, following significant differences can be observed in GA:

1. Genetic algorithms manipulate coded versions of the problem parameters instead of the parameters themselves, i.e. the search space is S instead of X itself. So, genetic algorithm finds solution approximately.
2. While almost all conventional methods search from a single point, genetic algorithm always operates on a whole population of points (strings-individuals). It improves the robustness of algorithm and reduces the risk of becoming trapped in a local stationary point.
3. Normal genetic algorithms do not use any auxiliary information about the objective function value such as derivatives. Therefore, they can be applied to any kind of continuous or discrete optimization problem.
4. Genetic algorithms use probabilistic transition operators while conventional methods for continuous optimization apply deterministic transition operators. More specifically, the way a new generation is computed from the actual one has some random.

Disadvantages

1. There is no guarantee that a Genetic algorithm will lead to a global optimum solution especially in cases where the populations have many subjects.
2. GA cannot assure constant optimization response times.

The above discussions clearly show that choice and application of data mining techniques or algorithms are not straight forward and require data mining knowledge as well as contextual knowledge (refer Appendix 25). As mentioned earlier it can be seen that choice of the technique or the algorithm, interpreting and understanding the knowledge discovered using the techniques or algorithms and producing and evaluating the prediction models are challenges that need to be addressed in every context. Thus in order to investigate how EDM can be implemented in HEIs using the data mining concepts discussed above there is a need to critically review EDM which follows next.

2.5.3 What is EDM?

EDM is a part of the field of data mining and is defined as the use of data mining techniques on educational data (Kaur et al. 2015). Prior to reviewing EDM in-depth, it is useful to understand the nature of EDM as a derivative of DM. From Figure 2.4 it can be seen that EDM is essentially dependent on DM concepts.

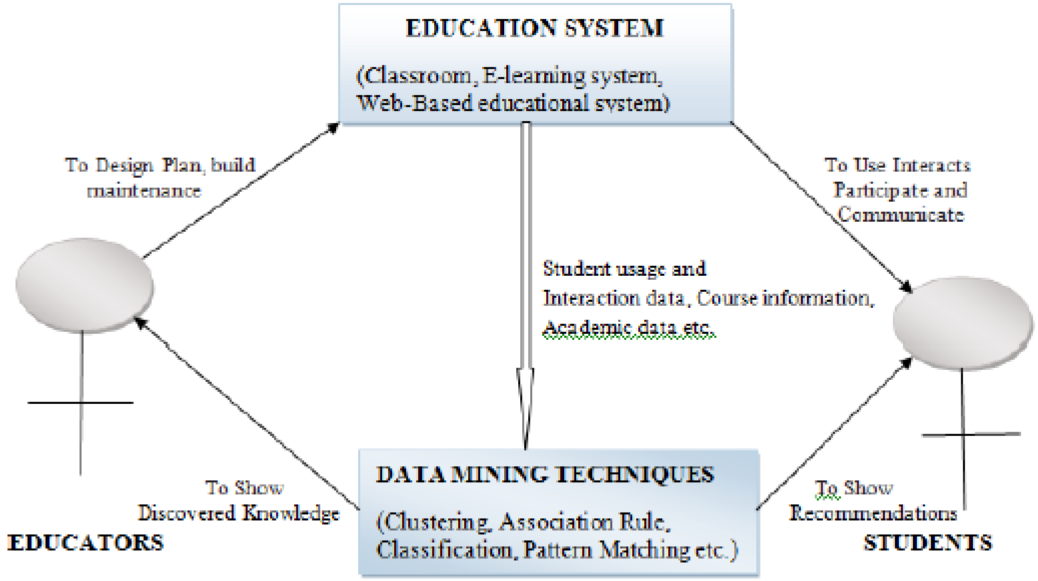

EDUCATION SYSTEM
(Classroom, E-learning system, Web-Based educational system)
To Design Plan, build maintenance
To Use Interacts Participate and Communicate
Student usage and Interaction data, Course information, Academic data etc.
To Show Discovered Knowledge
EDUCATORS
DATA MINING TECHNIQUES
(Clustering, Association Rule, Classification, Pattern Matching etc.)
To Show Recommendations
STUDENTS

Figure 2.4, Representation of EDM as a derivative of DM (Source: Kashyap and Chauhan, 2015)

While DM is the wider term used to describe any function that involves mining of data the term EDM is very specific to the field of education. For instance, EDM focuses on academic objectives (Kashyap and Chauhan, 2015). Some of the salient features of EDM that distinguishes EDM from the general concept of DM are that it (Chen et al. 2014):

- involves the analysis of structured data extracted from the course management system
- uses classroom technology
- involves educational decision-making using controlled online learning environments

In addition EDM is shown to have many phases as indicated in Figure 2.5.

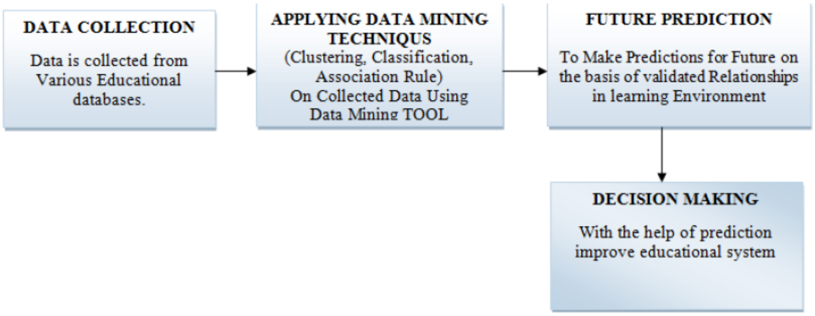

DATA COLLECTION
Data is collected from Various Educational databases.
APPLYING DATA MINING TECHNIQUS
(Clustering, Classification, Association Rule)
On Collected Data Using Data Mining TOOL
FUTURE PREDICTION
To Make Predictions for Future on the basis of validated Relationships in learning Environment
DECISION MAKING
With the help of prediction improve educational system

Figure 2.5, Phases of Educational Data Mining (Source: Kashyap and Chauhan, 2015)

Thus it can be seen that EDM as a specific data mining field is distinct although it can be brought under the overarching concept of DM and requires investigation that addresses those aspects that are particular to education. Furthermore, EDM is an emerging field (Nithya et al. 2016; Dwivedi & Singh, 2016) and many authors have recommended its use in decision making in the field of education including higher education that influence students, their learning activities, teaching activities and management aspects. Table 2.10 provides a selection of applications and uses of EDM in HEIs identified by different authors.

Table 2.10, Examples of applications and use of EDM

No.	**Education related activities**	**Authors**
1.	Better allocation of resources, predicting the performance of student, planning activities related to academic affairs and enhancing the effectiveness alumni activities. EDM pursues to find out patterns and make predictions that characterize learners' behaviours and achievements, domain knowledge content, assessments, educational functionalities, and applications.	Bidgoli (2003); Luan (2002).
2.	Additional insights can be gained from educational entities such as students, lecturers, staff, alumni, and managerial behaviour.	Delavari (2007)
3.	Allocate resources and staff more effectively, make better decisions on educational activities to improve students' success, increase students' learning outcome, increase student's retention rate, decrease students' drop-out rate, and reduce the cost of system processes.	Goyal & Vohra (2012); Delavari (2007).
4.	To identify at-risk students and bring out prediction models to identify with reasonable probability students who drop out based on truancy, disciplinary problems, changes in course performance and overall grades.	Davis (2012)
5.	Enables the use of new approaches to formative and predictive assessment. Educational administrators, students and teachers can get systematic feedback in real-time and use that material to improve academic performance.	Kharade & Wagh (2016)
6.	It is a novel DM application target for knowledge discovery, decisions-making, and recommendation	Vialardi-Sacin et al. (2009)
7.	EDM applications are gathered into eleven educational categories namely analysis and visualization of data, providing feedback for supporting instruction, recommendations for students, predicting students' performance, student modeling, detecting undesirable student behaviors, grouping students, social network analysis, developing concept maps, constructing courseware, and planning and scheduling.	Romero and Ventura (2010)

Surveys conducted on EDM (e.g. Romero and Ventura, 2010) have shown that it is an emerging field and that it is related to examining unique types of data and development of methods to examine them in the context of educational settings as well as employing those methods in gaining deeper understanding of students' learning process in those educational settings (Baker & Yacef, 2009). On the technological front, surveys conducted by Shu-Hsien et al. (2012) show that a number of techniques and methodologies used in data mining are widely used in EDM namely neural networks, algorithm architecture, dynamic prediction, analysis of system architecture, intelligent agent systems, modelling, knowledge-based systems, systems optimization and information systems. This indicates that EDM as a technique is versatile and different EDM techniques could be used for different requirements as is the case with data mining making EDM applicable to a variety of situations thus enhancing its ability to provide better outcomes. Thus both on the contextual and technological front, EDM appears to have a high potential in enabling an understanding of the hidden knowledge residing in educational datasets needed by HEIs to make the learning experience of students better. Furthermore, ever since its utility in education has been identified, it has attracted attention of researchers and much has been written about EDM and its application to education sector. Despite this, literature shows that EDM is still in its infancy (Kamber, 2011).

However what appears to have been achieved in the field of EDM in the literature up to now is only the beginning as some claim that EDM has greater potential that is yet to be tapped and can be effectively used in the HEI sector in improving student experience and decision making by both administrators and academia (Sclater et al. 2016). Considering its vast potential it is important to explore how EDM can be used to improve student performance, teaching and learning and decision making in HEIs.

2.5.4 Need for EDM in HEIs

Why EDM is needed to be used in HEIs when other data mining techniques are already present is an important question that needs to be answered. While EDM is considered to be a new and emerging technology, researchers have already come up with a number of models to predict student performance, academic and administrative decision making and a variety of other aspects that are of concern to students, teachers, managers of HEIs and parents (Agaoglu, 2016; Kaur et al. 2015). Although there is no doubt about EDM being part of the larger technique of data mining, the data mining techniques that need to be used in the field of education require the use of educational data that is distinctly different from other data and hence requires to be dealt with taking into account the characteristics of the education sector. For instance, Cunha and Miller (2012) explain that institutional performance is multi-dimensional in character and hence measuring the performance of an institution depends on multiple indicators accurate assessment of which is limited by the availability of data and the limitations of the currently available and commonly used analysing methods. In addition HEIs have varying capacity to use data as their culture and resource availability vary greatly that could reflect on the effective use of data. Besides, barriers and challenges plague HEIs in effectively using data for making decisions (Voorhees & Cooper, 2014). These aspects have affected the decision making process in HEIs such as making informed decisions based on rigorous assessment and analysis of relevant data (Menon et al. 2014). Adding to this problem is the growing volume of data that the HEIs have to deal with which in itself is a major problem as HEIs are not geared up to handle big data due to lack of effective data management systems. Moreover, data is currently stored in silos and such data may not be consistent with institutional wide data (Shacklock, 2016).

Despite difficulties and barriers that exist in HEIs that prevent the effective management of institutional data there is a growing interest in institutions to exploit the strengths of EDM as is evidenced by the adoption of EDM in institutions in US and Australia while institutions in UK are closely following the institutions in those two countries (Shacklock, 2016). However it must be borne in mind that much work needs to be done to overcome barriers to the use of EDM in institutions as those barriers can discourage institutions from adopting EDM. For instance Voorhees and Cooper (2014) have identified a host of barriers that affect institutions which include data management systems that are not updated, inadequacy of data, lack of adequate capacity to conduct analysis, recovery, treatment and re-examination of data, inability of human resources concerned with data analysis to manoeuvre through the organisational dynamics, pressure on research staff to comply with requirements and exacting workloads and increasing demands on IT staff. Needless to say that many of those barriers are daunting and may require innovation and leadership qualities on the part of HEIs to overcome them and implement EDM. An important attraction to achieve this is the benefits EDM is promising which include gaining competitive advantage and enhance student performance and satisfaction (Voorhees and Cooper, 2014).

The foregoing discussions have highlighted that while EDM, an offshoot of data mining, can be useful to HEIs, the following aspects pose challenges:

1. Discovering hidden knowledge from large volume data
2. Identifying methods to discover hidden knowledge
3. Determining how discovered knowledge could be used to make decisions to improve student learning experience.

The above challenges are formidable. More and more studies are being conducted on HEIs to determine which of the attributes of students can be used to enhance their learning experience (e.g. Liebowitz, 2017; Nikolovski et al. 2015). While EDM is still developing and being introduced in HEIs only now, contributions to this body of knowledge revolves around a number of areas including improvements in the algorithms that are being currently used to discover hidden knowledge and deal with large data (Kaur et al. 2015), comparing data mining methods, development of prediction models, understanding of discovered knowledge and data mining (KDDM) processes, improving current KDDM processes, discovering and analysing new knowledge that have not been hitherto extracted and investigated. In order to understand how these aspects can be addressed it is important to know the widely used data mining techniques and algorithms namely Classification, Regression, Association Rules mining, Genetic Algorithm, Clustering, Nearest Neighbors method, Decision Trees are used for information retrieval from educational databases (Kumar et al. 2017). Amongst the above this research focuses on Classification, Clustering, Regression, Prediction and Association Rule mining in order to investigate some of the challenges faced by HEIs pertaining to decision making and improvement in student learning experience. These aspects are discussed next.

2.6 Knowledge Discovery and Data Mining (KDDM) Process

KDDM processes are part of the overall data analysis concept, within which data mining is only one step. It is argued that just by knowing several algorithms used for data analysis one cannot accomplish a data mining (DM) project (Cios et al. 2007). Hence an overall process that helps in discovery of useful knowledge from data is needed. Thus data mining projects need to involve KDD processes. According to Fayyad et al. (1996) (also see Han & Kamber, 2012), a KDD process is made of iterative sequence methods (see Figure 2.4). KDD is the overall process that leads to discovery of useful knowledge from data. According to Rikhi (2015) KDD is a process that consists of an automatic extraction of non-obvious, hidden knowledge from large volumes of data. Literature shows that a number of KDD processes have been developed by researchers including KDD model (Fayyad et al. 1996a), Semma (Santos & Azevedo, 2005), CRISP-DM (Chapman et al. 2000) and Anand and Buchner (1998) (8 step model).

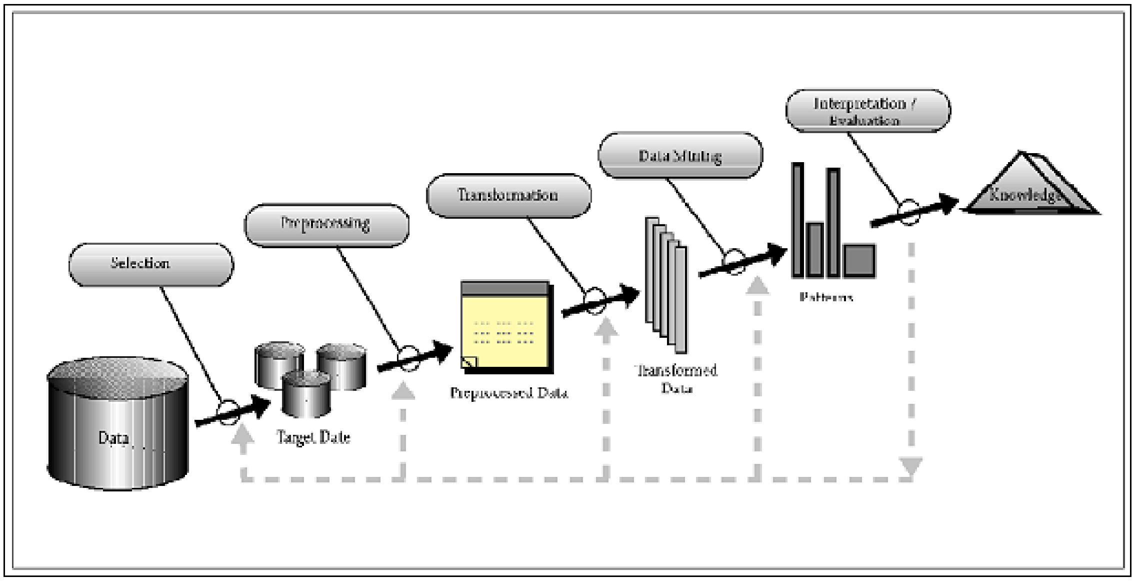


Figure 2.6, KDD process (*Source*: Fayyad et al. 1996)

The use of KDDM process for application to EDM has been attracting the attention of researchers and practitioners recently. While it is not clear from the literature which one of the many KDDM processes developed so far is the most suitable for application to HEIs, lack of a clear understanding of which one of the currently available KDDM processes could be most suitable to be applied to EDM is an important gap in the literature. There are no established KDDM processes that have been applied to large data belonging to HEIs, which in turn has created a lacuna in the literature due to which applying the concepts of DM and KDDM to improve HEI performance in terms of decision making and student performance has remained a challenge. To address this issue it is necessary to study the different KDDM process models that have been developed by researchers. Preliminary study showed that there are divergence of views amongst the researchers with regard to the concept of KDDM processes that could be implemented to discover useful knowledge to improve performance. A detailed literature on the KDDM process and the various models can be found in Appendix 23.

Table 2.11, Limitations of KDDM process models

Process Model	**Limitation**	**Limitation Identified By**	**Modified step or stage**	**New Concept**
KDD (Fayyad et al. (1996)) 5 step)	Lack of data collection step which is vital for the KDD techniques in some real applications such as information security and medical treatment Lack of domain knowledge leading to	Ruan, 2007 Redpath &	Data Collection step using previous mining results. It was added before data selection. Proposed an architecture	Inclusion of Data Collection step in data mining process to filter irrelevant data leading to better decision making (Ruan, 2007).

	decision making that maybe useful if such a knowledge is not part of the mined data.	Srinivasan, 2004	based on domain knowledge.	Introduction of Domain knowledge before the data selection (Redpath & Srinivasan, 2004).
	Unlike in other process models loop back to the second, third or fourth step are necessary due to prepared data which is not suitable for the mining process.	Kurgan & Musilek, 2006	Not addressed in the literature	Not addressed in the literature
	Lack of contextual information in the data	Vert et al. 2010	Not addressed in the literature	Not addressed in the literature
CRISP-DM	Insufficiency to handle multidimensional temporal data resulting in knowledge that cannot support decision making which is dependent on temporal issues	Catley et al. 2009	Phases 1 (business understanding), 2 (data understanding), 4 (data modelling), and 6 (deployment) were enhanced to suit temporal data.	Introduction of Intelligent Data Analysis architecture in the mining process for ensuring the mined knowledge to support decision making that need temporal aspects (Catley et al. 2009).
	The current CRISP-DM model is limited to address data that is free of human intervention	Li et al. 2009	Proposed model had two backbones of the model, namely data mining and applied intelligent system, three participation elements namely On-Line Analytical Processing (OLAP), six sigma, and domain knowledge.	On-Line Analytical Processing (OLAP), six sigma, and domain knowledge (Li et al. 2009).
	Lack of integrated process model.	Sharma et al. 2012	Identification of task–task dependencies (between tasks of the same phase and different phases) is the first step towards building an integrated process model	Building an integrated process model (Sharma et al. 2012)
	CRISP-DM suffers from absence of stages such as project management processes, integral processes (that assure project function completeness and quality) and organizational processes which are essential for data mining. This limitation can result in incomplete knowledge mined through the CRISP-DM process which in turn could affect the decision making.	Marbán et al. 2009	Not addressed in the literature	Not addressed in the literature
	Another limitation is that it is linear and sequential. Although feedback loops are represented, the sequential nature of the representation appears to exhibit an ordering of the knowledge For instance, data understanding and data preparation are given a place	(Rennolls & Al-Shawabkeh, 2008)	Not addressed in the literature	Not addressed in the literature

	between business understanding and modelling. But the representation of business understanding is essentially a prior model of the knowedge space and needs modelling techniques for its adequate representation. The current CRISP-DM model is limited to address data that is free of human intervention. Used widely for industrial purposes Lack of context in the data which adds potential knowledge to mined data	(Pan, 2010) (Catley et al. 2009) (Li et al. 2009) (Vert et al. 2010)	Not addressed in the literature Not addressed in the literature Not addressed in the literature Not addressed in the literature	Not addressed in the literature Not addressed in the literature Not addressed in the literature Context
Anand& Buchner (8 step) 1998	Does not accommodate for a step that is concerned with applying the discovered knowledge to work. Lack of contextual information in the data	(Kurgan and Musilek, 2006) (Vert et al. 2010)	Not addressed in the literature Not addressed in the literature	Not addressed in the literature Not addressed in the literature

The details provided in the Table 2.11 point out that researchers have been successful in overcoming some of the limitations identified in the literature although one major issue related to the contextual data has been found to be rarely addressed in the data mining literature.

In addition the foregoing information shows that none of the KDDM processes incorporate a contextual factor extraction and understanding stage and contextual data preparation stage that is needed to extract hidden knowledge from datasets characterised by contextual factors (refer Appendix 25). There is recognition amongst researchers who argue that existing KDDM processes generally do not mine or discover contextual data that are characterized by many attributes like temporal aspects (Vert et al. 2010). Although there are some papers that have discussed context awareness and context driven data mining, such discussions do not suggest how contextual factor knowledge could be linked to data mining process to support decision making in organisations including HEIs leading to achievement of business goals. If such an integration could be achieved the resulting KDDM model could be having a greater predictive power that is needed to make better decisions.

2.7 Discussion on different KDDM processes

The foregoing discussions have brought out the need to investigate EDM to discover the course taking pattern and predict the time-to-degree taken by the students, a major factor that affect student performance, student learning experience and decision making in HEIs. However in order to discover the pattern it has been argued in the previous section that there is a need for a KDDM process and simple datamining techniques may not be adequate. This is amply demonstrated by the taxonomy of KDDM processes provided in Appendix 23 and the experiment presented in Appendix 24. From Appendix 23 it can be seen that a number of KDDM processes have been used in practice by researchers including CRISP-DM, Generic model, Fayyad et al. model, Cabena et al. model and Cios et al. model. The taxonomy provided in Appendix 23 has defined the different steps involved in each one of those models. It can be seen that KDDM processes include additional steps when compared to simple datamining techniques. For instance literature shows that in a simple datamining project there is no need to include steps like business understanding and data understanding whereas without including those steps it is difficult to make business decisions using outcomes derived from datamining techniques. But the models provided in Appendix 23 although having the additional steps required for decision making, differ in many ways and each one has specific limitations. Those limitations are particularly glaring with regard to the use of contextual factors in the KDDM processes. Since contextual factors have been shown to play an important role in making decisions using the outcome of the mined data (see Section 2.3.3), it is necessary that those processes are reviewed again to find out ways to include contextual factors also in the process and address the limitation. However literature shows that none of the KDDM processes developed so far has been equipped to deal with contextual factors in their current form implying that there is a need to develop a method by which contextual factors could be included as part of the overall data mining process. While there are calls in the literature for including steps in the KDDM processes to mine data that takes into account contextual factors hidden in the dataset, a function KDDM process model has eluded the researchers. The need for including contextual factors arises due to the fact that in EDM, contextual factors have been argued to play a leading role in determining patterns using datamining techniques to make more accurate decisions. Since the focus of this research is to examine time-to-degree of students and its relationship to course taking patterns of students, the role of contextual factors for instance course difficulty and course difficulty pattern, cannot be ignored. While literature shows that many observable factors pertaining to students can be used to determine student performance, there is hardly any research that has examined student dataset for discovering unobservable factors hidden in the student dataset for instance course difficulty pattern. Considering the fact that contextual factors can play a significant role in datamining projects dealing with EDM and no specific KDDM process could be found in the literature that has addressed this gap, this research investigates a specific KDDM process chosen as an example to demonstrate the lack of facilities to extract patterns characterised by contextual aspects. Before investigating a KDDM process as an example the next step taken was to demonstrate the presence of such a gap in a widely used KDDM process namely CRISP-DM process. Thus the next section discusses the CRISP-DM to demonstrate the need for including a method to deal with contextual factors.

2.8 CRISP-DM:

While a few KDDM process models namely Fayyad et al. model, Cios et al. model and Anand and Buchner model have been used for datamining projects in the academic area (Kurgan & Musilek 2006), according to the literature CRISP-DM, an abbreviation for Cross Industry Standard Process for Data Mining, has been used only in the industry sector so far (Kurgan & Musilek 2006). Created in late 1996 by three forerunners of the then immature data mining market namely Daimler, SPSS and NCR, CRISP-DM describes the life cycle of data mining project as 6 phases and is shown in Figure 2.7.

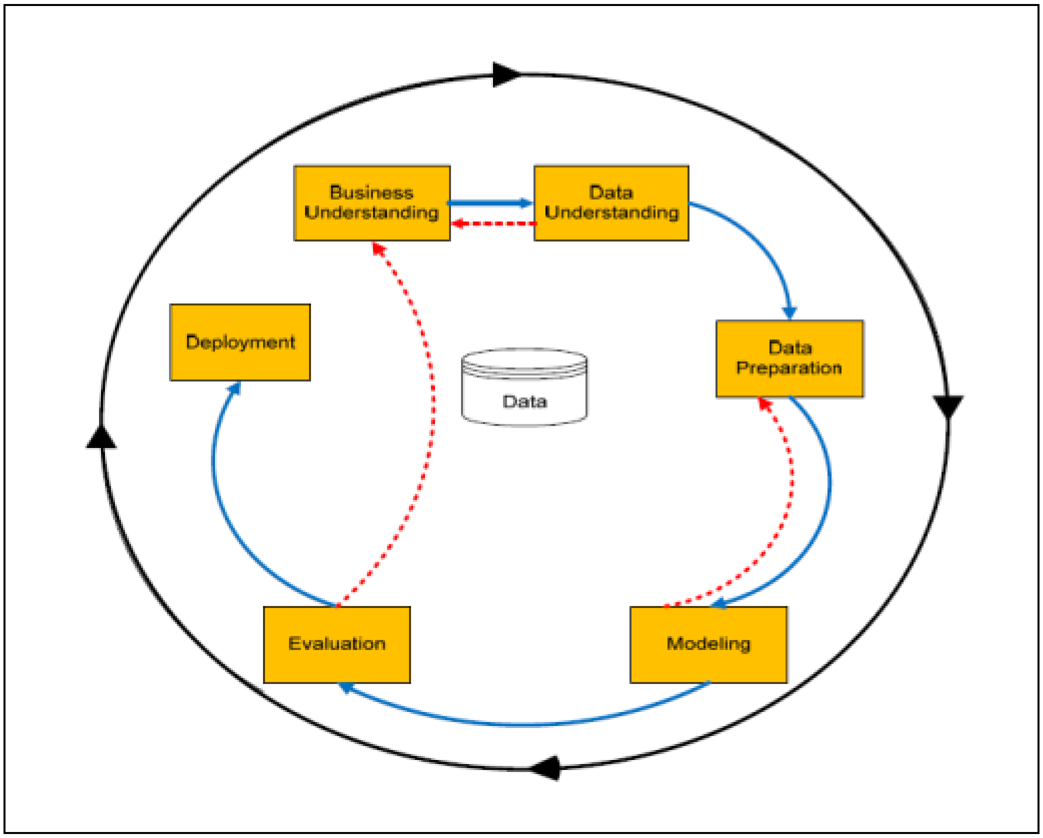
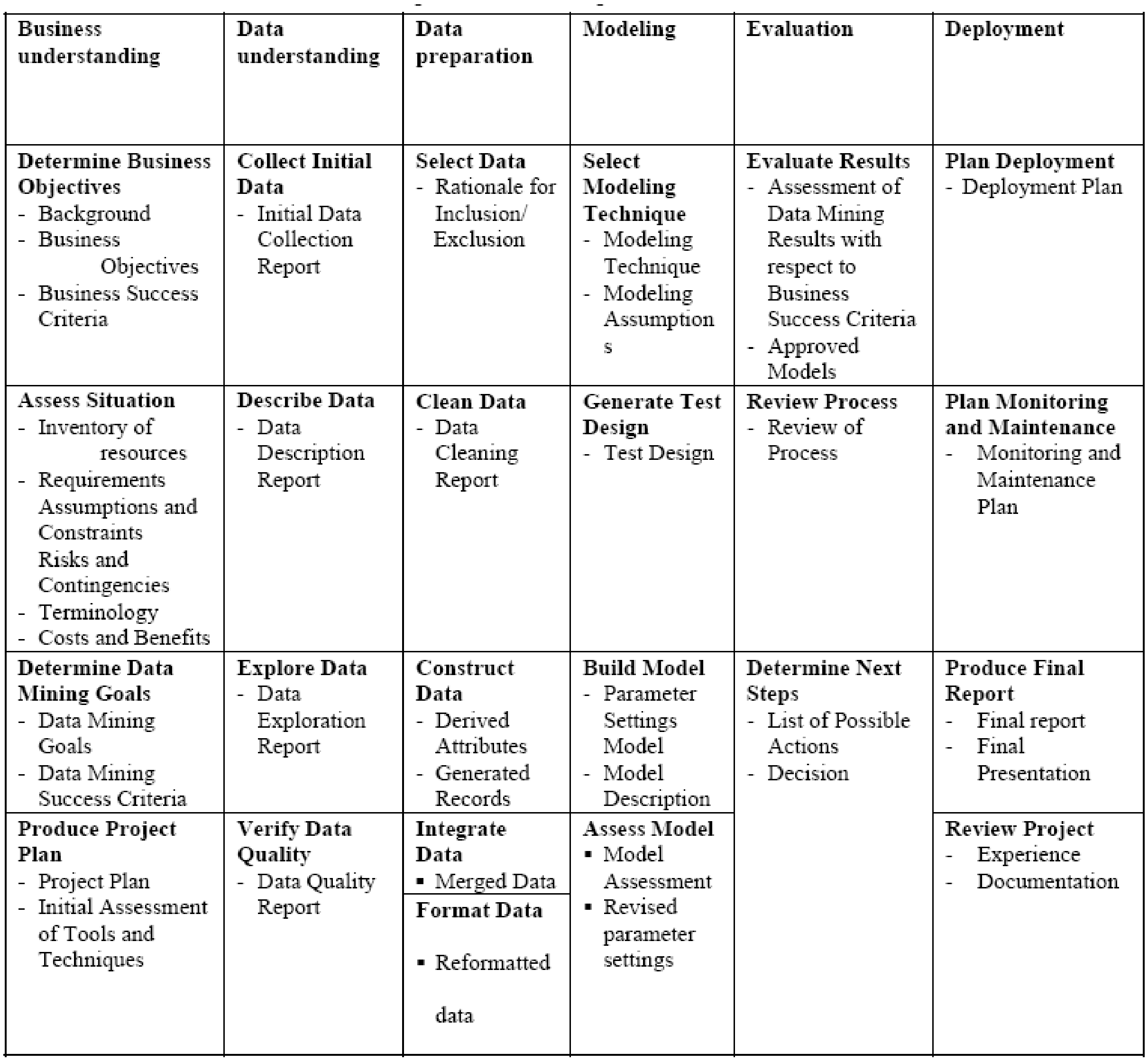


Figure 2.7, CRISP-DM process (Source: Chapman et al. 2000)

Table 2.12 provides an idea about the various steps used in CRISP-DM process.

Table 2.12, Phases, Tasks and Outputs - CRISP-DM process model



Business understanding	**Data understanding**	**Data preparation**	**Modeling**	**Evaluation**	**Deployment**
Determine Business Objectives - Background - Business Objectives - Business Success Criteria	**Collect Initial Data** - Initial Data Collection Report	**Select Data** - Rationale for Inclusion/ Exclusion	**Select Modeling Technique** - Modeling Technique - Modeling Assumptions	**Evaluate Results** - Assessment of Data Mining Results with respect to Business Success Criteria - Approved Models	**Plan Deployment** - Deployment Plan
Assess Situation - Inventory of resources - Requirements Assumptions and Constraints Risks and Contingencies - Terminology - Costs and Benefits	**Describe Data** - Data Description Report	**Clean Data** - Data Cleaning Report	**Generate Test Design** - Test Design	**Review Process** - Review of Process	**Plan Monitoring and Maintenance** - Monitoring and Maintenance Plan
Determine Data Mining Goals - Data Mining Goals - Data Mining Success Criteria	**Explore Data** - Data Exploration Report	**Construct Data** - Derived Attributes - Generated Records	**Build Model** - Parameter Settings Model - Model Description	**Determine Next Steps** - List of Possible Actions - Decision	**Produce Final Report** - Final report - Final Presentation
Produce Project Plan - Project Plan - Initial Assessment of Tools and Techniques	**Verify Data Quality** - Data Quality Report	**Integrate Data** ▪ Merged Data **Format Data** ▪ Reformatted data	**Assess Model** ▪ Model Assessment ▪ Revised parameter settings		**Review Project** - Experience - Documentation

The following main objectives were set to be achieved using CRISP-DM process

1. Enable business understanding.
2. Facilitate data understanding
3. Prepare data for mining.
4. Develop models using modelling stage and discover patterns.
5. Evaluate the results of the modelling stage by comparing those results with the business goals to be achieved using discovered patterns. If not achieved then again iterations take place until achieved.

6. Facilitate deployment of the model after evaluation is complete.

From the foregoing discussions it can be seen that contextual factors are not dealt with in the CFISP-DM process. According to the literature KDDM processes do not include contextual factors for instance course difficulty and lack contextual characterisation of the course taking pattern using contextual factors. In order achieve this KDDM processes including CRISP-DM process need to be modified as pointed out in the extant literature (Kurgan & Musilek 2006, Vert et al., 2010). In addition literature shows that ordinary datamining techniques are not sufficient enough to discover contextualised patterns including contextualised course taking pattern, (Vermunt,2005; Vert et al, 2010; Schilit, et al., 1994; Dey, 2001; Bolchini, et al., 2007). Thus on the one hand it can be seen that ordinary datamining techniques are not suitable to discover course taking patterns characterised by contextual factors, on the other existing KDDM process that could be used to discover contextualised course taking patterns of students are found inadequate to support the discovery of course taking patterns characterised by contextual factors. This is a major gap. In order to address this gap in this research CRISP-DM process has been selected based on the taxonomy of KDDM process provided in Appendix 23. Discussions on this aspect is provided in Section 3.4. Further to a discussion on the KDDM process models, it was necessary to know how the KDDM process models could be used for this research using relevant theories. This aspect is covered in the next section.

2.9 Synthesis of theories explaining KDDM process

As mentioned earlier the field of data mining is growing at a rapid pace and finds application in a number of areas making it multidisciplinary. Being multidisciplinary, theoretical underpinning of the concept spreads across disciplines. Artificial intelligence, database theory, data visualization, marketing, mathematics, operations research, pattern recognition, and statistics are some of the concepts and theories that have been identified as applicable to the concept of data mining in the extant literature (Wang & Han, 2016; Hirji, 2001). In addition Gardiner and Gillet (2015) identified rough set theory, association rule mining, emerging patterns, and formal concept analysis as theories that could be applied to ground knowledge discovery algorithms used in KDD process. Some of the widely used theories are explained next.

Artificial intelligence (AI) is a concept that enables the creation of systems that show the flexibility and versatility of human intelligence related to a wide range of cognitive domains, for instance planning, creativity, logical thinking, perception, language and learning. AI systems are purported to have the ability to transfer knowledge from one domain to another as well as learn from experience through interaction and from humans using their learning capabilities (National Science and Technology Council, 2016; p. 19). Database theory states that "*the order of attributes in a relation schema has significance to both semantics and operations. With this we simplify the employment of attributes to finite sets and write* $A = \{a1, a2, ... , a_q\}$, $q \epsilon N$ *for the ordered appearance in relation R. We notate* R_A *as shortform or* R_{A+D} *to identify a decision table* (Beer & Buhler, 2015, p. 151). Both AI and database theory find application in KDDM process as AI is found to be employed in the development of algorithms used in the KDDM process (Fisher et al. 2017) and database theory is found be employed in generating database tables and reports that are useful in decision making in KDDM process (Fayyad et al. 1996).

Another important theory that finds application is the rough set theory. According to the rough set theory sets of objects can be described by attribute values, the significance of the attributes could be analysed, dependencies between attributes could be found and decision rules could be produced (Pawlak, 1991). This theory is applicable straightaway to the data mining operations which are part of the KDD process. Similarly Formal Concept Analysis (FCA) (Ganter & Wille, 1998; Wille, 1992; 1982) offers a framework for conceptual clustering and finds application in KDD process (Zhang, 2004). FCA is a widely used framework that provides the theoretical underpinning for class hierarchy design and maintenance (Godin & Valtchev, 2005). According to Zhang (2004) the concept of FCA explains how the attributes are clustered grounded on the algebraic principle of Galois connection leading to the formation of a partially ordered set called concept lattice. It is further argued that clustering determines that collection of attributes that forms a coherent entity called a concept that is based on the philosophical criteria of unity between extension and intension. Zhang (2004) further explain that set of all objects that belong to a concept is called the extension while the set of attributes common to all those objects belonging to the concept is called the intension. Therefore a concept can be characterized by the property which is: a collection of attributes which agrees with the intension of its extension (Zhang, 2004).

Association rule mining (ARM) is considered to be a process that enables the discovery of associations between the objects in a database (Wang et al. 2007; Agrawal et al. 1993). ARM enables the finding of all association rules with support ≥ minsup and confidence ≥ minconf (Gardiner & Gillet, 2005). One of the algorithms that is based on ARM is the *a priori* algorithm (Agrawal et al. 1994). Furthermore, statistics is the basic theory that is used in KDD process.

Conventionally statistical theory stresses on mathematical formulation and validation of a methodology. Further it views simulations and empirical or practical evidence as a lower form of validation. Statistical theory has major implications to KDD process for instance nearest neighbors, clustering methods, association rules, feature extraction and visualization and many algorithms use statistical theory including genetic algorithms (Samundeeswari & Srinivasan, 2017).

While there are many other theories that find application in explaining and grounding the concept of KDD, all of them have not been discussed here. Only those that have been widely used in the KDD literature have been discussed. Further it is important to know that there are overlaps between theories and it is difficult to determine which one of the theories is more dominant than the other. Thus theoretical application to KDD must be done cautiously taking into account its trans-disciplinary nature.

2.10 Summary

This chapter has presented the current level of understanding of big data and its utility to HEIs. The discussion has brought out in detail the challenges faced in adopting and using big data. The arguments show that still it is not known whether the big data analysing processes currently developed can comprehensively be applied to different contexts, particularly the HEI context. In fact the review of the literature given above shows that big data as a concept is yet to find its use in educational sector as the sector is criticized to be ill-equipped to handle the big data (Sin & Muthu, 2015) and it is not known how big data could be used in the context of HEIs characterized by such contextual aspects as data quality, privacy, security and ethical issues. Hardly any theoretical model or algorithm that could be applied exclusively to support the HEIs in analysing the student datasets and extracting knowledge hidden in those datasets to achieve better performance has been developed (Daniel, 2015). Particularly analysis of student dataset pertaining to specific student performance factors including time-to-degree, CGPA and course taking pattern of students have not been studied applying the concepts of big data and data mining that is characterized by contextual factors including course difficulty indicating the limitations existing in the current literature. While the currently available data analysing processes and theories are seen to be plagued by limitations, such limitations need to be overcome as those limitations deter users of big data and researchers from exploiting the advantages offered by the analysis of big data.

Chapter 3 : Integration of EDM in CRISP-DM Process

3.1 Introduction

Two important aspects were raised in the previous chapter. One was about the problems faced by HEIs while applying data analysing processes namely data mining to analyse large datasets. The second was the lack of knowledge on the part of HEIs about factors namely time-to-degree, course taking pattern and contextual factors (course difficulty) embedded in the educational datasets, in which knowledge is hidden and which when discovered can be used to enhance the capability of the HEIs in achieving organisational goals but not exploited till date. These two aspects are investigated in this chapter. The previous chapter has discussed about data, data analysis, data mining, educational data mining and KDDM process that are essential to know how business problems could be solved using KDDM process, what limitations are encountered and how to overcome those limitations. Different DM models have been discussed knowledge about which is essential to generate the most appropriate model for a purpose.

As explained in Section 3.2 concepts of data mining and KDDM are found to be useful in mining data to discover hidden knowledge and improve the performance of organization. In Section 3.2, it was argued that currently available KDDM processes have not been applied to discover hidden knowledge in the datasets of HEIs and how the concept of EDM could be integrated in to a knowledge discovery process for improving the performance of HEIs and students. This was found to be a major gap in the literature. EDM is promising to uncover patterns and association rules amongst the variables being dealt with in HEIs that have not been observed until now but could significantly improve the performance of the HEIs as well as the students. This chapter reviews the some of the widely discussed KDD processes in the literature and analyses how specific student performance factors could be improved. This knowledge is expected to provide HEIs with a KDD process in which EDM concepts are integrated leading to knowledge that could enable HEIs to support the students better. An example of performance enhancement of HEIs could be to predict unknown values say student time-to-degree using variables in the dataset say course taking pattern of students (for instance, classification of students, regression, and anomaly detection) and describe human-understandable patterns and trends in the data (for instance, clustering of students, finding association rule amongst attributes of students, and summarization of the model) (Gorunescu, 2011). Towards this the chapter has identified the following equations and analyses the equations in-depth to know whether an EDM based KDD process can be developed to discover patterns and hidden knowledge and if so how. The equations that will be studied are:

CGPA = function of (number of courses, course taking pattern, course difficulty, time-to-degree) → 3.9

and

Time-to-degree = function of (number of courses, course taking pattern, course difficulty, CGPA) → 3.10

In choosing the KDDM process for testing the equations, CRISP-DM process was chosen which has been described. In addition changes to CRSIP-DM process have been suggested to enable accurate prediction of optimum CGPA and time-to-degree using course taking pattern and course difficulty level pattern. The chapter also discusses the Design Science (DS) Research methodology that has been chosen to ground the method used to modify and develop, test, evaluate and communicate about the artefact called modified CRISP-DM process. Further the experiments in this research were conducted in an anonymous university in Bahrain, where exclusive permission was obtained to investigate the educational dataset resident in a student registration system called ‘ADREG’. Details about ADREG are provided in Appendix 11.

3.2 Application of EDM concepts to HEIs

In this research four factors namely time-to-degree, course taking pattern and contextual factors (course difficulty and student potential) contributing to the performance of a HEI were investigated in order to know whether EDM concepts could be applied and the significance of applying EDM. These factors can be related in a number of ways to each other. For instance in Table 3.1 which has been created from the transcript of students studying in a private University in Bahrain it can be seen that the information concerning the set of courses in which each student is registered is related to the CGPA.

Table 3.1, Selected student data related to performance

Student ID	CGPA	Time-to-degree	Semester GPA	Semester PCR	Course Code	Semester
Stud1	3.78	3.5	3.67	15	ACCT 101, ARAB 101, ECON 101, ENGL 101, ENGL 102	1
Stud2	3.84	3.5	4	15	ARAB 101, ECON 101, ENGL 101, ITCS 101, MATH 103	1
Stud3	2.18	4	1.55	9	ARAB 101, ENGL 101, ITCS 101	1
Stud4	2.62	4	2.9	15	ARAB 101, ECON 101, ENGL 101, FREN 101, ITCS 101	1
Stud5	2.43	4	3.2	15	ACCT 101, ECON 101, ENGL 101, ITCS 101	1
Stud6	2.24	8.5	2.1	15	ARAB 101, ECON 101, ENGL 101, ITCS 101, MATH 103	1
Stud7	3.6	4	3.27	15	ARAB 101, ECON 101, ENGL 101, ITCS 101, MATH 103	1
Stud8	3.85	4	4	15	ACCT 101, ECON 101, ENGL 101, ITCS 101, MATH 103	1
Stud9	2.77	4	2.7	15	ACCT 101, ECON 101, ENGL 101, ITCS 101, MATH 103	1
Stud10	3.36	3.5	3.4	15	ARAB 101, ECON 101, ENGL 101, ITCS 101, MATH 103	1

In addition, Table 3.1 informs about the time-to-degree, the semester number and semester GPA. An inspection of the table leads to the following inference.

There is a relationship between set of courses, the number of courses and the CGPA although a formula to define the relationship amongst the three variables is not clearly emerging. For instance Stud1 has registered in 5 courses and scored a CGPA of 3.78 whereas Stud5 has registered in 4 courses and scored a CGPA of 2.43. The inference could be that

CGPA = function of (number of courses, course taking pattern) → (3.1)

There is a relationship between the set of courses and the time-to-degree but it is not clear whether a generalizable equation could be derived linking the two. For instance Stud1 has registered in 5 courses and graduated within 3½ years whereas Stud5 has registered in 4 courses and graduated within 4 years. The inference could be that

Time-to-degree = function of (number of courses, course taking pattern) → (3.2)

There is a relationship between the number of courses, set of courses, the CGPA and time-to-degree but it is not clear how to represent the relationship. For instance Stud1 has registered in 5 courses as a set comprising courses (ACCT 101, ARAB 101, ECON 101, ENGL 101, ENGL 102) and graduated within 3½ years with a CGPA of 3.78 whereas Stud5 has registered in 4 courses as a set comprising courses (ACCT 101, ECON 101, ENGL 101, ITCS 101) and scored a CGPA of 2.43. The inference could be that

Time-to-degree = function of (number of courses, course taking pattern, CGPA) → (3.3)

and

CGPA = function of (number of courses, course taking pattern, Time-to-degree) → (3.4)

There is a relationship between identical number of courses within the set of courses the students have registered in a semester, the courses that are different, the number of courses and the CGPA although specifying the relationship is not supported by any formula. For instance Stud1 has registered in 5 courses as a set comprising courses (ACCT 101, ARAB 101, ECON 101, ENGL 101, ENGL 102) and graduated within 3½ years with a CGPA of 3.78 whereas Stud5 has registered in 4 courses as a set comprising courses (ACCT 101, ECON 101, ENGL 101, ITCS 101) and scored a CGPA of 2.43. In this case both the students have the set of courses comprising (ACCT 101, ECON 101, ENGL 101) as common whereas the set of courses comprising (ARAB 101, ENGL 102, ITCS 101) as different. The inference that can be drawn is

**CGPA = function of (different courses, common courses, course taking pattern,
time-to-degree) → (3.5)**

and

**Time-to-degree = function of (different courses, common courses, course taking pattern,
CGPA) → (3.6)**

Equations 3.1 to 3.6 convey one important information which is there is some underlying knowledge that needs to be identified that can perhaps explain the relationship. One way of addressing this issue to identify a factor say contextual factor that may provide some basis to explain the relationship. For instance course difficulty could be one factor that may be the reason for the students to perform differently. The formula in Section 2.3.3 was applied to calculate the course difficulty for each course in which a student is registered as reflected in Table 3.2.

Table 3.2, Relationship between course difficulty and other student performance factors

Student ID	GPA	Time-to-degree	Semester GPA	Semester PCR	Course Code	Semester	Course Difficulty
Stud1	3.78	3.5	3.668	15	ACCT 101,ARAB 101,ECON 101,ENGL	1	5.57, 5.59, 5.62, 5.64,

					101, ENGL 102		5.62
Stud2	3.84	3.5	4	15	ARAB 101,ECON 101,ENGL 101,ITCS 101,MATH 103	1	6.17, 6.17, 6.18 6.19, 6.19
Stud3	2.18	4	1.55	9	ARAB 101,ENGL 101,ITCS 101	1	2.98, 2.93, 2.95
Stud4	2.62	4	2.934	15	ARAB 101,ECON 101,ENGL 101,FREN 101,ITCS 101	1	4.3, 4.29, 4.28, 4.27,4.23
Stud5	2.43	4	3.2	15	ACCT 101,ECON 101,ENGL 101,ITCS 101	1	3.84, 3.86, 3.87, 3.79
Stud6	2.24	8.5	2.134	15	ARAB 101,ECON 101,ENGL 101,ITCS 101,MATH 103	1	3.69, 3.74, 3.84, 3.88, 3.9
Stud7	3.6	4	3.268	15	ARAB 101,ECON 101,ENGL 101,ITCS 101,MATH 103	1	6.01, 6.03, 6.02, 6.02, 6
Stud8	3.85	4	4	15	ACCT 101,ECON 101,ENGL 101,ITCS 101,MATH 103	1	6.39, 6.36, 6.36, 6.37, 6.32
Stud9	2.77	4	2.732	15	ACCT 101,ECON 101,ENGL 101,ITCS 101,MATH 103	1	3.98, 4.06, 3.98, 4.02, 4.01
Stud10	3.36	3.5	3.4	15	ARAB 101,ECON 101,ENGL 101,ITCS 101,MATH 103	1	5.49, 5.5, 5.43 5.48, 5.43

Table 3.2 is showing surprising results and it is not easy to explain how this can happen. For instance student with code Stud8 has taken courses that are shown to be of highest difficulty amongst the 10 students taken as sample. However this student has scored the highest CGPA of 3.85. On the contrary student with code Stud1 has scored less CGPA of 3.78 although the course difficulty of the courses in which this student has registered is lower than that of the courses in which Stud8 has registered. That is to say despite the fact the course difficulty was comparatively lower, Stud1 scored lower CGPA than Stud8. The interpretation could be that there are one or more underlying factors that could not be observed and be the reason for this happening. What are those factors? It is not clear Similarly, Stud8 has taken longer time-to-degree than Stud1 although the difference in the CGPA between the two students is not much. Interestingly the difference in time-to-degree between the two students is only six months. This might have occurred due to some unobserved factor. That is to say, Stud8 could have achieved the same CGPA of 3.84 and graduated in 3.5 years if the unknown factor were to be discovered and used and if that factor were to be real. In another instance Stud3 has scored the lowest CGPA of 2.18 although the course difficulty figures of the courses in which the student has registered is much lower than that of Stud8 and number of courses in which Stud3 has registered is also lower (i.e. 4) than that of Stud8 (i.e. 5). Despite such an anomalous situation it is possible to think of the following relationships although at this point of time these relationships are only assumptions.

CGPA = function of (number of courses, course taking pattern, course difficulty) → (3.7)
and
Time-to-degree = function of (number of courses, course taking pattern,
course difficulty) → (3.8)

Or

CGPA = function of (number of courses, course taking pattern, course difficulty,
time-to-degree) → (3.9)
and
Time-to-degree = function of (number of courses, course taking pattern, course difficulty,
CGPA) → (3.10)

The arguments show that course difficulty is somehow related to the number of courses in which the student has registered, the course taking pattern, time-to-degree and CGPA although such a relationship is not clearly observed when one peruses the Table 3.2. A method needs to be found out that could enable discovery of the relationship.

In order to discover this relationship and arrive at a result and conclusion this research applies the concepts of EDM (see Section 2.5.3) that could enable the discovery of hidden knowledge and could be used to test the equations 3.1 to 3.10 so that it is possible to verify the relationships using the discovered knowledge. Such a decision to use EDM as a method to verify the relationships depicted in equations 3.1 to 3.10 emanates from the fact that manual verification of the equations will be virtually impossible when the number of students involved runs into hundreds of thousands and the dataset generated is huge. The use of a technique to analyse huge data like EDM is inevitable. EDM provides a way forward to discover hidden knowledge from the datasets and derive course taking patterns. In addition verification of the relationships using EDM could provide a definite way to improve the student learning performance and experience and also the decision making process in HEIs.

However gaps exist in the knowledge related to applying the concepts of EDM to verify the relationships. For instance from Section 2.5.4 the following limitations of EDM have been extracted.

1. Discovering hidden knowledge from large volume data is a challenge.
2. Identifying methods to discover hidden knowledge is difficult.
3. Determining how discovered knowledge could be used to make decisions to improve student learning experience.

That is to say, applying data mining concepts to EDM needs experimentation to check whether hidden knowledge from large volumes of education data could be discovered. Further, the outcome of the experimentation could be used to test the various relationships depicted in equations 3.1 to 3.10 using EDM techniques. To achieve this, an important problem need to be overcome. From Section 3.5 it is seen that data mining techniques could enable discovery of clusters of variables, classification of variables and association rules between variables in a dataset. It is necessary to check which of these techniques could be used to verify the relationships described in equations 3.1 to 3.10. In addition it is possible that already tested techniques such as algorithms which are used as part of the data mining process might not work if applied to EDM (see Section 2.5.2.4) due to those characteristics of the education sector that are different from those found in data related to industries (see Section 2.5.3). For instance if one uses genetic algorithm as part of the EDM to verify the relationship in equations 3.6 then limitations of genetic algorithm could be a bottleneck including lack of guarantee that it will lead to a global optimum solution especially in cases where the populations have many subjects. This could be a challenge in deriving the optimum time-to-degree if the population of students consists of different types of students for instance students of different nationalities. Thus unless experimented it won't be clear how general data mining concepts could be useful in EDM to discover knowledge or patterns, what barriers one could face and how a useful data mining technique could be identified as applicable to mine educational data in a variety of contexts. One way by which this could be tackled is to identify and use a specific knowledge discovery and data mining (KDDM) process.

3.3 KDDM framework to Integrate EDM

3.3.1 Analysis of equation 3.9

The main variable that needs to be determined in equation 3.9 is time-to-degree. This is expressed as number of years. This could vary usually between 3.5 years to 8 years. The other variable that needs to be determined is the CGPA. The student can score a maximum of 4.0 and a minimum of zero. The factor under investigation is the course taking pattern of students in each semester. The idea is to determine the pattern of courses that a student could take in a semester that could be related to CGPA and the time-to-degree. Course taking pattern is the set of courses a student will have to register in a semester or semesters and graduate through the programme. Example of course taking pattern of students is provided for a few individual students under the column "course code" in Table 3.3 below.

Table 3.3, CGPA and time-to-degree as a function of course taking pattern

Student ID	GPA	Time-to-degree	Semester GPA	Semester PCR	Course Code (course taking pattern)	Semester (Year 1)	Course Difficulty
Stud1	3.78	3.5	3.668	15	ACCT 101, ARAB 101, ECON 101, ENGL 101, ENGL 102	1	5.57, 5.59, 5.62, 5.64, 5.62, 5.66
Stud2	3.84	3.5	4	15	ARAB 101, ECON 101, ENGL 101, ITCS 101, MATH 103	1	6.17, 6.17, 6.18, 6.19, 6.19
Stud3	2.18	4	1.55	9	ARAB 101, ENGL 101, ITCS 101	1	2.98, 2.93, 2.95
Stud4	2.62	4	2.934	15	ARAB 101, ECON 101, ENGL 101, FREN 101, ITCS 101	1	4.3, 4.29, 4.28, 4.27, 4.23
Stud5	2.43	4	3.2	15	ACCT 101, ECON 101, ENGL 101, ITCS 101	1	3.84, 3.86, 3.87, 3.79,
Stud6	2.24	8.5	2.134	15	ARAB 101, ECON 101, ENGL 101, ITCS 101,MATH 103	1	3.69, 3.74, 3.84, 3.88, 3.9
Stud7	3.6	4	3.268	15	ARAB 101, ECON 101, ENGL 101, ITCS 101, MATH 103	1	6.01, 6.03, 6.02, 6.02, 6
Stud8	3.85	4	4	15	ACCT 101, ECON 101, ENGL 101, ITCS 101,MATH 103	1	6.39, 6.36, 6.36, 6.37, 6.32
Stud9	2.77	4	2.732	15	ACCT 101, ECON 101, ENGL 101, ITCS 101, MATH 103	1	3.98, 4.06, 3.98, 4.02, 4.01
Stud10	3.36	3.5	3.4	15	ARAB 101, ECON 101, ENGL 101, ITCS 101, MATH 103	1	5.49, 5.5, 5.43, 5.48, 5.43

From Table 3.3 it can be seen that course taking patterns of the courses in which the students have registered in semester 1 of year 1 can be related to the courses in which the students have registered. If only students register in courses there can be a pattern and in each semester there could be a set of courses that could be visualized as a pattern. This can be represented as

Course taking pattern = function of (courses) → (3.11)

Again arguing that the set of courses will consist of a specific number of courses it is possible to argue that course taking pattern is related to the number of courses. Thus equation 3.11 could be modified as:

Course taking pattern = function of (courses, number of courses) → (3.12)

Between any two students who belong to the same set and have begun their study at the same time in the anonymous university in which the research was conducted, there was a possibility that the course taking pattern was the same or different. Course taking pattern of any two students if it is the same it does not imply that their performance is the same for instance the grades scored by the students Stud8 and Stud9 in Table 3.3 who have registered in the same set of courses but have scored different CGPA. This could be due to some attribute of the courses called the contextual factors (see Section 2.3.3).The contextual factor that has been identified for investigation in this research is the course difficulty. The reason is that course difficulty is shown to be a function of the number of students and the grade scored by the students in a particular course (equation 2.1). Thus

Course difficulty = function of (course type, number of students, grade of students in the course) → (3.13)

Although the course difficulty of each course could be calculated individually for each student, it is not known whether the course difficulty of each course in the set of courses in which a student has registered could be influenced by the number of courses and hence the performance of the student in a semester. Thus there could be a possibility that the number of courses in which a student has registered in a semester (say ACCT 101, ARAB 101, ECON 101, ENGL 101, ENGL 102) could affect the course difficulty measure of each one of the courses an argument which is depicted as:

Course difficulty = function of (number of courses) (assumption) → (3.14)

Since course taking pattern and course difficulty are both related to number of courses and course difficulty is calculated for each course then it is possible to argue that for every course in the pattern of courses taken by a student there exists a pattern of course difficulty levels. Therefore it is possible to argue that:

Course difficulty level pattern = function of (course taking pattern,
number of courses) → (3.15)

Now number of courses taken by a student in a semester naturally is linked to the time-to-degree with higher the number of courses taken in a semester indicating shorter time-to-degree usually. However, even if a student registers in the maximum number of courses in a semester and hence in a year, there is a possibility that the student may not complete the degree within an optimum time. That is to say if a student has registered in six courses in a semester which is the maximum allowed number in a semester, there are possibilities that the student has scored a CGPA which is low and hence may like to repeat the courses. In this instance the time-to-degree will be affected. Thus scoring high CGPA is a major concern of students and majority of the students do not score high CGPA although it is hard to define what is a high CGPA any student could achieve. One of the possible reasons why students might not know the optimum CGPA they could score is the combination of courses they register in a semester. If the students do not know how difficult each course is then they would not know in what course they should register in a semester and in which order. Thus a new concept called course taking pattern related to course difficulty pattern could be thought of as a factor, knowledge about which could reveal what course taking pattern could help students achieve optimum CGPA and time-to-degree. Thus it is possible to posit:

CGPA = function of (course taking pattern, course difficulty) → (3.16)

Since course difficulty and course taking pattern are linked to the number of courses (equations 3.12 and 3.15) equation 3.16 could be modified as

CGPA = function of (course taking pattern, course difficulty, number of courses taken in a semester) → (3.17)

While it is known that time-to-degree could be affected by the number of courses registered in a semester, what is not known is whether time-to-degree will affect CGPA. Since course taking pattern can affect the time-to-degree and if students want to score high CGPA in shorter time-to-degree then those students will have to register in maximum number of courses in a semester. In this case the time-to-degree might force the students to register in certain courses without knowing the course difficulty level the combination of which could impact the students. For instance in Table 3.3 it can be seen that Stud8 and Stud9 have registered in the same set of courses and graduated in 4 years but Stud8 has scored a CGPA of 3.85 while Stud9 scored a CGPA of 2.77. It is possible to claim that Stud8 is more capable than Stud9 or it is possible to say that the set of courses Stud9 has registered in need to be analysed for the course difficulty measure and how the student has attempted to achieve a time-to-degree of 4 years at the cost of CGPA. Thus time-to-degree may affect CGPA in many students' cases. Thus taking to account the above arguments equation 3.17 can be rewritten again

CGPA = function of (course taking pattern, course difficulty, time-to-degree, number of courses taken in a semester) → (3.18)

There is an unknown situation depicted by equation 3.18. That is it is not known which course taking pattern, characterized by what difficulty level of courses, registered in which one of the semesters and in what time-to-degree can predict a CGPA. This needs investigation as knowledge about this can enable students to achieve optimum CGPA in an optimum time-to-degree using knowledge about course taking pattern, course difficulty pattern and the semester number in which students can register in those courses. Equation 3.18 is the same as equation 3.9 and the analysis give above shows that equation 3.9 is a possible condition that needs to be investigated.

3.3.2 Analysis of equation 3.10

Using the analysis conducted to establish equation 3.9 it can be seen that time-to-degree is affected by the number of courses in which a student has registered in a semester and the combination of courses chosen by the student without attaching importance to the difficulty level of the courses. That is to say number of courses per semester, the pattern of courses and the course difficulty pattern linked to the pattern of courses can be said to affect the time-to-degree of students. This relationship can be written as

Time-to-degree = function of (course taking pattern, course difficulty, number of courses taken in a semester) → (3.19)

Again as explained in the Section 3.2 many students would like to achieve a certain time-to-degree regardless of the CGPA they score (e.g. Stud9 in Table 3.1). However there are some students who would like to score a high CGPA at the cost of time-to-degree (e.g. Stud8 in comparison to Stud1). Stud8 in Table 2.2 has taken 4 years to graduate and scored a CGPA of 3.84 while Stud1 scored a CGPA of 3.78 but has graduated in 3.5 years. This argument shows that some time CGPA determines time-to-degree as in the case of Stud8. Using this argument equation 3.19 could be rewritten as

Time-to-degree = function of (course taking pattern, course difficulty, CGPA, number of courses taken in a semester) → (3.20)

There is an unknown situation depicted by equation 3.20. That is it is not known which course taking pattern, characterized by what difficulty level of courses, registered in which one of the semesters and what CGPA score can predict the time-to-degree. This needs investigation as knowledge about this can enable students to achieve optimum time-to-degree and optimum CGPA using knowledge about course taking pattern, course difficulty pattern and the semester number in which students can register in those courses. Equation 3.20 is the same as equation

3.10 and the analysis give above shows that equation 3.10 is a possible condition that needs to be investigated.

The purpose of establishing the relationships in equations 3.9 and 3.10 is twofold. One is that there is a possibility to enhance student learning experience in HEIs in regards to achieving optimum CGPA and time-to-degree by discovering the course taking pattern of students and linking that pattern to contextual factors e.g. course difficulty pattern of students. However it is not known how to discover this knowledge. Literature shows that course taking pattern can be discovered from the hidden knowledge in the education dataset. Data mining methods have been widely recommended to discover patterns and hidden knowledge from educational datasets and such knowledge could be used to predict events (Torgo, 2016). This leads to the second purpose which is to answer the question whether data mining methods could be used to discover course taking pattern from the educational dataset and predict the optimum CGPA and time-to-degree. In order to answer both the questions there was a need to know whether there is an underlying concept on which investigation about the data mining could be conducted. While the purpose is to answer the questions raised above, the primary focus is to predict optimum CGPA and time-to-degree using course taking pattern extracted from educational datasets by adopting a reliable and valid data mining method.

An important aspect that emerged in establishing the concept using which the data mining method could be identified to answer the questions was the relevance of course difficulty pattern, a contextual factor, that has a bearing on the prediction of optimum CGPA and time-to-degree. Thus how contextual factors need to be dealt with while data mining concepts are being used was another question that required to be answered. This research thus began with establishing the following relationships

a. CGPA as a function of course taking pattern of students, course difficulty pattern, time-to-degree and semester number (equation 3.9) and

b. time-to-degree as a function of course taking pattern of students, course difficulty pattern, CGPA and semester number (equation 3.10)

After establishing the basis, the discussions next proceeded to check whether educational data could be mined using data mining techniques. One example has been discussed in the Appendix 24 highlighting the limitations.

3.4 Reasons for Using CRISP-DM:

CRISP-DM, an abbreviation for Cross Industry Standard Process for Data Mining, is an industry standard and tool neutral KDDM process that was created in late 1996 by three forerunners of the then immature data mining market namely Daimler, SPSS and NCR. CRISP-DM describes the life cycle of data mining project as 6 phases and is shown in Figure 3.1.

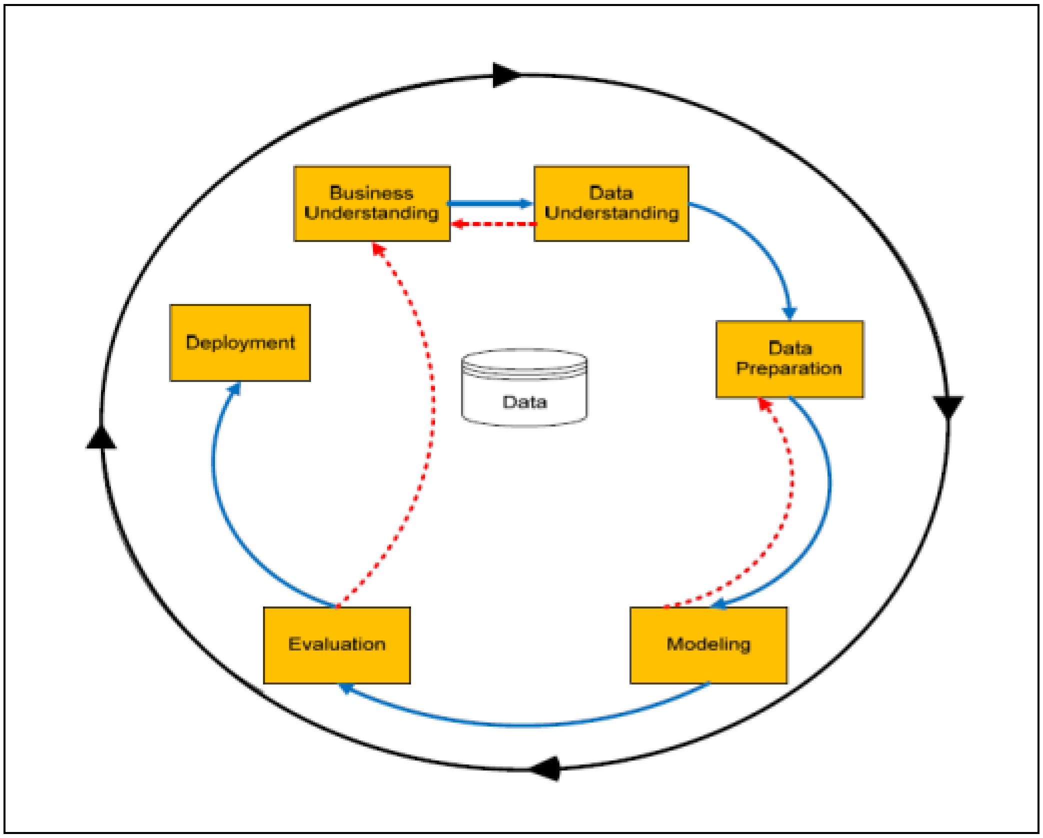
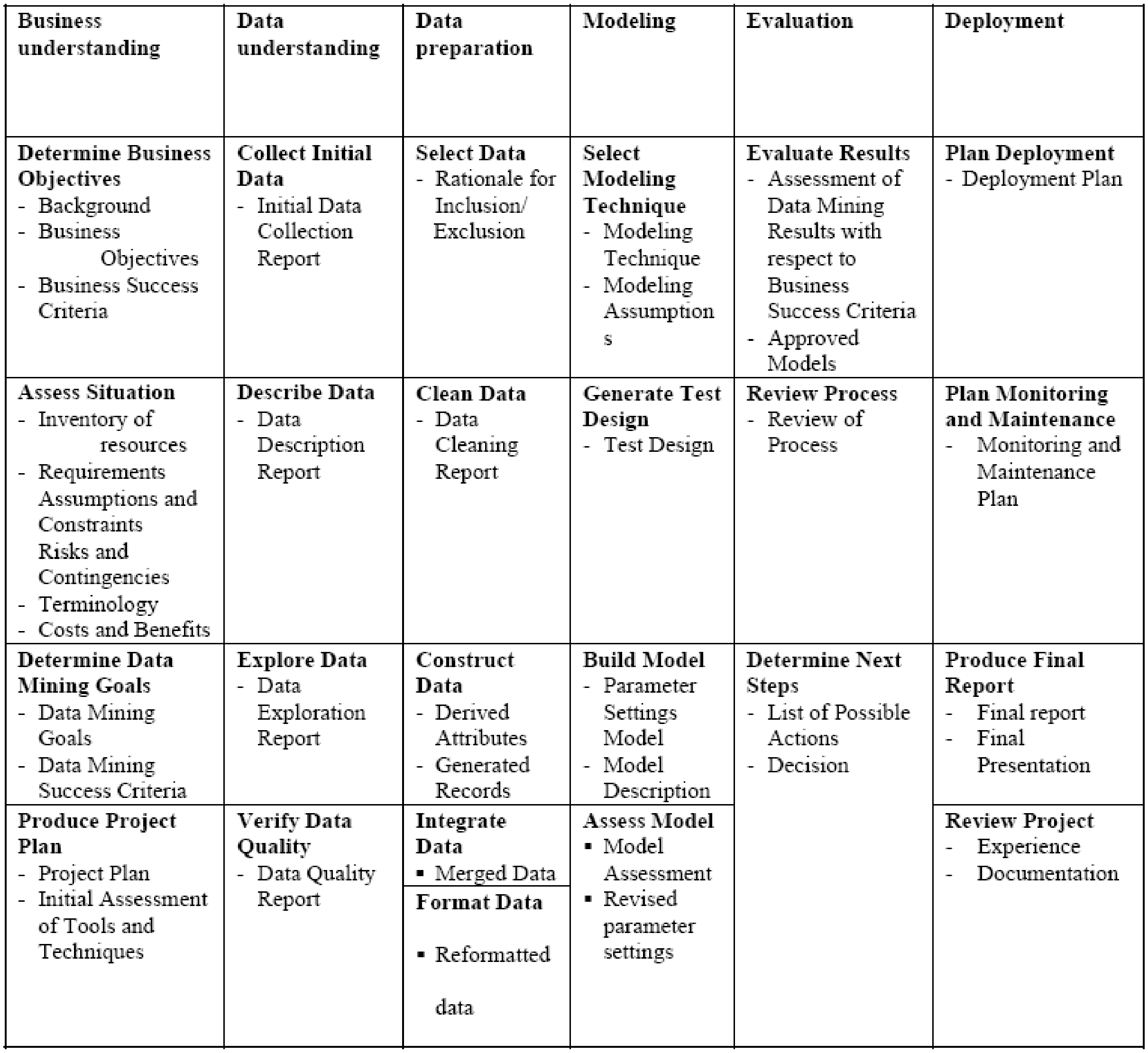


Figure 3.1, CRISP-DM process (Source: Chapman et al. 2000)

Table 3.4 provides an idea about the sub-steps involved in the various steps.

Table 3.4, Phases, Tasks and Outputs - CRISP-DM process model



Business understanding	Data understanding	Data preparation	Modeling	Evaluation	Deployment
Determine Business Objectives - Background - Business Objectives - Business Success Criteria	**Collect Initial Data** - Initial Data Collection Report	**Select Data** - Rationale for Inclusion/ Exclusion	**Select Modeling Technique** - Modeling Technique - Modeling Assumptions	**Evaluate Results** - Assessment of Data Mining Results with respect to Business Success Criteria - Approved Models	**Plan Deployment** - Deployment Plan
Assess Situation - Inventory of resources - Requirements Assumptions and Constraints Risks and Contingencies - Terminology - Costs and Benefits	**Describe Data** - Data Description Report	**Clean Data** - Data Cleaning Report	**Generate Test Design** - Test Design	**Review Process** - Review of Process	**Plan Monitoring and Maintenance** - Monitoring and Maintenance Plan
Determine Data Mining Goals - Data Mining Goals - Data Mining Success Criteria	**Explore Data** - Data Exploration Report	**Construct Data** - Derived Attributes - Generated Records	**Build Model** - Parameter Settings Model - Model Description	**Determine Next Steps** - List of Possible Actions - Decision	**Produce Final Report** - Final report - Final Presentation
Produce Project Plan - Project Plan - Initial Assessment of Tools and Techniques	**Verify Data Quality** - Data Quality Report	**Integrate Data** ▪ Merged Data **Format Data** ▪ Reformatted data	**Assess Model** ▪ Model Assessment ▪ Revised parameter settings		**Review Project** - Experience - Documentation

CRISP-DM process has been a cross industrial standard process and has been widely used in the industry. However its success in the field of HEIs is yet to be felt. Although CRISP-DM has been identified to have some limitations (see Table 7.10 of Appendix 23) the process offers a good basis to test the relationships depicted in equations 3.9 and 3.10. The choice of CRISP-DM provides the following clear advantages as well as objectives the users can achieve see Table 3.5.

Table 3.5, Advantages and objectives achieved in CRISP-DM (*Source*: Chapman et al. 2000)

Advantages
Ensure quality of knowledge discovery project results
Reduce skills required for knowledge discovery
Reduce costs and time
General purpose (i.e., stable across varying applications)
Robust (i.e., insensitive to changes in the environment)
Tool and technique independent
Tool supportable
Support documentation of projects
Capture experience for reuse
Support knowledge transfer and training

The above advantages overwhelm the limitations of CRISP-DM which makes it a more appropriate choice than others. "*CRISP-DM process 1.0: Step-by-step data mining guide*" developed by Chapman et al. (2000) was used in this research for adopting the CRISP-DM process which provided the step-by-step guide to implement the process for this study.

The following main objectives that were set to be achieved using CRISP-DM process in the experiments include broadly the following although the actual objectives are clearly outlined at the experimental stages explained in Chapters 5 and 6.

1. Data mining as pattern discovery process was not found suitable for this research as it was not linked to the business goals and business understanding. CRISP-DM enables this through the first stage (see Figure 4.3) namely business understanding. This step enables the identification of the task to be achieved in terms of the business requirements of the HEI. Thus business problems and goals to be achieved need to be understood accurately.

2. Educational dataset needs to be understood and integrated into the process prior to mining for discovering knowledge. In a usual data mining process there is hardly any control at this stage. The dataset given to the miner will be fed to the algorithm directly without understanding whether the dataset is an educational dataset or not and whether it will be useful to mine and generate results that lead to the achievement of the business goal. In contrast in the CRISP-DM process the data understanding stage and data preparation stages are very critical and complicated which ensure that appropriate dataset required to achieve the business goals are fed to the modelling algorithms. Thus dataset need to be understood carefully and prepared by appropriate formatting to be fed to the algorithms. However here again introduction of educational dataset into the CRISP-DM process needs to tested. Where

there are special features of the educational dataset that need to be taken into consideration while integrating the educational dataset, for instance mining contextual factors not observable (e.g. course difficulty), design science was used to explain how the integration could be achieved. This aspect is discussed in the next section.

3. The modelling stage of a common data mining stage and that of CRISP-DM process is very similar. However CRISP-DM provides a more accurate modelling due to the number of iterations that take place between this stage and the business goals stage to ensure accuracy of the models produced. The objective to be achieved here is to generate the most accurate model to answer the equations 3.9 and 3.10.

4. While in a common data mining process the project completes at this stage and the analysis of the model and the interpretation of the results are left to the data miner, in the CRIP-DM process there is a specific evaluation stage which compares the results of the modelling stage to the business goals to be achieved. If not achieved then again iterations take place until achieved. At this stage the objective to be achieved is to ensure that the pattern discovered and the relationships generated are checked for their validity with regard to the business goals to be achieved.

5. Finally the CRISP-DM process could be deployed after the evaluation is complete.

6. However one important limitation of all KDDM processes including CRISP-DM process that is lack of characterisation of the course taking pattern generated by the processes by contextual factors needed to be addressed. For this CRISP-DM process required modification. Design science was used for explaining both the modification and conducting the experiments leading to development of a modified CRISP-DM process. In addition process theory was used to support the introduction of the modification in the original CRISP-DM process.

After setting the basis and providing the rationale for drafting in the CRISP-DM process into this research the extensive testing of the CRISP-DM process and the development of the modification required to generate contextualised course taking pattern were carried out and explained in Chapters 5 and 6. Following the above the next step taken was to discuss the design science aspects that provided the guideline to develop the methodology to mine the educational dataset and interpret the results.

Process theory that was used to explain how the modification of the CRISP-DM process could be achieved is discussed under Section 5.5 in Chapter 5 where the actual modification has been discussed.

3.5 Application of Design Science Methodology:

Design Science (DS) is of significance in a field that is oriented to the development of successful artefacts particularly in the field of information systems. Design science is a research methodology that should meet the following goals:

a. it is consistent with prior literature
b. it provides a nominal process model for doing DS research, and
c. it provides a mental model for presenting and evaluating DS research in IS.

In addition literature shows that there are different DS methodologies that have been developed by researchers for instance, Peffers et al. (2007) (also see Hevner et al. 2004; Eekels & Roozenburg, 1991; Archer, 1984). There are guidelines developed by researchers to implement DS methodologies. This research adopts the widely used DS methodology developed Hevner et al. (2004). The guidelines and the application of those guidelines to the current research have been narrated in Table 3.6.

Table 3.6, Hevner's Guidelines

Guideline	Description	Application in the research
Guideline 1: Design as an Artefact	The final outcome of Design science research in Information Systems is to construct an artefact that is expected to solve a significant organizational problem. The developed artefact can be a construct, method, model or instantiation. The created artefact and the process of creation both form a part of the design science research process.	The artefact to be designed is the contextual KDDM process model that deals with the problem related to enhancement of student performance in HEIs and decision making pertaining student performance.
Guideline 2: Problem Relevance	The research problem should tackle the problems or issues faced by the organisation.	The problem of enhancing student performance (learning experience) and decisions to be made related to the enhancement are common problems found in HEIs. Any solution to minimise this problem is expected to significantly alter the situation. The modified

		KDDM artefact is expected to provide the solution to the problem.
Guideline 3: Design Evaluation	As the design is an iterative process, the evaluation phase offers essential feedback to the building phase which can then be used to improve the artefact so constructed. The evaluation methods that can be used are 1. Observational (through case studies and field studies) 2. Analytical (through static analysis, architecture analysis, optimization and dynamic analysis) 3. Experimental (through controlled experiments and simulation) 4. Testing (through functional or black box and structural or white box testing) 5. Descriptive (through informed arguments and scenario construction)	This research utilizes observational, analytical, testing and descriptive evaluation methods. Observational involved the study of students of an anonymous university used in the educational dataset as a case study. The attributes of students were studied to develop the dataset and process it to discover knowledge. Analytical – Statistical analysis were carried out to check the performance of the artefact and the quality of the data. Optimisation was carried out through iterations. Testing – it was carried out by subjecting the artefact to predefined testing procedure. Descriptive – this was used to define different scenarios and evaluate artefact against those assumed scenarios.
Guideline 4: Research Contributions	All design science research ought to provide one or more of the following contributions: **1.** The artefact that is designed itself that enables the answers of unsolved problems or solved problems in more efficient and effective ways; **2.** The expansion and enhancement of the knowledge base in the course of the development of novel, appropriately evaluated constructs methods, models or instantiations; and the methodologies in form of use of evaluation methods and proposal of new evaluation metrics.	1. This artefact is expected to contribute to research by enabling the HEIs to solve the unsolved problem of predicting optimum time-to-degree and CGPA using course taking patterns and course difficulty patterns of students. 2. The artefact also is expected to expand the CRISP-DM process to integrate educational dataset and contextual factor mining into the original CRISP-DM process thereby bringing out a modified CRISP-DM artefact.
Guideline 5: Research Rigor	Design science research requires the use of meticulous methods in both the building and assessment of artefacts. Rigor and relevance have to be balanced in the construction and assessment of the artefact. Knowledge of theoretical basics is necessary for building the artefact and the use of adequate assessment techniques as outlined in guideline 3 are needed for its assessment.	The research was rigorous in that it used the complex guidelines of developing and deploying a CRISP-DM process. This guideline has been developed taking into consideration the necessary theories, procedures, assessment techniques, evaluation and deployment of artefacts by Chapman et al. (2000).
Guideline 6: Design as a Search Process	Design is a search process to discover an effective solution to a problem.	An effective solution was to be found to solve equations 3.9 and 3.10 using modified CRISP-DM process as the artefact.
Guideline 7: Communication of Research	Design science research must be efficiently communicated to both professionals concerned with technology as well as management.	The outcomes of the research will be published as a PhD thesis that is appropriate to both technology and management professionals.

The DS methodology described above has been adopted in this research to integrate EDM into the CRISP-DM process model, modify the CRISP-DM process model, test the un-modified and modified CRISP-DM process models and evaluate the model. These steps have been implemented in Chapters 3 and 4. In addition the DS methodology goals were also checked to know whether they have been achieved while conducting and completing the experiments described in Chapters 3 and 4.

3.6 Summary

This chapter has developed the relationship between variables CGPA, time-to-degree, course taking pattern of students and course difficulty pattern using educational dataset. Equations 3.9 and 3.10 have been explained and established. To discover patterns data mining process was chosen as per the literature. However, an example of clustering data mining technique demonstrated that the technique was not adequate to generate course taking pattern accurately and course difficulty taking pattern (see Appendix 24). Literature pointed out that KDDM processes could help in generating accurate course taking pattern accurately and although contextual factors (e.g. course difficulty taking pattern) were outside the purview of KDDM processes. Some of the leading KDDM processes have been discussed and their strengths and limitations have been brought out. CRISP-DM process was chosen to integrate the educational dataset for mining purposes. The CRISP-DM process is argued to lack the ability to mine contextual factors. Hence modification of CRISP-DM process has been suggested to see whether contextualized course taking patterns could be discovered to predict optimum CGPA and time-to-degree. DS methodology has been chosen as the methodology that will be used in this research and has been discussed. Thus the chapter provides the basis for conducting experiments using CRISP-DM model to test the integration of educational data mining in the process, evaluate the CRISP-DM process to check whether contextual factors have been generated and whether it is possible to predict optimum CGPA and time-to-degree as a function of course taking pattern and course difficulty pattern.

Chapter 4 : Integration and Evaluation of EDM in CRISP-DM Process Using Clustering

4.1 Introduction:

At the end of Chapter 3 the rationale for choosing CRISP-DM process to determine the functions in equations 3.9 and 3.10 was given. This chapter uses CRISP-DM process to realise the functions in equations 3.9 and 3.10 and find out whether the KDD process is able to support EDM. This has been done in three steps. In Section 4.2 CRISP-DM specifications have been described to form a basis for testing the CRISP-DM process. In Section 4.3 the original CRISP-DM process developed by Chapman et al. (2000) was chosen to implement EDM in CRISP-DM and the outcomes analysed to know whether EDM can provide support to HEIs in terms of achieving these business goals will enable decision making to improve student learning experience outlined as below

4.1.1 Determine the best course taking pattern and course difficulty that yields the lowest time-to-degree and highest CGPA.

4.1.2 Establish a rule to link course taking pattern and course difficulty, time-to-degree, course difficulty and CGPA.

4.1.3 Predict the optimum time-to-degree and CGPA in terms of the course taking pattern and course difficulty.

Further this section analyses what are the shortcomings in the current CRISP-DM process. In Section 4.3 the integration of EDM in CRISP-DM model using clustering are discussed followed by association rules and classification techniques in Chapter 5. A modified CRISP-DM process has been developed and provided at the end of Chapter 5 which was assessed in Chapter 6 to know whether the modified CRISP-DM process overcomes the limitations found in the original CRISP-DM process discussed in Section 4.3.

Prior to describing the CRISP-DM process it is essential to set the testing criteria. They are:

4.1.4 The business goal to be achieved: Prediction of optimum time-to-degree and CGPA of students using hidden knowledge discovered in educational dataset using a KDDM process and make decisions to improve the student learning experience by providing a model or models that could be implemented in HEIs. In addition demonstrate how EDM is integrated into KDDM process for implementation in HEIs (see Section 3.2).

4.1.5 KDDM process used to discover hidden knowledge and integrate EDM: CRISP-DM process developed by Chapman et al. (2000).

4.1.6 Method used: Use of Design Science method to integrate EDM in the CRISP-DM process described in Section 3.5.

4.1.7 Data: The educational dataset made available by a system called ADREG described in Appendix 11.

4.1.8 Tool: The software tool that was used for performing the tasks of CRISP-DM was IBM SPSS Modeller 17.0 and for some mining purposes Weka 3.7.11.

The results obtained using the CRISP-DM process will be evaluated to know what short comings was there that prevent the use of the CRISP-DM process and how those shortcomings could be overcome.

4.2 CRISP-DM:

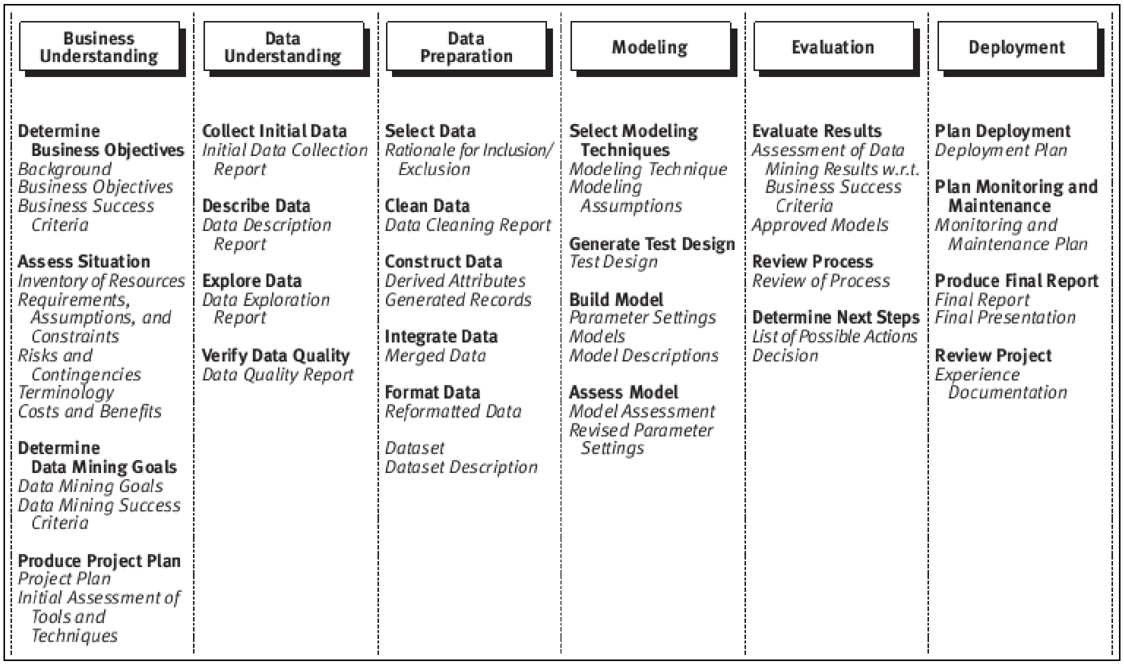
4.2.1 Introduction:

The CRISP-DM process used in this research was the one developed by Chapman et al. (2000).The steps used in this chapter follows the ones described in the document "*CRISP-DM process 1.0: Step-by-step data mining guide*" developed Chapman et al. (2000). The steps include

- Business understanding
- Data understanding
- Data preparation
- Modelling
- Evaluation
- Deployment

Each step has sub-steps which are depicted in Table 3.1. Each one of these steps are discussed in the following sections with regard to realising the functions in equation 3.9 and 3.10 as well as derive patterns to predict the optimum time-to-degree and CGPA of students and make decisions to achieve the business goal.

Table 4.1, Steps in CRISP-DM process



Business Understanding	Data Understanding	Data Preparation	Modeling	Evaluation	Deployment
Determine Business Objectives *Background* *Business Objectives* *Business Success Criteria* **Assess Situation** *Inventory of Resources* *Requirements, Assumptions, and Constraints* *Risks and Contingencies* *Terminology* *Costs and Benefits* **Determine Data Mining Goals** *Data Mining Goals* *Data Mining Success Criteria* **Produce Project Plan** *Project Plan* *Initial Assessment of Tools and Techniques*	**Collect Initial Data** *Initial Data Collection Report* **Describe Data** *Data Description Report* **Explore Data** *Data Exploration Report* **Verify Data Quality** *Data Quality Report*	**Select Data** *Rationale for Inclusion/ Exclusion* **Clean Data** *Data Cleaning Report* **Construct Data** *Derived Attributes* *Generated Records* **Integrate Data** *Merged Data* **Format Data** *Reformatted Data* *Dataset* *Dataset Description*	**Select Modeling Techniques** *Modeling Technique* *Modeling Assumptions* **Generate Test Design** *Test Design* **Build Model** *Parameter Settings* *Models* *Model Descriptions* **Assess Model** *Model Assessment* *Revised Parameter Settings*	**Evaluate Results** *Assessment of Data Mining Results w.r.t. Business Success Criteria* *Approved Models* **Review Process** *Review of Process* **Determine Next Steps** *List of Possible Actions* *Decision*	**Plan Deployment** *Deployment Plan* **Plan Monitoring and Maintenance** *Monitoring and Maintenance Plan* **Produce Final Report** *Final Report* *Final Presentation* **Review Project** *Experience Documentation*

4.2.2 Business Understanding:

The Business Understanding is the initial phase of the KDDM process model. Being the initial phase it exposes the data miner to the business or domain environment. This phase has four main tasks namely determine business objective, assess situation, determine data mining goals and produce project plan.

4.2.2.1 Determine Business Objective::

This task was split into three sub steps namely overview of business, compile business background and finally determine business objective. Initially the user gets an overview of the business, the key personnel involved, main products and expectation from datamining. Compiling business background included determination of organisation structure, describing problem area and proposing a current solution for the problem. Finally the business objective was determined and the business success criteria were established. In the present research the business objective has been set as described in Section 3.2 and using the discussions provided in Section 2.10 and the success criteria are described below which will help demonstrate the achievement of the business objective.

(A) **The business goal to be achieved:** Prediction of optimum time-to-degree and CGPA of students using hidden knowledge discovered in educational dataset using a KDD process and make decisions to improve the student learning experience by providing a model or models that could be implemented in HEIs. In addition demonstrate how EDM is integrated into KDDM process for implementation in HEIs (see Section 3.2). This can be achieved by extracting the functions in equations 3.9 and 3.10 and verifying their validity.

(B) **Business Success criteria**: see sections 4.1.1, 4.1.2 and 4.1.3.

4.2.2.2 Assess Situation:

This task was split into five sub-steps namely identify requirements, clarify assumptions, verify constraints, identify risks and contingencies and estimate costs. In the initial sub step the requirements in terms of personnel, hardware, and data sources were identified. Then the assumptions made in terms of data mining, requirements, success criteria were clarified. Then it was verified if all constraints were considered and if there was any constraint that has not been studied. Further it was checked if there were any risks or contingencies that had to be considered and action plans that were required to be prepared to address to those risks. Also any cost that would to be incurred during the KDDM process model were assessed and analysed.

Assessment of the situation involved an understanding of the business background in terms of available resources (personnel and material) prior to using CRISP-DM.

- *Personnel*: In this research the personnel involved were students, HEI decision makers, registrars and advisors.
- *Data*: Data of graduated students who have graduated since inception of the anonymous university which was involved in the experiment was expected to help in achieving the business goal mentioned. This is termed as educational data mining goal.
- *Risks*: No major risks involved as the experiments were conducted on past data. Anonymity of the students and the university has been maintained.

Sample Inventory of Resources

- **Hardware resource**: The main hardware required was a personal computer on which the data mining tool and SQL server were run. Typical specification used included a minimum 2 GB RAM and 500 GB of Hard disk capacity.
- **Data Sources**: A sample of the past dataset of graduated students consisting of the student code, CGPA, course codes of the courses against which the students had registered their names, order of the semester and time-to-degree (see sample dataset report in Appendix 13)
- **Personnel:** Data mining personnel (data miner), students (provided in the dataset).

4.2.2.3 Determine Educational Data Mining Goals:

In this research the term data mining task is termed as Educational Data mining tasks. This task was further split into three sub-tasks namely describing the type of data mining problem, documenting the technical goals and determining the data mining success criteria. First the type or the technique of data mining problem had to be identified namely clustering, classification, prediction or association. Once the technique was identified, then the technical goals for each data mining problem were identified. Specific unit of time was considered as a criterion because this criterion was not considered at this stage although the original CRISP-DM process recommends this factor to be included. The data mining goals are expected to vary according to the data mining technique chosen. The data mining success criteria were benchmarked using the data mining type determined as explained in section 2.6 of chapter 2.

4.2.2.4 Produce Project Plan:

The project plan is the master document detailing all the steps used in the data mining project. The plan detailed all the phases, the time line, risks expected in each phase and resources needed for the phase. Finally the benefit of using the KDDM process model and the help rendered to attain the business goal were made clear. This helped in determining the success of the KDDM process model in accomplishing the business goals.

Phase	Time line	Risk	Resource	Stat us	Remarks

Business Understanding	3 months		Personnel - Domain expert, Data Miner	Yes No	
Data Understanding	2 months	Lack of data	Personnel – Data Miner IBM SPSS Modeler PC with specification mentioned	Yes No	
Data Preparation	4 months	Data quality issues	Personnel – Data Miner IBM SPSS Modeler PC with specification mentioned	Yes No	
Modelling	2 months	Wrong models	Personnel – Data Miner IBM SPSS Modeler PC with specification mentioned	Yes No	
Evaluation	1 month		Personnel – Data Miner IBM SPSS Modeler PC with specification mentioned	Yes No	
Deployment	1 month		Personnel – Business decision maker	Yes No	

Table 4.2, Sample project plan

The benefit of the KDDM process model is to develop models that could be used in decision making to attain the business goal as mentioned in section 4.1.1 and 4.1.2. This helped in determining the success of the KDDM process model in accomplishing the business goals. The schematic of process model representing Business Understanding is shown below in Figure 4.1.

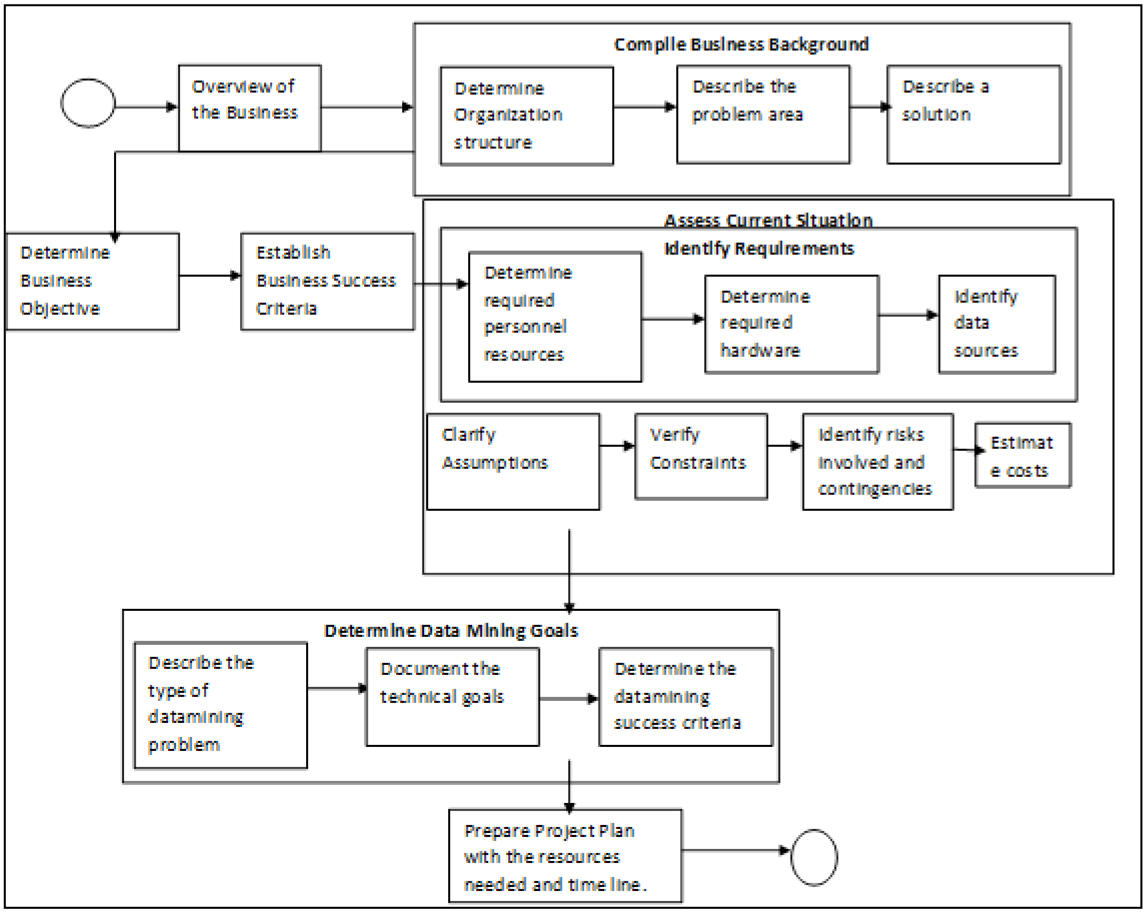

Overview of the Business
Compile Business Background
Determine Organization structure
Describe the problem area
Describe a solution
Determine Business Objective
Establish Business Success Criteria
Assess Current Situation
Identify Requirements
Determine required personnel resources
Determine required hardware
Identify data sources
Clarify Assumptions
Verify Constraints
Identify risks involved and contingencies
Estimat e costs
Determine Data Mining Goals
Describe the type of datamining problem
Document the technical goals
Determine the datamining success criteria
Prepare Project Plan with the resources needed and time line.

Figure 4.1, Schematic of the process model representing business understanding

4.2.3 Data Understanding Phase

The Data Understanding is the second phase of the CRISP-DM process model. This phase deals with data. This phase has four main tasks namely collect initial data, describe data, explore data and verify data quality. Each of the below steps have been explained in section 3.2 and chapter 4 under each technique clustering, association rule, classification and prediction.

(A) Collect Initial data: This task of collecting initial data was split into sub tasks namely determining if existing data sources were sufficient, identifying the need to purchase additional data and need for survey data. The first sub- task checked if the existing data was sufficient to address the business problem and objective. It also checked if additional data was needed to be obtained. Further it was verified if there was a need to collect data through survey which is needed for analysis. By the end of this task a Data Collection Report was prepared which listed the promising attributes from the data source and what attributes needed to be eliminated.

(B) **Describe data:** This task of describing data was split into sub tasks namely determine data size, determining value types and identifying coding schemes. The data size played a valid role in data mining as many techniques and algorithms depend on data size. The data types of the attributes were required to be considered or analysed so that they do not create trouble during modelling as some techniques/algorithms do not work well with certain data types. At the end of this task a Data Description Report was created that detailed the data quantity and data quality.

(C) **Explore data:** This task involves formulation of hypotheses, checking data errors and determining data summaries as sub tasks. Initially the hypotheses that could be formulated from the data were determined. The initial data errors which were caused by typos were determined using tables, charts and other visualisation tools. Data Summaries were used to find discrepancies in data. At the end of this task a Data Exploration Report was created that detailed the promising attributes that were used for the analysis and if there was any change in the data mining goals.

(D) **Verify data quality:** Missing data, coding inconsistency and measurement/data errors were sub- tasks of the main task verify data quality. This task enabled finding many aspects including missing data, data related to incorrect measurements and coding inconsistencies that involved non-standard units of measurement or value inconsistencies. Data Quality report detailed missing values, measurement errors and coding inconsistency. The schematic of process model representing Data Understanding is shown in the Figure 4.2.

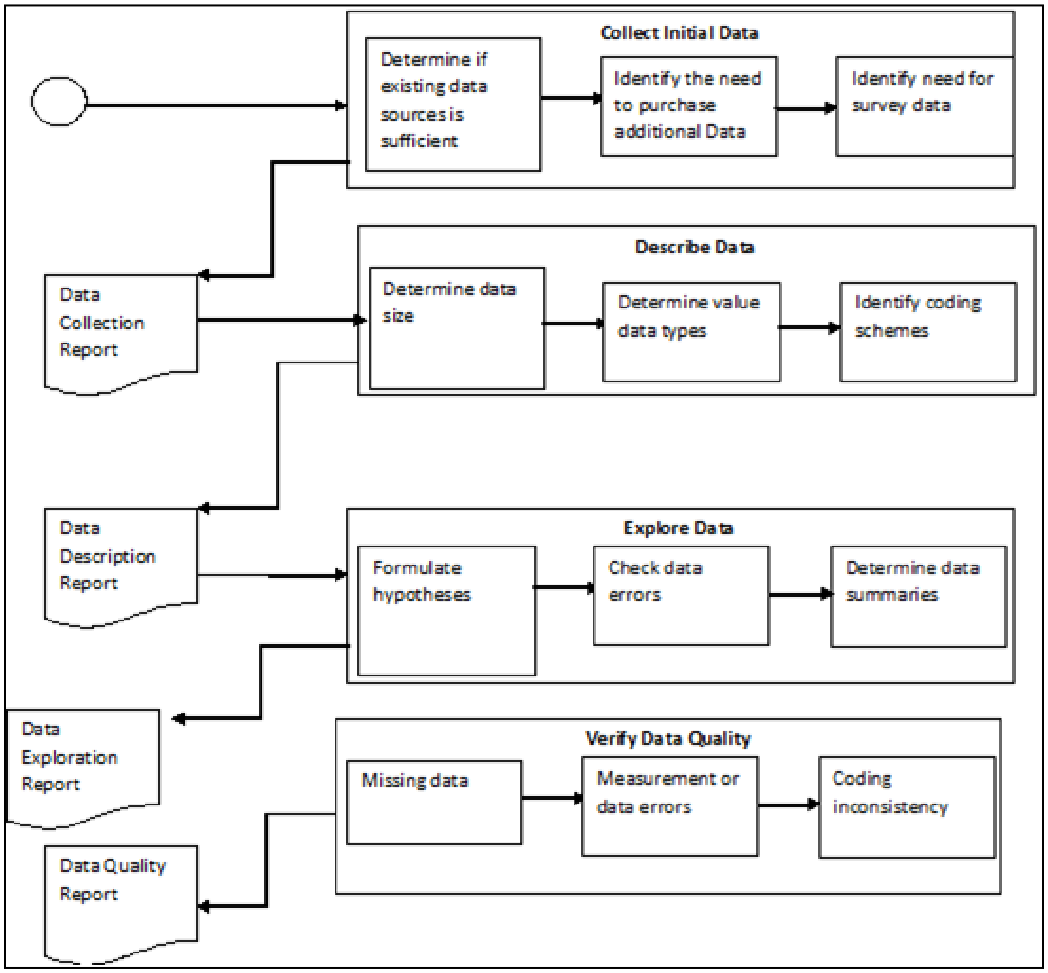


Figure 4.2, Schematic of the process model representing Data understanding

4.2.4 Data Preparation:

This phase covered all activities required to construct the final dataset from the initial raw data. Data preparation tasks were performed multiple times and not in any prescribed order. Tasks included table, record and attribute selection as well as transformation and cleaning of data required for modelling. These tasks must be completed for each one of the techniques identified in section 3.2.2. The steps involved in this phase are discussed below

4.2.4.1. Select Data: This step involved selection of data attributes and records from the initial data collected see section 3.2.3.

4.2.4.2. Clean Data: The Data Quality report generated in the data understanding stage was used as input for this stage. Cleaning phase has three sub-phases as described below:

- **Missing data identification: B**lank and null data were identified and eliminated.
- **Measurement or data error exclusion:** Data with errors see section 3.2.3were excluded.
- **Coding Inconsistency rectification:** Coding of attributes and variables was necessary to feed data into the data mining tool namely IBM SPSS Modeler 17.0. Inconsistency in the coding scheme if any was identified and rectified**.**

A Data Cleaning report was generated at the end of this stage.

4.2.4.3 Construct new data: This consisted of deriving new attributes and generate new records.

- **Derive new attributes:** There was a need to derive new attributes that were not already present in the dataset that should be derived at this step.
- **Generate new records:** If there is a need to generate new records that could be helpful for the analysis then it can be done at this stage. This stage was not needed in this research.

4.2.4.4 Integrate data: Integrate data involved the following 2 steps:

- **Merging Data** involves merging two data sets with similar records but different attributes. This was not required in this research.
- **Appending Data** involves integrating two or more data sets with similar attributes but different records. This was not required in this research.

4.2.4.5 Format data:

There was a need to format the data depending on the requirement of the algorithm chosen to be used in the modelling phase. For instance if a dataset consists of nominal values then clustering algorithm cannot be used and the dataset has to be formatted to suit the algorithm. This step has to be checked under experiments related to each one of the four data mining technique mentioned in section 3.2.2. The schematic of process model representing Data Preparation is shown in Figure 4.3.

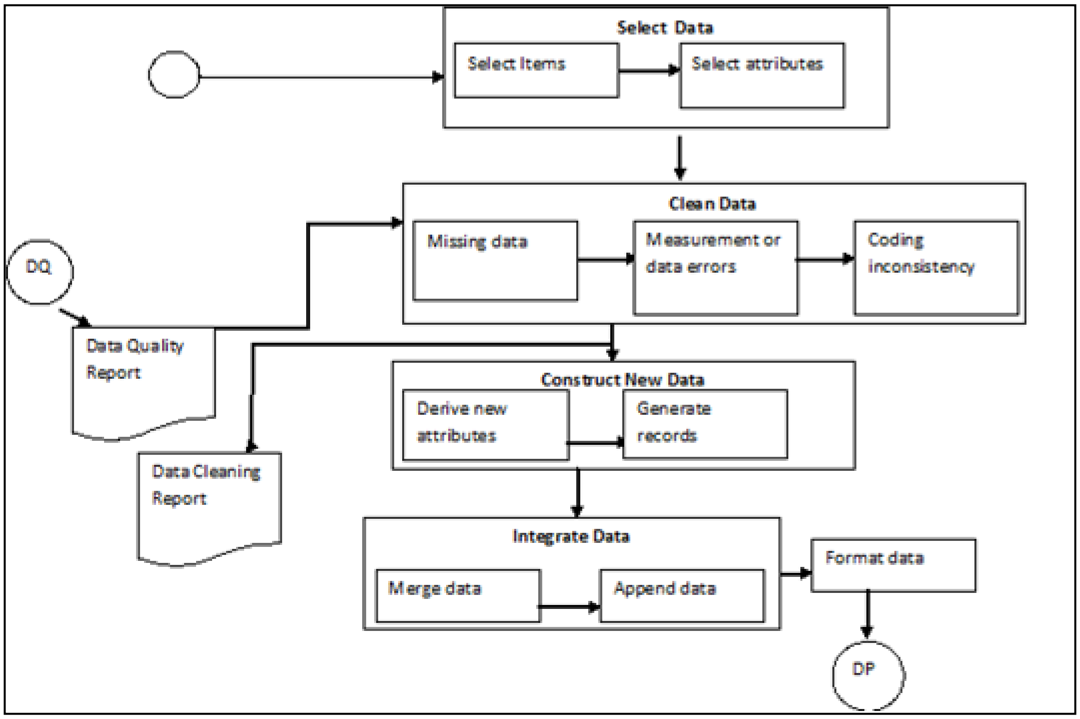


Figure 4.3, Schematic of the process model representing Data Preparation

4.2.5 Modelling

The modelling phase has four sub phases as discussed below.

(A) Select Modelling technique – The Modelling technique was selected by considering three things namely data types, data mining goal and specific modelling requirement. The modelling technique was chosen based on the data type of the targeted variable. The data mining goal was taken into consideration before selecting the modelling technique because the modelling technique will vary according to the data mining goal. For instance, when the data mining goal was set to gain insight into the data by finding hidden patterns, the technique used was clustering which is different from the technique that would be used if the data mining goal was set to find ranks or scores of some transactions. In addition any specific modelling requirement was also taken into account in the selection of modelling technique. For instance the datasets consisting of nominal values will require classification technique whereas numeric values will require regression technique. In this research clustering, classification, association rule and regression techniques were used (see Sections 2.5.2 of chapter 2).

(B) Generate Test Design: Before building the model the testing of the model was planned by generating a test design. There were two parts for generating the test design:

- **Determine criteria for "goodness" of a model:** The measurement of success of the model to be generated was determined here as any model generated has to be tested and evaluated in order to qualify for acceptance. The measurement varies across the different types of techniques and algorithms chosen. (See sections 2.5.2.1, 2.5.2.1.2 and 2.5.2.1.4 of chapter 2).

- **Define data on which criteria will be tested:** In order to apply the criteria referred in the previous section 'Determine criteria' it was necessary to define the data as one of the following 3 namely training dataset, test dataset and validation dataset. This definition depends on the data mining technique and hence has been discussed in detail under the specific data mining technique related experiments under section 4.3.

(C) Build Model: This step consisted of considering the parameter settings that were adjusted to control the modelling process. For instance, in clustering technique setting the number of clusters. Initially the model was built with default parameter settings of 2 clusters as set in the modeller and then modified during the course of the experiment to refine the model by increasing the number step by step till the optimum clusters were generated. Then the modelling algorithms was selected as mentioned in section 2.5.2 of chapter 2 were run to generate the models for instance clusters of students. The chosen algorithm mines the data and generates models that need to be documented (see section 4.3) related to each data mining technique that is used for assessment of the model to decide the best model amongst the lot as defined in the next section.

(D) Assess Model: In this step the data miner assessed the model and interpreted them according to the domain knowledge prevailing with the miner as well as the success criteria set in Section 3.2 of chapter 3 related to each data mining technique and desired test design formulated in Section 2.5.2. At this stage the model generated only is assessed and not the entire KDDM process**.**

The schematic of process model representing Modelling is shown in Figure 4.4.

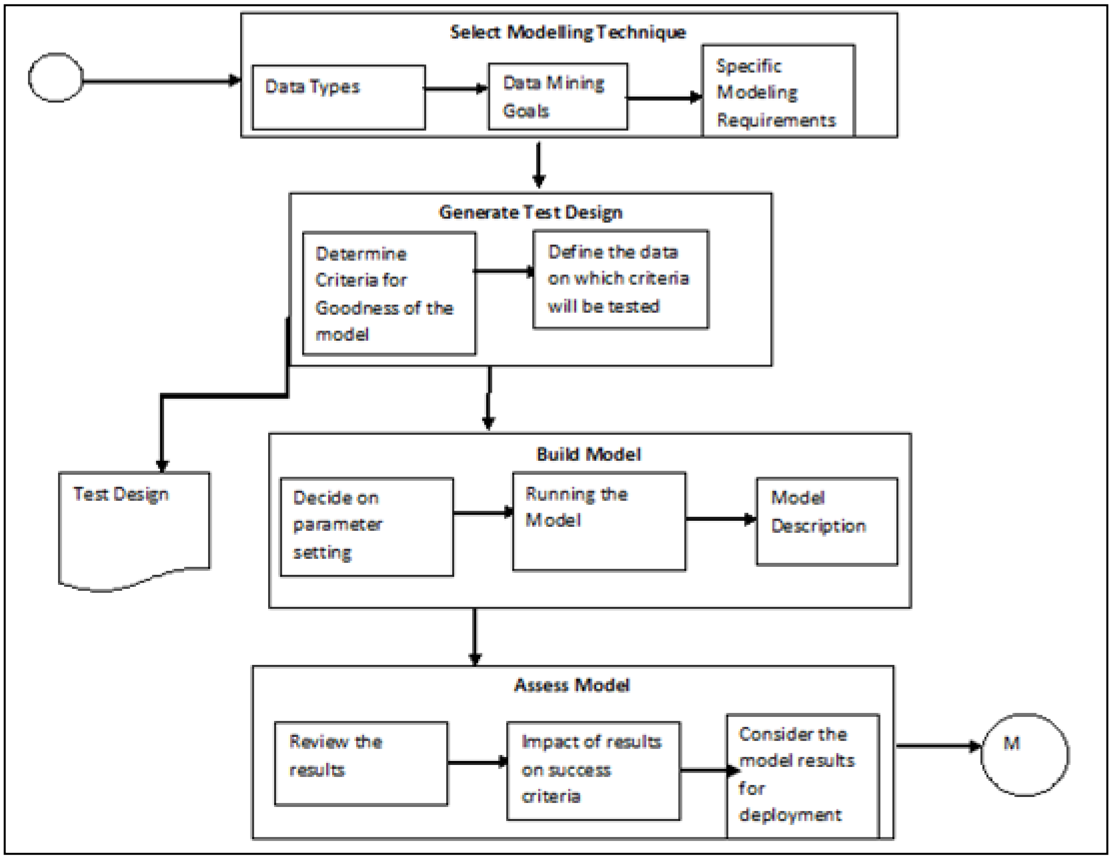

Select Modelling Technique
Data Types
Data Mining Goals
Specific Modeling Requirements
Generate Test Design
Determine Criteria for Goodness of the model
Define the data on which criteria will be tested
Test Design
Build Model
Decide on parameter setting
Running the Model
Model Description
Assess Model
Review the results
Impact of results on success criteria
Consider the model results for deployment
M

Figure 4.4, Schematic of the process model representing Modelling

4.2.6 Evaluation

The evaluation process consisted of three sub phases namely evaluation of results from modelling phase, reviewing the CRISP-DM process and determining the next steps.

(A) Evaluate Results: The evaluation of results was done by ranking the different models that were generated in order to deploy the highly ranked model. In addition evaluation involved testing whether the business goal could be achieved by the model(s) generated. Further it was checked if the findings from the models were unique and novel.

(B) Review Process: The review process included checking whether there was any scope for improving the model or results. The failures or mistakes in the review process were documented so that in future those failures or mistakes could be avoided. In addition the review included assessing whether the findings have raised more questions for which answers have to be explored.

(C) Determine the next steps: Decisions were made by the data miner at this stage to proceed with deployment of the results or refine the models if the results are not satisfactory or inaccurate.

The schematic of process model representing Evaluation is shown in Figure 4.5.

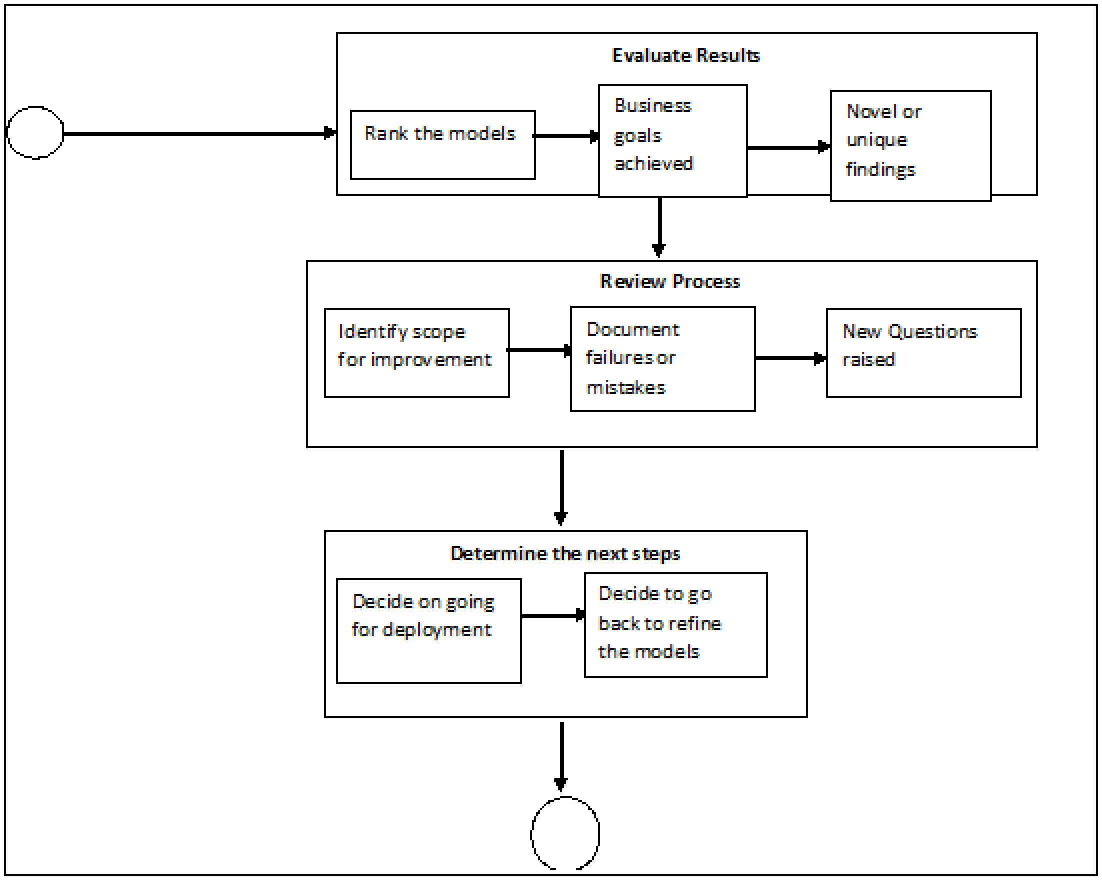


Figure 4.5, Schematic of the process model representing Evaluation

4.2.7 Deployment

The deployment phase has four stages namely deployment plan, monitoring and maintenance plan, production of final report and project review. Since the actual deployment is a complex process that requires a strategy for actual implementation of the results on real time basis by disturbing the organisation’s work flow, it was concluded that the deployment will be tested by simulating the performance of the CRISP-DM process for addressing different business problems. In addition there are recommendations by researchers which indicate that an accurately evaluated process is likely to be implemented successfully (Chapman et al. 2000). Based on these two criteria experiments were conducted on CRISP-DM process in this research. However for the sake of completeness of the study of CRISP-DM process the four stages mentioned above have been discussed next which need to be considered in an actual project.

(A) Deployment plan:

This stage considered deployment plan of the result. First, the results need to be summarized. A step-by-step plan for deployment needs to be prepared for each of deployable model. A plan has to be developed to disseminate this information to strategy or policy makers. Alternative deployment plans have to be considered also in case of any failure with the deployment plan. Identify any deployment problems and plan for contingencies.

(B) Monitoring and Maintenance Plan:

The monitoring and maintenance plan needs to be produced. The model expiry, what determines the model expiry and if the model can be reused if expired must be decided. A critical documentation must be prepared detailing this.

(C) Production of Final Report:

A final report detailing all the steps that have been left out in the previous steps needs to be documented. It could be used to broadcast the results. The report should include a thorough description of the original business problem, the process used to conduct data mining and cost of the project.

(D) Final Project Review:

A final review of the project is usually conducted to document the learning process, difficulties and overall impressions of the project.

4.3. Determination of the best course taking pattern and course difficulty that yields the lowest time-to-degree and highest CGPA using CRISP-DM process model 1

This section presents the results of the analysis of the CRISP-DM process model. The experiment was conducted based on the steps outlined in Section 4.2. As mentioned in Section 4.2 and the primary step that decided the conduct of the experiments was the assumption of the model that will be generated at the data mining stage by the data miner based on the business understanding of the miner which is aligned with the business goals of the organisation. This assumption led the data miner to conduct the experiment as outlined in Section 4.2. Three different models have been generated through experiments in this research to achieve each one of the business goals mentioned above. Each one of the experiments has provided distinct knowledge hidden in the dataset. In the absence of knowledge about how to make decisions to enhance the learning experience of student by predicting time-to-degree or CGPA or both (that is solving equations 3.9 and 3.10) in terms of the course taking pattern of students and other attributes it is necessary to generate different models and choose the most appropriate one satisfying the purpose. To begin with the following sections provide details of the experiments conducted on CRISP-DM model using clustering, association, classification and prediction techniques. Experiments were conducted following all the steps beginning with business understanding till evaluation of the CRISP-DM process model outlined in sections 4.2.1 to 4.2.7.

4.3.1 Business Understanding:

- **Determine business objectives –** To profile students by determining the best course taking pattern that yields the lowest time-to-degree and highest CGPA.
- **Assessment of situation –** The situation was assessed and found to be the same as outlined in Section 4.2.2.2.
- **Determination of Educational data mining (EDM) goals**
  - **a. Identification of data mining technique:** It was assumed that students could be clustered with common attributes namely course taking patterns, CGPA and time-to-degree to achieve the business goal. Therefore clustering technique was found to be the most suitable technique to achieve the business goal.
  - **b. Documenting the technical goals**: Clustering attributes based on specific measurement of number of clusters, cluster distribution; number of iterations, time taken to generate the model and accuracy in terms of percentage of pairs of tuples in the same cluster that share common label were used to achieve the technical goal of identifying the clusters. In addition Sum of Squared Error (SSE) was used as a measure of checking the quality of clusters only for evaluating k-means algorithm.

$$SSE = \sum_{i=1}^{K} \sum_{x \in C_i} dist^2(m_i, x)$$

x is a data point in cluster Ci and mi is the representative point for cluster Ci can show that mi corresponds to the center (mean) of the cluster

It must be noted that the cluster number (k value) that will be accepted when using k-means cluster is that which satisfies the condition that the SSE is the lowest amongst the cluster numbers tested. For example if the k-means cluster was used to test the data for k-values starting from 2 and ending with 10, then that result of the k-means cluster algorithm will be accepted which has the lowest SSE.

c. **Data mining goal success criteria:**

Ranking the clusters based on a comparison of the parameters namely number of clusters, cluster distribution; number of iterations, time taken to generate the model and accuracy in terms of percentage of pairs of tuples in the same cluster that share common label.

Data mining goal is to cluster the student records in terms of course taking patterns, CGPA and time-to-degree.

This step required iterations in order to determine the type of algorithm to be used and preparation of data.

4.3.2 Data Understanding:

- **Collect initial data -** Initial data is that dataset which consists of data about various attributes using which the data miner can create a final dataset that could be mined to enable achievement of the business goal. This data set was extracted from the ADREG system as mentioned in Appendix.11. Initially the dataset comprised 44 fields. Sample provided in Appendix 1 which includes the data collection report.

- **Describe data**

  Data size – 440 students.

  Determining value types - a combination of numeric, categorical and nominal values were present.

  Identifying coding schemes –alphanumeric coding was used to represent attributes uniquely.

  Sample Data Description report is provided in Appendix 2.

- **Explore data** – Initially two hypotheses were formulated by an inspection of the collected data (see data description report in Appendix 2). They are:

  - *Hypothesis A: CGPA can be expressed as a function of student course registration data by semester, time-to-degree and course difficulty by clustering.*

  - *Hypothesis B: Time-to-degree can be expressed as a function of student course registration data by semester, CGPA and course difficulty by clustering.*

  Although more hypotheses could be formulated this research has provided examples of only two hypotheses which are expected to be sufficient to demonstrate how the business goals were achieved using an EDM that was integrated into the CRISP-DM process model.

  The final hypothesis was formulated after exploration of data and an analysis of initial hypothesis. The analysis of the hypothesis showed that course registration data by semester could be treated as a pattern of courses and considered as a unique attribute. Besides as mentioned in chapters 2 and 3 it was assumed that course difficulty could be hidden in the dataset and could be discovered during data mining process and hence included as a function. In addition data exploration revealed that time-to-degree could be a function of the course taking pattern of students by semester and CGPA. Alternatively CGPA could also be considered as a function of course taking pattern and time-to-degree. These hypotheses have already been identified in chapter 2 in equations 3.9 and 3.10. The discussion in chapter 2 can be seen to match approximately to the data explored in this section.

  The data exploration reports are checked for distribution and relationships.

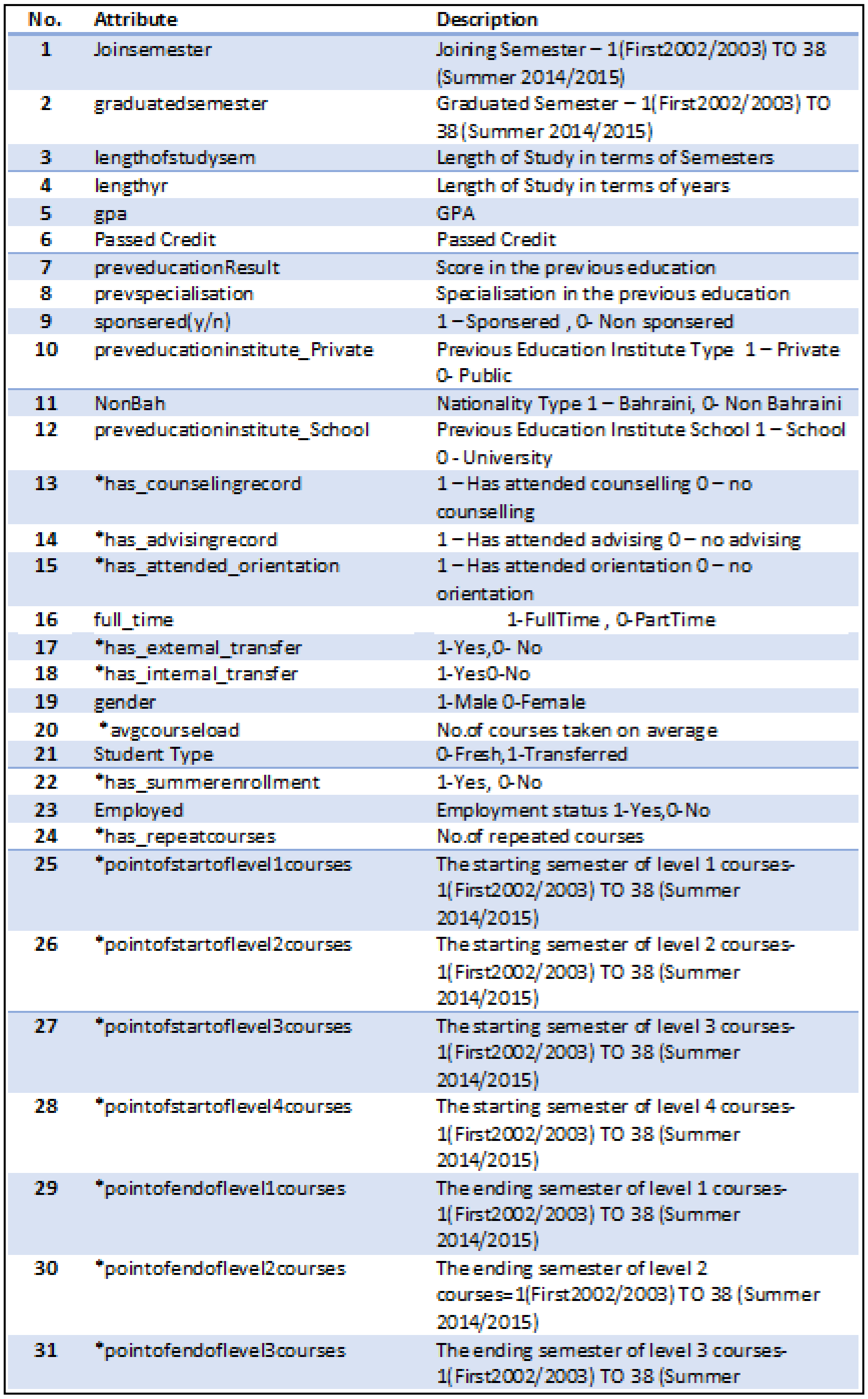
- **Verify data quality -** . The dataset was assessed for quality problems including missing values, extreme values, measurement or data error and coding inconsistencies. The outliers are removed from the dataset and are not used for the analysis. The data quality report consisting of the above is contained in Appendix 3.

A number of iterations had to be introduced in understanding data with regard to various steps.

4.3.3 Data Preparation:

- **Select data –**This dataset was extracted from the initial dataset described in section 4.3.2.The final dataset comprises 44 attributes which were used for mining. The dataset pertains to students belonging to 12 programmes and graduated during the period 2003 to 2014.The data size was 440. Data stored in various tables was joined in a single table in this stage.
- **Clean data –** The data was cleaned again by checking for missing values, measurement or data error and coding inconsistency and was found to the satisfactory and the variables were coded appropriately to enable the application of clustering algorithms.
- **Construct new data-** Some variables were extracted directly from the database since basically they were already present in the database. Some features prefixed with a * in the Table 5.3 were considered as new attributes as those attributes were derived through calculations from other attributes present in different tables.(See Table 4.3).
- **Integrate data –** This step was not required as no new dataset were either merged or appended.
- **Format data –** The data was formatted to suit the clustering algorithm by transforming all non-numeric data types to numeric data types.

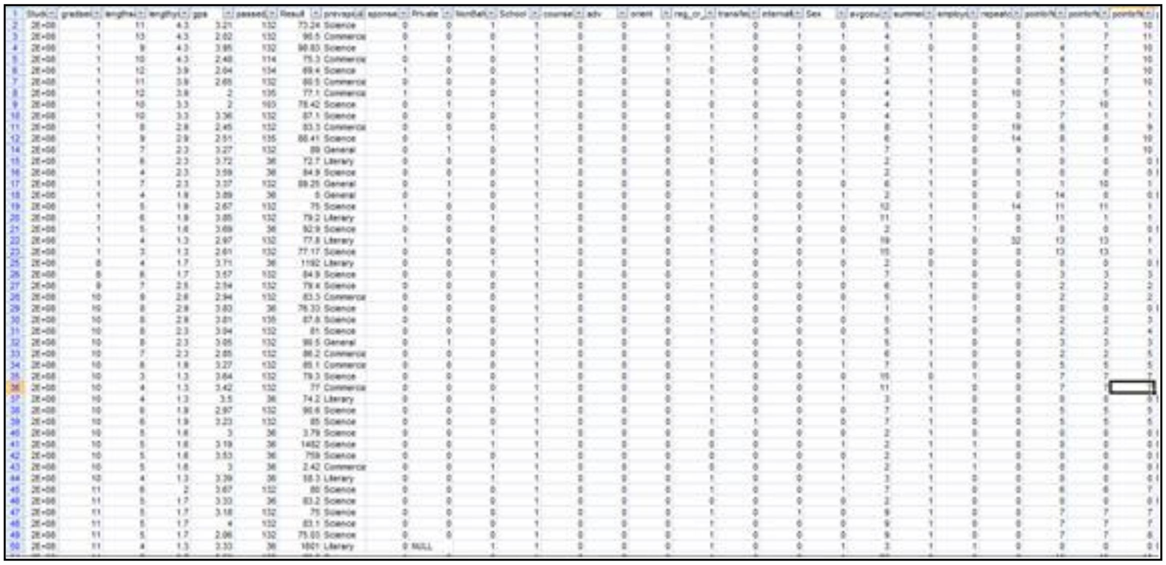
Table 4.3, Dataset used for clustering



No.	Attribute	Description
1	Joinsemester	Joining Semester – 1(First2002/2003) TO 38 (Summer 2014/2015)
2	graduatedsemester	Graduated Semester – 1(First2002/2003) TO 38(Summer 2014/2015)
3	lengthofstudysem	Length of Study in terms of Semesters
4	lengthyr	Length of Study in terms of years
5	gpa	GPA
6	Passed Credit	Passed Credit
7	preveducationResult	Score in the previous education
8	prevspecialisation	Specialisation in the previous education
9	sponsered(y/n)	1 –Sponsered , 0- Non sponsered
10	preveducationinstitute_Private	Previous Education Institute Type 1 – Private 0- Public
11	NonBah	Nationality Type 1 – Bahraini, 0- Non Bahraini
12	preveducationinstitute_School	Previous Education Institute School 1 – School 0 - University
13	*has_counselingrecord	1 – Has attended counselling 0 – no counselling
14	*has_advisingrecord	1 – Has attended advising 0 – no advising
15	*has_attended_orientation	1 – Has attended orientation 0 – no orientation
16	full_time	1-FullTime , 0-PartTime
17	*has_external_transfer	1-Yes,0- No
18	*has_internal_transfer	1-Yes0-No
19	gender	1-Male 0-Female
20	*avgcourseload	No.of courses taken on average
21	Student Type	0-Fresh,1-Transferred
22	*has_summerenrollment	1-Yes, 0-No
23	Employed	Employment status 1-Yes,0-No
24	*has_repeatcourses	No.of repeated courses
25	*pointofstartoflevel1courses	The starting semester of level 1 courses-1(First2002/2003) TO 38 (Summer 2014/2015)
26	*pointofstartoflevel2courses	The starting semester of level 2 courses-1(First2002/2003) TO 38 (Summer 2014/2015)
27	*pointofstartoflevel3courses	The starting semester of level 3 courses-1(First2002/2003) TO 38 (Summer 2014/2015)
28	*pointofstartoflevel4courses	The starting semester of level 4 courses-1(First2002/2003) TO 38 (Summer 2014/2015)
29	*pointofendoflevel1courses	The ending semester of level 1 courses-1(First2002/2003) TO 38 (Summer 2014/2015)
30	*pointofendoflevel2courses	The ending semester of level 2 courses=1(First2002/2003) TO 38 (Summer 2014/2015)
31	*pointofendoflevel3courses	The ending semester of level 3 courses-1(First2002/2003) TO 38 (Summer

32	*pointofendoflevel4courses	The ending semester of level 4 courses-1(First2002/2003) TO 38 (Summer 2014/2015)
33	*no_of_failures	The number of failures or F Grades
34	*no_of_withdrawals	The number of withdrawals from the courses or W Grades
35	*pointofstartofcorecourses	The starting semester of core courses-1(First2002/2003) TO 38 (Summer 2014/2015)
36	*pointofstartofcoreeleccourses	The starting semester of core elective courses-1(First2002/2003) TO 38 (Summer 2014/2015)
37	*pointofstartoffreeeleccourses	The starting semester of free elective courses-1(First2002/2003) TO 38 (Summer 2014/2015)
38	*pointofstartofhumcourses	The starting semester of humanity courses-1(First2002/2003) TO 38 (Summer 2014/2015)
39	*pointofendofcorecourses	The ending semester of core courses-1(First2002/2003) TO 38 (Summer 2014/2015)
40	*pointofendofcoreeleccourses	The ending semester of core elective courses-1(First2002/2003) TO 38 (Summer 2014/2015)
41	*pointofendoffreeeleccourses	The ending semester of free elective courses-1(First2002/2003) TO 38 (Summer 2014/2015)
42	*pointofendofhumcourses	The ending semester of humanity courses-1(First2002/2003) TO 38 (Summer 2014/2015)
43	*Student Class	OK,Good,Very Good,Excellent
44	Student Programme	Programme In which the student is enrolled

Table 4.4, Dataset that was input for clustering



1	Stud	gradsei	lengthsi	lengthyi	gpa	passed	Result	prevspecial	sponsoi	Private	NonBah	School	course	adv	orient	reg_cr_	transfei	internati	Sex	avgcou	summei	employi	repeatc	pointofs	pointofs	pointofs
2	2E+08	1	11	4.3	3.21	132	73.24	Science	0	0	1	1	0	0	1	1	0	1	0	5	1	0	0	1	1	10
3	2E+08	1	13	4.3	2.62	132	90.5	Commerce	0	0	0	1	0	0	1	1	0	1	1	4	1	0	5	1	7	11
4	2E+08	1	9	4.3	3.95	132	90.83	Science	1	1	1	1	0	0	0	1	0	0	0	6	0	0	0	4	7	10
5	2E+08	1	10	4.3	2.48	114	75.3	Commerce	0	0	0	1	0	0	1	1	0	1	0	4	1	0	0	4	7	10
6	2E+08	1	12	3.9	2.04	134	69.4	Science	1	0	0	1	0	0	1	0	0	0	1	3	1	0	0	5	8	10
7	2E+08	1	11	3.9	2.65	132	60.5	Commerce	0	0	0	1	0	0	0	1	0	0	0	4	1	0	0	5	7	10
8	2E+08	1	12	3.8	2	135	77.1	Commerce	1	0	0	1	0	0	0	1	1	0	0	4	1	0	10	1	5	1
9	2E+08	1	10	3.3	2	103	78.42	Science	0	1	1	1	0	0	0	0	0	0	1	4	1	0	3	7	10	1
10	2E+08	1	10	3.3	3.36	132	87.1	Science	0	0	0	1	0	0	0	1	0	0	0	4	1	0	0	7	1	1
11	2E+08	1	8	2.8	2.45	132	83.3	Commerce	0	0	0	1	0	0	0	1	1	0	1	6	1	0	10	6	8	9
12	2E+08	1	9	2.9	2.51	135	88.41	Science	0	1	1	1	0	0	0	1	1	0	1	6	1	0	14	6	8	10
14	2E+08	1	7	2.3	3.27	132	89	General	0	1	0	1	0	0	0	1	1	0	1	7	1	0	9	1	1	10
15	2E+08	1	8	2.3	3.72	36	72.7	Literary	0	0	0	1	0	0	0	0	0	0	1	2	1	0	1	9	6	0.1
16	2E+08	1	4	2.3	3.59	36	84.9	Science	0	0	0	1	0	0	0	1	0	0	1	2	1	0	0	6	6	0.1
17	2E+08	1	7	2.3	3.37	132	89.25	General	0	1	0	1	0	0	0	1	1	0	0	6	1	0	1	1	10	1
18	2E+08	1	4	1.9	3.89	36	5	General	0	1	0	1	0	0	0	1	0	0	1	2	1	0	0	14	6	0.1
19	2E+08	1	5	1.9	2.67	132	75	Science	1	0	0	1	0	0	0	1	1	0	1	12	1	0	14	11	11	1
20	2E+08	1	6	1.9	3.85	132	79.2	Literary	1	0	1	1	0	0	0	1	0	1	1	11	1	1	0	11	1	1
21	2E+08	1	5	1.6	3.69	36	92.9	Science	0	0	0	1	0	0	0	0	0	0	0	2	1	1	0	8	0	0.1
22	2E+08	1	4	1.3	2.97	132	77.8	Literary	1	0	0	1	0	0	0	1	1	0	0	10	1	0	32	13	13	1
23	2E+08	1	3	1.3	2.61	132	77.17	Science	0	0	0	1	0	0	0	1	0	0	1	15	0	0	0	13	13	1
25	2E+08	8	4	1.7	3.71	36	1192	Literary	0	0	1	1	0	0	0	1	0	0	0	2	0	0	0	0	0	0.1
26	2E+08	8	6	1.7	3.67	132	84.9	Science	0	0	0	1	0	0	0	1	0	1	1	7	1	0	0	3	3	3
27	2E+08	9	7	2.5	2.54	132	79.4	Science	0	0	0	1	0	0	0	1	0	0	0	6	1	0	0	2	2	2
28	2E+08	10	9	2.6	2.94	132	83.5	Commerce	0	0	0	1	0	0	0	1	0	0	0	5	1	0	0	2	2	2
29	2E+08	10	8	2.8	3.83	36	76.33	Science	0	0	0	1	0	0	0	1	0	0	1	1	1	1	0	0	0	0.1
30	2E+08	10	8	2.8	3.81	135	87.8	Science	0	0	0	1	0	0	0	1	0	0	0	5	1	0	0	2	2	3
31	2E+08	10	8	2.3	3.04	132	81	Science	0	0	0	1	0	0	0	1	0	0	0	5	1	0	1	2	2	4
32	2E+08	10	8	2.3	3.05	132	90.5	General	0	1	0	1	0	0	0	1	0	0	1	5	1	0	0	3	3	3
33	2E+08	10	7	2.3	2.85	132	86.2	Commerce	0	0	0	1	0	0	0	1	0	0	1	6	1	0	0	2	2	5
34	2E+08	10	8	1.8	3.27	132	85.1	Commerce	0	0	0	1	0	0	0	1	0	0	1	7	1	0	0	5	5	5
35	2E+08	10	3	1.3	3.64	132	79.3	Science	0	0	0	1	0	0	0	1	0	0	0	15	0	1	0	7	7	7
36	2E+08	10	4	1.3	3.42	132	77	Commerce	0	0	0	1	0	0	0	1	0	0	1	11	1	0	0	7	7	[illegible]
37	2E+08	10	4	1.3	3.5	36	74.2	Literary	0	0	1	1	0	0	0	1	0	0	1	3	1	0	0	0	0	0.1
38	2E+08	10	6	1.9	2.97	132	96.6	Science	0	0	0	1	0	0	0	1	0	0	0	7	1	0	0	5	5	5
39	2E+08	10	6	1.9	3.23	132	85	Science	0	0	0	1	0	0	0	1	1	0	0	7	1	0	0	5	5	5
40	2E+08	10	5	1.6	3	36	3.79	Science	0	0	1	1	0	0	0	0	0	0	0	2	1	0	0	0	0	0.1
41	2E+08	10	5	1.6	3.19	36	1482	Science	0	0	1	1	0	0	0	0	0	0	1	2	1	1	0	0	0	0.1
42	2E+08	10	5	1.6	3.53	36	759	Science	0	0	0	1	0	0	0	0	0	0	0	2	1	1	0	0	0	0.1
43	2E+08	10	5	1.6	3	36	2.42	Commerce	0	0	1	1	0	0	0	0	0	0	1	2	1	1	0	0	0	0.1
44	2E+08	10	4	1.3	3.39	36	58.3	Literary	0	0	1	1	0	0	0	1	0	0	1	3	1	0	0	0	0	0.1
45	2E+08	11	6	2	3.67	132	80	Science	0	0	0	1	0	0	0	1	0	0	1	7	1	0	0	6	6	7
46	2E+08	11	5	1.7	3.33	36	83.2	Science	0	0	0	1	0	0	0	0	0	0	0	2	1	0	0	0	0	0.1
47	2E+08	11	5	1.7	3.18	132	75	Science	0	0	0	1	0	0	0	1	0	1	0	9	1	0	0	7	7	7
48	2E+08	11	5	1.7	4	132	83.1	Science	0	0	0	1	0	0	0	1	0	0	0	9	1	0	0	7	7	7
49	2E+08	11	5	1.7	2.06	132	75.03	Science	0	0	0	1	0	0	0	1	0	0	0	9	1	0	0	7	7	8
50	2E+08	11	4	1.3	3.33	36	1601	Literary	0	NULL	1	1	0	0	0	1	0	0	1	3	1	1	0	0	0	0.1

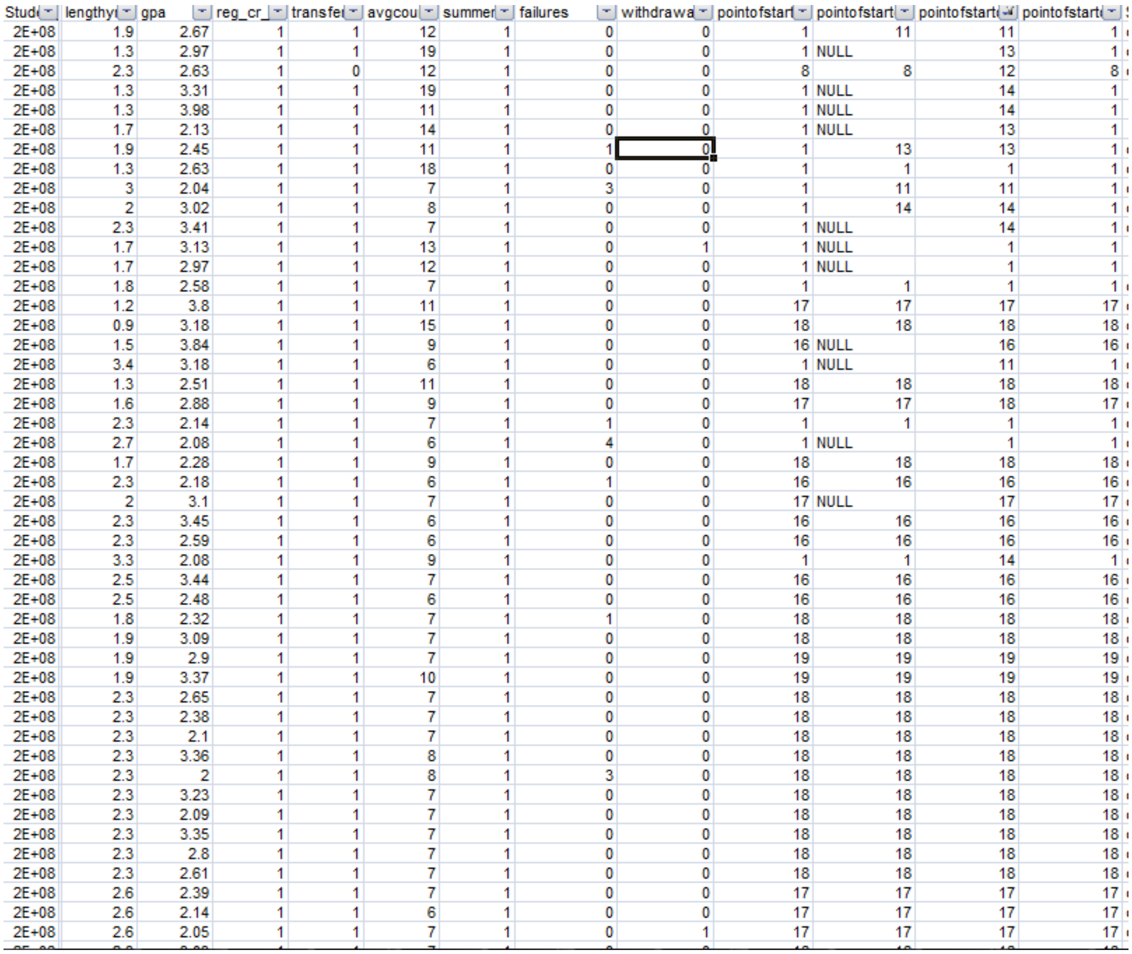
Based on the examination of Table 4.4 it was decided that the number of features that need to be use for mining be limited to 12 (See Table 4.5) because the other features when inspected carefully were not found to affect the time-to-degree or CGPA or course taking patterns and hence were not used as part of the dataset.

Table 4.5, Attributes that were selected for clustering analysis

No.	Attribute	Description
1	Lengthyr	Length of Study in terms of years
2	CGPA	CGPA
3	*avgcourseload	No.of courses taken on average
4	Student Type	0-Fresh,1-Transferred
5	*has_summerenrollment	1-Yes , 0-No
6	*no_of_failures	The number of failures or F Grades
7	*no_of_withdrawals	The number of withdrawals from the courses or W Grades
8	*pointofstartofcorecourses	The starting semester of core courses-1(First2002/2003) TO 38 (Summer 2014/2015)
9	*pointofstartofcoreeleccourses	The starting semester of core elective courses-1(First2002/2003) TO 38 (Summer 2014/2015)
10	*pointofstartoffreeeleccourses	The starting semester of free elective courses-1(First2002/2003) TO 38 (Summer 2014/2015)
11	*pointofstartofhumcourses	The starting semester of humanity courses-1(First2002/2003) TO 38 (Summer 2014/2015)
12	*Student Class	OK, Good, Very Good, Excellent

The input given to Weka Clustering is as follows which represents the fields described in Table 4.6.

Table 4.6, Input dataset given for Cluster modelling



Stud	lengthy	gpa	reg_cr_	transfe	avgcou	summer	failures	withdrawa	pointofstart	pointofstart	pointofstart	pointofstart
2E+08	1.9	2.67	1	1	12	1	0	0	1	11	11	1
2E+08	1.3	2.97	1	1	19	1	0	0	1	NULL	13	1
2E+08	2.3	2.63	1	0	12	1	0	0	8	8	12	8
2E+08	1.3	3.31	1	1	19	1	0	0	1	NULL	14	1
2E+08	1.3	3.98	1	1	11	1	0	0	1	NULL	14	1
2E+08	1.7	2.13	1	1	14	1	0	0	1	NULL	13	1
2E+08	1.9	2.45	1	1	11	1	1	0	1	13	13	1
2E+08	1.3	2.63	1	1	18	1	0	0	1	1	1	1
2E+08	3	2.04	1	1	7	1	3	0	1	11	11	1
2E+08	2	3.02	1	1	8	1	0	0	1	14	14	1
2E+08	2.3	3.41	1	1	7	1	0	0	1	NULL	14	1
2E+08	1.7	3.13	1	1	13	1	0	1	1	NULL	1	1
2E+08	1.7	2.97	1	1	12	1	0	0	1	NULL	1	1
2E+08	1.8	2.58	1	1	7	1	0	0	1	1	1	1
2E+08	1.2	3.8	1	1	11	1	0	0	17	17	17	17
2E+08	0.9	3.18	1	1	15	1	0	0	18	18	18	18
2E+08	1.5	3.84	1	1	9	1	0	0	16	NULL	16	16
2E+08	3.4	3.18	1	1	6	1	0	0	1	NULL	11	1
2E+08	1.3	2.51	1	1	11	1	0	0	18	18	18	18
2E+08	1.6	2.88	1	1	9	1	0	0	17	17	18	17
2E+08	2.3	2.14	1	1	7	1	1	0	1	1	1	1
2E+08	2.7	2.08	1	1	6	1	4	0	1	NULL	1	1
2E+08	1.7	2.28	1	1	9	1	0	0	18	18	18	18
2E+08	2.3	2.18	1	1	6	1	1	0	16	16	16	16
2E+08	2	3.1	1	1	7	1	0	0	17	NULL	17	17
2E+08	2.3	3.45	1	1	6	1	0	0	16	16	16	16
2E+08	2.3	2.59	1	1	6	1	0	0	16	16	16	16
2E+08	3.3	2.08	1	1	9	1	0	0	1	1	14	1
2E+08	2.5	3.44	1	1	7	1	0	0	16	16	16	16
2E+08	2.5	2.48	1	1	6	1	0	0	16	16	16	16
2E+08	1.8	2.32	1	1	7	1	1	0	18	18	18	18
2E+08	1.9	3.09	1	1	7	1	0	0	18	18	18	18
2E+08	1.9	2.9	1	1	7	1	0	0	19	19	19	19
2E+08	1.9	3.37	1	1	10	1	0	0	19	19	19	19
2E+08	2.3	2.65	1	1	7	1	0	0	18	18	18	18
2E+08	2.3	2.38	1	1	7	1	0	0	18	18	18	18
2E+08	2.3	2.1	1	1	7	1	0	0	18	18	18	18
2E+08	2.3	3.36	1	1	8	1	0	0	18	18	18	18
2E+08	2.3	2	1	1	8	1	3	0	18	18	18	18
2E+08	2.3	3.23	1	1	7	1	0	0	18	18	18	18
2E+08	2.3	2.09	1	1	7	1	0	0	18	18	18	18
2E+08	2.3	3.35	1	1	7	1	0	0	18	18	18	18
2E+08	2.3	2.8	1	1	7	1	0	0	18	18	18	18
2E+08	2.3	2.61	1	1	7	1	0	0	18	18	18	18
2E+08	2.6	2.39	1	1	7	1	0	0	17	17	17	17
2E+08	2.6	2.14	1	1	6	1	0	0	17	17	17	17
2E+08	2.6	2.05	1	1	7	1	0	1	17	17	17	17

A number of iterations had to be introduced in preparation of data with regard to various steps.

4.3.4 Modelling - Building a Clustering Model:

(A) Select Modelling technique: Two algorithms were selected namely k-means and EM clustering. K-means clustering is a widely used technique that is considered to be very simple to use when compared others. EM clustering was chosen as an alternative to compare the performance k-means clustering. EM clustering has a special feature to choose its own k-value unlike any other technique where the k-value must be fixed by the data miner. From prior research it was found that these two algorithms can be used to mine data and generate models.

(B) Generate Test Design:

- **Criteria to determine the goodness of the model:** These criteria were the same as mentioned Section 4.3.1 which includes verifying the sum squared error.

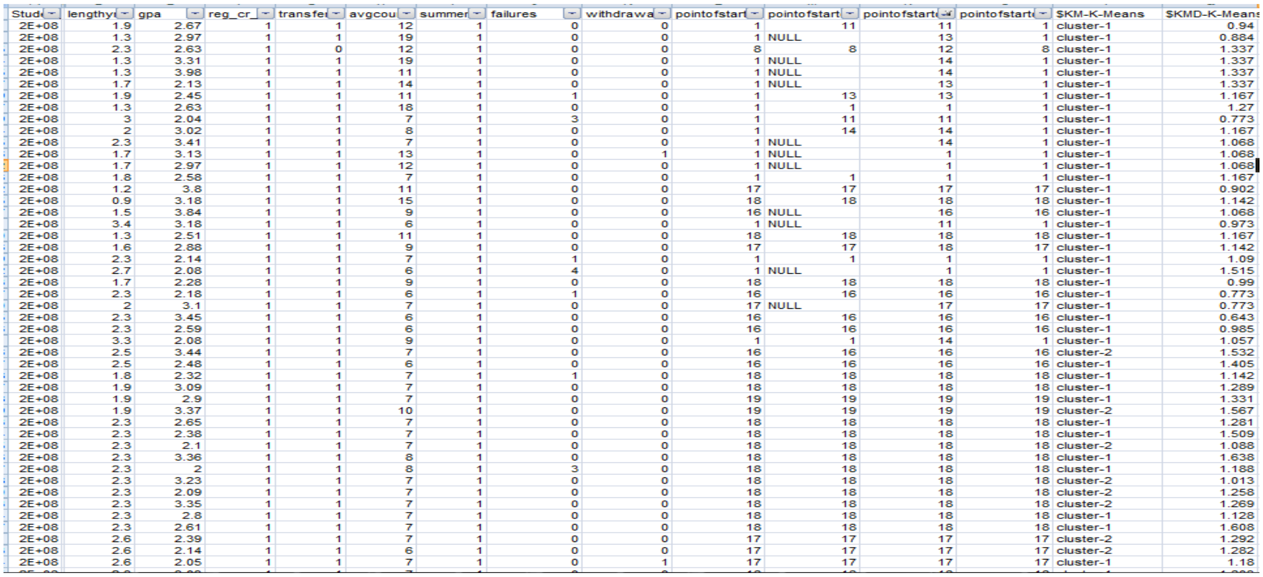
- **Definition of data on which the above criteria will be tested:** The dataset was split into two parts with 66% of the dataset called as training dataset and the remaining called as test dataset. This means that the model will be generated by the clustering technique using the training dataset while it will tested using the test dataset.

(C) Build Model:

- **Parameter Setting:** The number of clusters to be generated by the data mining technique was set as 2 for k-means and incremented by 1 until 10 clusters were generated. In the case EM clustering no such parameter setting was necessary as the algorithm defines the number by itself. The distance between data points which was measured by the algorithm chosen and it was measured using Euclidean distance. Euclidean distance measure is a widely used measure and was adapted as this distance provides an idea on how data points were clustered based on distance. Although there are other distance measurements for instance manhattan distance, literature shows that any one type of measure which provides support to model building is enough to be used and reported.

- **Generate Model**: After setting the parameters the model was generated using both k-means and EM clustering algorithms which is discussed next. A number of iterations had to be introduced in modelling the data with regard to various steps.

Table 4.7, Clusters created from modelling



Stude	lengthyı	gpa	reg_cr_	transfeı	avgcou	summer	failures	withdrawa	pointofstart	pointofstart	pointofstart	pointofstart	$KM-K-Means	$KMD-K-Means
2E+08	1.9	2.67	1	1	12	1	0	0	1	11	11	1	cluster-1	0.94
2E+08	1.3	2.97	1	1	19	1	0	0	1	NULL	13	1	cluster-1	0.884
2E+08	2.3	2.63	1	0	12	1	0	0	8	8	12	8	cluster-1	1.337
2E+08	1.3	3.31	1	1	19	1	0	0	1	NULL	14	1	cluster-1	1.337
2E+08	1.3	3.98	1	1	11	1	0	0	1	NULL	14	1	cluster-1	1.337
2E+08	1.7	2.13	1	1	14	1	0	0	1	NULL	13	1	cluster-1	1.337
2E+08	1.9	2.45	1	1	11	1	1	0	1	13	13	1	cluster-1	1.167
2E+08	1.3	2.63	1	1	18	1	0	0	1	1	1	1	cluster-1	1.27
2E+08	3	2.04	1	1	7	1	3	0	1	11	11	1	cluster-1	0.773
2E+08	2	3.02	1	1	8	1	0	0	1	14	14	1	cluster-1	1.167
2E+08	2.3	3.41	1	1	7	1	0	0	1	NULL	14	1	cluster-1	1.068
2E+08	1.7	3.13	1	1	13	1	0	1	1	NULL	1	1	cluster-1	1.068
2E+08	1.7	2.97	1	1	12	1	0	0	1	NULL	1	1	cluster-1	1.068
2E+08	1.8	2.58	1	1	7	1	0	0	1	1	1	1	cluster-1	1.167
2E+08	1.2	3.8	1	1	11	1	0	0	17	17	17	17	cluster-1	0.902
2E+08	0.9	3.18	1	1	15	1	0	0	18	18	18	18	cluster-1	1.142
2E+08	1.5	3.84	1	1	9	1	0	0	16	NULL	16	16	cluster-1	1.068
2E+08	3.4	3.18	1	1	6	1	0	0	1	NULL	11	1	cluster-1	0.973
2E+08	1.3	2.51	1	1	11	1	0	0	18	18	18	18	cluster-1	1.167
2E+08	1.6	2.88	1	1	9	1	0	0	17	17	18	17	cluster-1	1.142
2E+08	2.3	2.14	1	1	7	1	1	0	1	1	1	1	cluster-1	1.09
2E+08	2.7	2.08	1	1	6	1	4	0	1	NULL	1	1	cluster-1	1.515
2E+08	1.7	2.28	1	1	9	1	0	0	18	18	18	18	cluster-1	0.99
2E+08	2.3	2.18	1	1	6	1	1	0	16	16	16	16	cluster-1	0.773
2E+08	2	3.1	1	1	7	1	0	0	17	NULL	17	17	cluster-1	0.773
2E+08	2.3	3.45	1	1	6	1	0	0	16	16	16	16	cluster-1	0.643
2E+08	2.3	2.59	1	1	6	1	0	0	16	16	16	16	cluster-1	0.985
2E+08	3.3	2.08	1	1	9	1	0	0	1	1	14	1	cluster-1	1.057
2E+08	2.5	3.44	1	1	7	1	0	0	16	16	16	16	cluster-2	1.532
2E+08	2.5	2.48	1	1	6	1	0	0	16	16	16	16	cluster-1	1.405
2E+08	1.8	2.32	1	1	7	1	1	0	18	18	18	18	cluster-1	1.142
2E+08	1.9	3.09	1	1	7	1	0	0	18	18	18	18	cluster-1	1.289
2E+08	1.9	2.9	1	1	7	1	0	0	19	19	19	19	cluster-1	1.331
2E+08	1.9	3.37	1	1	10	1	0	0	19	19	19	19	cluster-2	1.567
2E+08	2.3	2.65	1	1	7	1	0	0	18	18	18	18	cluster-1	1.281
2E+08	2.3	2.38	1	1	7	1	0	0	18	18	18	18	cluster-1	1.509
2E+08	2.3	2.1	1	1	7	1	0	0	18	18	18	18	cluster-2	1.088
2E+08	2.3	3.36	1	1	8	1	0	0	18	18	18	18	cluster-1	1.638
2E+08	2.3	2	1	1	8	1	3	0	18	18	18	18	cluster-1	1.188
2E+08	2.3	3.23	1	1	7	1	0	0	18	18	18	18	cluster-2	1.013
2E+08	2.3	2.09	1	1	7	1	0	0	18	18	18	18	cluster-2	1.258
2E+08	2.3	3.35	1	1	7	1	0	0	18	18	18	18	cluster-2	1.269
2E+08	2.3	2.8	1	1	7	1	0	0	18	18	18	18	cluster-1	1.128
2E+08	2.3	2.61	1	1	7	1	0	0	18	18	18	18	cluster-1	1.608
2E+08	2.6	2.39	1	1	7	1	0	0	17	17	17	17	cluster-2	1.292
2E+08	2.6	2.14	1	1	6	1	0	0	17	17	17	17	cluster-2	1.282
2E+08	2.6	2.05	1	1	7	1	0	1	17	17	17	17	cluster-1	1.18

When the dataset was fed into Weka the following clusters were created. See table 4.7.

4.3.4.1 K-means Cluster Model

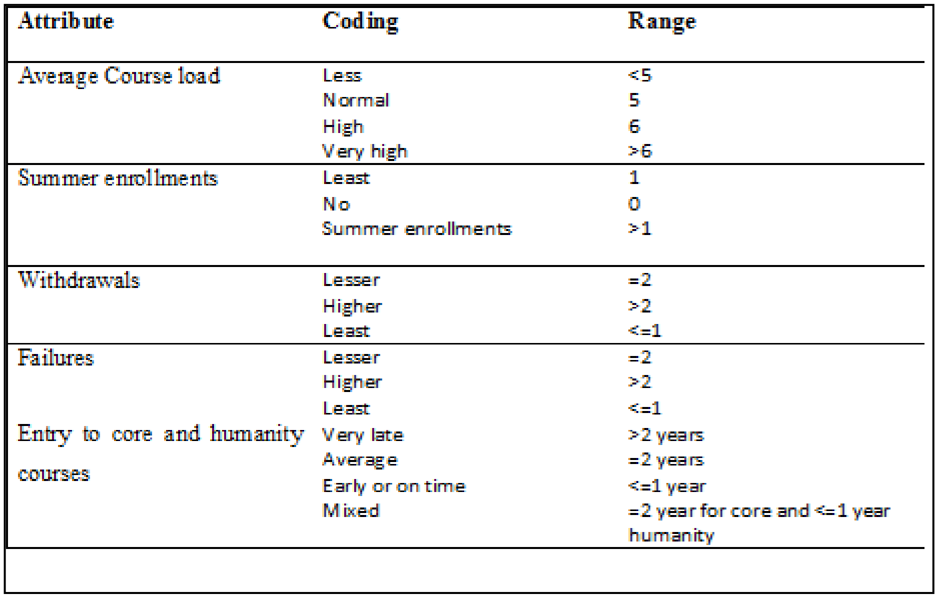
The dataset ready for mining was fed into Weka. K-means clustering algorithm was applied as per the test design set in the previous paragraph. The results of the test are given below.

When the dataset containing the limited number of attributes was fed into Weka the following clusters were created (see Tables 4.7 and 4.8).

Number of iterations: 4 (This number was found to produce the results in the shortest time when compared to other iterations. Iterations conducted that were less than 4 did not produce the best model in terms of SSE.). The attributes in Table 4.5 have been clustered and those attributes have been coded as below

Table 4.8, Coded Attributes that have been clustered

(*Source*: Sailesh et al. 2016 (publication of the student associated with the thesis))



Attribute	Coding	Range
Average Course load	Less	<5
	Normal	5
	High	6
	Very high	>6
Summer enrollments	Least	1
	No	0
	Summer enrollments	>1
Withdrawals	Lesser	=2
	Higher	>2
	Least	<=1
Failures	Lesser	=2
	Higher	>2
	Least	<=1
Entry to core and humanity courses	Very late	>2 years
	Average	=2 years
	Early or on time	<=1 year
	Mixed	=2 year for core and <=1 year humanity

In addition Weka has produced visualization that show points as instances that refer to the feature of the clusters. See visualizations provided in Fig. 4.6 and Fig. 4.7. The features shown in Fig. 4.6 and Fig. 4.7 are examples related to CGPA and time-to-degree. In clusters 0 and 1, students with CGPA ranging 2 and 4 were grouped. Similarly Fig. 4.7 shows the clustering of students pertaining to time-to-degree which showed the actual time taken by the students in completing the programme. The results indicate that students in cluster 1 took less time-to-degree than students in cluster 0. In addition further knowledge was discovered which included information about enrollment of students in summer session who did not repeat any courses and these factors contribute to the time-to-degree. Cluster 1 had students who had enrolled in summer and no repetitions of courses were found to take less time-to-degree when compared to those in cluster 0 where students who did not enrol to summer session and repeated courses.

The number of instances in clusters 0 and 1 are given below

Table 4.9, Clustered instances of k-means (*Source*: Sailesh et al. 2016 (publication of the student associated with the thesis))

	Cluster	
Cluster 0	283	64%
Cluster 1	157	36%

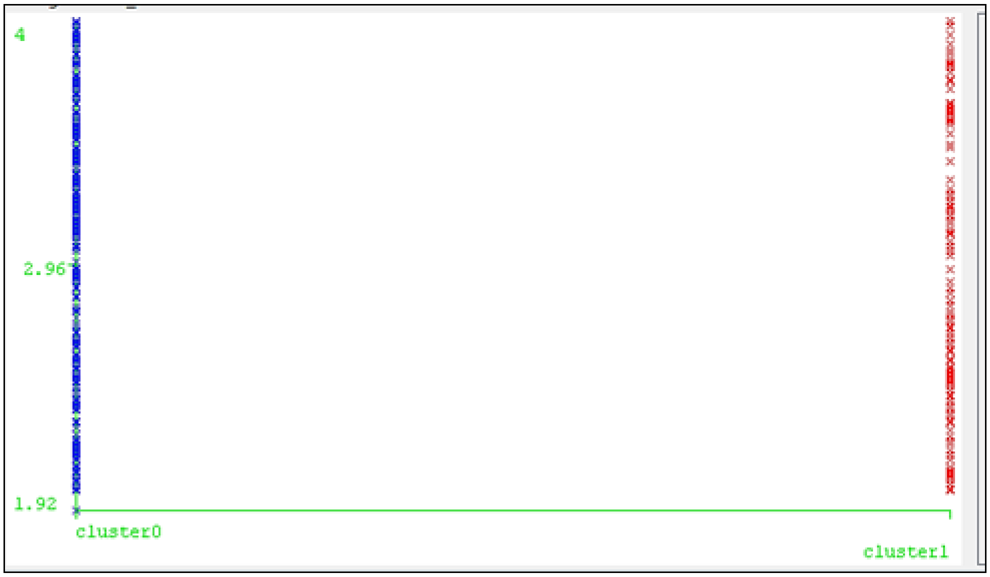


Figure 4.6, CGPA and Clusters

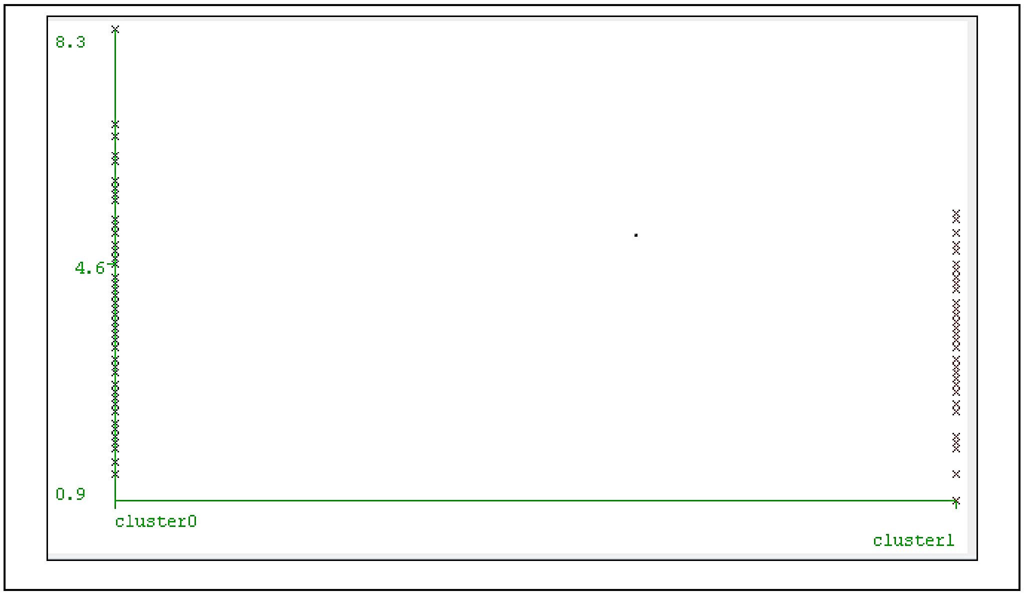


Figure 4.7, Time-to-degree and Clusters

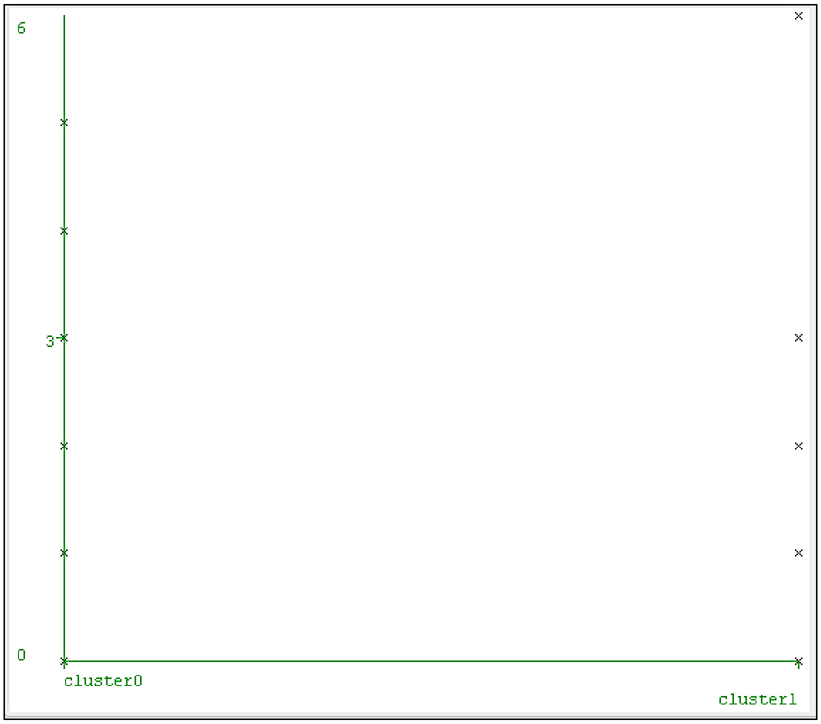


Figure 4.8, Number of withdrawals and clusters

From the examinations of clusters in Figure 4.6 and figure 4.7 the following contradiction were found that is cluster 1 showed that the time-to-degree of students is lesser, at the same time it was seen that one of the features associated with the cluster namely no. of withdrawals from courses per semester logically did not relate to lesser time-to-degree if the number is high. That is to say if one accepts the argument that higher number of withdrawals can lead to lower time-to-degree then there could be a meaningless situation where students could resort to withdrawals from courses after registration to reduce the time-to-degree. The situation is not acceptable because withdrawals imply accumulation of courses yet to be studied by the student and hence may require longer time-to-degree. This anomalous result provided by k-means appears to have occurred due to an inefficient mechanism of selecting the number of clusters. A closer examination of the results revealed that the features profiled in cluster 1 had lower correlation with each other. In order to verify this anomalous situation another algorithm was used to check the clustering of students namely EM clustering to verify whether the problem was due to algorithm or dataset.

The mining of the data using k-means algorithm was conducted by choosing the number of k as 2 in the beginning and incremented by 1 until 10 to generate 10 clusters. The result showed that the clusters generated by assigning k=2 could only be accepted as these 2 clusters were shown to have the lowest SSE of 26.34 out of the 9 cluster reports produced by Weka for all values of k ranging from 2 to 10 (See Table 4.10).

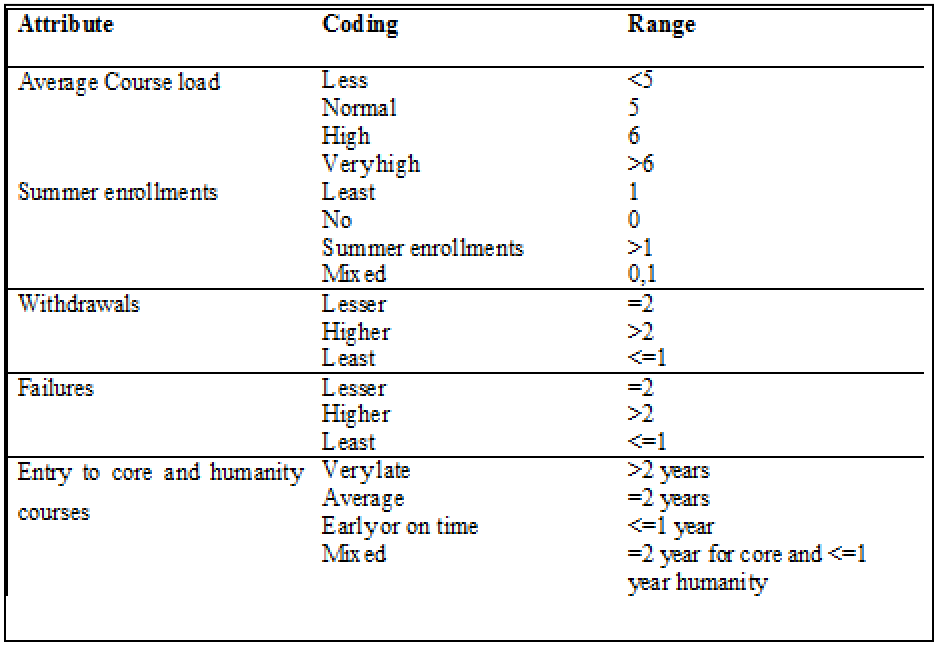
Table 4.10, k-means cluster results for different k values

Algorithm	**No. of cluster**	**No. of iterations**	**Cluster distribution**	**Sum of Squared Errors (SSE)**	**Time taken to build the model**
k-means	2	4	Cluster 0 283 (64%) Cluster 1 157 (36%)	26.34	0.17 seconds
k-means	3	5	Cluster 0 283 (64%) Cluster 1 100 (22.7%) Cluster 2 57 (12.9%)	45.000	0.25 seconds
k-means	4	8	Cluster 0 183 (41.59%) Cluster 1 100 (23%) Cluster 2 57(12%) Cluster 3 100 (23%)	600.000	0.35 seconds
k-means	5	12	Cluster 0 100 (23%) Cluster 1 100 (23%) Cluster 2 57(12%) Cluster 3 100 (23%) Cluster 4 83 (18.86%)	800.000	0.45 seconds
k-means	6	15	Cluster 0 50 (11.36%) Cluster 1 100 (23%) Cluster 2 57(12%) Cluster 3 100 (23%) Cluster 4 83 (18.86%) Cluster 5 50 (11.36%)	900.000	0.56 seconds
k-means	7	16	Cluster 0 50 (11.36%) Cluster 1 100 (23%) Cluster 2 57(12%) Cluster 3 50 (23%) Cluster 4 83 (18.86%) Cluster 5 50 (11.36%) Cluster 6 50 (23%)	1000.000	0.65 seconds
k-means	8	19	Cluster 0 50 (11.36%) Cluster 1 50 (23%) Cluster 2 57(12%) Cluster 3 50 (23%) Cluster 4 83 (18.86%) Cluster 5 50 (11.36%) Cluster 6 50 (23%) Cluster 7 50 (23%)	1200.000	0.78 seconds
k-means	9	20	Cluster 0 50 (11.36%) Cluster 1 50 (23%) Cluster 2 57(12%) Cluster 3 50 (23%) Cluster 4 83 (18.86%) Cluster 5 50 (11.36%) Cluster 6 50 (11.36%) Cluster 7 25 (5.6%) Cluster 8 25 (5.6%)	1500.000	0.86 seconds
k-means	10	22	Cluster 0 50 (11.36%) Cluster 1 50 (23%) Cluster 2 57(12%) Cluster 3 50 (11.36%) Cluster 4 83 (18.86%) Cluster 5 50 (11.36%) Cluster 6 25(5.6%) Cluster 7 25 (5.6%)	2000.000	0.97 seconds

			Cluster 8 25 (5.6%) Cluster 9 25 (5.6%)		

The student attributes that emerged from the clusters are coded as shown in Table 4.11.

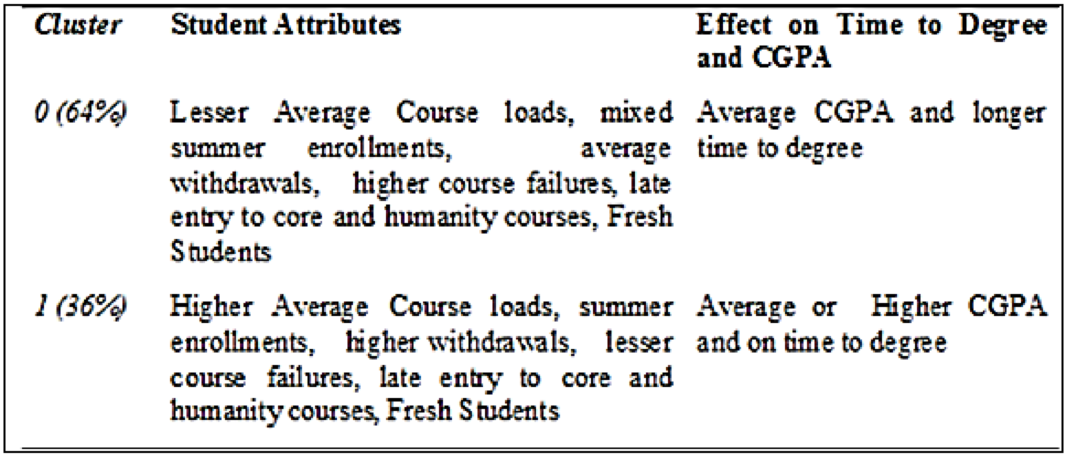
Table 4.11, Coding of Student Attributes that emerged from the clusters (*Source*: Sailesh et al. 2016 (publication of the student associated with the thesis))



Attribute	Coding	Range
Average Course load	Less	<5
	Normal	5
	High	6
	Veryhigh	>6
Summer enrollments	Least	1
	No	0
	Summer enrollments	>1
	Mixed	0,1
Withdrawals	Lesser	=2
	Higher	>2
	Least	<=1
Failures	Lesser	=2
	Higher	>2
	Least	<=1
Entry to core and humanity courses	Verylate	>2 years
	Average	=2 years
	Earlyor on time	<=1 year
	Mixed	=2 year for core and <=1 year humanity

The final set of clusters accepted as a result of rejecting the different clusters in Table 4.10 is provided in a new Table 4.12 below.

Table 4.12, Final set of clusters (*Source*: Sailesh et al. 2016 (publication of the student associated with the thesis))



Cluster	Student Attributes	Effect on Time to Degree and CGPA
0 (64%)	Lesser Average Course loads, mixed summer enrollments, average withdrawals, higher course failures, late entry to core and humanity courses, Fresh Students	Average CGPA and longer time to degree
1 (36%)	Higher Average Course loads, summer enrollments, higher withdrawals, lesser course failures, late entry to core and humanity courses, Fresh Students	Average or Higher CGPA and on time to degree

4.3.4.2 EM Clustering Model

Like in the case of k-means clustering algorithm, the dataset ready for mining is fed to Weka tool and in place of k-means, EM clustering algorithm is chosen and applied. While the test conditions remain the same, the number of clusters that was to be generated was fixed the algorithm itself. This was not the case with k-means algorithm where the number of clusters must be decided by the data miner.

The results of the model generated by EM clustering are given below.

Clustered Instances (*Source*: Sailesh et al. 2016 (publication of the student associated with the thesis))

0	53 (12%)
1	64 (15%)
2	35 (8%)
3	95 (22%)
4	43 (10%)
5	31 (7%)
6	36 (8%)
7	34 (8%)
8	4 (1%)
9	45 (10%)

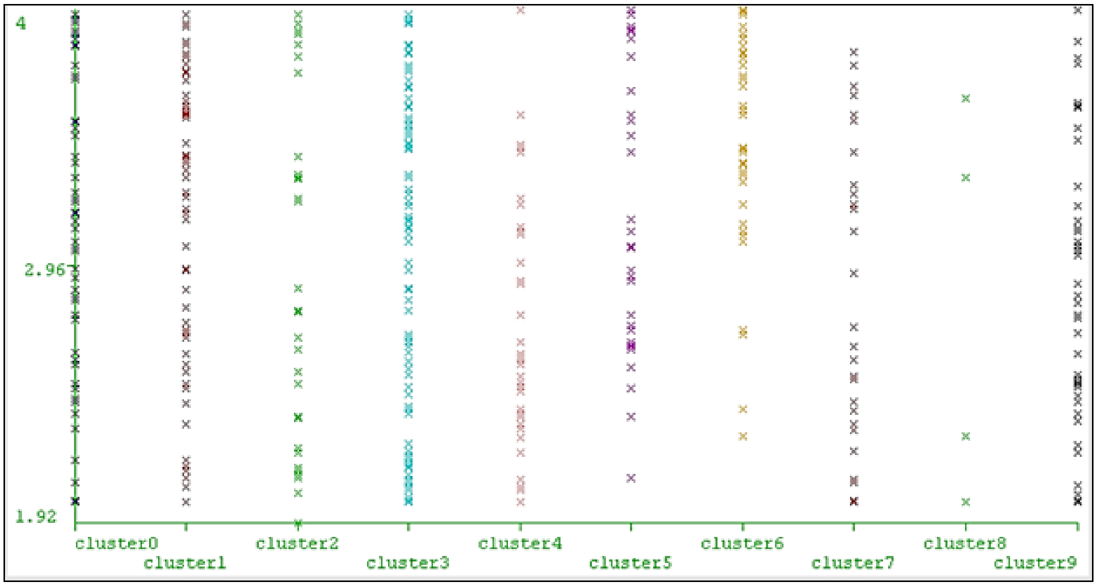


Figure 4.9, CGPA and EM Clusters

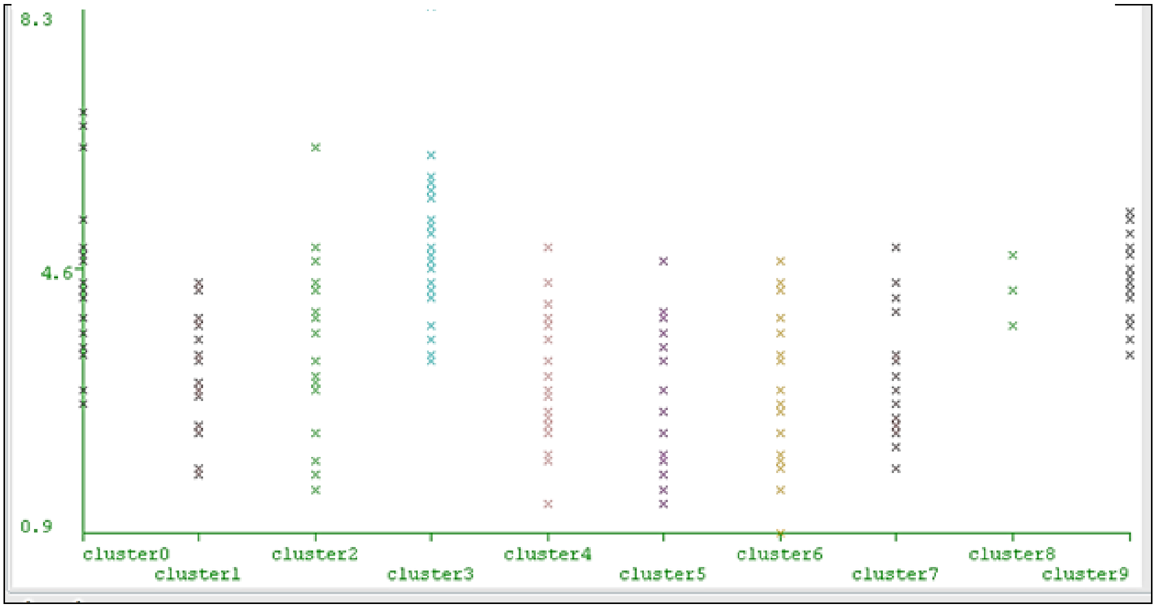
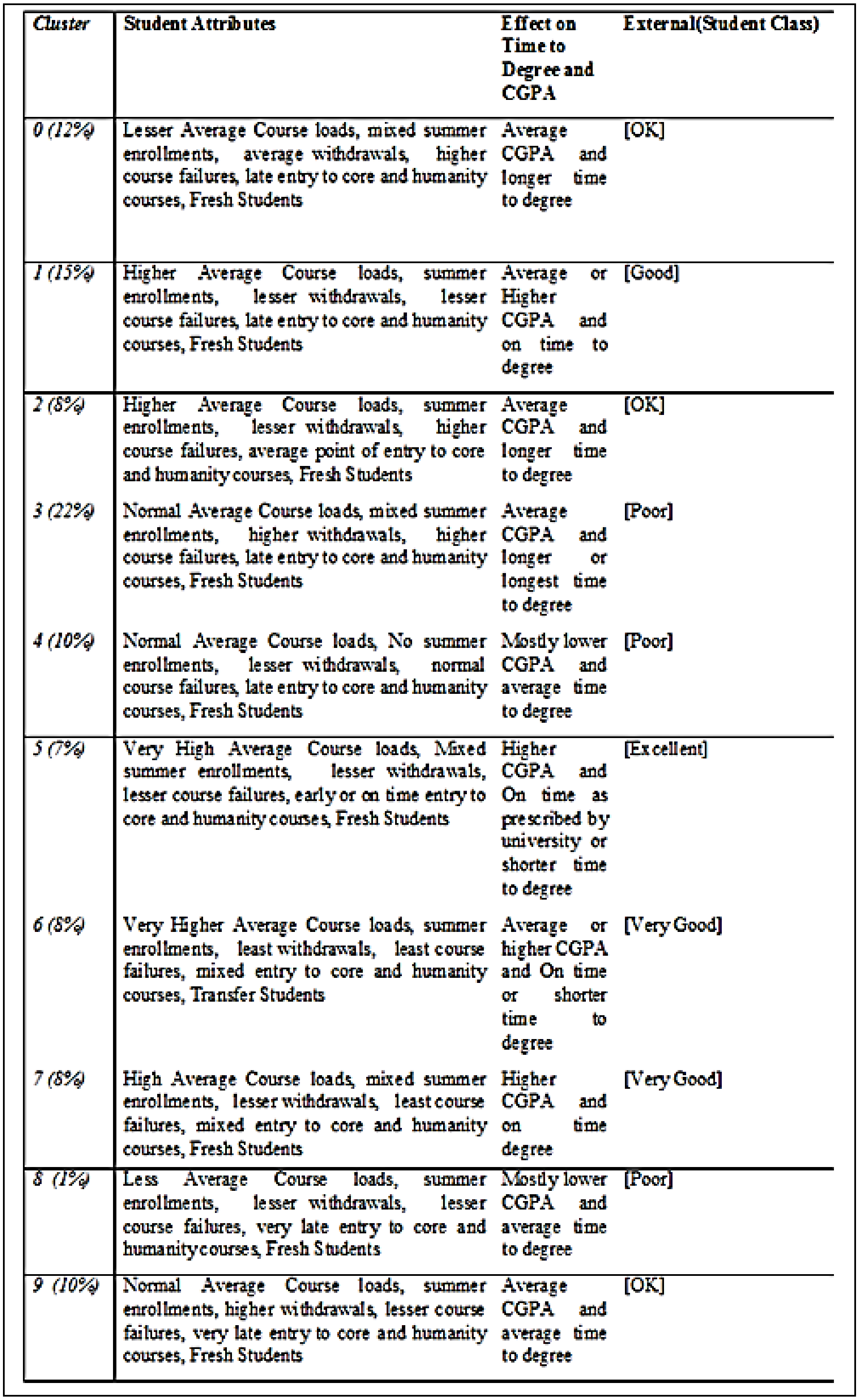


Figure 4.10, Time-to-Degree and EM Clusters

The 10 clusters generated by EM clustering algorithm is shown in Table 4.13.

Table 4.13, EM Clusters ((*Source*: Sailesh et al. 2016 (publication of the student associated with the thesis))



Cluster	Student Attributes	Effect on Time to Degree and CGPA	External(Student Class)
0 (12%)	Lesser Average Course loads, mixed summer enrollments, average withdrawals, higher course failures, late entry to core and humanity courses, Fresh Students	Average CGPA and longer time to degree	[OK]
1 (15%)	Higher Average Course loads, summer enrollments, lesser withdrawals, lesser course failures, late entry to core and humanity courses, Fresh Students	Average or Higher CGPA and on time to degree	[Good]
2 (8%)	Higher Average Course loads, summer enrollments, lesser withdrawals, higher course failures, average point of entry to core and humanity courses, Fresh Students	Average CGPA and longer time to degree	[OK]
3 (22%)	Normal Average Course loads, mixed summer enrollments, higher withdrawals, higher course failures, late entry to core and humanity courses, Fresh Students	Average CGPA and longer or longest time to degree	[Poor]
4 (10%)	Normal Average Course loads, No summer enrollments, lesser withdrawals, normal course failures, late entry to core and humanity courses, Fresh Students	Mostly lower CGPA and average time to degree	[Poor]
5 (7%)	Very High Average Course loads, Mixed summer enrollments, lesser withdrawals, lesser course failures, early or on time entry to core and humanity courses, Fresh Students	Higher CGPA and On time as prescribed by university or shorter time to degree	[Excellent]
6 (8%)	Very Higher Average Course loads, summer enrollments, least withdrawals, least course failures, mixed entry to core and humanity courses, Transfer Students	Average or higher CGPA and On time or shorter time to degree	[Very Good]
7 (8%)	High Average Course loads, mixed summer enrollments, lesser withdrawals, least course failures, mixed entry to core and humanity courses, Fresh Students	Higher CGPA and on time degree	[Very Good]
8 (1%)	Less Average Course loads, summer enrollments, lesser withdrawals, lesser course failures, very late entry to core and humanity courses, Fresh Students	Mostly lower CGPA and average time to degree	[Poor]
9 (10%)	Normal Average Course loads, summer enrollments, higher withdrawals, lesser course failures, very late entry to core and humanity courses, Fresh Students	Average CGPA and average time to degree	[OK]

Table 4.14, Comparison of different EM Clustering iterations

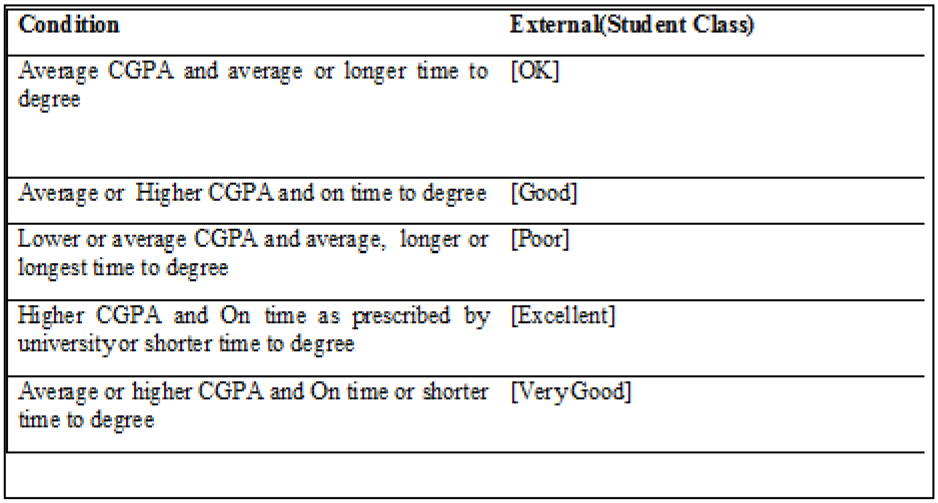
Algorithm	No. of iterations	Time taken to build the model
EM	4	0.17 seconds
EM	5	0.25 seconds
EM	8	0.35 seconds
EM	12	0.45 seconds

Like the discussions in the previous section it can be seen that Figures 4.9 and Fig 4.10 provide the clusters that have been generated using EM clustering algorithm. In those figures data points are clustered and the performance of the algorithm is tabulated in table 4.14 which gives an idea about the quality of performance. The data points clustered in Figures 4.9 and 4.10 can be identified with many of the features given as input (see Table 4.5). The result of the clustering is provided as a report in Table 4.13. The results show that 10 clusters can be derived with each cluster having data points with varying features. These features have been characterized for evaluation in the same way as mentioned in Section 4.2.6 and Table (attribute and codes) 4.11. Unlike the clusters provided by k-means algorithm the clusters generated by EM clustering technique do not exhibit any contradiction in terms of the features of the data points and their relationship with time-to-degree. At this point it was essential the performance of two algorithms is compared based on success criteria. Accordingly, the performance was evaluated and reported in the next section.

4.3.5. Evaluation

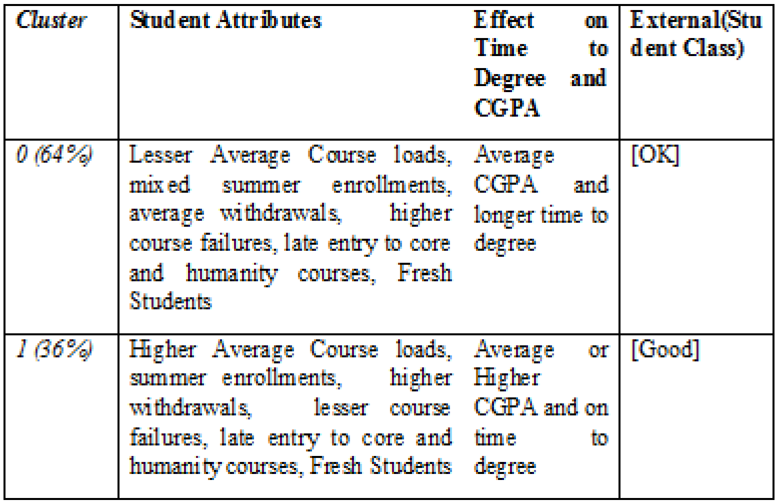
Evaluation of the performance of k-means and EM algorithms showed differences in the clustering and the features of the data points within the cluster (See Tables 4.10 and 4.14). In addition if the goal is to characterize the clusters a strategy must be employed by which it is possible to define a set of external characterization features and evaluate the results obtained through clustering process against those external characterization features. In such instances external characterization features are normally those not used in the modelling process. In this research an external feature namely student class was used. This is related to class labelling individual instances of the cluster. For example in this research Student class label was categorized as Poor, OK, Good, Very Good and Excellent (see Table 4.15). Following this labelling evaluation was done as given in tables (with student class label) Tables 4.15 and 4.16 which provides ranking as well as comparison of the performance of the two algorithms.

Table 4.15, Categorisation of student class labels ((*Source*: Sailesh et al. 2016 (publication of the student associated with the thesis))



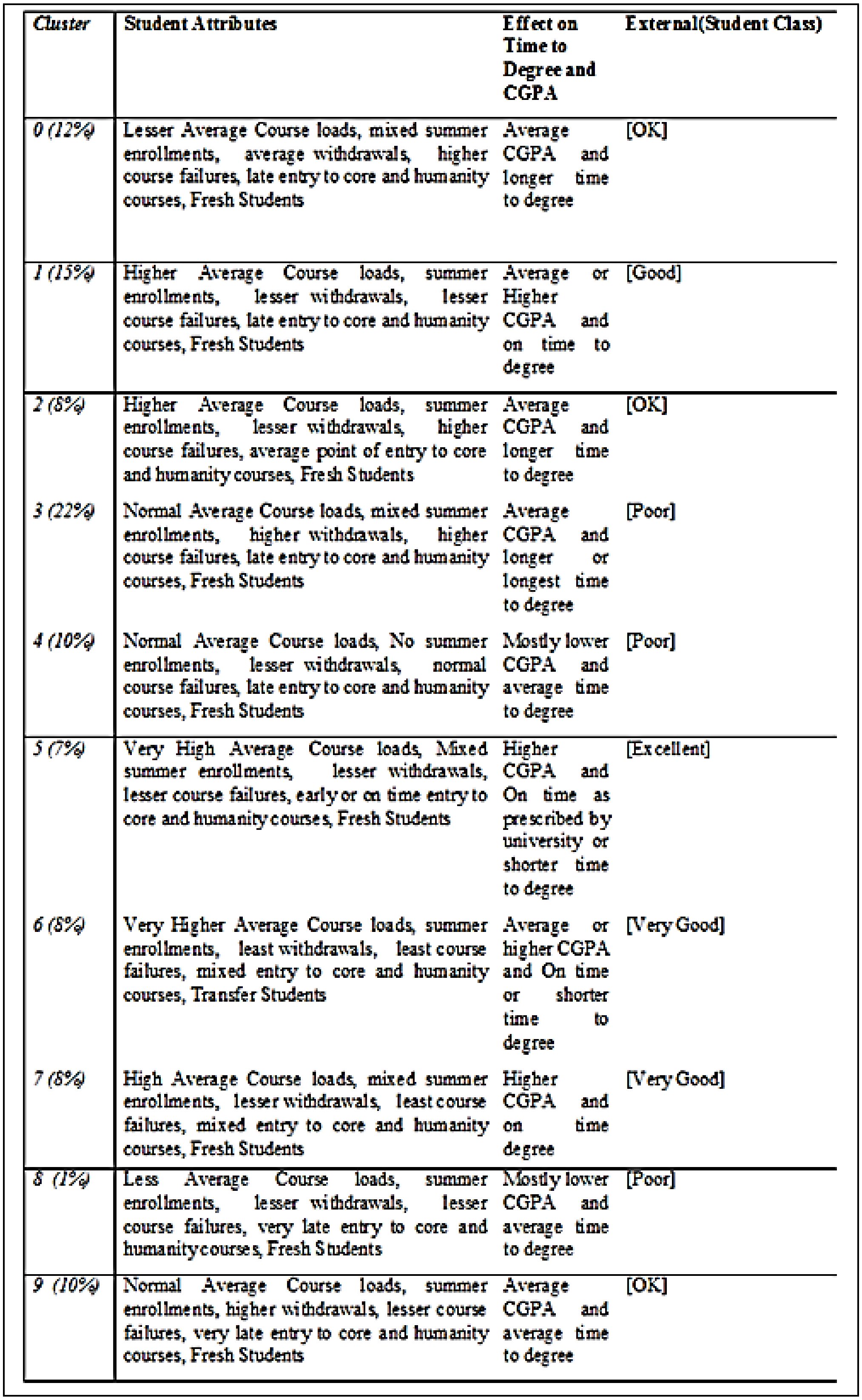
Condition	External(Student Class)
Average CGPA and average or longer time to degree	[OK]
Average or Higher CGPA and on time to degree	[Good]
Lower or average CGPA and average, longer or longest time to degree	[Poor]
Higher CGPA and On time as prescribed by university or shorter time to degree	[Excellent]
Average or higher CGPA and On time or shorter time to degree	[Very Good]

Table 4.16, k-means clusters with class labels ((Source: Sailesh et al. 2016 (publication of the student associated with the thesis))



Cluster	Student Attributes	Effect on Time to Degree and CGPA	External(Student Class)
0 (64%)	Lesser Average Course loads, mixed summer enrollments, average withdrawals, higher course failures, late entry to core and humanity courses, Fresh Students	Average CGPA and longer time to degree	[OK]
1 (36%)	Higher Average Course loads, summer enrollments, higher withdrawals, lesser course failures, late entry to core and humanity courses, Fresh Students	Average or Higher CGPA and on time to degree	[Good]

Table 4.17, EM clustering Student profiles with student attributes and effect CGPA and time-to-degree ((*Source*: Sailesh et al. 2016 (publication of the student associated with the thesis))



Cluster	Student Attributes	Effect on Time to Degree and CGPA	External(Student Class)
0 (12%)	Lesser Average Course loads, mixed summer enrollments, average withdrawals, higher course failures, late entry to core and humanity courses, Fresh Students	Average CGPA and longer time to degree	[OK]
1 (15%)	Higher Average Course loads, summer enrollments, lesser withdrawals, lesser course failures, late entry to core and humanity courses, Fresh Students	Average or Higher CGPA and on time to degree	[Good]
2 (8%)	Higher Average Course loads, summer enrollments, lesser withdrawals, higher course failures, average point of entry to core and humanity courses, Fresh Students	Average CGPA and longer time to degree	[OK]
3 (22%)	Normal Average Course loads, mixed summer enrollments, higher withdrawals, higher course failures, late entry to core and humanity courses, Fresh Students	Average CGPA and longer or longest time to degree	[Poor]
4 (10%)	Normal Average Course loads, No summer enrollments, lesser withdrawals, normal course failures, late entry to core and humanity courses, Fresh Students	Mostly lower CGPA and average time to degree	[Poor]
5 (7%)	Very High Average Course loads, Mixed summer enrollments, lesser withdrawals, lesser course failures, early or on time entry to core and humanity courses, Fresh Students	Higher CGPA and On time as prescribed by university or shorter time to degree	[Excellent]
6 (8%)	Very Higher Average Course loads, summer enrollments, least withdrawals, least course failures, mixed entry to core and humanity courses, Transfer Students	Average or higher CGPA and On time or shorter time to degree	[Very Good]
7 (8%)	High Average Course loads, mixed summer enrollments, lesser withdrawals, least course failures, mixed entry to core and humanity courses, Fresh Students	Higher CGPA and on time degree	[Very Good]
8 (1%)	Less Average Course loads, summer enrollments, lesser withdrawals, lesser course failures, very late entry to core and humanity courses, Fresh Students	Mostly lower CGPA and average time to degree	[Poor]
9 (10%)	Normal Average Course loads, summer enrollments, higher withdrawals, lesser course failures, very late entry to core and humanity courses, Fresh Students	Average CGPA and average time to degree	[OK]

A number of iterations had to be introduced in evaluation with regard to various steps. From the examination of the Tables 4.16 and 4.17 findings were derived which are discussed next.

4.3.6 Findings

The output of k-means algorithm could not be used for evaluation because the clusters derived had contradicting features as explained in Section 4.3.4.1. The clusters showed instance of students who were shown to have lower time-to-degree while those students had been recorded as having repeated courses many times. In this situation it is improbable that those students could have graduated with shorter time-to-degree which is a contradiction. As a next step performance of EM algorithm was evaluated using Table 4.17. In this table, the first column indicates cluster number (0 to 9) second column denotes the set of features extracted by the algorithm from amongst the 12 attributes listed in Table 4.5.

The third column provides the effect of student attributes on CGPA and time-to-degree. The fourth column shows the categorization student class label listing provided in Table 4.17. An examination of the table revealing a unique cluster under student class labels namely “Excellent” (cluster 5) and “Good” (cluster 1) the same cannot be said of the remaining 3 Student Class labels “Poor”, “Very Good” and “OK”. For instance cluster numbers 6 and 7 were categorized under student class labels “Very Good”. Similarly cluster numbers 0, 2 and 9 were categorized under the class label “OK” and finally cluster numbers 3, 4 and 8 were categorized under student class label “Poor”.

The above statements could be used for a variety of purposes and decision making. For instance, from student angle it could serve the purpose of gaining knowledge to improve their performance with regard to achieving optimum time-to-degree and CGPA. If a student is in the cluster 4 then the outcome of the student performance is characterized by the attribute “Mostly lower CGPA and average time-to-degree” See Table 4.17. This student could be encouraged to perform better by focused advising and monitoring. On the other hand a student clustered under cluster 5 characterized by the attribute “Higher CGPA and on time as prescribed by university or shorter time-to-degree” can be encouraged by recognizing the performance through awards and rewards like scholarships so that the student could be identified early in the academic career. Such identification could enable the HEIs to continuously monitor the performance of the student to maintain the performance. So towards this, HEIs could make decisions with regard to student learning factors such as assessment methods, changing curricula and funding aspects. In addition, the mining process has extracted important knowledge as follows.

4.3.6.1 Time-to-degree and CGPA has been clustered along different attributes indicating that the 12 attributes identified in the dataset fed into the CRISP-DM process (see Table 4.17) have a relationship to CGPA and time-to-degree. For instance using cluster 5 it is possible to visualize a function as follows:

CGPA = function of (Mixed summer enrollments, very high average course loads, lesser course failures, lesser withdrawals, Fresh Students, early or on time entry to core courses and early or on time entry to humanity courses, time-to-degree) → 4.1

In this clustering the attributes namely '*very high Average Course loads'* (this attributes falls under the category average course load in Table 4.11), '*mixed summer enrollments'* (this attributes falls under the summer enrollments in Table 4.11) and '*early or on time entry to core courses*' (this attributes falls under the category point of entry to core courses in Table 4.11) and '*early or on time entry to humanity courses'* (this attributes falls under the category point of entry to humanity courses in Table 4.11) represent a part of the student course registration data by semester attribute identified in the hypotheses HA and HB. That is to say equation 4.1 can be re-written as

CGPA = function of (student course registration data by semester, lesser withdrawals, lesser course failures, Fresh Students, time-to-degree) → 4.2

Taking the argument which states that student course registration data by semester can considered as equivalent to course taking pattern of student's equation 4.2 could be rewritten as

CGPA = function of (course taking pattern, lesser withdrawals, lesser course failures, Fresh Students, time-to-degree) →4.3

Similarly using cluster 5 it is possible to visualize another function as follows:

Time-to-degree = function of (Mixed summer enrollments, Very High Average Course loads, lesser course failures, lesser withdrawals, Fresh Students, early or on time entry to core and humanity courses, CGPA) → 4.4

Using the same arguments given under equation 4.1, it is possible to rewrite equation 4.4 as

Time-to-degree = function of (course taking pattern, lesser withdrawals, lesser course failures, Fresh Students, CGPA) → 4.5

From equations 4.3 and 4.5 it can be inferred that course taking pattern is part of the function that could determine CGPA and time-to-degree. Although the same argument could be extended to the other attributes namely lesser withdrawals (this attribute falls under the category number of withdrawals in Table 4.11), lesser course failures (this attribute falls under the category number of failures in Table 4.11) and fresh students (this attribute falls under the category student type) it can be said that these three attributes are unlikely to impact CGPA. However time-to-degree could be affected by lesser withdrawals (this attribute falls under the category number of withdrawals in Table 4.11), lesser course failures (this attribute falls under the category number of failures in Table 4.11) but not fresh students (this attribute falls under the category student type). How these three attributes affect time-to-degree apart from course taking pattern is an interesting knowledge that appears to be hidden in the dataset, which should separately investigated. Again it can be seen that student type (freshman or transferred) is unlikely to affect both CGPA and time-to-degree.

4.3.6.2 The important finding that emerges is that attributes namely '*very High Average Course loads, Mixed summer enrollments, early or on time entry to core courses and early or on time entry to humanity courses'* can be considered as course taking pattern is knowledge hidden in the dataset that has been brought out by data mining. This is an important discovery as course taking pattern is usually represented in the form of a set of courses and no relevance is attached to any of these attributes. Thus it can be argued that while these attributes could represent course taking pattern it is not clear how they can be denoted clearly as a sequence or set of courses, knowledge about which needs to be extracted. Further although there is lack of clarity on how these can be denoted clearly as course taking pattern, it is possible to derive an answer to this question. For instance if the attribute 'v*ery high average course loads'* is taken into consideration, it is possible to raise a question what are those courses that are denoted by the attribute. An answer to this question straightaway leads to the set of courses in which the student has registered which is nothing but the course taking pattern. This is knowledge hidden in the mined cluster that has to be discovered further. Similar arguments can be made with regard to the other three attributes namely *‘Mixed summer enrollments, early or on time entry to core courses and early or on time entry to humanity courses’*. When asked the question what are the courses that constitute ‘*mixed summer enrollments, early or on time entry to core courses and early or on time entry to humanity courses*’, then a set of courses will emerge which is nothing but the course taking pattern. This is knowledge hidden in the cluster that needs to be extracted. This is an important limitation that needs to be addressed as in clustering this knowledge is not clear.

However the other attributes that cannot be linked to time-to-degree and CGPA in a similar fashion as portrayed in equation 3.5 are point of entry of core elective and point of entry of free elective because these two attributes are not found as part of any cluster. The final equations that can be formed using the 10 clusters generated by EM cluster are:

CGPA = function of (course taking pattern, time-to-degree) →4.6

Time-to-degree = function of (course taking pattern, number of withdrawals, number course failures, CGPA) →4.7

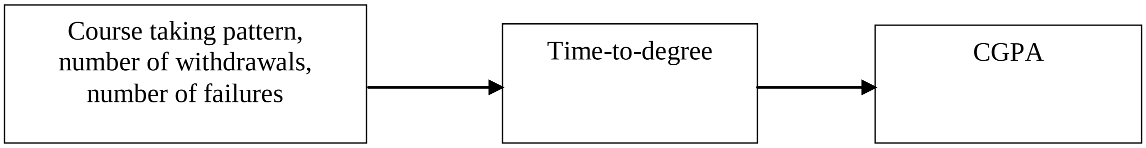

Course taking pattern, number of withdrawals, number of failures
Time-to-degree
CGPA

Figure 4.11, Probable relationship between time-to-degree, CGPA and course taking pattern - 1

4.3.6.3 The next important thing that emerges is that the functions in equations 4.6 and 4.7 can be depicted as Figure 4.12 below. This Figure 4.11 shows that course taking pattern affects both time-to-degree and CGPA. Whereas the other attributes namely number of withdrawals and number of failures affect time-to-degree in association with course taking pattern. In addition it is also clear that time-to-degree and CGPA are interrelated as they are part of all the clusters. This implies that CGPA and time-to-degree could determine each other. A finding that is not reported or perceived in the HEIs. This hidden relationship between CGPA and time-to-degree raises an important question which is: Is it worthwhile for a student to graduate at a shorter time-to-degree with lower CGPA or graduate with a higher CGPA taking longer time-to-degree? This is a paradoxical situation difficult to answer as any answer would have a contradiction with regard to the next step the student could take. For instance if a student scores a CGPA of 2.5 and graduates in 3.5 years whether this student should be considered to be a better student than a student who scores a CGPA of 3.5 and graduates in 4 years. It is difficult to answer this question because for the first student the objective may be to gain an employment whereas for the second student the objective maybe to go for higher studies. So the relationship between CGPA and time-to-degree is not very clear and appears to be nebulous. This aspect could not be clarified by clustering and is a limitation.

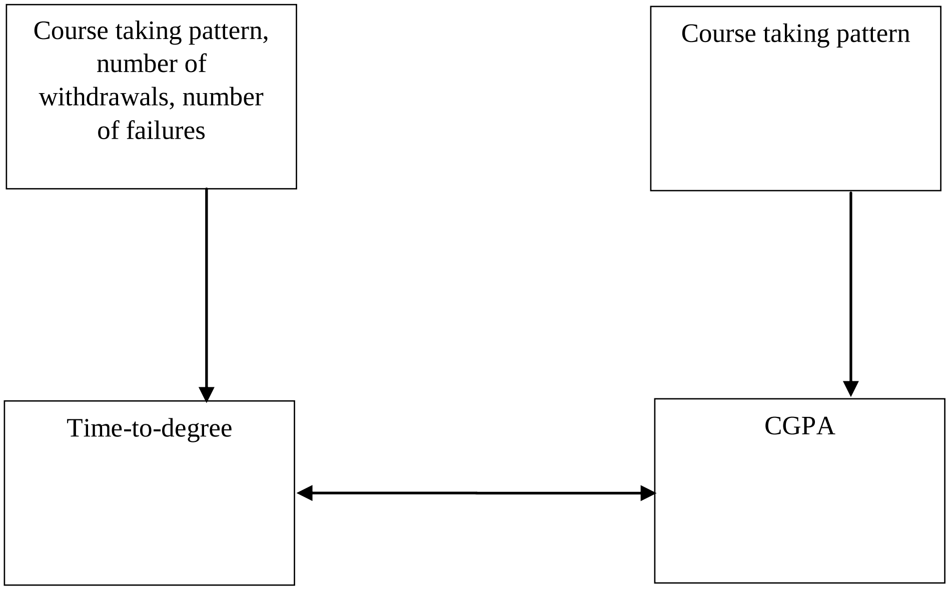

Course taking pattern, number of withdrawals, number of failures
Course taking pattern
Time-to-degree
CGPA

Figure 4.12, Probable relationship between time-to-degree, CGPA and course taking pattern – 2

4.3.6.4 In another instance, an examination of those multiple clustering under the same student class labels (For e.g. Cluster 3 and 4) shows that the CGPA and time-to-degree attributes remain the same but the set of student attributes vary. This provides an opportunity to address those different students attributes to be unified under a single student class category. For instance, in clusters 3 and 4 the student attributes that have differences include withdrawals and course failures. If the withdrawals and course failures are brought under control then a single cluster can be achieved. But there is another point that is emerging is how to improve the level of student category to approach the excellent category. The cluster 5 which shows that the cluster of students under the category excellent have the features that commonly point towards those students having achieved excellent results (E.g. Higher CGPA and shorter time-to-degree). For instance, features mainly "*very high average course loads, mixed summer enrollments, lesser withdrawals, lesser course failures, early or on time entry to core and humanity courses, and fresh Students*" can be brought under this cluster. If these student attributes were addressed with regard to students who are profiled under remaining clusters in a way that those students could achieve or approach the category "Excellent" through better academic support then there is a possibility to predict that students under clusters 0 to 4 and 6 to 9 could achieve the attributes namely 'higher CGPA and on time as prescribed by university or shorter time-to-degree (class label Excellent)'.

While the above explanations could be extended to all the student attributes under examination the only evidence that provides an instance of the attribute that could be related to course taking pattern is entry to different types of courses namely core or humanity courses. For instance, cluster 5 shows students who have registered in core and humanities courses at entry points of time earlier than the normally expected entry point of time in their study at the university. Entry points can be the beginning of a semester or summer session. This implies that a pattern can be assigned to the student course registration behaviour. For example, if a student under cluster 5 registers for a course in the first year in humanities (e.g. English level 2) that is normally prescribed by the university to be studied in the second year, then the set of courses registered by the student in year 1 becomes the pattern of courses that could lead to lower time-to-degree and higher CGPA. This shows that the course registration pattern can be characterized by entry point registration of courses. The same argument can be extended to the normal course registration pattern expected of a student specified by the university. Thus it can be seen that a clear relationship can be established between the course taking patterns and time-to-degree using cluster analysis. Thus the function that could be described using the above arguments is

Time-to-degree = f (course taking pattern, CGPA) → 4.8

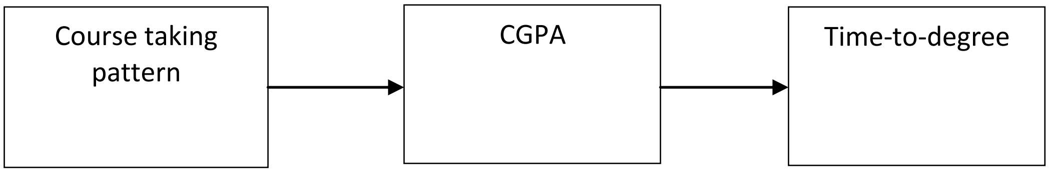


Figure 4.13: Probable relationship between time-to-degree, CGPA and course taking pattern - 3

It must be noted that equation 4.8 is already seen to be a subset of equation 4.7.

This knowledge about the course taking patterns provides a unique opportunity regarding student learning in terms of the following: (a) student can be advised to stick to optimum time-to-degree prescribed by university with regard to core courses and humanities, (b) universities can now assign students to specific groups or sections depending on their course registration pattern and provide them adequate academic support that may produce better results, (c) regarding curriculum analysis HEIs might review the pattern of time-to-degree, CGPA and course taking pattern to enable student learning to achieve better results by redesigning the curriculum as curricula are found to have impact on the student performance. Redesign of curriculum could be in terms of content and learning outcomes, (d) knowledge related to time-to-degree, CGPA and course taking patterns can be used on deciding on the student learning assessments. For instance faculties may decide to emphasize specific assessment type based on the course taking pattern of successful students and provide support to less successful students without changing quality. For instance assessments of students usually take place in terms of written examination, quizzes, assignments, projects, seminars, group discussions and practicals. In this case if a student clustered under the category 'Excellent' is found to have done well in projects and seminars when compared to other students. Then taking this as example other students could be supported by the faculty to improve their performance in those assessment methods.

4.3.6.5. Finally it must be noted that equation 5.6 satisfies the hypotheses HA only partially as the functions mentioned in that hypothesis has been partially realized with course taking pattern and time-to-degree found to have a relationship with CGPA and not course difficulty. Similarly it must be noted that equation 4.7 satisfies the hypothesis HB only partially as the functions mentioned in that hypothesis has been partially realized with course taking pattern and CGPA found to have a relationship with time-to-degree and not course difficulty.

4.3.6.6 Another finding is that the clustering technique performed as detailed in Section 4.3.4 although the business goal was not achieved. While the CRISP-DM process was found to perform as expected, when using the clustering technique, the model generated by the technique as part of the process did not lead to the achievement of the business goal in terms of discovery of course taking patterns to determine CGPA and time-to-degree.

4.3.6.7 The most significant finding of this research at this point is the importance of those student attributes other than course taking patterns found in the clusters (see Table 4.17), to CGPA and time-to-degree. Without addressing those features it may be difficult to clearly argue that course taking patterns only impact CGPA and time-to-degree. This needs further study using EDM concepts.

Based on the above discussions it can be concluded that the business goal has been partially achieved.

4.3.7 Limitations:

4.3.7.1 The business goal defined by equations 3.9 and 3.10 has only been partially achieved through the function provided in equations 3.6 and 3.7. For instance, the main function to be achieved requires the establishment of a relationship between time-to-degree on the one hand and course taking pattern, course difficulty and CGPA on the other. However equation 5.7 shows that time-to-degree is a function of course taking pattern, number of withdrawals, number of failures and CGPA. Course difficulty is not found to be an attribute that affects time-to-degree. This is due to the fact that the dataset could not be mined by the process to discover course difficulty as an attribute. In fact course difficulty has not been used as an attribute in any HEI. And hence there is no specific data collected against this attribute. Thus there is a need to investigate how this attribute can be discovered from the dataset if it is hidden in the dataset. This is another process that could be termed as contextual factor mining that needs to a part of any KDDM process without which it is not possible to discover patterns that are characterized by contextual aspects.

4.3.7.2 Another important limitation is the number of clusters that need to be analysed to determine those useful to make decisions. This problem arises primarily due to the large number of attributes that need to be considered while analysing educational data. This has a potential to lead the data miner commit errors due to subjective decision making involved in including or excluding attributes mined. If errors occur, then the clusters extracted could show wrong attributes as affecting the time-to-degree. Contextual factors could to some extent alleviate this difficulty an argument that needs to be verified.

4.3.7.3 A major constraint in using the current CRISP-DM process is the choice of the algorithm needed to discover patterns. Since each algorithm depends on the quality of dataset that is mined and also the data mining goal. It is difficult to conclude which one of the algorithms is most suitable for achieving the data mining goal. There is a need to carry out repeated experiments with different algorithm taking into account the parameter settings specific to those algorithms before finally deciding on the algorithm that could be used in the KDDM process. This is a lengthy and laborious process that can lead to longer duration of time in analysing data making the process inefficient.

4.3.7.4 Clustering does not provide relationship within a cluster because of which it is not possible to determine how each one of the attributes in a cluster could interact with the other. Lack of knowledge of this interaction can result in complications in decision making. Finally, clustering does not provide the exact element of a pattern like set of course (e.g. ACCT 101, ARAB 101, ECON 101, ENGL 101, ENGL 102 which represent courses namely Accounting I, Composition for native speakers of Arabic I, Principles of Microeconomics, Academic English I and Academic English II respectively. These courses and course codes are particular to the university whose data has been used for experiment in this research). Instead it provides information indirectly about the elements in the cluster. For instance the attribute average course load indicates the set courses in which a student had registered on average in a semester which is a broad indication of the pattern but not the specific course registered in by the student in the pattern. This is limitation as it is not possible to get an accurate knowledge of the pattern.

4.3.7.5 EDM has been integrated in CRISP-DM process model.

4.4 Summary

This experiment has been able to clearly demonstrate the usefulness of CRISP-DM process to bring out hidden knowledge in large data that could be useful in decision making related to student learning in HEIs. A relationship has been established between course taking patterns and time-to-degree. Clustering has been proved as a tool to bring out such a relationship. In addition specific attributes have been analysed. Student attributes have been identified by clustering and students have been profiled under ten clusters. The outcomes have been evaluated with the support of external characterization with five student class labels. Using this rating it is possible to assign future students to specific clusters identified in Table 4.17. The contribution lies in showing how CRISP-DM can be used to extract pattern of a set of courses of students, registered in a semester in a particular programme in a university that can be used to determine the time-to-degree for a given CGPA by profiling students using clustering technique. However such a pattern was not possible to be discovered using just the data mining technique described in Appendix 24.

This method promises to change the way students could be categorized and taught according to their profiles and offers a new way by which students could be grouped based on their course taking pattern to provide focused support by instructors in all the semesters thus enabling the students to achieve an optimum time-to-degree. Until now most of the research outcomes found in the KDDM literature deal with only one course and a few semesters to determine the time-to-degree whereas no specific method is found that could enable the data mining of a set of courses to derive a pattern of courses that could be used to determine the time-to-degree for a given CGPA and profile the students.

Chapter 5 : Integration and Evaluation of EDM in CRISP-DM Process Model using Association rules and Classification

5.1 Introduction:

The previous chapters detailed the need for modifying KDDM process model to have contextual knowledge processing stage(s) for extracting course taking patterns. This chapter presents the experimental results of association rules and genetic algorithm of CRISP-DM the KDDM process model. The algorithms and datasets detailed in the previous chapters are experimented using the KDDM process model. The sections of this chapter details the experiments conducted on KDDM model using association and classification techniques. The further sections analyses and evaluates the results got from the techniques and algorithms chosen as described in section 4.4. All the steps from Business understanding till evaluation of the KDDM model is carried out with the modelling stage on the selected techniques.

5.2 Determination of association rules that links courses registered by students, time-to-degree, course difficulty and CGPA using CRISP-DM process model 2

5.2.1 Business Understanding:

5.2.1.1 Determine business objectives – Establish a rule to link course taking pattern, time-to-degree, course difficulty and CGPA.

5.2.1.2 Assessment of situation – The situation was assessed and found to be the same as outlined in Section 4.2.2.2.

5.2.1.3 Determination of Educational data mining (EDM) goals –

- **Identification of data mining technique:** There could be a rule linking course taking patterns, course difficulty, CGPA and time-to-degree to achieve the business goal. Therefore association rule technique was found to be the most suitable technique to achieve the business goal (See section 2.5.2.2 of chapter 2)
- **Documenting the technical goals**: The interestingness of the rules can be evaluated using the following measures (See section 2.5.2.2 of chapter 2)

  Support(s) = Fraction of transactions that contain both X and Y

  Confidence (c) = Measures how often items in Y appear in transactions that contain X, confidence

For every rule the above measures support and confidence are calculated. Rules with 100% confidence were considered as interesting (See section 2.5.2.2 of chapter 2).

- **Data mining goal success criteria:**

  Generating rules which are interesting with 100% confidence

  Generating rules which are interesting with high support value.

A number of iterations had to be introduced in business understanding with regard to various steps.

5.2.2 Data Understanding:

- **Collect initial data -** Similar to section 4.3.2.
- **Describe data –** Similar to section 4.3.2.
- **Explore data**. Similar to section 4.3.2.
- **Verify data quality –** Similar to section 4.3.2.

A number of iterations had to be introduced in understanding data with regard to various steps.

5.2.3 Data Preparation:

(A) Select data –This dataset was extracted from the initial dataset described in section 5.2.2. To apply association rule technique, the data type could be categorical and nominal which is different from clustering. All numeric values were converted into nominal values. Thus the final dataset extracted comprised 59 attributes which were used for mining. It should be noted here that the number of attributes in the data set extracted for association rule is higher than the ones extracted in the case of clustering where only 44 attributes were initially extracted. The reason for this is that clustering does not function with nominal data types whereas association rule could function.

The dataset pertains to students belonging to 12 programmes and graduated during the period 2003 to 2014.The data size was 1292. Data stored in various tables was joined in a single table in this stage. However association rule mining reports produced by the software were reported for only one programme namely Bachelor's Degree in Accounting and Finance which is provided as an example only. Other programmes could be mined in similar fashion. SQL queries were used to extract course sequence number (i.e. c1, c2….c56). Reproducing all the results would highly voluminous consuming too much of the resources needed for producing this thesis. (See table 6.1).

(B) Clean data – similar to section 4.3.3.

(C) Construct new data- similar to section 4.3.3.

(D) Integrate data – similar to section 4.3.3.

(E) Format data – The data was formatted to suit the association rule algorithm in terms of rows and columns (See table 5.1).

A number of iterations had to be introduced in preparing the data with regard to various steps.

Table 5.1, Dataset for Association Rules

No.	Attribute	Description
1	Student id	Student identification number
2	GPA	Cumulative GPA of student at graduation
3	Length	Time-to-degree in terms of Semesters
4	C1	Course no.1
5	C2	Course no.2
6	C3	Course no. 3
7	C4	Course no.4
8	C5	Course no.5
9	C6	Course no. 6
10	C7	Course no.7
11	C8	Course no.8
12	C9	Course no.9
13	C10	Course no.10
14	C11	Course no.11
15	C12	Course no.12
16	C13	Course no.13
17	C14	Course no.14
18	C15	Course no.15
19	C16	Course no.16
20	C17	Course no.17
21	C18	Course no.18
22	C19	Course no.19
23	C20	Course no.20
24	C21	Course no.21
25	C22	Course no.22
26	C23	Course no.23
27	C24	Course no.24
28	C25	Course no.25
29	C26	Course no.26
30	C27	Course no.27
31	C28	Course no.28
32	C29	Course no.29
33	C30	Course no.30
34	C31	Course no.31
35	C32	Course no.32
36	C33	Course no.33
37	C34	Course no.34
38	C35	Course no.35
39	C36	Course no.36
40	C37	Course no.37
41	C38	Course no.38
42	C39	Course no.39
43	C40	Course no.40
44	C41	Course no.41
45	C42	Course no.42
46	C43	Course no.43
47	C44	Course no.44
48	C45	Course no.45
49	C46	Course no.46
50	C47	Course no.47
51	C48	Course no.48
52	C49	Course no.49
53	C50	Course no.50
54	C51	Course no.51
55	C52	Course no.52
56	C53	Course no.53
57	C54	Course no.54
58	C55	Course no.55
59	C56	Course no.56

5.2.4 Modelling:

(A) Select Modelling technique: Apriori algorithm is a widely used algorithm that is considered to be useful in data mining technique pertaining to association rules. From prior research it was found that this algorithm could be used to mine data and generate models.

(B) Generate Test Design:

- **Criteria to determine the goodness of the model:** These criteria were the same as mentioned Section 5.2.1 which includes measuring the interestingness of a rule in terms of support and confidence. Although other measures including lift and conviction are used by researchers to measure interestingness, literature shows that support and confidence are the most widely used measures of interestingness. Therefore support and confidence are the two measures used in association rule mining in this research.
- **Definition of data on which the above criteria will be tested:** The full dataset was used to test the above criteria.

(C) Build Model:

- **Parameter Setting:**

  The parameter settings used in building the model were set in terms of maximum set of rules, minimum condition support, minimum confidence, minimum lift, minimum rule support (See table 5.2). The Table 5.2 is partially extracted from IBM SPSS modeller.

Table 5.2, Parameter setting for association rules

Build Settings

Maximum Number of Rules	1,000
Minimum Condition Support	.05
Minimum Confidence	.10
Minimum Rule Support	.05
Minimum Lift	2.00
Maximum Number of Items in a Rule	10
Maximum Number of Items in a Condition	6
Maximum number of Items in a Prediction	4
Use only True Value for Flag Fields	True
Allow Rules without Conditions	False
Evaluation Measure Sorting the Rules	Confidence

- **Generate Model:**

  Once the minimum parameter values were set, the software tool was run and the model generated is reported in Tables 5.3 and 5.4.

Table 5.3, Most frequent item sets

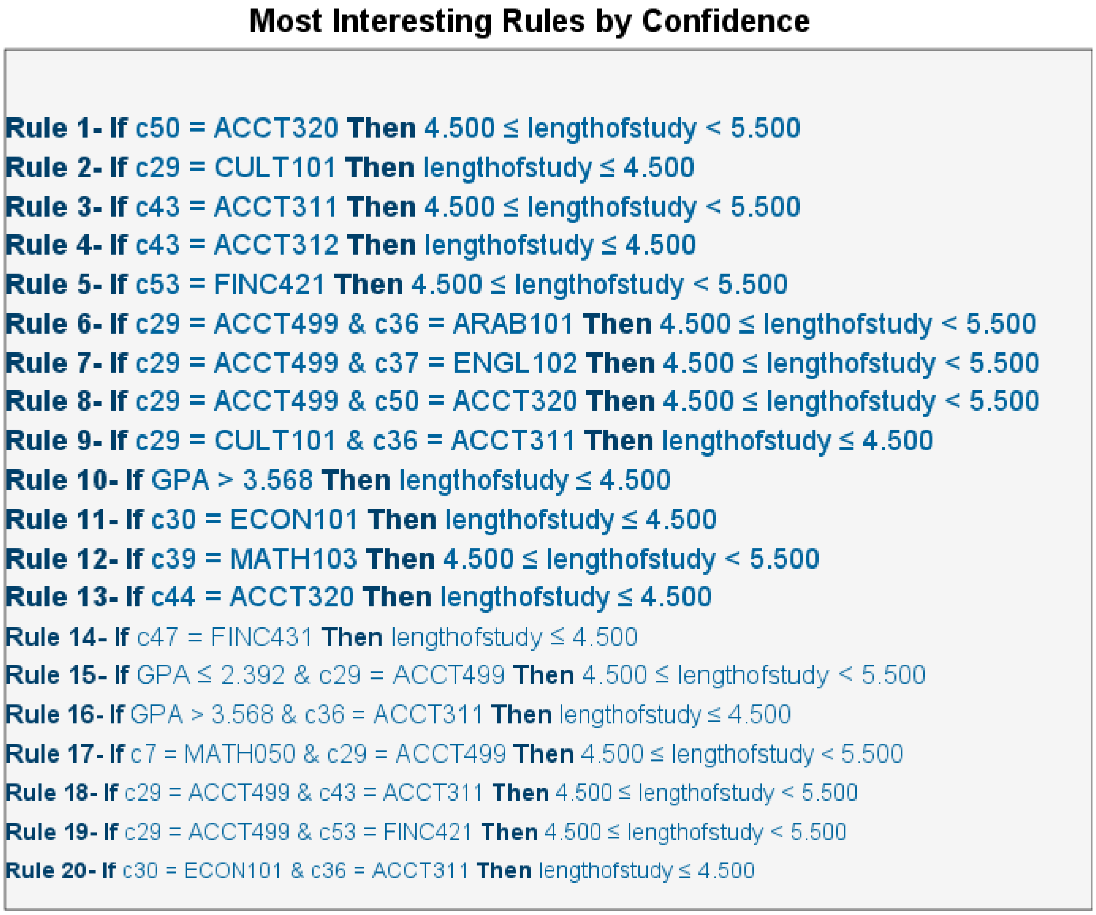
Information for Most Frequent Items			
Item name	Records (%)	Conditions (%)	Predictions (%)
lengthofstudy ≤ 4.500	38.46	.00	45.52
4.500 ≤ lengthofstudy < 5.500	36.54	.00	42.92
c29 = ACCT499	32.69	20.52	.00
2.392 ≤ GPA < 2.784	28.85	3.54	6.84
c23 = ACCT402	28.85	11.32	.00
GPA ≤ 2.392	26.92	1.89	3.07
c6 = MATH050	26.92	3.77	.00
c36 = ACCT311	26.92	20.28	.00
c16 = ACCT321	26.92	6.60	.00
2.784 ≤ GPA < 3.176	25.00	1.18	3.54
c36 = ARAB101	19.23	5.19	.00
c7 = MATH050	17.31	4.48	.00
c50 = ACCT499	17.31	8.73	.00
c43 = ACCT301	17.31	3.07	.00
c46 = FINC320	15.38	3.30	.00
c23 = ACCT403	15.38	.94	.00
c50 = ACCT201	15.38	1.65	.00
c50 = ACCT320	15.38	7.31	.00
5.500 ≤ lengthofstudy < 6.500	13.46	.00	.94
c7 = MATH103	13.46	.47	.00
c9 = ECON102	13.46	.71	.00
c37 = ENGL102	13.46	4.72	.00
c39 = FINC210	13.46	8.49	.00
c29 = ARAB101	13.46	1.42	.00
c6 = ITCS101	11.54	1.89	.00
c53 = FINC421	11.54	3.77	.00
c30 = ENGL102	11.54	1.65	.00
c45 = BANK302	11.54	4.25	.00
c9 = ENGL201	11.54	.24	.00
c36 = ACCT499	11.54	.24	.00
c37 = BANK220	11.54	1.65	.00
c43 = ACCT312	11.54	3.07	.00
c43 = ACCT311	11.54	4.95	.00
c43 = ACCT499	11.54	1.65	.00
c24 = FINC320	11.54	.71	.00
c29 = CULT101	11.54	5.66	.00
c30 = ECON101	9.62	3.30	.00
3.176 ≤ GPA < 3.568	9.62	.24	.71
GPA > 3.568	9.62	1.89	3.30
c47 = FINC431	9.62	.94	.00
c51 = ACCT403	9.62	1.89	.00
c24 = BANK302	9.62	.71	.00
c44 = ACCT320	9.62	2.83	.00
c53 = INTR400	7.69	7.08	.00
6.500 ≤ lengthofstudy < 7.500	7.69	.00	2.36
c47 = FINC421	7.69	8.96	.00
c39 = MATH104	7.69	1.89	.00
c39 = MAGT121	7.69	2.83	.00

c50 = ACCT312	7.69	1.89	.00
c23 = BFRM498	7.69	8.02	.00
c44 = ACCT403	7.69	2.36	.00

Table 5.4, Most interesting rules by confidence

				Sorted By Confidence (%)	Other Evaluation Statistics			
Rank	**Rule ID**	**Condition**	**Prediction**		**Condition Support (%)**	**Rule Support (%)**	**Lift**	**Deployability (%)**
1	1	c50 = ACCT320	4.500 ≤ lengthofstudy < 5.500	100.00	15.38	15.38	2.74	.00
2	2	c29 = CULT101	lengthofstudy ≤ 4.500	100.00	11.54	11.54	2.60	.00
3	3	c43 = ACCT311	4.500 ≤ lengthofstudy < 5.500	100.00	11.54	11.54	2.74	.00
4	4	c43 = ACCT312	lengthofstudy ≤ 4.500	100.00	11.54	11.54	2.60	.00
5	5	c53 = FINC421	4.500 ≤ lengthofstudy < 5.500	100.00	11.54	11.54	2.74	.00
6	6	c29 = ACCT499 c36 = ARAB101	4.500 ≤ lengthofstudy < 5.500	100.00	11.54	11.54	2.74	.00
7	7	c29 = ACCT499 c37 = ENGL102	4.500 ≤ lengthofstudy < 5.500	100.00	11.54	11.54	2.74	.00
8	8	c29 = ACCT499 c50 = ACCT320	4.500 ≤ lengthofstudy < 5.500	100.00	11.54	11.54	2.74	.00
9	9	c29 = CULT101 c36 = ACCT311	lengthofstudy ≤ 4.500	100.00	11.54	11.54	2.60	.00
10	10	GPA > 3.568	lengthofstudy ≤ 4.500	100.00	9.62	9.62	2.60	.00
11	11	c30 = ECON101	lengthofstudy ≤ 4.500	100.00	9.62	9.62	2.60	.00
12	12	c39 = MATH103	4.500 ≤ lengthofstudy < 5.500	100.00	9.62	9.62	2.74	.00
13	13	c44 = ACCT320	lengthofstudy ≤ 4.500	100.00	9.62	9.62	2.60	.00
14	14	c47 = FINC431	lengthofstudy ≤ 4.500	100.00	9.62	9.62	2.60	.00
15	15	GPA ≤ 2.392 c29 = ACCT499	4.500 ≤ lengthofstudy < 5.500	100.00	9.62	9.62	2.74	.00
16	16	GPA > 3.568 c36 = ACCT311	lengthofstudy ≤ 4.500	100.00	9.62	9.62	2.60	.00
17	17	c7 = MATH050 c29 = ACCT499	4.500 ≤ lengthofstudy < 5.500	100.00	9.62	9.62	2.74	.00
18	18	c29 = ACCT499 c43 = ACCT311	4.500 ≤ lengthofstudy < 5.500	100.00	9.62	9.62	2.74	.00

19	**19**	**c29 = ACCT499 c53 = FINC421**	**4.500 ≤ lengthofstudy < 5.500**	**100.00**	**9.62**	**9.62**	**2.74**	**.00**
20	**20**	**c30 = ECON101 c36 = ACCT311**	**lengthofstudy ≤ 4.500**	**100.00**	**9.62**	**9.62**	**2.60**	**.00**



Most Interesting Rules by Confidence

Rule 1- If c50 = ACCT320 **Then** 4.500 ≤ lengthofstudy < 5.500
Rule 2- If c29 = CULT101 **Then** lengthofstudy ≤ 4.500
Rule 3- If c43 = ACCT311 **Then** 4.500 ≤ lengthofstudy < 5.500
Rule 4- If c43 = ACCT312 **Then** lengthofstudy ≤ 4.500
Rule 5- If c53 = FINC421 **Then** 4.500 ≤ lengthofstudy < 5.500
Rule 6- If c29 = ACCT499 & c36 = ARAB101 **Then** 4.500 ≤ lengthofstudy < 5.500
Rule 7- If c29 = ACCT499 & c37 = ENGL102 **Then** 4.500 ≤ lengthofstudy < 5.500
Rule 8- If c29 = ACCT499 & c50 = ACCT320 **Then** 4.500 ≤ lengthofstudy < 5.500
Rule 9- If c29 = CULT101 & c36 = ACCT311 **Then** lengthofstudy ≤ 4.500
Rule 10- If GPA > 3.568 **Then** lengthofstudy ≤ 4.500
Rule 11- If c30 = ECON101 **Then** lengthofstudy ≤ 4.500
Rule 12- If c39 = MATH103 **Then** 4.500 ≤ lengthofstudy < 5.500
Rule 13- If c44 = ACCT320 **Then** lengthofstudy ≤ 4.500
Rule 14- If c47 = FINC431 **Then** lengthofstudy ≤ 4.500
Rule 15- If GPA ≤ 2.392 & c29 = ACCT499 **Then** 4.500 ≤ lengthofstudy < 5.500
Rule 16- If GPA > 3.568 & c36 = ACCT311 **Then** lengthofstudy ≤ 4.500
Rule 17- If c7 = MATH050 & c29 = ACCT499 **Then** 4.500 ≤ lengthofstudy < 5.500
Rule 18- If c29 = ACCT499 & c43 = ACCT311 **Then** 4.500 ≤ lengthofstudy < 5.500
Rule 19- If c29 = ACCT499 & c53 = FINC421 **Then** 4.500 ≤ lengthofstudy < 5.500
Rule 20- If c30 = ECON101 & c36 = ACCT311 **Then** lengthofstudy ≤ 4.500

Figure 5.1, Most interesting rules by confidence

As a condition association rule mining generates models in terms of attributes alongside the discussion given in Section 2.5.2.2 of Chapter 2. Association rule generates relationships based on the occurrence of the most frequent sets in the dataset that have minimum support. For instance in Figure 5.1 the relationships generated by IBM Modeller using a priori algorithms are given and some examples of those relationships are explained below.

Rule 1 – If c50 = ACCT320 then 4.5<= lengthofstudy < 5.5 (confidence 100% , rule support 15.38 and condition support 15.38)

Rule 7 – if c29= ACCT499 and c37 = ENGL102 then 4.5<= lengthofstudy < 5.5 (confidence 100% , rule support 11.54 and condition support 11.54)

Rule 10 – if GPA > 3.568 then lengthofstudy <= 4.5 (confidence 100%, rule support 9.62 and condition support 9.62)
Rule 16 – If GPA > 3.568 and c36 = ACCT311 then lengthofstudy ≤ 4.5 (confidence 100%, rule support 9.62 and condition support 9.62).

Rule 1 is valid because the confidence is reported to be 100% and confidence measure is supported by two factors condition support and rule support. The minimum condition for accepting a rule is that confidence should be 100% when the support is highest (that is when both condition support and rule support are the highest and equal). It must be noted that the ratio of condition support to rule support i.e. (condition support / rule support) will be 1 for all accepted rules and confidence is equal to [(condition support / rule support) x 100%]. This shows that confidence is defined by the ratio of condition support to rule support. In addition it must be noted that a 100% confidence indicates a 100% interestingness which is a condition mentioned in the literature for accepting any association rule (See section 2.5.2.2).

Rule 1 shows that the course (ACCT320 – Intermediate Cost Accounting; this is defined is the study plan of the dataset of the university where this research was conducted) and its order defined in the transcript as (c50) act as antecedents to the consequent namely length of study (4.5<= lengthofstudy < 5.5). This means that all those students who have achieved a time-to-degree that lies in the range 4.5 to 5.5 years have registered in ACCT320 and this course was the 50th course amongst the set of 59 courses the students could register in the programme in Bachelor's degree in Accounting and Finance. The attribute c50 needs to be interpreted for its meaning which is related to the semester or semesters in which the students had registered while studying ACCT320, a knowledge that can be obtained by perusing the transcript of the students. Perusal of the transcript yielded the result that c50 was offered in semester numbers 9 and 10. This information provides knowledge on the set of courses the students would have registered in leading the extraction of knowledge as a pattern of courses manually. This is a major limitation of association rule mining as it will be virtually impossible to extract this knowledge manually if the dataset contains data of thousands of students. Although it is possible to generate the most frequent course or courses in which the students have registered and hence generate patterns it will not be easy to address thousands students using a manual process.

Another important inference that could be derived is that the most frequent length of study lies in the range 4.5 and 5.5 which excludes the time-to-degree of students beyond 5.5 years although such students exist. In addition, it can be seen that this rule alone cannot provide a complete rule to determine the consequent. For instance, if there are students not mined by association rule algorithm whose frequency of occurrence of time-to-degree is 5 such students cannot be identified exclusively leading to a situation where accuracy of knowledge suffers.

Table 5.5, Analysis of association rules

Example of anonymous students whose case is referred here	**Rule**	**Inference**	**Interpretation**	**Example of pattern of courses of one student registered in a particular semester**
Student A (semester 9)	Rule 1 – If c50=ACCT320 then 4.5<= lengthofstudy < 5.5	Students are registered in ACCT 320 in semester 9 or 10 as the 50th course.	If ACCT320 is linked with those courses in semester 9 or 10 in which each student has registered then a pattern of courses would emerge for each student and should be extracted by a manual process.	ACCT 320, ACCT 312, FINC 322, FINC 323
Student B (semester 10)				ACCT 320, ACCT 312, ACCT341, ACCT321
Student C (Semester 6)	Rule 7 – if c29= ACCT499 and c37 = ENGL102 then 4.5<= lengthofstudy < 5.5	ENGL102 is registered in semester 6 or 7 as 37th course and ACCT499 is registered in semester 4 or 5 as 29th course.	If ENGL102 is linked with other courses registered along with other courses then the pattern of courses should emerge by manual process	ENGL102, BANK221, FINC211, MAKT201, STAT202
Student D (Semester 7)				ENGL102, BANK221, BANK302, BANK220,BANK320
Student E (Semester 4)				______
Student F (Semester 5)				ACCT 404, ACCT499
___	Rule 10 – if GPA > 3.568 then lengthofstudy <= 4.5	CGPA>3.568 when timetodegree is lesser than 4.5	________	__________
Student G (Semester 6)	Rule 16 – If GPA>3.568 and c36 = ACCT311 then lengthofstudy <= 4.5	ACCT 311 is registered in semester 6 or 7 as 36th course	If ACCT311 is linked with other courses registered along with other courses then the pattern of courses should emerge by manual process	ACCT311, BANK302, FINC322, FINC323
Student H (Semester 7)				ACCT 311, BANK302, FINC322, FINC323

Similar arguments can be provided for other rules (Rules 7, 10 and16). So the result of analysis of these rules is presented in Table 5.5.

The model generated by the CRISP-DM process is represented in the Figure 5.1. A number of iterations had to be introduced in understanding data with regard to various steps.

This model consists of 20 association rules which need to be evaluated for the patterns generated and whether those patterns are interesting. In addition, whether these association rules are able to provide support to test the hypothesis HA and HB will be evaluated next.

5.2.5 Evaluation

Evaluation is done by objective interestingness measure that uses statistics derived from data to determine whether a pattern is interesting in terms of support and confidence measures and analyse the results of the modeling to test the hypotheses HA and HB. As far as interestingness of patterns are concerned only 20 rules generated by Apriori algorithm were found to be interesting (interestingness 100%). However in order to derive meaningful patterns the 20 association rules were evaluated. Evaluation showed the following

Table 5.6, Relationship of precedent and consequent when length of study is less than 4.5

No.	**Rule no.**	**Precedent**		**Semester no. extracted from the transcript of the students**		**Consequent (years)**
		Course no.	**Course code**			
1.	Rule11	C30	ECON101	5 or 6	➡	Length of study ≤4.5
	Rule 20	C30	ECON101, ACCT311	5 or 6		
2.	Rule 14	C47	FINC431	9 or 10		
3.	Rule 2, Rule 9	C29	CULT101	6 or 7		
4.	Rule 4	C43	ACCT312	9 or 10		
5.	Rule 9	C29, C26	CULT 101, ACCT 311	6 or 7		
6.	Rule 13	C44	ACCT 320	9 or 10		

Table 5.7, Relationship of precedent and consequent when time-to-degree is between 4.5 and 5.5

No.	**Rul e no.**	**Precedent**		**Semester no. extracted from the transcript of the students**		**Consequent (years)**
		Course no.	**Course code**			
1.	1	C50	ACCT320	11 OR 12	➡	Length of study ≥4.5 and < 5.5
2.	3	C43	ACCT311	9 OR 10		
3.	5	C53	FINC421	11 OR 12		
4.	6	C29 C36	ACCT499 ARAB101	6 OR 7 7 OR 8		
5.	7	C29 C37	ACCT499 ENGL102	6 OR 7 7 OR 8		
6.	8	C29 C50	ACCT499 ACCT320	6 OR 7 11 OR 12		
7.	12	C39	MATH103	9 OR 10		
8.	17	C7 C29	MATH050 ACCT499	1 OR 2 6 OR 7		
9.	18	C29 C43	ACCT499 ACCT311	6 OR 7 9 OR 10		
10.	19	C29 C53	ACCT499 FINC421	6 OR 7 11 OR 12		

5.2.5.1. Only 18 courses out of 57 were extracted and patterns generated as precedents (see Table 5.6 and 5.7).

5.2.5.2. Only two time-to-degree measures (length of study) namely ≤ 4.5 years and 4.5 ≤ lengthofstudy < 5.5 years were extracted as consequents (see Tables 5.6 and 5.7).

5.2.5.3. The key courses that formed part of a pattern of a minimum of two courses are CULT101, ECON101, ACCT320, ACCT311, ACCT499 and FINC421 for which two different times to degree related in the mined result (see Tables 5.6 and 5.7). For instance the courses CULT101, ECON101 and ACCT320 can be related a time-to-degree that is ≤ 4.5 years whereas the courses ACCT499 and FINC421 can be related to a time-to-degree this 4.5 ≤ lengthofstudy < 5.5 years. However the course ACCT311 has a unique situation as it is found to be part of a pattern with CULT101 and ECON101 that can be related to a time-to-degree of ≤ 4.5 years whereas as part of ACCT499 can be related to a time-to-degree of 4.5 ≤ lengthofstudy < 5.5 years. This implies that if a student has registered in the above courses in a way as shown in Table 5.6 and 5.7 then the time-to-degree could probably be determined in accordance with the table. That is to say if a student registers in CULT101 as a course bearing number C29 which indicates that the course is registered in semester 6 or 7 (based on the transcript) then the same student if registers in the course ACCT311 as a course bearing number C36 which indicates the semester number 7 or 8 then the time-to-degree could probably be determined as ≤ 4.5 years. Similar arguments could be made for all the courses that are forming a pattern.

5.2.5.4. As far as the individual courses not paired with other courses are concerned (e.g. FINC431) it can be seen from Tables 5.6 and 5.7 that the student could register in those courses in semesters indicated in the transcript by the course number to achieve a probable time-to-degree that is ≤ 4.5 years or 4.5 ≤ lengthofstudy < 5.5 years. For instance a student could register in the course FINC431 (course number C47) in the semester 9 or 10 and the probable time-to-degree could be ≤ 4.5 years whereas with regard to the course FINC421 (course number C53) a student could register in the semester 11 or 12 and the probable time-to-degree would be 4.5 ≤ lengthofstudy < 5.5 years. In either case it is not certain that the student will complete the course within the time-to-degree found through association rule as the time-to-degree not only depends on the registration pattern of the courses of the student but also the other courses not reported by the association rule. While there is some knowledge that has been discovered, but that knowledge is not providing an accurate pattern of courses in which a student could register in a specific semester. This is a limitation.

5.2.5.5. Tables 5.6 and 5.7 show that there is a relationship between CGPA and time-to-degree (Rule 10). However this relationship is valid only for all CGPA figures greater than 3.568 and time-to-degree ≤ 4.5. There is no course taking pattern associated with this. But the relationship shows that CGPA acts as the antecedent of time-to-degree, which is new knowledge discovered.

5.2.5.6. There are courses that can be linked to the CGPA and together they act as antecedent to time-to-degree (Figure 5.1). Rule 15 shows that a student who has registered in a course ACCT499 with a course number of C29 (semester 6 or 7 as seen in the transcript) and has scored a CGPA ≤ 2.392 is likely to graduate with a time-to-degree of ≤ 4.5 years. However Rule 16 shows that a student who has registered in the course ACCT 311 with a course number C36 (semester 6 or 7 as in seen in the transcript) and has scored a CGPA ≥ 3.568 is likely to graduate with a time-to-degree of 4.5 ≤ lengthofstudy < 5.5 years. This aspect shows that there are situations where lower CGPA with a certain course taking pattern could result in shorter time-to-degree and such students despite scoring low CGPA might have been benefited in ways including gaining employment. Similarly a student scoring a higher CGPA with a particular course taking pattern by taking a longer time-degree-to graduate could lose time and might lag behind in comparison those students who have graduated taking shorter time-to-degree scoring lower CGPA. In such a situation it may so happen that there is a need to decide whether lower CGPA with shorter time-to-degree is preferable to higher CGPA with longer time-to-degree. This kind of a circumstance is paradoxical and requires deeper investigation to decide on which of the two conditions needs to be improved. In either case it can be seen that some factor other than those investigated might have a role to play. That is to say the student who scores lower CGPA but graduates with a shorter time-to-degree if encouraged might score a higher CGPA but there is no certainty. If some characteristic of the course or the teacher or the student or the institution influences the student then even after encouragement the student might still score the same CGPA. In this situation it may be necessary to identify a factor that might be linked to the student or the course which could lead help the student to get better results. For instance if a contextual factor namely course difficulty is introduced then course difficulty as an attribute might provide a way by which the student could be encouraged to score a better CGPA with shorter time-to-degree. Similar arguments could be extended to the case of the student who is seen to take along a longer time-to-degree with higher CGPA. However association rule mining does not discover any knowledge that can be considered as a contextual factor. This is a serious limitation.

5.2.5.7. From the foregoing discussion it can be seen that the following functions could be realized:

Based on Association Rules 1, 2, 3, 4, 5, 6, 7, 8, 9, 11, 12, 13, 14, 17, 18, 19, 20 it is possible to posit the following function:

Time-to-degree = function of (course taking pattern, course number,

semester number) →5.1

Based on Association Rules 15, 16 it is possible to posit the following function:

Time-to-degree = function of (course taking pattern, CGPA, course number,

semester number) → 5.2

Finally based on Association Rule 10 it is possible to posit the following function:

Time-to-degree = function of (CGPA) → 5.3

An inspection of equations 5.1, 5.2 and 5.3 shows that equation 5.2 can be accepted as the most comprehensive association rule amongst the three. When compared with equations 3.9 and 3.10, equation 5.2 shows 3.10 is partially satisfied and equation 3.9 is not supported. That is to say hypothesis HA is not accepted while HB is partially accepted. A number of iterations had to be introduced in evaluation with regard to various steps.

5.2.6 Findings

5.2.6.1. Association rule mining has provided new knowledge about student attributes that need to be included as part of the set to enable discovery of the course number sequence, semester number sequence and the time-to-degree.

5.2.6.2. The rule shows that course sequence number and semester number should treated as running number and these sequences are unique to each student.

5.2.6.3. It is possible to predict the semester number of a student as a consequent of course sequence number which can be depicted as:

Semester number = function (course sequence number) → 5.4

5.2.6.4. Knowledge about the course sequence number and hence semester number could inform the set of courses that affect the time-to-degree.

5.2.6.5. Association rule mining has also brought out that CGPA determines time-to-degree which is open to debate.

5.2.6.6. In another instance association rule mining shows that CGPA can be grouped alongside course sequence number and semester number.

5.2.6.7. While some of these findings are contradictory, for instance course sequence number is shown to be determining semester number whereas semester number is appearing to be independent of course sequence number. That is to say there is no way a student can be compelled to register in a course in a specific semester as the student has the option to choose some courses that could fall under several semester meaning that student could choose that course and register in any of the semesters. For example C29 (ACCT499) could be a course a student could register in semesters 6 or 7 or 8 or 9 or 10 leaving it open for the student to choose the semester for registering in the course. This contradiction needs to be addressed. The attributes in the current dataset did not provide any knowledge on how to address the contradiction. Perhaps there are other hidden factors related to the course or a semester which is not known and plays a role in creating this contradiction. There is a need to understand this and discover the factor or factors that could be used to address this contradiction. One factor that is likely to play a role is the contextual factor as contextual factors (course difficulty)are argued to be affecting student attributes as well course attributes (see Section 2.3.3 of chapter 2).

5.2.6.8. Association rule cannot discover the exact time-to-degree or CGPA. It can only discover the range even though the initial dataset has specific data about these attributes.

5.2.6.9. Association rule mining algorithms cannot discover course taking pattern particular to a semester or deal with infrequent occurrences of events. This has a major implication to find the most optimum set of courses that could form a pattern to determine the optimum time-to-degree.

5.2.6.10. Outcomes of association rule mining suffer from the fact that only some rules are discovered as significant from amongst hundreds rules. In the current instance while the algorithm brought out a total of 424 rules the numbers of acceptable rules were only 20.

5.2.6.11. Association rule mining does not include all the courses. In the current list of 58 courses the mining has found rules for only 18 courses which gives rise to the question whether other courses are significant to time-to-degree or not. This is a limitation.

5.2.6.12. As far as CRISP-DM process it is evident that at the Data preparation stage there is a need to add the course number sequence and semester number sequence to the dataset if the association between the attributes in the dataset if the association rule to be found out to determine course taking pattern and time-to-degree.

5.2.6.13. Another finding is that the association rule technique performed as detailed in Section 2.6.2 although the business goal was not achieved. While the CRISP-DM process was found to perform as expected when using the association rule technique, the model generated by the technique as part of the process did not lead to the achievement of the business goal in terms of discovery of course taking patterns to determine CGPA and time-to-degree.

5.2.6.14. The pattern of course number sequence and semester number sequence provides an opportunity for the faculty to locate courses in which students might generally be weak in performance. In such a case it is possible for the faculty to enhance teaching quality to support the students to overcome their weakness. For instance students registered in a course number C36 (semester number 7 or 8) is shown to have a CGPA ≥ 3.568 and graduated in ≤ 4.5 years' time-to-degree (Rule 16). The faculty concerned with this course could focus on the quality of teaching to ensure that CGPA level achieved by the students in semesters 7 or 8 is enhanced to say ≥ 3.84 which is an A- grade without affecting the time-to-degree. This can contribute to enhance the teaching quality.

5.2.6.15. EDM has been integrated in CRISP-DM process model.

Based on the above discussions it can be concluded that the business goal has been partially achieved.

5.2.7 Limitations

5.2.7.1 The main limitation of association rule mining is that it cannot discover a set of courses that could form a course taking pattern in specific semesters leaving a gap in the understanding of how a course taking pattern can be visualised. For instance in Table 3.1 (Chapter 3) an example of course taking pattern is indicated as (ACCT 101, ARAB 101, ECON 101, ENGL 101, ENGL 102) and is shown to be linked to semester 1. Association rule mining does not provide a pattern and by semester in this manner.

5.2.7.2 Association rule mining cannot provide knowledge about infrequently occurring events. This results in an incomplete knowledge which cannot be used in improving student learning performance.

5.2.8 Summary

Association rule mining experiment demonstrates the usefulness of CRISP-DM process to discover hidden knowledge in large data that could be used in decision making related to student learning experience although in a limited manner. An association rule has been created to relate time-to-degree to course taking pattern. Association rule shows that certain courses could determine certain range of time-to-degree. Another knowledge discovered is the nature of CGPA as an antecedent of time-to-degree. Further two new attributes related to courses namely course number sequence and semester number sequence have been discovered that are useful in generating association rule between time-to-degree and course taking pattern. This knowledge could be used to discover those courses that can be manually grouped together to form a pattern by semester although it is tedious. However the course number and semester number sequence offer knowledge to categorise courses by semester in order to optimize on the time-to-degree. For instance if the course number sequence of majority of the students show a pattern, then instructors could use that knowledge to understand why so many students either score low or high and analyse what factor can be addressed to improve the quality of education offered for instance improve teaching in particular courses where a majority of students have weak performance in terms of CGPA. Association rules used in a majority of prior research efforts have not investigated the entire period of study of the students who graduated from HEIs, instead addressed only specific semester or a few semesters. This has a limitation of incomplete knowledge about the association rule. However this research has investigated the entire period of study for applying association rule technique which provides a more complete knowledge about the association rules and has a better chance to be utilized in HEIs.

5.3 Prediction of optimum time-to-degree and CGPA in terms of the course taking pattern and course difficulty using CRISP-DM process model 3:

5.3.1 Business Understanding:

(A) **Determine business objectives –** To predict the optimum time-to-degree and CGPA taking into account course taking patterns and to classify courses and students using discovered knowledge.

(B) **Assessment of situation –** The situation was assessed and found to be the same as outlined in Section 4.2.2.2.

(C) **Determination of Educational data mining (EDM) goals –**

- **Identification of data mining technique:** It was assumed that there could be a relationship between course taking patterns, CGPA and time-to-degree which could enable the prediction of CGPA and time-to-degree, to achieve the business goal. Prediction related tasks in data mining literature are shown to be using Genetic algorithm and considered to be the most suitable technique to achieve the business goal (Bajpai and Kumar, 2010).
- **Documenting the technical goals**:

  The concept of GA is to create numerous generations of chromosomes (patterns) prior to finding the best solution. Firstly, these chromosomes are acquired from initial generation using the process of selection, crossover and mutation. Further, they (patterns) are selected based on their performance to participate in crossover. Then, evaluation of the performance is done using fitness function (FF) which is a performance measure, which assesses the relevance of the prediction. In the GA process, the fitness function is used in two operators namely selection and mutation. In the fitness functions used for the discovery of classification rules the following factors from the confusion matrix are often used:

The precision-recall measure was used where every group can be viewed as a result of a specific class. Precision is a fraction of correctly grouped instances and recall is the fraction of correctly retrieved instances out of all matching instances.

Precision (exactness) - % of tuples that the classifier labelled as positive are actually positive

Precision = (TP)/(TP + FP)

where TP=True Positive, TN=True Negative, FP=False Positive and FN-False Negative. Perfect precision score that can be achieved is a maximum of 1.

For any instance X,

True positive (TP): the actual class is X and the predicted class is also X.

• False positive (FP): the actual class is X, but the predicted class is not X.

• True negative (TN): the actual class is not X and the predicted class is also not X.

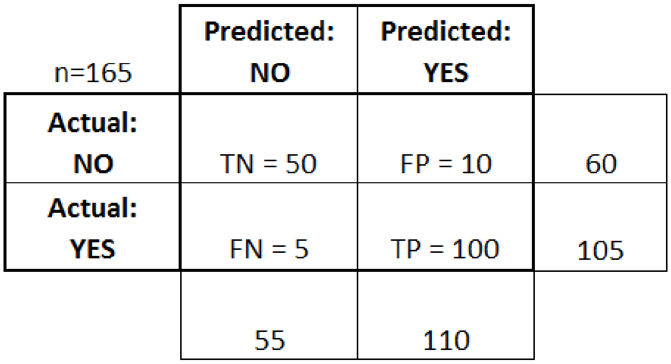
• False negative (FN): the actual class is not X, but the predicted class is X.

Recall (completeness) - % of positive tuples the classifier label is positive

Recall = (TP) / (TP+FN)

Perfect recall score that can be achieved is a maximum of 1.0 (Lin and Ho, 2002).

A table used to describe the performance of a classification model over a set of test data for which the true values are known is called confusion matrix. A sample confusion matrix is shown in figure 5.2.



n=165	Predicted: NO	Predicted: YES	
Actual: NO	TN = 50	FP = 10	60
Actual: YES	FN = 5	TP = 100	105
	55	110	

Figure 5.2, Sample confusion matrix

F-measure is a measure of a test's accuracy. The F-measure can be interpreted as a weighted average of the precision and recall, where an F-measure reaches its best value at 1 and worst at 0.

- **Data mining goal success criteria:**

Predict time-to-degree and CGPA by extracting course taking patterns.

A number of iterations had to be introduced in understanding the business with regard to various steps.

5.3.2 Data Understanding:

- **Collect initial data -** similar to section 4.3.2.
- **Describe data –** similar to section 4.3.2.
- **Explore data**. Similar to section 4.3.2.
- **Verify data quality –** similar to section 4.3.2.

A number of iterations had to be introduced in understanding data with regard to various steps.

5.3.3 Data Preparation:

(A) Select data –

This dataset was extracted from the initial dataset described in section 5.3.2.The final dataset comprises 8 attributes namely Student ID, Programme, CGPA, time-to-degree, semester passed credits, semester GPA, course code and semester which were used for mining (See Table 5.8). The dataset pertains to 1292 students belonging to 12 programmes and graduated during the period 2003 to 2014 with each student having 56 records. The data size was (1292 x 56) = 72352 records. Data stored in various tables was joined in a single table in this stage. Table 5.9 shows examples of the records of some students. (See table 5.8).

Table 5.8, Attributes used for Genetic algorithm

Attribute	Description	Example
Student ID	Student Identification Number	Stud1
Programme	Programme name	BSAF
CGPA	Cumulative GPA	3.78
Time-to-degree	Years taken to complete the programme	3.5
Semester Completed Credits	Credits completed in the semester	15
Semester GPA	GPA scored in the semester	3.668
Course code	Course code registered	ACCT101
Semester	Semester number	1

(B) **Clean data –** refer section 4.3.3.

(C) **Construct new data-** refer section 4.3.3.

(D) **Integrate data –** refer section 4.3.3.

(E) **Format data –** The data was formatted to suit the genetic algorithm in terms of rows and columns (See Table 5.9).

A number of iterations had to be introduced in preparing data with regard to various steps.

Table 5.9, Data set used for genetic algorithm

Student_ID	CGPA	Time-to-degree	Semester GPA	Semester Completed Credits	Course_Code	Semester
Stud1	3.78	3.5	3.668	15	ACCT 101	1
Stud1	3.78	3.5	3.668	15	ARAB 101	1
Stud1	3.78	3.5	3.668	15	ECON 101	1
Stud1	3.78	3.5	3.668	15	ENGL 050	1
Stud1	3.78	3.5	3.668	15	ENGL 101	1
Stud1	3.78	3.5	3.668	15	ENGL 102	1
Stud1	3.78	3.5	3.668	15	ITCS 101	1
Stud1	3.78	3.5	3.668	15	MATH 050	1
Stud1	3.78	3.5	3.668	15	MATH 103	1
Stud1	3.78	3.5	3.734	15	ENGL 202	2
Stud1	3.78	3.5	3.734	15	FINC 210	2
Stud1	3.78	3.5	3.734	15	FREN 101	2
Stud1	3.78	3.5	3.734	15	MAGT 221	2
Stud1	3.78	3.5	3.734	15	STAT 101	2
Stud1	3.78	3.5	3.945	18	BANK 310	3
Stud1	3.78	3.5	3.945	18	ECON 421	3
Stud1	3.78	3.5	3.945	18	FINC 321	3
Stud1	3.78	3.5	3.945	18	FINC 410	3
Stud1	3.78	3.5	3.945	18	MAGT 421	3
Stud1	3.78	3.5	3.945	18	MAKT 220	3
Stud1	3.78	3.5	4	6	ACCT 311	4
Stud1	3.78	3.5	4	6	ACCT 499	4
Stud1	3.78	3.5	3.833333	18	ACCT 201	5
Stud1	3.78	3.5	3.833333	18	ECON 102	5
Stud1	3.78	3.5	3.833333	18	ENGL 201	5
Stud1	3.78	3.5	3.833333	18	ITCS 121	5
Stud1	3.78	3.5	3.833333	18	MAGT 121	5
Stud1	3.78	3.5	3.833333	18	MATH 104	5
Stud1	3.78	3.5	3.61	18	ACCT 301	6
Stud1	3.78	3.5	3.61	18	BANK 201	6
Stud1	3.78	3.5	3.61	18	ECON 302	6
Stud1	3.78	3.5	3.61	18	FINC 310	6
Stud1	3.78	3.5	3.61	18	FREN 102	6
Stud1	3.78	3.5	3.61	18	ITMA 201	6
Stud1	3.78	3.5	3.868	15	ACCT 321	7
Stud1	3.78	3.5	3.868	15	BANK 320	7

Stud1	3.78	3.5	3.868	15	FINC 421	7
Stud1	3.78	3.5	3.868	15	ITMA 401	7
Stud1	3.78	3.5	3.868	15	MAKT 320	7
Stud1	3.78	3.5	3.556667	9	BANK 302	8
Stud1	3.78	3.5	3.556667	9	FINC 320	8
Stud1	3.78	3.5	3.556667	9	MAGT 321	8
Stud1	3.78	3.5	4	12	ACCT 401	9
Stud1	3.78	3.5	4	12	ECON 401	9
Stud1	3.78	3.5	4	12	INTR 400	9
Stud2	3.84	3.5	4	15	ARAB 101	1
Stud2	3.84	3.5	4	15	ECON 101	1

5.3.4 Modelling:

(A) Select Modelling technique: Genetic algorithm was used to generate the course taking pattern and predict time-to-degree and CGPA. The pseudo code of the GA obtained from the literature (Holland, 1975) (see Section 2.5.2.5) was used to generate the course taking pattern of students from the dataset. In order to implement the pseudo code another tool was required using which a new code was written alongside the steps given in the pseudo code. The tool used to write the code was C# embedded in Visual Studio (2012). The GA with new code was implemented to generate the course taking pattern of students from the dataset. This process had to be used in developing the GA because currently available GA versions did not generate the course taking pattern. After developing the code and the modifying the GA the had to be evaluated to check whether it is able to handle the dataset fed into the algorithm and if so whether it is able to generate the course taking pattern of students. This was a tedious process as the code developed to make algorithm run had to be modified through iterations. After several iterations, the modified showed that there is a possibility to generate course taking pattern. The modification of the coding had to take into account the attributes of the students and the number of students to ensure that the model generated by the algorithm is having a high probability of achieving the business objective. After reaching this stage the GA was actually used to test the dataset as per the parameters set to achieve the data mining goal.

(B) Generate Test Design:

- **Criteria to determine the goodness of the model:** These criteria were the same as mentioned Section 5.3.1. which includes measuring the precision and recall.

- **Definition of data on which the above criteria will be tested:** The full dataset was used to test the above criteria.

(C) Build Model:

- **Parameter Setting:**

  The experiment consists of two steps. The first step was a set of experiments conducted to find the best set of parameters. These parameters included population size (*ps*), chromosome length *(cl)*, crossover probability (*cp*), mutation probability (*mp*), and the termination criteria. The second step is to apply GA to generate course taking patterns to predict time-to-degree and CGPA and test its actual performance where the output was evaluated using the recall and precision measures.

- Further based on the discussions given in Section 2.5.2.5 the parameters set and tabulated as shown in Table 5.10.

Table 5.10, Parameter setting for Genetic Algorithm

Parameter Description	**Value**
Population size	100
Maximum number of generations	50
Chromosome length	60
Crossover rate	0.7
Mutation rate	0.7

A number of iterations had to be introduced in modelling data with regard to various steps.

(D) Generate Model:

Before producing the final model, a series of experiments were conducted by adjusting the cross-over, mutation, population size and chromosome length values in order to develop the GA to generate a course taking pattern that has a meaningful link to each student by ID and other attributes. The development of GA involved modification of the coding of the original GA which was specific to generation of course taking pattern as found in Table 5.11. Modification of coding was a complex process involved in the development of the GA for a specific purpose namely generation of course taking pattern. The final model thus generated is provided in Table 5.11 with student records and fields (CGPA, time-to-degree and semester) (see Table 5.9) which has columns student identification (student_id), time-to-degree (len), Cumulative CGPA (GPA), semester GPA (SGPA) and course taking patterns (course).

Table 5.11, Course taking patterns generated from Dataset without contextual factors Extracted from Student data

student_id	programme_	len	CGPA	SGPA	course	Pattern number 1
Stud1	BSAF	4.5	3.33	2.932	ACCT 301 ,FINC 310 ,FINC 321 ,FREN 101 ,STAT 202	Pattern1
Stud2	BSAF	4.5	3.06	2.4	ACCT 301 ,FINC 310 ,FINC 321 ,FREN 101 ,STAT 202	Pattern1
Stud3	BSAF	5	2.44	2.666	ACCT 312 ,BANK 220 ,CULT 101 ,ENGL 202 ,ITMA 201	Pattern2
Stud4	BSAF	6	2.28	2.4	ACCT 312 ,BANK 220 ,CULT 101 ,ENGL 202 ,ITMA 201	Pattern2
Stud5	BSAF	4.5	3.1	3.334	ACCT 311 ,ACCT 321 ,ENGL 202 ,FINC 321 ,STAT 202	Pattern3
Stud6	BSAF	4.5	3.49	3.468	ACCT 311 ,ACCT 321 ,ENGL 202 ,FINC 321 ,STAT 202	Pattern3
Stud7	BSAF	4.5	3.75	3.6	ACCT 301 ,ACCT 311 ,ARAB 102 ,FINC 421 ,ITMA 201	Pattern4
Stud8	BSAF	4.5	3.41	2.8	ACCT 301 ,ACCT 311 ,ARAB 102 ,FINC 421 ,ITMA 201	Pattern4
Stud9	BSAF	4	3.4	3.668333	ACCT 321 ,ARAB 201 ,BANK 220 ,FINC 431 ,ITCS 121 ,PHOT 101	Pattern5
Stud10	BSAF	4	3.46	3.723333	ACCT 321 ,ARAB 201 ,BANK 220 ,FINC 431 ,ITCS 121 ,PHOT 101	Pattern5
Stud11	BSAF	4.5	2.55	2.168333	ACCT 321 ,ACCT 402 ,BANK 302 ,ENGL 201 ,FINC 421 ,PHOT 101	Pattern6
Stud12	BSAF	4.5	2.33	2.333333	ACCT 321 ,ACCT 402 ,BANK 302 ,ENGL 201 ,FINC 421 ,PHOT 101	Pattern6
Stud13	BSAF	5	2.5	1	ACCT 403 ,BANK 302 ,ENGL 201 ,FINC 320 ,FINC 421	Pattern7
Stud14	BSAF	5.5	2.36	0.8	ACCT 403 ,BANK 302 ,ENGL 201 ,FINC 320 ,FINC 421	Pattern7
Stud15	BSAF	4	3.57	3.732	ACCT 301 ,CULT 102 ,ENGL 202 ,FINC 320 ,ITCS 121	Pattern8
Stud16	BSAF	4	3.84	3.666	ACCT 301 ,CULT 102 ,ENGL 202 ,FINC 320 ,ITCS 121	Pattern8
Stud17	BSAF	4	3.6	3.934	ACCT 402 ,ACCT 403 ,BANK 302 ,ENGL 201 ,VDEO 101	Pattern9
Stud18	BSAF	4	3.22	3.202	ACCT 402 ,ACCT 403 ,BANK 302 ,ENGL 201 ,VDEO 101	Pattern9
Stud19	BSAF	4	3.42	3.398	ACCT 312 ,ACCT 320 ,ACCT 341 ,BANK 302 ,FINC 310	Pattern10
Stud20	BSAF	4	3.23	3.2	ACCT 312 ,ACCT 320 ,ACCT 341 ,BANK 302 ,FINC 310	Pattern10
Stud21	BSAF	3	3.88	3.778333	ACCT 404 ,BANK 302 ,ECON 301 ,FINC 320 ,FINC 421 ,STAT 202	Pattern11
Stud22	BSAF	3	3.53	3.556667	ACCT 404 ,BANK 302 ,ECON 301 ,FINC 320 ,FINC 421 ,STAT 202	Pattern11

The Table 5.11 is an extraction from the larger table produced by the Genetic algorithm. Hence only sample set of students and courses have been provided along with other attributes CGPA, SGPA, Student ID, time-to-degree and course.

From Table 5.11 it can be seen that the model generated shows some relationship between course taking pattern, time-to-degree, SGPA and CGPA. Thus it is possible to generate the following functions which

CGPA = function (SGPA, course taking pattern, time-to-degree) → 5.5

Time-to-degree = function (SGPA, course taking pattern, CGPA) → 5.6

The Table 5.11 shows 11 patterns Pattern1 to Pattern11 by the algorithm. An inspection of the table shows that there are students with the same course taking pattern but varying CGPA e.g. Pattern1. Another aspect observed is that even for those students whose course taking pattern is the different the time-to-degree is same e.g. Pattern1, Pattern3 and Pattern6. These observations need to be understood along with other hidden knowledge that could be discovered from the table. This was made possible by the evaluation step provided next.

5.3.5 Evaluation:

The model generated by GA is table 5.11. This table has attributes as a function of students. That is to say rows represents students and column represent attributes. The input to the CRISP-DM process is a dataset prepared to be mined with 8 attributes see Table 5.8 and 1292 students. The model generated by the CRISP-DM process is a similar table as that of the input table except there was no course taking pattern which was not found in the dataset fed as found in the model See table 5.11. This is a major discovery. The course patterns generated have been named as Pattern 1,Pattern 2 and so on. The first inference that could be derived is that CRISP-DM process has generated knowledge about course taking patterns hidden in the dataset. This confirmed that the process is working.

Table 5.12, Confusion Matrix

	Predicted condition		
	Total population	**Prediction positive**	**Prediction negative**
True condition	Condition positive = 1200	1100 (TP)	100 (FN)
	Condition negative = 92	75 (FP)	17 (TN)

Table 5.13, Detailed Accuracy

True Positive Rate	**False Positive Rate**	**Precision**	**Recall**	**F-Measure**	**Prediction of time-to-degree**
0.916	0.083	0.821	0.984	0.971	Accurate

Prediction of time-to-degree in terms of course taking patterns – the result shows that it is accurate meaning that the course taking pattern determines the time-to-degree this is corroborated by the Tables 5.12 and 5.13. For detailed depiction of time-to-degree is provided in Appendix A.

The next step in the evaluation involves the testing of the performance of the algorithm and the process in terms of data mining goals set in section 5.3.1. The measures examined: F-measure was measured as 0.971 see table 5.13 the closer this figure is to 1 that much accurate is the performance of GA. The precision was measured as 0.821 see table 5.13 the closer this figure is to 1 that much accurate is the performance of GA. The recall was measured as 0.984 see table 5.13 the closer this figure is to 1 that much accurate is the performance of GA.

From the above report it can be inferred that the algorithm was performing properly and hence the process.
After checking the testing parameters successfully the evaluation proceeded to examine the attributes and instances of the Table 6.11. From Table 5.11 it can be seen that the following functions can be derived.

Time-to-degree = function (course taking pattern, CGPA, SGPA, student ID) → 5.7

CGPA = function (course taking pattern, Time-to-degree, SGPA, student ID) → 5.8

If one compares the Table 5.11 with the equation 5.7 it can be see that when course taking pattern varies then the time-to-degree varies in some cases but does not vary in some cases. However if one compares the Table 5.11 with the equation 5.8 it can be see that when course taking pattern varies then the CGPA varies in all cases and even cases where the course taking pattern is the same the CGPA is not the same. This is a unique finding. Thus it is not possible to clearly argue that course taking patterns alone affect the time-to-degree or CGPA. On the other hand when one changes CGPA in equation 5.7 then also it can be seen that time-to-degree remains constant in some cases and remains constant in some cases. But what emerges is that when the attribute time-to-degree is checked there it is seen that some students have graduated in 3 years while some other students have graduated in 6 years with varying combination of courses and CGPA. For instance if one peruses the rows under pattern P11, it can be seen that students Stud 21 and Stud22 have graduated in 3 years and have taken the same set of courses but their CGPA is not the same. Thus if pattern P11 is taken as the most optimum result which shows that a certain course taking pattern indicates the best possible CGPA and shortest time-to-degree. Here the attraction is the time-to-degree which is seen to be the shortest amongst the lot.

So the following questions can be raised

Does course taking pattern really determine the time-to-degree and CGPA?

Does time-to-degree affect CGPA or vice versa?

The answer to the question is that when the outcome of the data in Table 5.11 is perused there appears to be some relationship between the course taking pattern and time-to-degree. For instance while Stud21 has taken a time-to-degree of 3 years to graduate with a course taking pattern indicated by P11, Stud17 has taken a time-to-degree of 4 years to graduate with a course taking pattern P9. When the students Stud21 and Stud17 compared then the following aspects emerge.

The course taking patterns are different in a particular semester number

The number of courses taken by Stud21 is 6 whereas the number of courses taken by Stud17 is 5.

The Stud21 has scored the highest CGPA of 3.88 while the Stud17 has scored a CGPA of 3.6. That is to say despite taking a longer time-to-degree and lower number of courses Stud17 has not achieved the highest CGPA and lower time-to-degree. It is possible to argue that in the case of Stud17 the course taking pattern might have affected the student while in the case of Stud21 it might not have been. Or in both the cases the course taking pattern might have affected the students with one student achieving the highest CGPA in a shorter time-to-degree due to one particular course taking pattern and the other adopting a different course taking pattern could not achieve the highest CGPA and shorter-time-to degree. If this argument is used then there is a possibility to fix the course taking pattern (Pattern 11) of Stud21 as the reference point and the other students can be compared to this reference. Another interesting aspect is that Stud21 has registered in the maximum of 6 courses in that semester (maximum allowed load in a semester in the University where this research was conducted) whereas Stud17 has registered in only 5 courses which is not the maximum. Thus it emerges that some attribute of a student or the course or both might be contributing to this.

However if the performance of Stud21 is analysed then it can be seen that it can be taken as reference and the other students could be encouraged to achieve it taking into consideration the course taking pattern adopted by Stud21, the number of courses taken by the student in the semester, the CGPA scored and some other attribute or attributes of the student or the course that needs to be discovered. As far as the other attribute or attributes are concerned, the performance of Stud22 gives some information. For instance the course taking pattern and time-to- degree of Stud22 is exactly same as that of Stud21 but the CGPA is different (3.53). It can be argued that either Stud22 had less potential than Stud21 or the courses were more difficult to handle for this student than Stud21 and hence could not score a high CGPA. This can be corroborated with the performance of students Stud15 and Stud16 which is very similar to the performance of the students Stud21 and Stud22. That is Stud15 and Stud16 have registered in the same number of courses (5 in number) and the same pattern of courses P8 but Stud16 has scored a CGPA of 3.84 while Stud15 has scored a CGPA of 3.57. The same arguments as those posited for Stud21 and Stud22 are valid here, that is to say either Stud15 had less potential than Stud16 or the courses were more difficult to handle for this student than Stud15 and hence could not score a high CGPA. This needs to be investigated. This is new knowledge discovered.

However the most important knowledge discovery that emerges is when one compares Stud 21 and Stud16 as both have scored a CGPA which is in the proximity of each other but there is a one year gap in the time-to-degree between the two with Stud16 taking 4 years as the time-to-degree and Stud21 taking 3 years for time-to-degree. That is Stud 16 has scored a CGPA of 3.84 while Stud21 has scored a CGPA of 3.88 a difference of 0.04 for a time-to-degree difference of one year. The reason for this could be: course taking pattern, number of courses registered in the semester, student attribute and course attribute. The latter two attributes are unknown yet. This comparison yields knowledge that although Stud16 appears to have the potential to score high CGPA, perhaps the pattern of courses taken by this student, the number of courses taken in a semester and another unknown attribute of the course could have contributed to the higher time-to-degree. This comparison clearly points out that course taking pattern could be a significant factor that affects time-to-degree and CGPA. This argument when taken in conjunction with the performance of Stud22 it can be argued that apart from student potential an attribute of the course namely course difficulty referred under Section 2.3.3 (of chapter 2) may have a role to play. Thus keeping the student potential as constant it may be worthwhile to vary the course difficulty and check whether the performance of both Stud22 and Stud16 could approach that of Stud21 in terms of the time-to-degree and CGPA. This knowledge hidden in the dataset is clearly not appearing to be easy to discover.

This brings the evaluation to the case of other students found the table. Similar arguments could be extended their performance with one exception that not all students may be able to approach the performance of Stud21 in terms of course taking pattern, number of courses per semester, CGPA and time-to-degree. This argument led to the examination of student who have scored lower CGPA and taken longer-time-to-degree with varying course taking pattern and number of courses per semester. For instance Stud 14 has scored the lowest CGPA of 2.36 and has taken longer time-to-degree (5.5 years). Although this student has taken 5 courses in the semester with a pattern P7, the number years to degree could have been higher because of many other reasons like registration in lower number of courses in semesters, course taking pattern, withdrawals from the semester, student potential, course difficulty and the like. This case needs to be examined in greater detail not just course taking pattern and time-to-degree.

Thus it can be seen that in most of the cases reported in the Table 5.11 except for the case of Stud22 and Stud16 the performance of all other students might have been affected by course taking pattern, course difficulty, student potential and some other factors. Thus it is reasonable to conclude that course taking pattern along with student attributes or course attributes or both do seem to contribute as one of the major factors that could determine the time-to-degree and CGPA. In addition the cases of Stud16, Stud21 and Stud22 when compared show that course taking pattern may play a limited role in regards to the time-to-degree and CGPA as the students approach performance levels close to a CGPA of 4.0. In contrast course taking pattern may play a significant role in regards to the time-to-degree and CGPA at performance levels much less than that of Stud21 for instance Stud 11 who has scored a CGPA of 2.55 and has taken 4.5 years to degree.

With regard to the question does time-to-degree affect CGPA or vice versa the results reported in Table 5.11 show that CGPA and time-to-degree do not vary according to any specific rule or formula or pattern. For instance for the same time-to-degree students Stud 21 and Stud 22 have scored different CGPA. This is also seen in a number of cases of other students in the table. Thus there is no specific relationship that could be used to predict time-to-degree in terms of CGPA or vice versa. However it is possible to use the two in conjunction to determine the course taking pattern or number of courses a student can register in a semester or course difficulty or student potential with the latter two not having been investigated yet. That is to say when course taking pattern is considered to affect the time-to-degree it is necessary to consider CGPA alongside and vice versa. This is needed because it is not possible to improve CGPA without taking into account the time-to-degree and vice versa as optimum performance of students is only possible to be determined if both are taken into account, which is new knowledge. For instance the ideal performance of a student could be scoring a CGPA of 4.0 taking three years to degree.

The discussions lead to the following inference:

Time-to-degree = function of (course taking pattern, number of courses registered per semester)→ 5.9

CGPA = function of (course taking pattern, number of courses registered per semester) →5.10

If one brings in the concept of course difficulty of a course (contextual factor) as affecting the courses and hence the course taking pattern then it may be worthwhile to revise the equations 5.1 and 5.2 as follows:

Time-to-degree = function of (course taking pattern, course difficulty, number of courses registered per semester) → 5.11

CGPA = function of (course taking pattern, course difficulty, number of courses registered per semester) → 5.12

These functions may reveal further hidden knowledge in the dataset and better prediction of time-to-degree and CGPA. Thus the two equations could be combined as follows

Function (time-to-degree, CGPA) = function of (course taking pattern, course difficulty, number of courses registered per semester) →5.13

Thus the evaluation shows that the business goals were only partially achieved. A number of iterations had to be introduced in understanding data with regard to various steps.

5.3.6 Findings

The evaluation has led to a number of findings discussions about which follow.

5.3.6.1 The functions 3.9 and 3.10 have only been partially realized. That is to say CGPA is determined by only course taking patterns and number of courses and not time-to-degree or course difficulty. Similarly time-to-degree is determined by only course taking patterns and number of courses and not CGPA or course difficulty.

5.3.6.2 The extent to which CGPA and time-to-degree vary with respect to variation in course taking pattern is not known. This is a limitation.

5.3.6.3 Course difficulty as a contextual factor was not mined by the algorithm but may have relationship with CGPA and time-to-degree which needs to be investigated.

5.3.6.4 It is possible to predict CGPA and time-to-degree together in terms of course taking pattern and number of courses.

5.3.6.5 The direction of change in CGPA and time-to-degree to determine the optimum performance should be the opposite. That is the highest CGPA should be scored when the student graduates with the shortest time-to-degree. That is there is a kind of inverse variation observed between CGPA and time-to-degree. This is new knowledge.

5.3.6.6 Using this knowledge it is possible to identify students with lower CGPA and taking longer time-to-degree and provide additional academic support to enhance CGPA and reduce the time-to-degree by altering course taking pattern to match that of the best student or students.

5.3.6.7 The genetic algorithm developed can be used to generate course taking pattern of students as a model of classification technique in terms of a set of courses clearly defining the exact course in the set for each student. This is a major contribution as a GA algorithm to generate course taking pattern of students and predict either time-to-degree or CGPA does not exist until now.

5.3.6.8 The classification technique was found to generate the best model when compared to clustering or association rule technique and enabling the CRISP-DM KDDM process to perform nearly as defined in Section 3.2 with some deviation. For instance the CRISP-DM process does not entail developing new algorithms to discover knowledge and deploy. In this research it was necessary to develop a GA that could be used to fulfil the business objective in terms of generating the course taking pattern.

5.3.6.9 There is no relationship between CGPA and time-to-degree and hence it is not possible to determine one with the other. This is a limitation and contradicts the findings derived in section 4.2.6 where it is argued that CGPA determines time-to-degree. From the above it is concluded that hypotheses HA and HB were partially achieved (see equations 4.9 and 4.10).

5.3.6.10 EDM has been integrated in CRISP-DM process model.

5.3.7 Limitations

5.3.7.1 The main limitation is the need to develop new algorithm to achieve the business goals in terms of course taking pattern to determine CGPA and time-to-degree. Already existing GA algorithms do not readily allow the generation of patterns so developing a new GA is a tedious and complex tasks.

5.3.7.2 Without an appropriate algorithm the CRISP-DM process will not perform as expected.

5.3.7.3 The prediction of CGPA and time-to-degree is not accurate in terms of course taking pattern.

5.3.7.4 There is a need to mine dataset to discover contextual factors in the absence of which the prediction of CGPA and time-to-degree may not be accurate.

5.3.7.5 The CRISP-DM process in the current form is unlikely to enable the user to discover accurate knowledge to achieve the business goal. Although partially it does.

5.3.8 Summary

The classification technique used in the CRISP-DM process was able to generate course taking pattern of students more accurately. However there was a need to develop a new GA. The CRISP-DM process with the newly developed GA is able to generate a model that provided knowledge to relate course taking pattern of students to CGPA and time-to-degree. However since prediction of the classified courses as pattern was only partially successful it was felt that there could be additional factors that were not mined by the GA which could have been responsible for the limitation arising in predicting the CGPA and time-to-degree. It was suggested that student potential and course difficulty could be considered as contextual factors affecting the predictive functioning of the algorithm. Since at this stage it is only suggested at the concept level further investigation is needed to whether contextual factors could be integrated into the dataset in the CRISP-DM process to verify whether it is possible to predict CGPA and time-to-degree in terms of course taking pattern. However as far as the CRISP-DM process was concerned it has been established that process could generate course taking pattern as a set of courses using the classification technique only. This in turn provides a basis for further investigation to see whether contextual factor could be extracted to know their influence on course taking pattern and hence CGPA and time-to-degree.

5.4 Comparison of the results of the data mining techniques used in CRISP-DM process

Table 5.14, Comparison of Data Mining techniques used in CRISP-DM

Data mining Techniques **Factors for comparison**	**Clustering technique**	**Association rule technique**	**Classification using genetic algorithm**	**Remarks**
Knowledge discovered	1. Student course registration data by semester can be considered as equivalent to course taking pattern of students. 2. The discovered student attributes can be considered as course taking pattern. 3. CGPA and time-to-degree could determine each other.	1.The key courses identified if registered in the way shown could lead to similar time-to-degree 2. The relationship that CGPA acts as the antecedent of time-to-degree 3. There are courses that can be linked to the CGPA and together they act as antecedent to time-to-degree 4. The rule shows that course sequence number and semester number should treated as running number and these sequences are unique to each student.	1. Optimum performance of students is only possible if time-to-degree and CGPA both are considered. 2. It is possible to predict CGPA and time-to-degree together in terms of course taking pattern and number of courses. 3. There is a kind of inverse variation observed between CGPA and time-to-degree.	Knowledge discovered by clustering and association rule is more useful to other aspects of education delivery than predicting time-to-degree and CGPA in terms of course taking pattern. However knowledge discovered by using GA could predict time-to-degree and CGPA.
Usefulness of	(a) student can be advised to stick	1. Students can be advised	Using this knowledge it is	Knowledge discovered by

discovered knowledge	to optimum time-to-degree prescribed by university with regard to core courses and humanities, (b) HEIs can now assign students to specific groups or sections depending on their course registration pattern and provide them adequate academic support that may produce better results, (c) redesign curriculum to get better results(d) knowledge related to time-to-degree, CGPA and course taking patterns can be used to decide on the student learning assessments	to register to key courses in the way it was prescribed in the rules. 2. The other courses that are linked could be discovered and used to advise students.	possible to identify students with lower CGPA and taking longer time-to-degree and provide additional academic support to enhance CGPA and reduce the time-to-degree by altering course taking pattern to match that of the best student or students.	clustering is more useful for profiling students. Knowledge discovered by association rule is more useful to advise students. Knowledge discovered by GA is more useful for prediction of time-to-degree and CGPA in terms of course taking pattern.
Course taking pattern generation	Partial generation of course taking pattern in terms of student attributes that can be considered as course taking pattern.	Partial generation of course taking pattern in terms of key courses.	Complete generation of course taking pattern	Course taking pattern was generated as a set of courses only by classification using GA.
Achievement of business goal	Partial achievement of business goal.	Partial achievement of business goal	Partial achievement of business goal	The three techniques were useful in achieving the business goals partially only. However out of the three techniques the classification provided the closest solution in terms of generating the course taking pattern as a set of courses to predict time-to-degree and CGPA directly whereas the remaining two produced only a relationship between course taking pattern and time-to-degree and CGPA indirectly.
Limitations	1. The business goal 3.9 and 3.10 were partially achieved. 2. Contextual factor (course	1. Association rule cannot discover the exact time-to-degree or CGPA	1. Contextual factor (course difficulty) was not discovered in the process.	Amongst the limitations of the three techniques the limitations of classifications

	difficulty) was not discovered in the process. 3. Analysis of clusters with many attributes to make decisions could be erroneous. 4. Relationship of attributes within the clusters is not known.	2. It cannot discover a set of courses. 3. Contextual factor (course difficulty) was not discovered in the process.	2. The prediction of CGPA and time-to-degree is not accurate in terms of course taking pattern. 3. The need to develop new algorithm to achieve the business goals in terms of course taking pattern to determine CGPA and time-to-degree.	affect the KDD process the least in terms of achieving the business goal. Further the three techniques did not produce any knowledge related to contextual factors which may play a role in accurately predicting the time-to-degree and CGPA in terms of course taking pattern.

From the Table 5.14 it is clear in order to predict time-to-degree and CGPA in terms of course taking pattern the classification technique is the one which is closest to produce a model that uses a set of courses as pattern. Thus it can be concluded that the classification technique used in CRISP-DM process is the one that would be used for predicting the time-to-degree and CGPA in terms of course taking pattern. In addition in order to remove the limitations in producing an accurate model that could be used to predict the time-to-degree and CGPA in terms of course taking pattern this research went one step further. That is a contextual factor related to the courses namely course difficulty was proposed to be included in the CRISP-DM process to check whether it can really enable an accurate prediction of the time-to-degree and CGPA in terms of course taking pattern. The inclusion of a contextual factor was thought of based on the theoretical arguments presented in chapter 2 and chapter 3. Including contextual factors in the CRISP-DM process involves a complex method. Thus while it was concluded at this stage that CRISP-DM process indeed is helpful in achieving the business goal set for the KDD process, it was not possible to entirely achieve the goal and hence further investigations were carried out. Two aspects were considered. One was checking whether the limitations of the CRISP-DM process outlined in Table 2.11 (Chapter 2) contributed to it. Another was to study the theoretical aspects that could provide the basis for introducing the change in the process. These aspects are discussed next.

5.5 Enhancement of CRISP-DM process to include contextual factor for more accurate prediction of time-to-degree and CGPA

An important conclusion that can be arrived at from an inspection of Table 4.14 is that the hypotheses 3.9 and 3.10 have been only partially achieved. One of the reasons that have been explained in the precious section is that additional attributes related to courses may have to be added to the educational dataset. This attribute is termed as the contextual attribute. In Section 2.3.3 contextual factors have been discussed in detail and one factor that has been identified as affecting courses and probably course taking pattern based on the literature is course difficulty. Extracting course difficulty hidden in the educational dataset and integrating the course difficulty data into the dataset is a complex process. In addition if this complex process has to be integrated into the CRISP-DM process to generate course taking pattern characterized by course difficulty and verify whether hypothesis 3.9 and 3.10 have been achieved or not. Achievement of hypothesis 3.9 and 3.10 requires modification in the process of the original CRISP-DM process an argument supported by process theory.

According to process theory (Mahr, 1982), an outcome of a process cannot only be determined by variation in the necessary condition for instance input but additional ingredients related to the outcome. For instance, if course taking pattern can be generated from the dataset then the dataset with its attributes becomes a necessary condition to generate course taking pattern. However if the course taking pattern has to predict time-to-degree the dataset with the current attributes by itself will not be a sufficient condition to enable course taking pattern to predict time-to-degree. Additional ingredients like contextual factors may be needed to determine the sufficient condition. Using this theory it can be argued that if the current CRISP-DM process has to include the contextual factor and generate a context based course taking pattern then the existing CRISP-DM process needs be modified to satisfy the necessary and sufficient conditions. Necessary condition is providing the dataset that is contextualised. Sufficient condition is to identify the stage at which contextualization of the course taking pattern could be contextualized using external input.

The CRISP-DM process shown in figure 4.3 was used in order to introduce contextual factor and generate patterns with course difficulty attributes. The classification technique was the one chosen for developing a CRISP-DM process that can produce contextually characterized pattern as clustering and association rule techniques were not found suitable was they had limitation introducing patterns (see table 4.14). The genetic algorithm used was the one described in section 2.5.2.5. The initial model of CRISP-DM was altered in a simple fashion as shown in Figure 5.3.

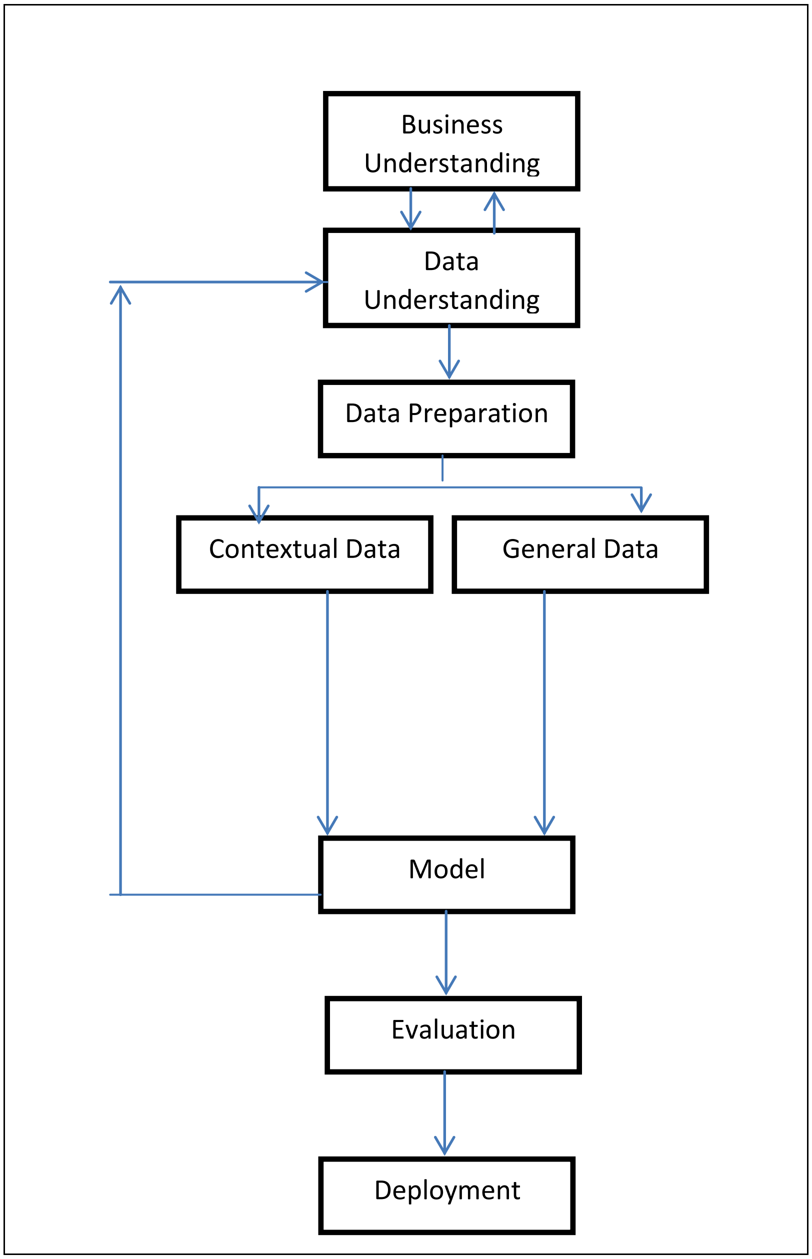


Figure 5.3, Modified CRISP-DM model - 1

In Figure 5.3 it can be seen that contextual data has been extracted from the general student data and separately fed into modelling stage along with the general student data. This modification was affected by a limitation.

The limitation is separating context data from general student data and feeding the dataset as given in Figure 5.3 to the modelling stage is difficult as the modelling requires a single file and available software tools can take only one file.

In order to overcome this limitation the CRISP-DM process in Figure 5.3 was modified further as shown in Figure 5.4.

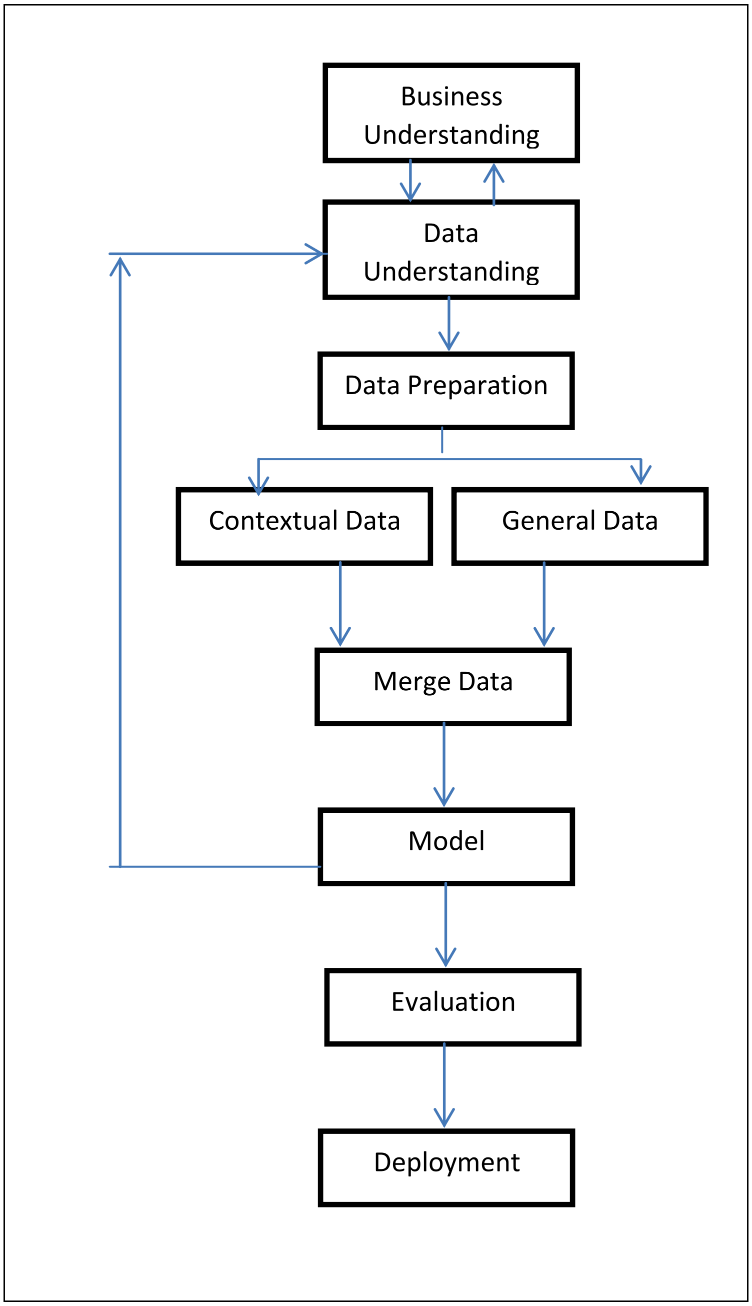


Figure 5.4, Modified CRISP-DM model 2

5.5.1. The course taking pattern and course difficulty by pattern were not produced separately. This will require manual intervention.

5.5.2 If there is a difficulty at the modelling stage and the results have to be feedback to the data preparation stage the results will go to both contextual and general data stages. This could result in an error of choosing the right set of patterns to be analysed. E.g. when the patterns generated is fed back to both contextual and general it is possible that the miner could erroneously deal with course taking pattern as contextual data pattern and vice versa. This could result in grave errors. This limitation must be solved.

5.5.3 The business goals may not be interpreted properly at the evaluation stage is the course taking pattern and contextual pattern are grouped together without distinction due to manual intervention business goals could be interpreted wrongly.

Further to the above discussions the final model to be tested in Chapter 6 is given in Figure 5.5.

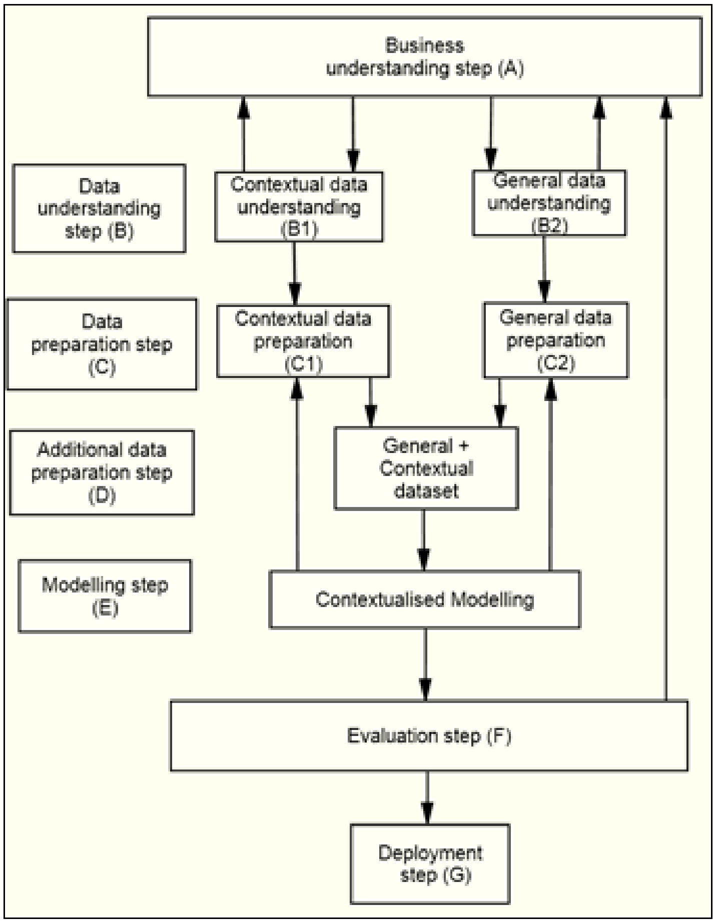

Business understanding step (A)
Data understanding step (B)
Contextual data understanding (B1)
General data understanding (B2)
Data preparation step (C)
Contextual data preparation (C1)
General data preparation (C2)
Additional data preparation step (D)
General + Contextual dataset
Modelling step (E)
Contextualised Modelling
Evaluation step (F)
Deployment step (G)

Figure 5.5, Final Modified CRISP-DM model

5.6 Summary

The outcome of this chapter provided a complete testing and evaluation of the CRISP-DM process in order to integrate EDM into the process and develop the best model that could be used to discover hidden knowledge from the dataset. The discovered models were generated by three techniques namely clustering, association rules and classification. The CRISP-DM process model template provided by Chapman et al. 2000 was used in the exact manner defined in the document developed by those authors titled CRISP-DM.pdf. Each one of the steps have been matched to the current research and has been outlined in section 4.2. This section provided the CRISP-DM process tailored for testing the functions 3.9 and 3.10. The clustering, association rule and classification techniques were aligned with the general steps in section 4.2. The results produced by the 3 different techniques were evaluated and a comparison provided in table 6.14. The comparison showed that classification technique produced the closest model that could enable the discovery of course taking patterns to predict time-to-degree and CGPA. However it was also found the results produced by classification technique could only partially satisfy functions in 3.9 and 3.10. Thus the chapter proceeded to check how this could be tackled as the main limitation was the lack of an accurate prediction of time-to-degree and CGPA in terms of course taking pattern and course difficulty. Course difficulty could not be discovered using any techniques although in chapter 3 it was argued that course difficulty could enable accurate prediction of time-to-degree and CGPA as contextual factor.

Based on further investigation and theoretical support the CRISP-DM process was further undertaken to be modified. The investigations showed that the CRISP-DM process needs to be modified in a step by step manner as described in section 4.2. The final CRISP-DM process model thus developed was used to conduct experiments to verify the functions 3.9 and 3.10 in chapter 6.

Chapter 6 : Enhancement and Experimentation of CRISP-DM

6.1 Introduction

In chapters 4 and 5, the integration of EDM in KDDM process namely CRISP-DM has been achieved through a demonstration of experiments in terms of developing 3 models namely clustering, association rule and classification. However the integration when tested to determine the time-to-degree of students and CGPA using the course taking pattern of students in semesters did not provide clear prediction of time-to-degree and CGPA in terms of course taking pattern. It was argued that some factors related to courses hidden in the dataset could be the reason for the ambiguity in the prediction of time-to-degree and CGPA using course taking pattern and evaluates the validity of equations 3.9 and 3.10. Unobservable factors hidden in the data called contextual factors (e.g. course difficulty) were thought of as one of the reasons that could result in this lack of clarity in prediction. The reasons for such an assumption have been explained in chapter 5. This chapter checks or tests if such an assumption is valid by introducing modification in the CRISP-DM process tested in chapter 6 using genetic algorithm for classification. The steps provided in section 3.2 have been followed step by step in this chapter to test the modified CRISP-DM process model provided in Figure 5.5 modified CRISP-DM.

6.2 Explanation of the modification introduced in the original CRISP-DM model

It was proposed in Section 2.2 of chapter 2 that the course taking pattern of students could be related to the contextual factors related to the course. The contextual factor thus selected for investigation was course difficulty. As explained the Section 3.2, course difficulty is an attribute of each course but is unique ***to*** each student. For instance in Table 3.6 the course difficulty figures for the Stud1 for the ***five*** courses (ACCT 101, ARAB 101, ECON 101, ENGL 101, ENGL 102) in which the student had registered are (5.57, 5.59, 5.62, 5.64, 5.62) respectively. Similarly the course difficulty figures for Stud8 for the five courses (ACCT 101, ECON 101, ENGL 101, ITCS 101, and MATH 103) in which the student had registered are (6.39, 6.36, 6.36, 6.37, and 6.32). It can be seen that the course difficulty figures for the course ACCT 101 for Stud1 is 5.57 whereas for STUD8 is 6.39 which shows that the course difficulty attribute is not only course specific but also student specific. In addition like the course taking pattern, there is a course difficulty pattern that is emerging for each set of course taking pattern for each student. This pattern of course difficulty for each student when linked to the course taking pattern of the student then there could be knowledge hidden that could inform how the course taking pattern could be linked to the time-to-degree and CGPA. For instance Stud 1 has taken 3.5 years to graduate whereas Stud8 has taken 4 years to graduate. The course difficulty measured for Stud8 was higher than that of Stud1. Stud1 and Stud8 had registered in 5 courses each. In order to know whether the performance of the two students was different or the same, CGPA was compared. CGPA of Stud1 was 3.78 whereas that of Stud8 was 3.84. It can be interpreted that the course taking pattern of Stud1 has resulted in lower course difficulty measures and shorter time-to-degree whereas the course taking pattern of Stud8 has resulted in higher course difficulty measure and longer time-to-degree. The difference in the CGPA between two students is only 0.06 with Stud 8 scoring higher CGPA of 3.84. Therefore it can be seen that the course taking pattern of Stud1 appears to yield an optimum time-to-degree and CGPA. That is if the course taking pattern of Stud8 was adjusted to that of Stud1 then Stud8 might graduate in a time-to-degree that is less than 4 years. This argument needs to be tested. Based on the above the following equation was constructed.

This contextual attribute needs to be extracted from the data through mining and then linked to course taking pattern which in turn could be related to time-to-degree. That is to say in order to determine the time-to-degree more accurately it is essential to extract course difficulty and link it to course taking pattern to determine the time-to-degree. This assumption led to the design and development of the modified CRISP-DM model using which course taking pattern and the course difficulty pattern were generated to determine time-to-degree and CGPA using a single process. Using a single process it was possible to integrate two educational datasets, one general and the other contextual to generate patterns of two interrelated variables and to make more accurate predictions.

6.3 Explanation of the integration of the contextual and general education dataset

From Figure 4.4 it is clear that the raw education dataset has been subjected to processes of data understanding and data preparation in parallel because there is a need to mine both contextual data and raw data at the same time. From an inspection of Figure 4.4 it can be seen that the data preparation for generating course taking pattern and the data preparation for generating course difficulty pattern are different. The difference lies in the fact that course difficulty data needs to be extracted from the raw data using an SQL query process for each course for each student by semester and linked to each course of a student accurately. This needs to be done in a way that satisfies the equation 2.1 that provided the formula to generate course difficulty pattern.   As far as the course taking pattern data preparation was concerned the SQL query used does not require any formula to be used. Hence the SQL query developed for both the situations are not the same and hence different paths are needed to prepare the dataset for generating patterns for both the course taken by student and the course difficulty by semester.

As can be seen from the discussion above now two datasets will be generated which need to be processed to generate two different patterns as part of the KDD process. The KDD processes that dealt with both the contextual dataset and course taking pattern dataset (termed as general dataset in this research) are the same but need to happen in parallel. As explained in Section 2.11 CRISP-DM process was chosen as the KDD process for conducting the experiment to introduce the contextualisation of the general dataset although nowhere in the CRISP-DM process literature it has been highlighted that there could be a way by which contextual factors could be processed. In the literature (see Section 2.6) it is widely believed that KDD processes are not capable of handling contextual aspects. This was demonstrated taking the case of CRISP-DM process in Chapters 5 and 6. The experiment in this research, this aspect was tested in way depicted in the modified CRISP-DM process developed in this research (Figure 4.4). The assumptions and steps used in the modified CRISP-DM process are very similar to those explained in Chapters 3 and 4. That is this research uses exactly the same steps outlined for implementing the CRISP-DM process to integrate the EDM but tests the contextualisation of the general dataset using the contextual factor mining concepts or theoretical support provided by Vert et al. (2010). Thus the new CRISP-DM should demonstrate that the model produced at the data mining stage indeed is contextualised and the concepts supporting the contextualisation are satisfied using a separate pseudo code, computer program and set of parameters identified in Chapter 3 (section 3.3). How the new modified CRISP-DM performed is now explained in the next section.

6.4 Prediction of optimum time-to-degree and CGPA in terms of the course taking pattern and course difficulty using modified CRISP-DM process model:

6.4.1 Business understanding:

(A) Determine business objectives -Predicting optimum time-to-degree and CGPA in terms of course taking pattern of students and course difficulty to enhance student learning experience.

(B) Assessment of situation – similar to section 5.3.1

(C) Determination of data mining goals –

Identification of data mining technique: It was assumed that there could be a relationship between course taking patterns, course difficulty, CGPA and time-to-degree which could enable the prediction of CGPA and time-to-degree, to achieve the business goal. As mentioned in section 5.3.1 prediction related tasks in data mining literature are shown to be using Genetic algorithm and considered to be the most suitable technique to achieve the business goal (Bajpai & Kumar, 2010).

Documenting the technical goals: Similar to section 5.3.1

Data mining goal success criteria:

Predict time-to-degree and CGPA by extracting course taking patterns and course difficulty pattern.

6.4.2 General Data Understanding:

- **Collect initial data -** Refer section 5.3.2.
- **Describe data –** Refer section 5.3.2.
- **Explore data**. Refer section 5.3.2.
- **Verify data quality -** Refer section 5.3.2.

6.4.3 Contextual Data Understanding:

Here the event is registration to courses and getting grades. The contextual factors identified for the business problem is mentioned in brackets under each category.

(A) Dimensions of context.

- time – the span of time and characterization of time for an event (semester of registering the course)
- space – the spatial dimension (Class size of the course)
- impact – the relative degree of the effect of the event on surrounding events (GPA, time-to-degree)
- similarity – the amount by which events could be classified as being related or not related (semester GPA,CGPA).

(B) Information criticality factors (ICF).

- time period of information collection (last 5 years graduate data was taken removing the old graduates)
- criticality of importance,
- impact (semester GPA, semester passed credits)
- ancillary damage of miss classification (CGPA and time-to-degree)
- spatial extent data set coverage (entire HEI)
- Proximity to spatially or conceptually to other related data sets (CGPA).

(C) Quality of the data

- currency, how recently was the data collected, is the data stale and smells bad (last 10 years' data of graduate students)
- ambiguity, when things are not clear cut –(same Course difficulty figure occurring for two different students registering in the same course)
- contradiction, what does it really mean when conflicting information comes in different sources (higher course difficulty leading to shorter time-to-degree or higher CGPA)
- Truth, how it can be known this is really the truth and not an aberration (higher course difficulty leading to longer time-to-degree or lower CGPA).

6.4.4 General Data Preparation:

(A) Select data

This dataset was extracted from the initial dataset described in section 6.4.2.The final dataset comprises 8 attributes namely Student ID, Programme , CGPA, time-to-degree, semester passed credits, semester GPA, course code and semester which were used for mining (See table 5.1). The dataset pertains to 1292 students belonging to 12 programmes and graduated during the period 2003 to 2014 with each student having 56 records. The data size was (1292 x 56) = 72352 records. Data stored in various tables was joined in a single table in this stage. Table 6.2 shows examples of the records of some students. (See table 6.1).

Table 6.1, General Dataset fields for Genetic Algorithm

Attribute	Description	Example
Student ID	Student Identification Number	Stud1
Programme	Programme name	BSAF
CGPA	Cumulative GPA	3.78
Time-to-degree	Years taken to complete the programme	3.5
Semester Completed Credits	Credits completed in the semester	15
Semester GPA	GPA scored in the semester	3.668
Course code	Course code registered	ACCT101
Semester	Semester number	1

(B) Clean data – refer section 5.3.3.

(C) Construct new data- refer section 5.3.3.

(D) Integrate data – similar to section 5.3.3.

(E) Format data – The data was formatted to suit the genetic algorithm in terms of rows and columns (See table 6.2).

Table 6.2, Data set used for genetic algorithm

student_id	len	gpa	semester	sem_gradepoints	course_code
Stud1	4	2.17	1	1.89	ENGL 050
Stud1	4	2.17	1	1.89	HIST 121
Stud1	4	2.17	1	1.89	ITCS 101
Stud1	4	2.17	1	1.89	MAGT 121
Stud1	4	2.17	1	1.89	MATH 052
Stud1	4	2.17	2	2.198	ACCT 301
Stud1	4	2.17	2	2.198	ENGL 102
Stud1	4	2.17	2	2.198	MAKT 201
Stud1	4	2.17	2	2.198	MATH 104
Stud1	4	2.17	2	2.198	STAT 101
Stud1	4	2.17	3	1.723333	ACCT 321
Stud1	4	2.17	3	1.723333	ACCT 403
Stud1	4	2.17	3	1.723333	CULT 102
Stud1	4	2.17	3	1.723333	ENGL 201
Stud1	4	2.17	3	1.723333	FINC 320
Stud1	4	2.17	3	1.723333	ITMA 201
Stud1	4	2.17	4	3.556667	ACCT 341
Stud1	4	2.17	4	3.556667	ARAB 102
Stud1	4	2.17	4	3.556667	FINC 431
Stud1	4	2.17	5	2.064	ACCT 101
Stud1	4	2.17	5	2.064	ARAB 101
Stud1	4	2.17	5	2.064	ECON 101
Stud1	4	2.17	5	2.064	ENGL 101
Stud1	4	2.17	5	2.064	MATH 103
Stud1	4	2.17	6	1.332	ACCT 311
Stud1	4	2.17	6	1.332	BANK 220
Stud1	4	2.17	6	1.332	ENGL 201
Stud1	4	2.17	6	1.332	FINC 210
Stud1	4	2.17	6	1.332	STAT 202

Stud1	4	2.17	7	2.22	ACCT 402
Stud1	4	2.17	7	2.22	BANK 302
Stud1	4	2.17	7	2.22	CULT 101
Stud1	4	2.17	7	2.22	ENGL 202
Stud1	4	2.17	7	2.22	FINC 321
Stud1	4	2.17	7	2.22	FINC 421
Stud1	4	2.17	8	3.33	ACCT 499
Stud1	4	2.17	8	3.33	INTR 400
Stud1	4	2.17	9	2.333333	ACCT 201
Stud1	4	2.17	9	2.333333	ECON 102
Stud1	4	2.17	9	2.333333	ITCS 121
Stud1	4	2.17	10	2.553333	ACCT 312
Stud1	4	2.17	10	2.553333	ACCT 320
Stud1	4	2.17	10	2.553333	FINC 310
Stud1	4	2.17	11	1.666667	BFRM 498
Stud1	4	2.17	11	1.666667	ECON 301
Stud1	4	2.17	11	1.666667	ECON 421

6.4.5 Contextual Data Preparation:

This step involved the extraction of course difficulty data of students and courses in which the students had registered as mentioned in section 6.4.4. This stage prepares the dataset to be of high quality defined as per the steps given below.

(A) Select contextual data

Table 6.3, Contextual Dataset fields for Genetic Algorithm

Attribute	**Description**	**Example**
Student ID	Student Identification Number	Stud1
Course code	Course code registered	ACCT101
Semester	Semester number	1
Course difficulty See section 2.3.3	$Difficulty_c = \frac{\sum_{t \in BE_c} \sum_{j=1}^{m_t} G_{j,t} * W_t}{\sum_{t \in BE_c} W_t * m_t}$	0.5678

(B) Clean contextual data – similar to section 5.3.3.

(C) Construct new contextual data- similar to section 5.3.3.

(D) Integrate contextual data – similar to section 5.3.3.

(E) Format contextual data – The data was formatted to suit the genetic algorithm in terms of rows and columns (See table 6.4).

Table 6.4, Contextual data for genetic algorithm

student_id	course_code	semester	difficulty
Stud1	ENGL 050	1	0.199833
Stud1	HIST 121	1	0.59593
Stud1	ITCS 101	1	0.634503
Stud1	MAGT 121	1	0.555763
Stud1	MATH 052	1	0.198760
Stud1	ACCT 301	2	0.642895
Stud1	ENGL 102	2	0.638115
Stud1	MAKT 201	2	0.571839
Stud1	MATH 104	2	0.571082
Stud1	STAT 101	2	0.55403
Stud1	ACCT 321	3	0.701872
Stud1	ACCT 403	3	0.712938
Stud1	CULT 102	3	0.764843
Stud1	ENGL 201	3	0.572327
Stud1	FINC 320	3	0.638547
Stud1	ITMA 201	3	0.600862
Stud1	ACCT 341	4	0.905313
Stud1	ARAB 102	4	0.828498
Stud1	FINC 431	4	0.614638
Stud1	ACCT 101	5	0.592099
Stud1	ARAB 101	5	0.69595
Stud1	ECON 101	5	0.57861
Stud1	ENGL 101	5	0.637785
Stud1	MATH 103	5	0.588106
Stud1	ACCT 311	6	0.631473
Stud1	BANK 220	6	0.578573
Stud1	ENGL 201	6	0.572327
Stud1	FINC 210	6	0.587196
Stud1	STAT 202	6	0.603671
Stud1	ACCT 402	7	0.701123
Stud1	BANK 302	7	0.557765
Stud1	CULT 101	7	0.748719
Stud1	ENGL 202	7	0.61227
Stud1	FINC 321	7	0.612641
Stud1	FINC 421	7	0.640831
Stud1	ACCT 499	8	0.908492
Stud1	INTR 400	8	0.052632
Stud1	ACCT 201	9	0.590262
Stud1	ECON 102	9	0.568344
Stud1	ITCS 121	9	0.653765
Stud1	ACCT 312	10	0.718198
Stud1	ACCT 320	10	0.647537
Stud1	FINC 310	10	0.615152
Stud1	BFRM 498	11	0.702107
Stud1	ECON 301	11	0.632647
Stud1	ECON 421	11	0.617569

6.4.6 Additional Data Preparation:

The data prepared in step General Data Preparation (section 6.4.4) and Contextual Data Preparation (section 6.4.5) was merged and the resulting dataset is shown in Table 5.5. This step was newly conceived to introduce the contextual factor dataset and merged with general data set to generate a unified dataset that is contextualised using the quality parameters that is essential for determining the contextual nature of the factor. Adding this step to the CRISP-DM process is new and modifies the character of the process nature of the original CRISP-DM process in line with the process theory mentioned in section 5.5. The complexity involved in introducing this step using an excel tool and writing a macro to generate the unified dataset. In addition another step to verify whether the dataset is contextualised with the introduction of course difficulty data was to be verified for which a pseudo code not tested with respect to genetic algorithm was written and a programme developed in C# and visual studio which ensured that the dataset was indeed contextualised using course difficulty as the contextual factor. This is a novel approach to ensure that the results obtained through the chosen KDDM process is able to enable the users to achieve the business goals without error and accurately. This contextualised dataset was fed into the modelling stage.

Table 6.5, Merged contextual dataset for Genetic algorithm

student_id	Len	gpa	semester	sem_gradepoints	course_code	difficulty
Stud1	4	2.17	1	1.89	ENGL 050	0.199833
Stud1	4	2.17	1	1.89	HIST 121	0.59593
Stud1	4	2.17	1	1.89	ITCS 101	0.634503
Stud1	4	2.17	1	1.89	MAGT 121	0.555763
Stud1	4	2.17	1	1.89	MATH 052	0.198760
Stud1	4	2.17	2	2.198	ACCT 301	0.642895
Stud1	4	2.17	2	2.198	ENGL 102	0.638115
Stud1	4	2.17	2	2.198	MAKT 201	0.571839
Stud1	4	2.17	2	2.198	MATH 104	0.571082
Stud1	4	2.17	2	2.198	STAT 101	0.55403
Stud1	4	2.17	3	1.723333	ACCT 321	0.701872
Stud1	4	2.17	3	1.723333	ACCT 403	0.712938
Stud1	4	2.17	3	1.723333	CULT 102	0.764843
Stud1	4	2.17	3	1.723333	ENGL 201	0.572327
Stud1	4	2.17	3	1.723333	FINC 320	0.638547
Stud1	4	2.17	3	1.723333	ITMA 201	0.600862
Stud1	4	2.17	4	3.556667	ACCT 341	0.905313
Stud1	4	2.17	4	3.556667	ARAB 102	0.828498
Stud1	4	2.17	4	3.556667	FINC 431	0.614638
Stud1	4	2.17	5	2.064	ACCT 101	0.592099
Stud1	4	2.17	5	2.064	ARAB 101	0.69595
Stud1	4	2.17	5	2.064	ECON 101	0.57861
Stud1	4	2.17	5	2.064	ENGL 101	0.637785
Stud1	4	2.17	5	2.064	MATH 103	0.588106
Stud1	4	2.17	6	1.332	ACCT 311	0.631473
Stud1	4	2.17	6	1.332	BANK 220	0.578573
Stud1	4	2.17	6	1.332	ENGL 201	0.572327
Stud1	4	2.17	6	1.332	FINC 210	0.587196
Stud1	4	2.17	6	1.332	STAT 202	0.603671
Stud1	4	2.17	7	2.22	ACCT 402	0.701123
Stud1	4	2.17	7	2.22	BANK 302	0.557765
Stud1	4	2.17	7	2.22	CULT 101	0.748719
Stud1	4	2.17	7	2.22	ENGL 202	0.61227
Stud1	4	2.17	7	2.22	FINC 321	0.612641
Stud1	4	2.17	7	2.22	FINC 421	0.640831
Stud1	4	2.17	8	3.33	ACCT 499	0.908492
Stud1	4	2.17	8	3.33	INTR 400	0.052632
Stud1	4	2.17	9	2.333333	ACCT 201	0.590262
Stud1	4	2.17	9	2.333333	ECON 102	0.568344
Stud1	4	2.17	9	2.333333	ITCS 121	0.653765
Stud1	4	2.17	10	2.553333	ACCT 312	0.718198
Stud1	4	2.17	10	2.553333	ACCT 320	0.647537
Stud1	4	2.17	10	2.553333	FINC 310	0.615152
Stud1	4	2.17	11	1.666667	BFRM 498	0.702107
Stud1	4	2.17	11	1.666667	ECON 301	0.632647
Stud1	4	2.17	11	1.666667	ECON 421	0.617569

6.4.6.1 Finding the presence or absence of course difficulty using a specially designed algorithm

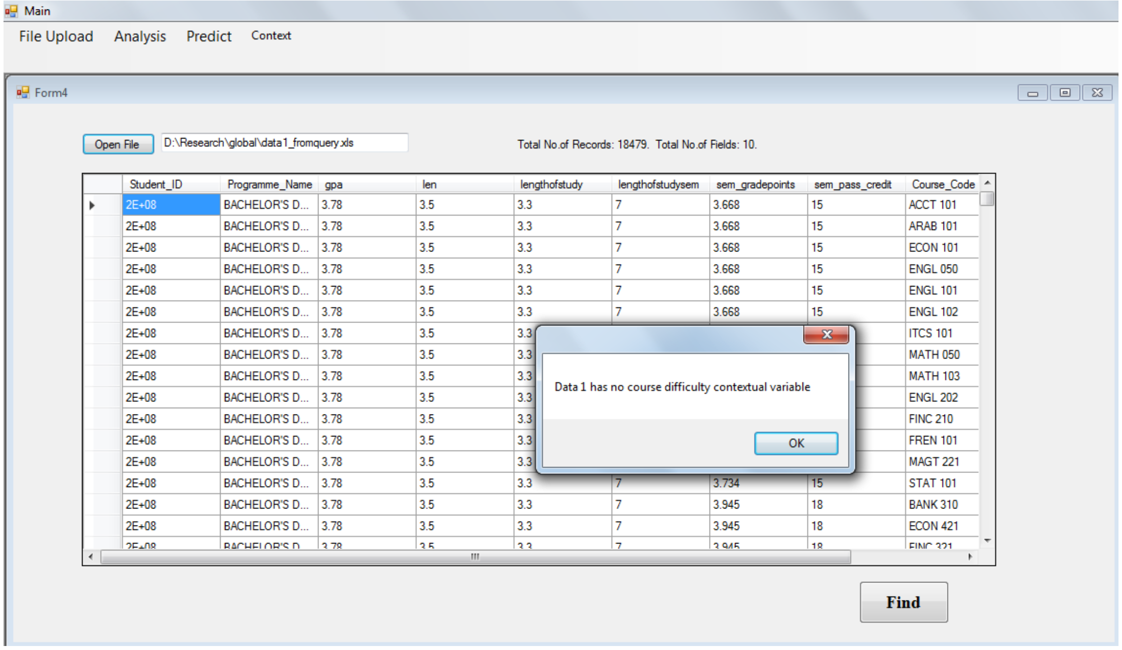
Based on the formula given by Vialardi et al. (2011) to compute the course difficulty, a pseudo code was written in line with the guidelines provided by Vert et al. 2010. The pseudo code is given below

Pseudo code to find context in datasets

```
INPUT Datasets D
For (j=1 to j<=totdatasets;j++)
Begin
INPUT Dataset Dj
INPUT Outputvar or Predictorvar (OF) of Dj
Int ct = 0;
// for each field F in the Dataset Dj
For (i=1;i<=totfields;i++)
// for each record r in the Dataset Di
For (r=1 ;r<=totrecs;r++)
Begin
//check if every field F has dimensions of context in all the records
//check when field F is not OF (output field)
If Fir <> OFir then
// findDOC – find dimensions of context
//findICF – find information criticality factors
//findQual – find data quality
If findDOC(Fir) = true and findICF(Fir) = true and findQual(Fir) = true then
ct = ct + 1;
else
continue;
else
continue;
End
End
```

```
if ct >=1 then
//dataset has context
//display ct as rank of dataset based on number of context variables
Return(Dj has context with rank ct)
Else
//dataset has no context variables
Return(Dj has no context)
End
```

The program developed based on the pseudocode using C# and visual studio was executed to check for context in dataset and screenshots are provided below in figures 6.1 and 6.2 provide useful information while Figure 6.1 is a screenshot of the output which indicates that the dataset fed as input does not contain contextual factor using a specially designed algorithm . This test confirms that the program developed to test presence or absence of contextual factor is functioning.



D:\Research\global\data1_fromquery.xls

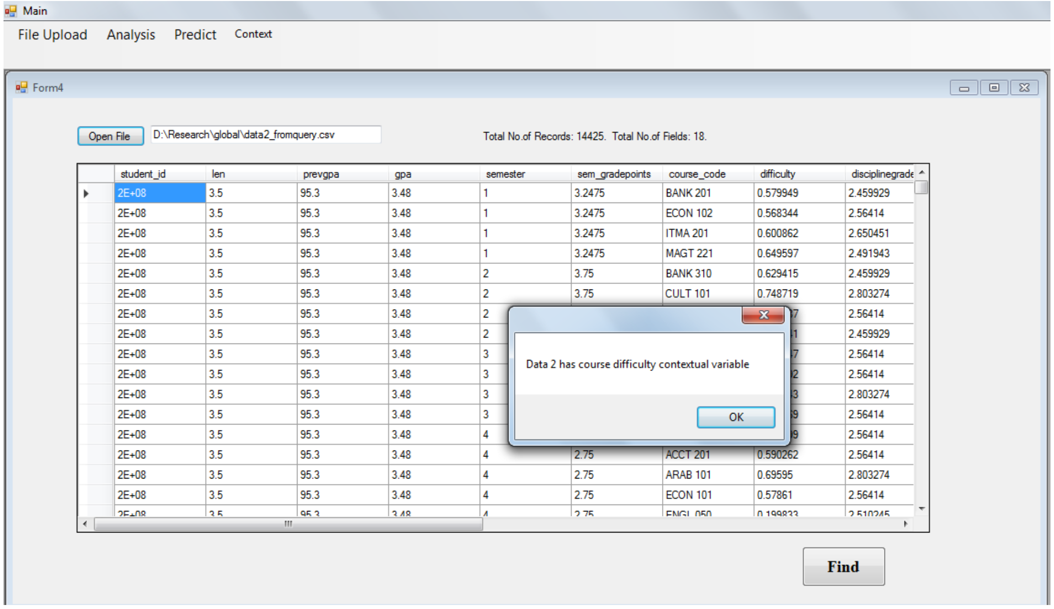
Total No.of Records: 18479. Total No.of Fields: 10.

Student_ID	Programme_Name	gpa	len	lengthofstudy	lengthofstudysem	sem_gradepoints	sem_pass_credit	Course_Code
2E+08	BACHELOR'S D...	3.78	3.5	3.3	7	3.668	15	ACCT 101
2E+08	BACHELOR'S D...	3.78	3.5	3.3	7	3.668	15	ARAB 101
2E+08	BACHELOR'S D...	3.78	3.5	3.3	7	3.668	15	ECON 101
2E+08	BACHELOR'S D...	3.78	3.5	3.3	7	3.668	15	ENGL 050
2E+08	BACHELOR'S D...	3.78	3.5	3.3	7	3.668	15	ENGL 101
2E+08	BACHELOR'S D...	3.78	3.5	3.3	7	3.668	15	ENGL 102
2E+08	BACHELOR'S D...	3.78	3.5	3.3				ITCS 101
2E+08	BACHELOR'S D...	3.78	3.5	3.3				MATH 050
2E+08	BACHELOR'S D...	3.78	3.5	3.3				MATH 103
2E+08	BACHELOR'S D...	3.78	3.5	3.3				ENGL 202
2E+08	BACHELOR'S D...	3.78	3.5	3.3				FINC 210
2E+08	BACHELOR'S D...	3.78	3.5	3.3				FREN 101
2E+08	BACHELOR'S D...	3.78	3.5	3.3				MAGT 221
2E+08	BACHELOR'S D...	3.78	3.5	3.3	7	3.734	15	STAT 101
2E+08	BACHELOR'S D...	3.78	3.5	3.3	7	3.945	18	BANK 310
2E+08	BACHELOR'S D...	3.78	3.5	3.3	7	3.945	18	ECON 421
2E+08	BACHELOR'S D...	3.78	3.5	3.3	7	3.945	18	FINC 321

Data 1 has no course difficulty contextual variable

OK

Figure 6.1, checking presence of contextual factor in Data1



D:\Research\global\data2_fromquery.csv

Total No.of Records: 14425. Total No.of Fields: 18.

student_id	len	prevgpa	gpa	semester	sem_gradepoints	course_code	difficulty	disciplinegrade
2E+08	3.5	95.3	3.48	1	3.2475	BANK 201	0.579949	2.459929
2E+08	3.5	95.3	3.48	1	3.2475	ECON 102	0.568344	2.56414
2E+08	3.5	95.3	3.48	1	3.2475	ITMA 201	0.600862	2.650451
2E+08	3.5	95.3	3.48	1	3.2475	MAGT 221	0.649597	2.491943
2E+08	3.5	95.3	3.48	2	3.75	BANK 310	0.629415	2.459929
2E+08	3.5	95.3	3.48	2	3.75	CULT 101	0.748719	2.803274
2E+08	3.5	95.3	3.48	2			7	2.56414
2E+08	3.5	95.3	3.48	2			1	2.459929
2E+08	3.5	95.3	3.48	3			7	2.56414
2E+08	3.5	95.3	3.48	3			2	2.56414
2E+08	3.5	95.3	3.48	3			3	2.803274
2E+08	3.5	95.3	3.48	3			9	2.56414
2E+08	3.5	95.3	3.48	4			9	2.56414
2E+08	3.5	95.3	3.48	4	2.75	ACCT 201	0.590262	2.56414
2E+08	3.5	95.3	3.48	4	2.75	ARAB 101	0.69595	2.803274
2E+08	3.5	95.3	3.48	4	2.75	ECON 101	0.57861	2.56414
2E+08	3.5	95.3	3.48	4	2.75	ENGL 050	0.199833	2.510245

Data 2 has course difficulty contextual variable

Figure 6.2, Checking presence of contextual factor in Data2

6.4.7 Modelling:

(A) Select Modelling technique: Genetic algorithm was used to generate the course taking pattern, course difficulty and predict time-to-degree and CGPA. As far as course difficulty was concerned the data in the merged dataset was coded using a nominal scale. This coding was more useful in interpretation than numeric values.

(B) Generate Test Design:

- **Criteria to determine the goodness of the model:** similar to 5.3.4
- **Definition of data on which the above criteria will be tested:** The full dataset characterised by contextual factor namely course difficulty was used to test the above criteria.

(C) Build Model:

- **Parameter Setting: refer 5.3.4**

Based on the discussions given in Section 5.3.4 the parameters were set and tabulated as shown in Table 6.6.

Table 6.6: Parameter setting for Genetic to process contextual dataset

Parameter Description	**Value**
Population size	150
Maximum number of generations	75
Chromosome length	75
Crossover rate	0.7
Mutation rate	0.7

(D) Generate Model:

Before producing the final model, a series of experiments were conducted by adjusting the cross-over, mutation, population size and chromosome length values in order to develop the GA to generate a course taking pattern and course difficulty pattern. The development of GA involved modification of the coding of the original GA which was specific to generation of course taking pattern as found in Table 6.7. Modification of coding was a complex process involved in the development of the GA for the specific purposes of generation of course taking pattern and course difficulty pattern. The final model has columns student identification (student_id), time-to-degree (len), Cumulative CGPA (GPA), semester GPA (SGPA), difficulty and course taking patterns (course).

Table 6.7, Course taking pattern and course difficulty pattern for contextual data

Student ID	Timetodegree	CGPA	SGPA	difficulty	course	Pattern
stud1	4.5	3.33	2.932	Difficult ,Difficult ,Difficult ,Average ,Difficult	ACCT 301 ,FINC 310 ,FINC 321 ,FREN 101 ,STAT 202	Pattern1
stud2	4.5	3.06	2.4	Difficult ,Difficult ,Difficult ,Average ,Difficult	ACCT 301 ,FINC 310 ,FINC 321 ,FREN 101 ,STAT 202	Pattern1
stud3	5	2.44	2.666	Average ,Difficult ,Average ,Difficult ,Difficult	ACCT 312 ,BANK 220 ,CULT 101 ,ENGL 202 ,ITMA 201	Pattern2
stud4	6	2.28	2.4	Average ,Difficult ,Average ,Difficult ,Difficult	ACCT 312 ,BANK 220 ,CULT 101 ,ENGL 202 ,ITMA 201	Pattern2
stud5	4.5	3.1	3.334	Difficult ,Average ,Difficult ,Difficult ,Difficult	ACCT 311 ,ACCT 321 ,ENGL 202 ,FINC 321 ,STAT 202	Pattern3
stud6	4.5	3.49	3.468	Difficult ,Average ,Difficult ,Difficult ,Difficult	ACCT 311 ,ACCT 321 ,ENGL 202 ,FINC 321 ,STAT 202	Pattern3
stud7	4.5	3.75	3.6	Difficult,Difficult,Easy,Difficult,Difficult	ACCT 301 ,ACCT 311 ,ARAB 102 ,FINC 421 ,ITMA 201	Pattern4
stud8	4.5	3.41	2.8	Difficult,Difficult,Easy,Difficult,Difficult	ACCT 301 ,ACCT 311 ,ARAB 102 ,FINC 421 ,ITMA 201	Pattern4
stud9	4	3.4	3.668333	Difficult ,Difficult ,Average ,Easy	ACCT 321 ,ARAB 201 ,BANK 220 ,FINC 431 ,ITCS 121 ,PHOT 101	Pattern5
stud10	4	3.46	3.723333	Difficult ,Difficult ,Average ,Easy	ACCT 321 ,ARAB 201 ,BANK 220 ,FINC 431 ,ITCS 121 ,PHOT 101	Pattern5
stud11	4.5	2.55	2.168333	Difficult,Average,Difficult,Easy,Difficult,Easy	ACCT 321 ,ACCT 402 ,BANK 302 ,ENGL 201 ,FINC 421 ,PHOT 101	Pattern6
stud12	4.5	2.33	2.333333	Difficult,Average,Difficult,Easy,Difficult,Easy	ACCT 321 ,ACCT 402 ,BANK 302 ,ENGL 201 ,FINC 421 ,PHOT 101	Pattern6
stud13	5	2.5	1	Average,Difficult,Easy,Difficult,Difficult	ACCT 403 ,BANK 302 ,ENGL 201 ,FINC 320 ,FINC 421	Pattern7
stud14	5.5	2.36	0.8	Average,Difficult,Easy,Difficult,Difficult	ACCT 403 ,BANK 302 ,ENGL 201 ,FINC 320 ,FINC 421	Pattern7
stud15	4	3.57	3.732	Difficult ,Average ,Difficult ,Difficult ,Average	ACCT 301 ,CULT 102 ,ENGL 202 ,FINC 320 ,ITCS 121	Pattern8
stud16	4	3.84	3.666	Difficult ,Average ,Difficult ,Difficult ,Average	ACCT 301 ,CULT 102 ,ENGL 202 ,FINC 320 ,ITCS 121	Pattern8
stud17	4	3.6	3.934	Average ,Average ,Difficult ,Easy ,Easy	ACCT 402 ,ACCT 403 ,BANK 302 ,ENGL 201 ,VDEO 101	Pattern9
stud18	4	3.22	3.202	Average ,Average ,Difficult ,Easy,Easy	ACCT 402 ,ACCT 403 ,BANK 302 ,ENGL 201 ,VDEO 101	Pattern9
stud19	4	3.42	3.398	Average ,Average ,Easy ,Difficult ,Difficult	ACCT 312 ,ACCT 320 ,ACCT 341 ,BANK 302 ,FINC 310	Pattern10
stud20	4	3.23	3.2	Average ,Average ,Easy ,Difficult ,Difficult	ACCT 312 ,ACCT 320 ,ACCT 341 ,BANK 302 ,FINC 310	Pattern10
stud21	3	3.88	3.778333	Difficult,Difficult,Difficult,Difficult,Difficult,Difficult	ACCT 404 ,BANK 302 ,ECON 301 ,FINC 320 ,FINC 421 ,STAT 202	Pattern11
stud22	3	3.53	3.556667	Difficult,Difficult,Difficult,Difficult,Difficult,Difficult	ACCT 404 ,BANK 302 ,ECON 301 ,FINC 320 ,FINC 421 ,STAT 202	Pattern11

6.4.8 Evaluation:

Table 6.8, Confusion matrix of GA for contextual data

	Predicted condition		
	Total population	**Prediction positive**	**Prediction negative**
True condition	Condition positive = 1210	1110 (TP)	100 (FN)
	Condition negative = 82	65 (FP)	17 (TN)

Table 6.9, Detailed Accuracy

True Positive Rate	**False Positive Rate**	**Precision**	**Recall**	**F-Measure**	**Prediction of time-to-degree**
0.917	0.050	0.944	0.917	0.931	Accurate

The first step in the evaluation involved the testing of the performance of the algorithm and the process in terms of data mining goals set in section 6.4.1. The measures examined: F-measure was measured as 0.944 see table 6.9 the closer this figure is to 1 that much accurate is the performance of GA. The precision was measured as 0.944 see table 6.9 the closer this figure is to 1 that much accurate is the performance of GA. The recall was measured as 0.917 see table 6.9 the closer this figure is to 1 that much accurate is the performance of GA.

Evaluation involved examining the course taking pattern by assigning the course difficulty level for each registered course and linking the pattern of courses to time-to-degree and assessing the CGPA scored as shown in the Table 6.7 for semester 3. This table includes sample of students of the programme Bachelor's degree in Accounting and Finance and who have been admitted and registered in the programme in the same year. The course difficulty level for each course was measured using a five point scale (very difficult, difficult, average, easy and very easy). The assumption here is that higher the number of registered courses of a student then lower should be the time-to-degree and lower will be the CGPA if the course difficulty level measure is the highest ('very difficult'). For instance, stud 6 in Table 6.7 is shown to have registered in 5 courses. Out of these, the course difficulty measure is identified as 'Difficult' for four courses (ACCT311, ENGL202, FINC321, STAT202) with the remaining 1 course (ACCT321) identified as 'Average'. His CGPA is found to be 3.49. It is reasonable to expect that this student would take longer time to graduate because he was able to complete the total number of courses

(44) in 4.5 years at an average of 5 courses a semester and a CGPA around the same figure of 3.49. The course difficulty level has perhaps not allowed the student to register in more courses in a semester, say 6. If ‘stud6’ had registered in six courses in a semester on average, then the student could have graduated in four years. Thus course taking pattern and the corresponding course difficulty measure of a course can be assumed to affect the time-to-degree. Alternatively if ‘stud6’ had registered in 6 courses whose course difficulty measure were to be a combination of ‘Easy’,’ Average’ and ‘Difficult’ (say a course taking pattern of 6 courses namely ACCT 402, ACCT 320, ENGL 201, ARAB 102, BANK 302, STAT 202) with course difficulty measured as ‘Average’ ‘Average’, ‘Easy’, ‘Easy’, ‘Difficult’, and ‘Difficult’ then there is a possibility that ‘stud6’ could have scored either the same CGPA of 3.49 or higher and taken less time-to-degree (4 years). The student in this case is unlikely to score a CGPA less than 3.49. The reason is that the course difficulty measures of the imaginary set of courses are lower than those courses the student has actually registered in.

Thus it can be seen that course taking pattern when associated with the course difficulty measurement makes it possible to predict to some extent the optimum time-to-degree and an optimum CGPA. That is, it is possible to expect that when students register in courses whose difficulty level is defined as “difficult” then if the number of such courses in the set of registered courses in a semester is high then there is a possibility that the student could score lower CGPA but is not necessary. Similarly when the number of courses in which a student registers is 6 then the time-to-degree is expected to be lower than usual (usual indicates 4 years). However registering in 6 courses could increase the difficulty level for some students. Hence it is possible to decide on the set of courses that could make the difficulty level to be comfortable for the students so that the students are able to score a good GPA and graduate in a shorter time-to-degree. However where students find the difficulty level as high and still are advised to register in 6 courses, then it becomes necessary on the part of the University to provide additional support to those students so that they score optimum GPA depending on the difficulty level calculated for the student and graduate in shorter time. The above arguments can be demonstrated as follows taking the help of the mined data provided in Table 6.7.

If one inspects the rest of the students in Table 6.7 then it can be seen that the remaining students can be classified under time-to-degree 4 (8 students), 4.5 (8 students), 5 (2 students), 5.5 (1 student) and 6 (1 student). While inspecting the records of students who have taken a time-to-degree of 5, 5.5 and 6 it is possible to assess their performance easily as the number of such students is less. But the students under the time-to-degree category 4 and 4.5 attract attention. It can be seen that majority of these students have scored CGPA higher than 3 but their time-to-degree is higher. This could be due to the following reasons.

The pattern of courses and the course difficulty pattern measures could be a reason. For instance students Stud1 and Stud2 have registered in courses whose patterns are (ACCT 301, FINC 310, FINC 321, FREN 101, STAT 202) and the course difficulty level patterns are (Difficult, Difficult, Difficult, Average, Difficult) respectively. Both students have achieved a time-to-degree of 4.5 years. Their CGPA stood at 3.33 and 3.04 respectively. If the students had taken a course taking pattern (ACCT 301, FINC 310, FINC 321, FREN 101, PHOT 101, ITCS 121) with course difficulty pattern equal to Difficult, Difficult, Difficult, Average, Easy, Average) then they would have registered in 6 courses in that semester. Because of this their chances to reduce the time-to-degree and scoring higher CGPA will be better. This can happen because the number of courses whose difficulty level is 'difficult' has come down to 3 from 4 (ACCT 301, FINC 310, FINC 321) and the course difficulty levels of other courses (FREN 101, PHOT 101, ITCS 121) were measured as (Average, Easy, Easy). If similar logic is applied to registration of courses of Stud1 and Stud2 in every semester, then those students might have graduated in shorter time-to-degree and scored better CGPA. That is to say there emerges a pattern of courses and a pattern of course difficulty levels that would accurately inform the students to determine the set of courses and the semester in which they need to be registered in. This knowledge is unique for each student and could be mined and tested. This knowledge could help in better advising of the students to enable them achieve optimum time-to-degree. Some examples are provided in this regard next.

For instance, when the entire student dataset of 1292 graduated students of the anonymous university where the research was conducted was taken, the data related to Bachelor's Degree in Accounting and Finance students was extracted and mined to discover knowledge regarding course taking pattern and time-to-degree. The data pertaining to a total of 706 students was mined. Out of those students only 9 students had achieved the shortest time-to-degree of 3 years (Table 6.10).

Table 6.10, 9 students with shortest time-to-degree

Student	Time-to-degree	CGPA	Very Difficult	Difficult	Average	Easy	Very Easy
Stud21	3	3.88	10	15	14	5	0
Stud22	3	3.53	15	15	9	5	0
Stud25	3	2.78	15	16	8	5	0
Stud24	3	2.54	15	17	7	5	0
Stud29	3	2.5	15	17	7	5	0
Stud26	3	2.21	16	12	11	5	0
Stud23	3	3.1	17	13	9	5	0
Stud28	3	3	17	13	7	7	0
Stud27	3	2	18	13	10	3	0

From Table 6.10, it can be seen that 9 students achieved the shortest time-to-degree of 3 years. All of them have graduated by completing 44 courses. Their CGPA is not the same. Their overall course difficulty level pattern is not the same except for two students (Stud24 and Stud29). One thing that broadly emerges is that the students have achieved optimum time-to-degree but not optimum CGPA. This could be due to the pattern of courses those students have registered in and the pattern of course difficulty levels. For instance Stud21 has an overall course difficulty pattern of 10 'very difficult', 15 'difficult', 14 'average' and 5 'easy' level courses. This appears to be the optimum course difficulty pattern that a student has achieved to graduate in the shortest time-to-degree. The Table—informs that any student who has registered in more number of courses whose course difficulty level is measured as 'very difficult' than Stud21, has scored less than optimum CGPA. For instance Stud22 has registered in 15 courses whose course difficulty levels were measured as very difficult. The CGPA of this student is 3.53 which is lower than that of Stud21 whose CGPA is 3.88. So if Stud22 had balanced the course taking pattern with a different course difficulty pattern somewhat similar to Stud21 then Stud22 might have scored higher CGPA which would have approached that of Stud21 or even gone higher. That is to say if Stud22 had registered in less number of courses measured with course difficulty level of very difficult (that is less than 15) then there would have been a chance for Stud22 to have scored higher CGPA. That is to say lower the numbers of courses measured as having the course difficulty level of 'very difficult' higher the chances of scoring high CGPA. Similar explanation could be given to the other students found in that Table 6.7 with regard to determining the optimum CGPA.

However some more new knowledge could be uncovered from Table 6.7. For instance apart from identifying the number of courses that could be categorised as 'very difficult' and using that information to find the optimum CGPA, the combination of course difficulty levels namely 'very difficult', 'difficult' and ‘average’ could also provide some information. For instance when Table 6.7 was reorganised in a way as shown in the Table 6.11 below, with CGPA presented in the descending order, then it can be seen that some students (e.g. Stud24 and Stud29) have scored lower CGPA (2.54 and 2.5) respectively although those students have taken less number of courses that could be categorised under the course difficulty level ‘very difficult’ (15 courses).

Table 6.11, Students with different difficulty levels

Student	Time-to-degree	CGPA	Very Difficult	Difficult	Average	Easy	Very Easy
Stud21	3	3.88	10	15	14	5	0
Stud22	3	3.53	15	15	9	5	0
Stud23	3	3.1	17	13	9	5	0
Stud28	3	3	17	13	7	7	0
Stud25	3	2.78	15	16	8	5	0
Stud24	3	2.54	15	17	7	5	0
Stud29	3	2.5	15	17	7	5	0
Stud26	3	2.21	16	12	11	5	0
Stud27	3	2	18	13	10	3	0

The reason for this could be that the combination of courses taken by these students had higher number of courses whose difficulty level were measured as ‘difficult’ (17 courses) and ‘average’ (7) alongside 15 courses categorised under ‘very difficult’. This pattern might have played a role. This is evident when one compares the course difficulty pattern of the courses taken by Stud28 who has scored a CGPA of 3 and has registered in 17 courses measured as ‘very difficult’ but only 13 courses measured as ‘difficult’. In comparison to Stud28, Stud24 and Stud29 have scored a lower CGPA of 2.54 and 2.5 respectively although they have registered in lesser number of courses measured as ‘very difficult’ (15) but have registered in 15 other courses measured as ‘difficult’. The number of courses measured as ‘difficult’ is more in the case of Stud24 and Stud29 whereas it is less in the case of Stud28. The reason why Stud24 and Stud29 who have registered in 15 courses measured as ‘very difficult’ and have scored lower CGPA than Stud28 who has registered in 17 courses measured as ‘very difficult’ could be the role of courses measured as ‘difficult’ which is higher in the case of Stud24 and Stud29 than Stud28. Thus not only the course difficulty level measure ‘very difficult’ appears to play a role in the CGPA scored by the students but also the measures namely ‘difficult’ and ‘average’. Based on these arguments it is possible to say that course taking patterns that have a combination of courses with course difficulty patterns ‘very difficult’, ‘difficult’ and ‘average’ are seen to determine the CGPA. That is to say course taking patterns when explained with course difficulty level patterns provide accurate prediction of optimum time-to-degree and CGPA. This conclusion can be arrived at based on the information given in Tables 6.7 and 6.11.

Although the above conclusion is based on the case of only 9 students, it must be understood that those nine students have achieved the optimum time-to-degree of 3 years amongst 706 students and in one case the highest CGPA of 3.88. Evaluation of any other mined report involving students who have taken longer time-to-degree cannot be considered as optimum. The above conclusions can be extended to other students who have taken longer time-to-degree and scored lesser CGPA. It is possible to evaluate each student's course taking pattern and determine the course difficulty pattern and find out through data mining how those students could be encouraged to achieve shorter time-to-degree and optimum CGPA using the outcome of the evaluation of course taking pattern together with the course difficulty pattern. It is interesting to note that the 9 students evaluated have taken the same number of 6 courses in each semester which indicates that when students register in the maximum number of courses (6) per semester in which they are allowed to register then the number of courses per semester vanishes as a factor that could determine the time-to-degree or CGPA. However in case of students who have registered in less than 6 courses this factor can be a determinant of time-to-degree and CGPA as lower number of courses per semester obviously points to longer time-to-degree.

However this conclusion was arrived at based on the course taking pattern generated by the modified CRISP-DM process that has mined 44 courses studied over the entire programme duration of each student. In order to have more accurate prediction of optimum time-to-degree and CGPA another evaluation was conducted at the semester level using Table 6.12. Semester wise course taking patterns of Stud21 and Stud22 were compared using the mining report generated by modified CRISP-DM process.

Table 6.12, Semester wise course taking patterns of Stud21 and Stud22

Sem	Student	(Course code, Course difficulty)						SGPA	CGPA
Semester1	Stud21	ACCT 101, Difficult	ARAB101, Average	ECON101, Average	ENGL101, Difficult	MATH103, Easy	STAT 101, Easy	3.93	3.93
	Stud22	ACCT 101, Difficult	ARAB101, Average	ECON101, Average	ENGL101, Difficult	MAKT201, Very difficult	STAT 101, Easy	3.53	3.53
Semester 2	Stud21	BANK220, Difficult	FINC 210, Very Difficult	ACCT201, Difficult				3.91	3.91
	Stud22	BANK220, Difficult	FINC 210, Very Difficult	ACCT201, Difficult				3.49	3.5
Semester 3	Stud21	ACCT 404, Difficult	BANK 302, Difficult	ECON 301, Difficult	FINC 320, Difficult	FINC 421, Difficult	STAT 202, Difficult	3.778	3.88
	Stud22	ACCT 404, Difficult	BANK 302, Difficult	ECON 301, Difficult	FINC 320, Difficult	FINC 421, Difficult	STAT 202, Difficult	3.556	3.55
Semester 4	Stud21	ENGL102, Average	ITCS 101, Easy	ITMA201, Very Difficult				3.78	3.88
	Stud22	MAKT101, Very difficult	ITCS 101, Easy	ITMA201, Very Difficult				3.4	3.5
Semester 5	Stud21	ACCT 312, Very Difficult	ACCT320, Difficult	BANK320, Average	ACCT 321, Average	ACCT 341, Very Difficult	FINC321, Average	3.78	3.8
	Stud22	ENGL102, Average	ACCT 312, Very Difficult	ACCT 341, Very Difficult	FINC321, Average	ACCT 321, Average	BANK320, Average	3.53	3.52
Semester 6	Stud21	ACCT 402, Very Difficult	ACCT 403, Very difficult	FINC 320, average	ECON 421, Average	ACCT 401, Average	FINC431, difficult	3.83	3.86
	Stud22	ACCT 402, Very Difficult	ACCT 403, Very difficult	FINC 320, Average	ECON 421, Average	ACCT 401, Average	FINC431, difficult	3.4	3.5
Semester 7	Stud21	ENGL201, Very difficult	MATH 104, Difficult	HIST121, Easy				3.78	3.87
	Stud22	ENGL201, Very difficult	ITMA301, Very Difficult	FINC 328, Very Difficult				3.06	3.38
Semester 8	Stud21	ETHC 391, Average	BFRM 498, Average	ARAB102, Very difficult	CULT101, Average	ENGL 202, Very difficult	ECON102, Difficult	3.78	3.87
	Stud22	ETHC 391, Average	BFRM 498, Average	ARAB102, Very difficult	CULT101, Average	ENGL 202, Very difficult	ECON102, Difficult	3.61	3.44
Semester 9	Stud21	CULT102, Difficult	PHOT 101, Average	VDEO 101, Easy				4	3.95
	Stud22	CULT102, Difficult	MAGT 121, Very Difficult	ITCS122, Very difficult				3.06	3.38
Semester 10	Stud21		ACCT499, Very difficult					3.95	3.88
	Stud22		FINC 499, Very difficult					3.45	3.53

Table 6.13 shows that in the first semester when the course taking pattern of Stud21 and Stud22 are compared the following results are derived.

Table 6.13, First Semester course taking patterns of Stud21 and Stud22

Sem	Student	(Course code, Course difficulty)						SGPA	CGPA
Semester1	Stud21	ACCT 101, Difficult	ARAB101, Average	ECON101, Average	ENGL101, Difficult	MATH103, Easy	STAT 101, Easy	3.93	3.93
	Stud22	ACCT 101, Difficult	ARAB101, Average	ECON101, Average	ENGL101, Difficult	MAKT201, Very difficult	STAT 101, Easy	3.53	3.53

The main difference is in the fourth course. While Stud21 has opted for MATH103 whose course difficulty level is measured as ‘easy’, Stud22 has opted for MAKT201 whose course difficulty level was measured as ‘very difficult’. The result was Stud22 scored a lower semester GPA of 3.53 in comparison to Stud21 who scored a semester GPA of 3.93. Similar results can be seen with regard to semesters 7 and 9 (Table 5.14).

Table 6.14, Semester 7 and 9 course taking patterns of Stud21 and Stud22

Sem	Student	(Course code, Course difficulty)						SGPA	CGPA
Semester 7	Stud21	ENGL201, Very difficult	MATH 104, Difficult	HIST121, Easy				3.78	3.87
	Stud22	ENGL201, Very difficult	ITMA301, Very Difficult	FINC 328, Very Difficult				3.06	3.38
Semester 9	Stud21	CULT102, Difficult	PHOT 101, Average	VDEO 101, Easy				4	3.95
	Stud22	CULT102, Difficult	MAGT 121, Very Difficult	ITCS122, Very difficult				3.06	3.38

These examples show that there is clear evidence that the course taking pattern of students can predict optimum time-to-degree and CGPA with 3 years being the optimum time-to-degree and 3.88 being the optimum CGPA. In order to make wider generalisations across the students who have taken longer time-to-degree the same arguments can be extended. For instance if one compares Stud21 with Stud1 it is possible to argue that Stud1 can be encouraged to register in 6 courses that have a combination of courses whose course difficulty level measures are a mixture of ‘difficult’, ‘average’ and ‘easy’. How many courses should fall under each one of the course difficulty level measure category is left to the choice of the student or the adviser. The adviser could fall back on the outcome of the CRISP-DM process to decide on the combination of courses that the student could register in each semester. Next special coaching can be given to students like Stud1 and help them cope with the load of 6 courses to achieve optimum time-to-degree and CGPA.

From the above discussions it can be seen that

(Optimum CGPA, Optimum Time-to-degree) = function of (course taking pattern, course difficulty, semester number) → (6.1)

The foregoing results and discussions clearly point out that the optimum time-to-degree and the highest CGPA can be determined by a set of 6 courses in a pattern (maximum allowed by a university in a semester) in every semester with varying course difficulty level measures. When this result is considered as optimum, then the following decisions could be facilitated in the HEIs.

A. Accurately determine the optimum time-to-degree and CGPA of each student using course taking pattern and course difficulty level pattern.

B. Grouping students like ‘stud21’ and ‘stud22’ to enable them to graduate in 3 years.

C. Grouping students like ‘stud6’ who have registered in only 5 courses and analyse their performance in terms of their ability to score the same CGPA or higher to enable them to achieve a performance similar to ‘stud21’ and ‘stud22’.

D. Grouping students like ‘stud17’ and ‘stud18’who have taken 4 courses and analyse their performance in terms of their ability to score the same CGPA or higher to enable them to achieve a performance similar to ‘stud21’ and ‘stud22’.

E. Assess the performance of students who have registered in 4 or 5 courses in a semester at an early stage in their academic career in the university using modified CRISP-DM to predict their optimum time-to-degree with higher CGPA based on the performance of past students. Use this knowledge improve the performance of those students to achieve optimum time-to-degree and CGPA.

F. The term optimum time-to-degree and CGPA needs careful understanding. While the discussions above point out that the optimum time-to-degree is three years, the same cannot be applied uniformly across all students. For instance the optimum time-to-degree for some students could be 4 years as they may not be able to register in 6 courses in a semester or register in summer sessions due to certain limitations (e.g. working students). Similarly some other students may have difficulty in paying fees and hence may not be able to graduate in 3 years and register in 6 courses per semester. In such cases also optimum time-to-degree could be worked out using the modified CRISP-DM process to determine the courses those students could register in each semester as a pattern.

G. In addition a large number of students do not register in more than four or five courses and do not register in summer sessions due to reasons such as their ability to cope with academic load. In such cases it is possible to identify the courses using the course difficulty level pattern and group students according to their ability. Then those students could be given special coaching leading to better performance in terms of time-to-degree and CGPA using knowledge of the course taking pattern and course difficulty level pattern.

6.4.8.1 Review of the modified CRISP-DM process

The business goal has been partially achieved although the main assumption that course taking pattern and course difficulty could determine time-to-degree and CGPA has been established. In addition all the steps used in testing a CRISP-DM model have been verified and found to be achieved. Thus the modified CRISP-DM process has been proven to be functioning as per the specified parameter after the changes introduced in the process.

6.4.8.2 Determination of next steps

The results obtained with regard to predicting time-to-degree and CGPA using course taking pattern and course difficulty level pattern clearly indicate the modified CRISP-DM process has the potential to be deployed in the HEIs to improve student learning experience and support decision making. Thus the next steps involved prior to the deployment include:

- prediction of optimum time-to-degree and CGPA as a function of course taking pattern and course difficulty level pattern of students using the CRISP-DM process and hypothetical situations.

- Enhancement of time-to-degree of some students who have taken longer time-to-degree when compared to a student who has been identified as having scored the optimum CGPA (3.88) and time-to-degree (3 years) by simulation of course taking pattern based on a reference model generated using CRISP-DM process and CGPA.

6.4.8.3 Prediction of optimum time-to-degree and CGPA as a function of course taking pattern and course difficulty level pattern of students

In Section 5.3.6 it was already established that the optimum time-to-degree is 3 years. The question is whether this time-to-degree and CGPA could be predicted. This was tested by taking the case of Stud22. The reference optimum time-to-degree and CGPA were chosen based on the best student who achieved the shortest time-to-degree and highest CGPA. This was achieved by Stud21 who graduated in 3 years (the shortest by far) and scored the highest CGPA of 3.88. The course taking pattern of Stud 21 thus became the reference for all students to either emulate or at the least approach it.

Thus the course taking pattern of Stud22 was assessed and compared with that of Stud21 semester by semester. Table 6.15 provides the comparison.

Table 6.15, Comparison of Stud21 and Stud22 course taking pattern and course difficulty pattern for all semesters

Semester	Student	Course code						SGPA	CGPA
Semester1	Stud21	ACCT 101, Difficult	ARAB101, Average	ECON101, Average	ENGL101, Difficult	**MATH103, Easy**	STAT 101, Easy	3.93	3.93
	Stud22	ACCT 101, Difficult	ARAB101, Average	ECON101, Average	ENGL101, Difficult	**MAKT201, Very difficult**	STAT 101,Easy	3.53	3.53
Semester 2	Stud21	BANK220, Difficult	FINC 210, Very Difficult	ACCT201, Difficult				3.91	3.91
	Stud22	BANK220, Difficult	FINC 210, Very Difficult	ACCT201, Difficult				3.49	3.5
Semester 3	Stud21	ACCT 404, Difficult	BANK 302, Difficult	ECON 301, Difficult	FINC 320, Difficult	FINC 421, Difficult	STAT 202, Difficult	3.778	3.88
	Stud22	ACCT 404, Difficult	BANK 302, Difficult	ECON 301, Difficult	FINC 320, Difficult	FINC 421, Difficult	STAT 202, Difficult	3.556	3.55
Semester 4	Stud21	**ENGL102, Average**	ITCS 101, Easy	ITMA201, Very Difficult				3.78	3.88
	Stud22	**MAKT101, Very difficult**	ITCS 101, Easy	ITMA201, Very Difficult				3.4	3.5
Semester 5	Stud21	ACCT 312, Very Difficult	ACCT320, Difficult	BANK320, Average	ACCT 321, Average	ACCT 341, Very Difficult	FINC321, Average	3.78	3.8
	Stud22	ENGL102, Average	ACCT 312, Very Difficult	ACCT 341, Very Difficult	FINC321, Average	ACCT 321, Average	BANK320, Average	3.53	3.52
Semester 6	Stud21	ACCT 402, Very Difficult	ACCT 403, Very difficult	FINC 320, average	ECON 421, Average	ACCT 401, Average	FINC431, difficult	3.83	3.86
	Stud22	ACCT 402, Very Difficult	ACCT 403, Very difficult	FINC 320, average	ECON 421, Average	ACCT 401, Average	FINC431, difficult	3.4	3.5
Semester 7	Stud21	ENGL201, Very difficult	**MATH 104, Difficult**	**HIST121, Easy**				3.78	3.87
	Stud22	ENGL201, Very difficult	**ITMA301 Very Difficult**	**FINC 328 Very Difficult**				3.06	3.38
Semester 8	Stud21	ETHC 391, Average	BFRM 498, Average	ARAB102. Very difficult	CULT101, Average	ENGL 202, Very difficult	ECON102, Difficult	3.78	3.87
	Stud22	ETHC 391, Average	BFRM 498, Average	ARAB102, Very difficult	CULT101, Average	ENGL 202, Very difficult	ECON102, Difficult	3.61	3.44
Semester 9	Stud21	CULT102, Difficult	**PHOT 101, Average**	**VDEO 101, Easy**				4	3.95
	Stud22	CULT102, Difficult	**MAGT 121, Very Difficult**	**ITCS122, Very difficult**				3.06	3.38
Semester 10	Stud21		**ACCT499, Very difficult**					3.95	3.88
	Stud22		**FINC 499, Very difficult**					3.45	3.53

While comparing it can be seen that Stud21 course taking pattern was different from that of Stud22. So the course taking pattern of Stud21 where hypothetically adjusted as shown in Table 6.16. This table was created based on the assumption that some courses in which Stud22 had registered, if changed to a pattern similar to that of Stud21 then Stud22 might approach a CGPA score achieved by Stud21 (CGPA of 3.88).

Table 6.16, Hypothetical adjustment of courses for Stud22 to reach higher CGPA

Student	Course code						SGPA	CGPA	Remarks
Stud22	ACCT 101, Difficult	ARAB101, Average	ECON101, Average	ENGL101, Difficult	MAKT201, Very difficult	STAT 101, Easy	3.53	3.53	
Stud22-simulated	ACCT 101, Difficult	ARAB101, Average	ECON101, Average	ENGL101, Difficult	**MATH103, Easy**	STAT 101, Easy	**3.8**	**3.8**	**Assuming student gets A-**
Stud22	BANK220, Difficult	FINC 210, Very Difficult	ACCT201, Difficult				3.49	3.5	
Stud22-simulated	BANK220, Difficult	FINC 210, Very Difficult	ACCT201, Difficult				3.49	**3.65**	
Stud22	ACCT 404, Difficult	BANK 302, Difficult	ECON 301, Difficult	FINC 320, Difficult	FINC 421, Difficult	STAT 202, Difficult	3.556	3.55	
Stud22-simulated	ACCT 404, Difficult	BANK 302, Difficult	ECON 301, Difficult	FINC 320, Difficult	FINC 421, Difficult	STAT 202, Difficult	3.556	**3.61**	
Stud22	MAKT101, Very difficult	ITCS 101, Easy	ITMA201, Very Difficult				3.4	3.5	
Stud22-Simulated	**ENGL102, Average**	ITCS 101, Easy	ITMA201, Very Difficult				**3.9**	**3.769**	**Assuming student gets A-**
Stud22	ENGL102, Average	ACCT 312, Very Difficult	ACCT 341, Very Difficult	FINC321, Average	ACCT 321, Average	BANK320, Average	3.53	3.52	
Stud22-simulated	ENGL102, Average	ACCT 312, Very Difficult	ACCT 341, Very Difficult	FINC321, Average	ACCT 321, Average	BANK 320, Average	3.53	**3.655**	
Stud22	ACCT 402, Very Difficult	ACCT 403, Very difficult	FINC 320, average	ECON 421, Average	ACCT 401, Average	FINC431, Difficult	3.4	3.5	
Stud22-simulated	ACCT 402, Very Difficult	ACCT 403, Very difficult	FINC 320, average	ECON 421, Average	ACCT 401, Average	FINC431, difficult	3.4	3.6	
Stud22	ENGL201, Very difficult	ITMA301, Very Difficult	FINC 328, Very Difficult				3.06	3.38	
Stud22-simulated	ENGL201, Very difficult	**MATH 104, Difficult**	**HIST121, Easy**				**3.9**	**3.75**	**Assuming student gets A-**
Stud22	ETHC 391, Average	BFRM 498, Average	ARAB102, Very difficult	CULT101, Average	ENGL 202, Very difficult	ECON102, Difficult	3.61	3.44	
Stud22-simulated	ETHC 391, Average	BFRM 498, Average	ARAB102, Very difficult	CULT101, Average	ENGL 202, Very difficult	ECON102, Difficult	3.61	**3.68**	
Stud22	CULT102, Difficult	MAGT 121, Very Difficult	ITCS122, Very difficult				3.06	3.38	
Stud22-simulated	CULT102, Difficult	**PHOT 101, Average**	**VDEO 101, Easy**				**3.9**	**3.79**	**Assuming student gets A-**
Stud22		FINC 499, Very difficult					3.95	3.88	
Stud22-simulated		**ACCT499, Very difficult**					**3.9**	**3.84**	

To begin with seven courses of Stud22 were identified that were measured as ‘very difficult’. They were MAKT201, MAKT101, ITMA301, FINC 328, MAGT 121, ITCS122 and FINC 499 (Table 6.15). The corresponding courses in which Stud21 had registered in the same semesters as that of Stud22 where identified. They were MATH103 (course difficulty level: ‘Easy’), ENGL102 (course difficulty level: ‘Average’), MATH 104 (course difficulty level: ‘Difficult’), HIST121 (course difficulty level: ‘Easy’), PHOT 101 (course difficulty level: ‘Average’), VDEO 101 (course difficulty level: ‘Easy’) and ACCT499 (course difficulty level: ‘Very difficult’) (Table 6.16). In the CRISP-DM process the courses taken by Stud22 were modified with these seven courses. That is the courses MAKT201, MAKT101, ITMA301, FINC 328, MAGT 121, ITCS122 and FINC 499 (Table 6.15) were replaced with MATH103, ENGL102, MATH 104, HIST121, PHOT 101, VDEO 101 and ACCT499 (Table 6.16). Then in order to calculate the CGPA, there was a need to assume the grades Stud22 might score in MATH103, ENGL102, MATH 104, HIST121, PHOT 101, VDEO 101 and ACCT499 (Table 6.16) individually. From the history of the student it was found that the Stud22 had scored over 3.84 (equivalent to the grade ‘A’ which is the highest grade that is being awarded in the anonymous university in which the research was conducted; any numeric grade that falls in the range 3.84 to 4.00 will be awarded letter grade ‘A’) semester GPA in more than half the number of courses in the programme. Stud22 is seen to have scored already B+ (3.53) (B+ is a letter grade equivalent to a score of GPA that falls in the range 3.33-3.67) (Table 6.15). Now if any alteration is made to the course taking pattern of Stud22 then it must result in a higher CGPA greater than 3.53. The next higher grade could be thought of as the result. The higher grade should fall between 3.67 and 3.84 (equivalent to the letter grade of ‘(A-)’). So Stud22 was assumed to score a CGPA that will earn a letter grade of ‘(A-)’. Thus in the dataset the grades for all the seven new courses were changed to 3.67. Then the CGPA and time-to-degree were computed. The overall CGPA rose to 3.84 (which approach the letter grade ‘A’). That is to say if Stud22 had registered in MATH103, ENGL102, MATH 104, HIST121, PHOT 101, VDEO 101 and ACCT499 (Table 6.16) instead of MAKT201, MAKT101, ITMA301, FINC 328, MAGT 121, ITCS122 and FINC 499 (Table 6.15) in specific semesters concerned and maintained a performance that would have fetched at the least ‘(A-)’ grade in the courses MATH103, ENGL102, MATH 104, HIST121, PHOT 101, VDEO 101 and ACCT499, then it is seen that the CGPA dramatically changes. Incidentally the time-to-degree still remains the most optimum (3 years).

The following aspects must be borne in mind:

- This result was not reached by arbitrarily changing the seven courses mentioned above. But the courses were changed one by one. That is to say MAKT201 was replaced first with MATH103. The CGPA was checked by running the CRISP-CM process without changing the other six courses. The result was CGPA increased marginally. Next MAKT101 was replaced with ENGL102. Now two courses have been changed. The CRISP-CM process was run without changing the other five courses. The CGPA improved further but not substantially. This iterative process was run until the seventh course was reached when the CGPA was seen as 3.84.

- It must be noted that the assumption Stud22 will maintain a performance that will enable the student to reach this CGPA. If not the CGPA may not improve. The assumption that the student will do better is based on the comparison of the course difficulty level of the two sets of courses, one in which already the student had registered and the other hypothetically introduced. That is the seven original courses were all measured as 'very difficult' whereas the new seven courses were measured as ('Easy', 'Average', 'Difficult', 'Easy', 'Average', 'Easy' and 'Very difficult'). The comparison shows that in the original course taken pattern the seven courses were measured as 'very difficult' whereas in the hypothetical set there is only one course measured as 'very difficult' with three courses measured as 'easy' and two courses measured as 'average' increasing the probability that the student will score at the least a grade point average of 3.67 in each. So there is a high probability that this may happen.

- There will be question whether this could lead to a perfect course taking pattern that could lead Stud22 a maximum CGPA of 4.0. Although this is ideally possible but in reality this may not happen as the performance of the students is unpredictable. This brings into focus the ability of the student to score well which is another contextual factor. This contextual factor may determine the actual performance of the student and hence whether the perfect score of 4.0 could be achieved by any student based on the knowledge gained using the modified CRISP-DM process.

The above arguments clearly demonstrate there is a good possibility that the mined data can provide information to advisers and students to predict the CGPA and time-to-degree from the knowledge discovered through the modified CRISP-DM process in terms of course taking pattern and course difficulty pattern. A similar exercise was conducted to predict the time-to-degree which is given next.

6.4.8.4 Enhancement of time-to-degree of students who have taken longer time-to-degree as a function of course taking pattern and course difficulty level pattern of students

Similar to the case of Stud22 another student Stud7 (see Table 6.7) who graduated with a time-to-degree of 4.5 was chosen for predicting an improved time-to-degree taking the course taking pattern, CGPA and time-to-degree of Stud21 as reference. The analysis of the course taking pattern of Stud7 revealed that the student registered in different semesters as follows (Table 6.17):

Table 6.17, Analysis of course taking pattern of Stud7

	Semester number											
	1	2	3	4	5	6	7	8	9	10	11	12
Number of courses	4	5	2	5	5	5	5	2	6	4	1	1

From Table 6.17 it can be seen that Stud7 has not been consistent in regards to the number of courses registered per semester. This inconsistency is the main reason for the student to have taken longer time-to-degree.

Table 6.18, Hypothetical adjustment of courses for Stud7 to reach optimum time-to-degree

Semester	Student	Course code, Course difficulty						SGPA	CGPA	Credits completed	Time-to-degree (Years)
Semester 1	Stud7		ARAB101, Average		ENGL101, Difficult	MATH103, Easy	ITCS101, Easy	3.83	3.83	12	
	Stud7-simulated	**ACCT 101, Difficult**	ARAB101, Average	**ECON101, Average**	ENGL101, Difficult	MATH103, Easy	ITCS101, Easy	3.83	3.83	**18**	
Semester 2	Stud7	ACCT 101, Difficult	ECON101, Average	ENGL102, Average	HIST121, Easy	MATH 104, Difficult		3.87	3.85	24	
	Stud7-simulated	ACCT 201, Difficult	ECON102, Difficult	ENGL102, Average	HIST121, Easy	MATH 104, Difficult	BANK220, Difficult	3.87	3.85	36	
Semester 3	Stud 7	ACCT 201, Difficult	ECON102, Difficult	This row indicates the summer session in the first year of study of the student.				4	3.88	30	
	Stud7-simulated	STAT202, Difficult	CULT101, Easy	MAGT121, Very difficult	This row indicates the summer session in the first year of study of the student.			4	3.88	45	
Semester 4	Stud 7	BANK220,Difficult	FINC210, Very Difficult	ITCS 121, Difficult	MAGT 121, Very Difficult	VDEO 101, Easy		3.6	3.79	45	
	Stud7-simulated	FINC210, Very difficult	ARAB 201, Average	BANK 302, Difficult	ENGL 201, Difficult	FINC 310, Difficult	STAT 101, Easy	3.6	3.79	63	
Semester 5	Stud 7	ARAB 201, Average	BANK 302, Difficult	ENGL 201, Difficult	FINC 310, Difficult	STAT 101, Easy		3.4	3.7	60	
	Stud 7-simulate	ACCT 301, Difficult	ACCT 311, Difficult	ENGL 202, Very difficult	FINC 421, Difficult	ITMA201, Very Difficult	ACCT 312, Very Difficult	3.4	3.7	81	
Semester 6	Stud 7	ACCT 301, Difficult	ACCT 311, Difficult	ARAB102. Very difficult	FINC 421, Difficult	ITMA201, Very Difficult		3.6	3.68	75	
		This row indicates semester 1 of year 3 in the original course taking pattern of the student									
	Stud7-simulated	ACCT 312, Very Difficult	ACCT 320, Difficult	FINC431, Difficult	This row indicates the introduction of the summer session in the second year of study of the student. This student did not opt for summer session in 2nd year.					90	
Semester 7	Stud 7	ACCT 312, Very Difficult	ACCT 320, Difficult	FINC431, Difficult	MAKT201, Very Difficult	STAT 202, Difficult		3.87	3.71	90	
	Stud7-simulated	ACCT320, Difficult	BANK302, Difficult	ENGL215, Very Difficult	FINC321, Average	ITCS 121, Difficult	MAKT201, Very Difficult	3.87	3.89	108	
Semester 8	Stud 7	ACCT 401, Difficult	ECON 301, Very Difficult					3.5	3.7	96	
	Stud 7	**ECON 301, Very Difficult**	**ACCT 321, Average**	**ACCT402, Very Difficult**	**ACCT 403, Very difficult**	**ECON 421, Average**	**FINC320, Average**		3.89	126	
Semester 9	Stud 7	ACCT 321, Average	ACCT402, Very Difficult	ACCT 403, Very difficult	ECON 421, Average	ENGL 202, Very difficult	FINC320, average	3.72	3.7	114	
		This row indicates semester 2 of year 4 in the original course taking pattern of the student									
	Stud 7-simulate	ACCT499, Difficult	INTR 400, Difficult	This row indicates the introduction of the summer session in the third year of study of the student. This student did not opt for summer session in 3rd year. Student graduates at this point.				3.67	3.85	**132**	**3 years**
Semester 10	Stud 7	BANK302, Difficult	CULT102,Difficult	FINC321, Average	INTR400, Difficult			3.67	3.75	126	
	Stud 7-simulate	x	x	x	x			3.83	3.84		
Semester 11	Stud 7	ACCT499, Difficult						3.45	3.53		
	Stud 7-simulate	x									
Semester 12	Stud 7	ACCT499, Difficult						4	3.75	132	4.5 years
	Stud 7-simulate	x									

Although there could be many reasons for this, had the student been advised to register in at the least 5 courses in each semester and 2 courses in summer, this student would have graduated in less than 4 years. Taking this argument as the basis and hypothetically if one alters the course taken pattern of Stud7 in line with that of Stud21, then it is possible to know through the knowledge discovered using the modified CRISP-DM process whether the student could have graduated in less than 4.5 years. Thus taking into account the course difficulty level of the courses in which Stud7 had registered, the number of courses per semester and the pattern of courses already there, a new course taking pattern was developed using a computer programme which iterated semester by semester and compared the data of Stud7 with that of Stud21. The assumption was that the CGPA scored by Stud7 will remain constant and no change would occur due to any change in the pattern of courses. The second assumption was that the semester GPA would be adjusted for each course introduced as part of the new pattern of courses by the computer programme to keep the CGPA constant. The third assumption was that the student performance in terms of the CGPA will remain the same as before even after the introduction of new pattern of courses. Table 6.18 provides the comparison of Stud7's existing course taking pattern and the proposed course taking pattern (in black background and white fonts) and the time-to-degree. This was achieved through a process of iteration which included

- Changing courses measured as 'very difficult' semester by semester
- Introducing additional courses in each semester to make the number of courses as 6
- Checking the pattern of courses thus changed with that of Stud21 which is the reference
- Adjusting the course taking pattern in each semester in such a way that in no semester Stud7 has to register in more than two courses measured as difficult
- Introduction of summer session and
- Calculating the semester GPA of the newly introduced courses in every semester to ensure that the CGPA is maintained constant.

Then it was seen that a total of 36 changes have been introduced in the original course taking pattern of Stud7 in addition to the introduction of two additional summer sessions (see Table 6.18). In fact Stud7 registered in only one summer session. With such a major change in the course taking pattern, the student data was mined and it was discovered that the student could graduate in 3 years maintaining the same CGPA. It must be noted that increasing the number of courses to six in a semester and introducing additional summer sessions while may point towards the natural reduction in the time-to-degree it is not possible to predict this automatically. The reason for this is that the pattern of courses and their difficulty levels are shown to have an effect on the performance of the students and many students are not automatically encouraged to register in 6 courses in a semester. Many students have the concern that 6 courses in a semester and summer sessions may tax them affecting their performance and CGPA. In such a situation if a course taking pattern is determined that has a combination of courses whose course difficulty levels are distributed then more students could be encouraged to register in 6 courses a semester and summer sessions also. The Table 6.18 then will then be transformed as (see Table 6.19):

Table 6.19, Hypothetical registration of courses of stud7 to optimise time-to-degree

	Semester number											
	1	2	3	4	5	6	7	8	9	10	11	12
Number of courses: Originally registered	4	5	2	5	5	5	5	2	6	4	1	1
Rearranged registration of courses	6	6	3	6	6	3	6	6	2	x	x	x

While the experiment was conducted using a hypothetical situation which took the reference of Stud21 as the one indicating optimum time-to-degree, the results show that if a student is provided the required academic advise and support, then it is possible to achieve the most optimum time-to-degree for any student. It is significant to note that although the course taking pattern of Stud7 was adjusted taking the course taking pattern of Stud21, in not even one semester does the course taking pattern of Stud7 is exactly similar to that of Stud21. Thus it is possible to argue that the course taking pattern of Stud21 is not the only pattern of courses that could be taken by students, but there can be other combinations. In addition although Stud7 was not shown to improve the CGPA due to the change in course taking patterns, this assumption is not likely to be true in real situation. It could be higher or lower than 3.75 as the performance of Stud7 cannot be considered to be a constant and could vary due to a number of reasons. Moreover, the assumption that the student could register in six courses could be void if the student is not capable of handling more courses in a semester in which case the above arguments may need to be set aside. Furthermore, registering summer sessions or additional courses in a semester requires the student to pay fees and if the student is not able to pay the required fees then the above arguments may not be useful.

The above discussions while showing the potential of the modified CRISP-DM process in predicting time-to-degree and CGPA using course taking pattern of students and course difficulty level through the method of simulating hypothetical situation provide a basis for deploying the modified CRISP-DM process in HEIs. Thus as explained in section 3.2.7 the deployment of modified CRISP-DM process would be planned, deployed, monitored and maintained to achieve the business goal.

6.4.9 Findings:

The equations 3.9 and 3.10 were only partially achieved. That is to say, the equations 3.9 and 3.10 need to be re-specified. From equation 6.1 it can be seen that the following relationships depicted 3.9 and 3.10 are not valid:

CGPA = function of (course taking pattern, course difficulty,
time-to-degree, semester number) → **(3.9)**
and
Time-to-degree = function of (course taking pattern, course difficulty,
CGPA, semester number) → (3.10)

Equation 3.9 and 3.10 need to be rewritten as

CGPA = function of (course taking pattern, course difficulty, Semester number) → (6.2)
and
Time-to-degree = function of (course taking pattern, course difficulty,
semester number) → (6.3)

Equations 6.2 and 6.3 could be combined and rewritten as in equation 6.1

(Time-to-degree, CGPA) = function of (course taking pattern, course difficulty, semester number) → (6.1)

From equation 6.1 it is possible infer that the time-to-degree and CGPA can be predicted using course taking pattern of students, the course difficulty measure and the semester number. The following relationships were not found to be valid.

CGPA = function of (time-to-degree)
Time-to-degree = function of (CGPA)

6.4.9.1 If the course taking pattern of students are uncovered and characterised by contextual factors that is course difficulty using a KDD process like modified CRISP-DM process then it is possible to determine the optimum time-to-degree and CGPA of students.

6.4.9.2 Time-to-degree and CGPA are independent of each other and are predicted by course taking pattern, course difficulty and semester number.

6.4.9.3 Modified CRISP-DM process has been found to function as per the specified parameters and has been tested to show that a way has been found to integrate contextual factors into the KDDM process. Thus fills gap in the KDD literature.

6.4.9.4 A new algorithm has been developed to mine both course taking pattern as well as course difficulty pattern which introduce innovation in the original CRISP-DM process leading to the development of a contextualised data mining stage in the KDDM process.

6.4.9.5 A new stage linking business understanding stage to contextual data understanding and preparation has been introduced to extract knowledge about course difficulty. In addition a merging section has been introduced to merge general data and contextual data to be fed into the modelling stage of the CRISP-DM process.

6.4.9.6 The CRISP-DM process produced models in the form course taking patterns and course difficulty patterns which could be used for decision making.

6.4.9.7 Based on the evaluation many hidden things related to prediction of time-to-degree and CGPA were discovered. A variety of combinations of course taking pattern and course difficulty pattern were examined to explore how different types of students could be benefited by knowing accurately the courses they have to register in, semester by semester and accordingly creating their own pattern of courses in which they want to register.

6.4.9.8 EDM has been successfully integrated as demonstrated the generation of patterns related to course taken by students and relative course difficulty.

6.4.9.9 The hypotheses HA and HB were verified and found to be partially achieved.

6.4.9.10A simulation of the hypothetical situation was carried out to predict the CGPA and time-to-degree which was used to demonstrate how the modified CRISP-DM process could be deployed in HEIs. Since actual deployment of the process is virtually impossible simulation was carried out. The examples of the hypothetical situation are highly probable to happen in real life. Thus this research is said to have achieved the modified CRISP-DM process in accordance with the documentation of the literature.

6.4.10 Limitations:

1. The contextual factor tested in this research is course difficulty. However there could be other contextual factors, for instance student potential that may have a role in the prediction of time-to-degree and CGPA which were not discovered.
2. It is necessary to modify the GA if a new contextual factor needs to be introduced.
3. If contextual factors are not properly understood by the data miner then the patterns generated may not have proper meaning with regard to the contexts and hence decision making may suffer.

6.5 Summary:

This chapter has demonstrated the performance and testing of the modified CRISP-DM process in terms of integrating EDM into the CRISP-DM process for predicting time-to-degree and CGPA using course taking patterns. The performance was found to satisfy the minimum required values to be achieved with regard to precision, recall and F-measure. In addition, the modified CRISP-DM process has successfully performed to generate the course taking pattern from the merged data characterised by course difficulty which was selected the contextual factor. In addition, actual examples were taken to discuss the ability of the process to discover knowledge to predict optimum CGPA and time-to-degree in terms of course taking pattern and course difficulty pattern. Thus this chapter provides the basis to discuss the findings in the following chapter.

Chapter 7 : Discussion and Conclusions

7.1 Introduction

This chapter discusses the outcome of the experiments conducted to test the KDDM process model used to address the research questions. In the chapters 4, 5 and 6 the CRISP-DM model and its modified version were tested and evaluated as per the chosen guideline (Chapman et al. 2000). In addition the methodology used to design the artefacts namely CRISP-DM process integrated with EDM and modified CRISP-DM process integrated with contextualized EDM was grounded in the design science. The outcomes from testing and evaluating the artefacts are discussed in this chapter alongside the conclusions. While the discussion section addresses the research questions set for this research, the conclusions provide the contributions of this research to knowledge, theory, method and practice.

7.2 Discussion

RQ1: Is there an unobservable relationship between course taking pattern, optimum CGPA and optimum time-to-degree hidden in the educational dataset of undergraduate students of a higher education institution? If there exists a relationship between the three factors then is there an unobservable contextual factor course difficulty level pattern hidden in the educational dataset of undergraduate students of a higher education institution that affect the relationship between course taking pattern, optimum CGPA and optimum time-to-degree? Using KDDM process is it possible to establish the relationship between course taking pattern, course difficulty level pattern, optimum CGPA and optimum time-to-degree by discovering the unobservable relationship mentioned above?

This research question was addressed as 3 parts namely

RQ1 part1 - Is there an unobservable relationship between course taking pattern, optimum CGPA and optimum time-to-degree hidden in the educational dataset of undergraduate students of a higher education institution?

RQ1 part2 - If there exists a relationship between the three factors then is there an unobservable contextual factor course difficulty level pattern hidden in the educational dataset of undergraduate students of a higher education institution that affect the relationship between course taking pattern, optimum CGPA and optimum time-to-degree?

RQ1 part3 - Using KDDM process is it possible to establish the relationship between course taking pattern, course difficulty level pattern, optimum CGPA and optimum time-to-degree by discovering the unobservable relationship mentioned above?

Each one of these parts is addressed next.

7.2.1 RQ1 part1

This research was conducted in anonymous university in Bahrain. As mentioned in Section 1.7 of Chapter 1 the university offers both undergraduate and graduate programs. American credit system is followed in the university. Students have to study 44 courses (also called subjects) each of 3 credits usually over four years at the rate of 11 courses each year on average. Students can register in a minimum of 4 courses or a maximum of 6 courses in the first and semesters of each year and a minimum of 1 course and a maximum of 3 courses during summer. Summer session is not mandatory. Students could withdraw from a semester and continue further. Thus a student could complete a minimum of 8 courses in a year without registering in the summer or a maximum of 15 including summer in a year. A student who is registering in a minimum of 8 courses can logically complete the 44 courses in 5.5 years without withdrawing or repeating any course. Similarly a student can also register in 15 courses each year and graduate in 4 years. The students were assessed using numerical grades and awarded equivalent letter grades (see Table 1.2). The student data resides in the educational dataset in a computer system called ADREG (see Appendix.11). As mentioned in Appendix the complete data of the students, right from the application stage to graduation is maintained in the ADREG system. Research question (RQ1) was answered by applying a simple computer technique in Microsoft Excel initially to randomly check whether any course taking pattern could be generated and related to CGPA and time-to-degree for a set of 10 anonymous students (See table 3.1).

From the discussions given in Section 3.2 of chapter 3 and Table 3.1 it can be seen that a basic function relating course taking pattern to CGPA and time-to-degree was developed. The two functions developed are re written here for convenience.

CGPA = function of (different courses, common courses, course taking pattern, time-to-degree) → 7.1

and

Time-to-degree = function of (different courses, common courses, course taking pattern, CGPA) → 7.2

From equations 3.5 and 3.6 it is clear that course taking pattern is related to CGPA and time-to-degree. It must be noted here that equations 3.5 and 3.6 provide the probable relationship between course taking pattern, CGPA and time-to-degree without indicating whether the CGPA and time-to-degree are optimum. At this stage to check whether CGPA and time-to-degree could be optimum or not to answer this KDDM process had to be employed and educational dataset was systematically processed to discover hidden knowledge (unobservable phenomenon) and course taking pattern was generated to know whether there exists a relationship between course taking pattern, optimum CGPA and optimum time-to-degree. In order to process the educational data to determine the optimum CGPA and optimum time-to-degree a data mining technique was applied (see Appendix 24). Course taking pattern could not be generated based on the discussions and rationale provided under sections 5.4 and 5.5 respectively CRISP-DM process was applied to generate course taking pattern and relate it to optimum CGPA and time-to-degree (see chapters 4 and 5). This finding is in line with some of the arguments found in the literature. For instance literature suggests the use of knowledge discovery and data mining (KDDM) for generating course taking pattern (Kovacic, 2010) which is supported by the findings. After carrying out experiments thoroughly the CRISP-DM process generated models to check whether course taking pattern is related to optimum time-to-degree and CGPA. It was found that two techniques namely clustering and association rules could provide only knowledge that indicated a possible indirect relationship between course taking pattern and optimum CGPA and time-to-degree (Sections 4.3 and 5.2). This finding is partially supported by the literature. For instance Zeidenberg (2011) used clustering techniques to understand the course taking pattern of community college students in order to determine the programs of study. The findings of Zeidenberg (2011) did not cover association rules and arguments suggesting the use or otherwise of Association rules as part of KDDM were not found in the literature. Findings in the literature could not exactly be compared to the findings of this study and studies that have produced similar results are hard to find in the literature. However some similarity could be found in the studies conducted by Kovacic (2010) and Bahr (2010) who have shown that registering in particular courses is linked to time-to-degree. Thus it can be concluded that this finding is a unique contribution to the relevant literature and can be considered as a novelty.At this stage while the research outcomes showed that there is a concept called course taking pattern and it is related to CGPA and time-to-degree in order to know whether there existed a concept called optimum CGPA and optimum time-to-degree further investigations were needed. The difficulty was that educational datasets do not provide knowledge on how to optimize the CGPA and time-to-degree as a function of course taking pattern. At this stage RQ1 – part 1 the question was only partially answered (See equations 3.5 and 3.6).

Further investigations into how optimum CGPA and time-to-degree could be derived from the dataset were carried out using RQ1-part2.

7.2.2 RQ1-part2

RQ1 part2 - If there exists a relationship between the three factors then is there an unobservable contextual factor course difficulty level pattern hidden in the educational dataset of undergraduate students of a higher education institution that affect the relationship between course taking pattern, optimum CGPA and optimum time-to-degree?

From the literature it is seen that events are driven by contexts many times and a concept called contextual meta data needs to be understood to gain previously unknown insight into the event and contextually derived knowledge (Vert et al. 2011). Context driven processing is directed by the environment and meanings explaining the event. Thus as explained in Section 3.2 one such contextual factor was thought to have relationship to course taking pattern which if discovered could provide an answer to generate optimum CGPA and time-to-degree. One contextual factor described in the literature that could have a relationship to course taking pattern was course difficulty (see Section 3.2). This led to the inquiry into contextual factors driving the processing of data leading to the next part of the question related to the relationship between the contextual factor namely course difficulty level pattern, course taking pattern, optimum time-to-degree and optimum CGPA. The course difficulty level pattern adding few new stages for discovering patterns of courses with different difficulty levels and link those levels to courses and course taking pattern. This was verified manually by calculating course difficulties for the courses taken by the 10 students in Table 3.1 using the equation to calculate the course difficulty level provided in section 2.3.3. Table 3.2 provides the details about the course difficulty level pattern and its relationship to course taking pattern, CGPA and time-to-degree. As can be seen from the Table

3.2 and the explanations given thereof equations 3.7 and 3.8 could be derived which clearly point out that CGPA and time-to-degree are related to course difficulty level pattern which are provided here for convenience.

CGPA = function of (number of courses, course taking pattern,
course difficulty) → 7.3

and

Time-to-degree = function of (number of courses, course taking pattern,
course difficulty) → 7.4

Further as explained in the same section 3.3 equations 3.5, 3.6, 3.7 and 3.8 were consolidated and 3.9 and 3.10 were derived and provided below for reference.

CGPA = function of (number of courses, course taking pattern, course difficulty,
time-to-degree) → 7.5

and

Time-to-degree = function of (number of courses, course taking pattern, course
difficulty, CGPA) →7.6

At this point also it can be seen that there is a relationship between course taking pattern, course difficulty level pattern, CGPA and time-to-degree. There is hardly any study that has tried to find a relationship between course taking pattern, contextual factor (course difficulty level pattern), CGPA and time-to-degree over all semesters of a programme in which a student registered. However there are studies that show that there is a relationship between certain course registration pattern in one semester and time-to-degree (Volkwein & Lorang, 1996; Adelman,2006; Cabrera et al.,2005) which is a very basic finding that does not imply a course pattern of the sort defined in this research. For instance in this research the course taking pattern of a student in a semester is defined (ACCT101, BANK101, ITCS101, ENGL101, MATH101). However according Adelman (2006) one course namely mathematics affects the time-to-degree of students when registered in different semesters. One course does not become a pattern. Thus this finding is new and contributes to the relevant body of knowledge. However here again it can be seen that even adding course difficulty level pattern to the dataset does not indicate whether it is possible to extract optimum CGPA and time-to-degree hidden in the dataset. Thus at this point again it can be seen that RQ1- part2 is only partially achieved. To discover the optimum CGPA and time-to-degree as a function of course taking pattern and course difficulty level pattern, further investigations were conducted by processing the educational dataset to extract hidden knowledge using KDDM process as explained in Section RQ1-Part1. As explained CRISP-DM process was employed to discover contextual factor (course difficulty level). Experiments using clustering, association rule and classification did not generate any pattern related to course difficulty level. This aspect was verified and addressed by answering RQ1- Part3.

7.2.3 RQ1-Part3

Using KDDM process is it possible to establish the relationship between course taking pattern, course difficulty level pattern, optimum CGPA and optimum time-to-degree by discovering the unobservable relationship mentioned above?

From the literature it can be seen that KDDM processes usually do not generate patterns associated with contextual information (Vert et al.2010, Vajrikar et al. 2003). Lack of contextual information in the patterns discovered could lead to underestimation of performance, for instance lack of knowledge about optimum CGPA and time-to-degree. This finding is supported in literature for instance the findings of Vajirkar et al. (2003) who argued that no KDDM process model including CRISP-DM process model generates patterns containing contextual information. This research tested the same and found that it cannot generate. Therefore there was necessity to develop a KDDM artefact that could discover patterns associated with contextual information. In this research a modified CRISP-DM process model was developed to address the issue of generating course difficulty level patterns along with course taking pattern thus contextualize the patterns discovered to predict optimum CGPA and time-to-degree. Design science methodology (Table 4.9) was used to develop, test and evaluate the model. Guidelines provided by Hevner et al. (2004) were followed. The development of the model has been described in Section 5.5. The developed model is depicted in Figure 5.5.

The modified CRISP-DM model was tested based on the guidelines provided by Chapman et al. (2000). The comparison between the unmodified and modified CRISP-DM process is given in Table 7.1.

Table 7.1, Comparison of Steps between unmodified and modified CRISP-DM

Model	Step1	Step 2	Step 3	Step4	Step5	Step6	Limitations
CRISP-DM	1.Understanding of business objectives and requirements, which are converted into a DM problem definition.	2 Identification of data quality problems, data exploration, and selection of interesting data subsets	3 Preparation of the final dataset, which will be fed into DM tool(s), and includes data and attribute selection, cleaning, construction of new attributes, and data transformations	-Calibration and application of DM methods to the prepared data	-Evaluation of the generated knowledge from the business Perspective	-Presentation of the discovered knowledge in a customer-oriented way. Performing deployment, monitoring, maintenance, and writing final report	There is lack of guideline or tool on how to approach this step and complete it without challenge (Step1). Lack of clarity on the list of all data quality checks to be performed and on the ways to address or overcome data quality issues (Step2). The prepared data might not be suitable for the DM method as the data is not formatted to suit it (Step3). Feedback loop is not discussed in detail in the documentation(Step4) Lack of documentation on guidelines of evaluation (Step5). Lack of feedback loop from deployment to business understanding (Step6). Lack of Contextual Processing.

Modified CRISP-DM	1.Understanding of business objectives and requirements, which are converted into a DM problem definition	2a. General Data Understanding- This step is similar to Data understanding of CRISP-DM process 2b. Contextual Data Understanding – this step is used to understand contextual factors.	3a. General Data Preparation- This step is same as CRISP-DM Step3. 3b. Contextual Data Preparation - In this step contextual factors were generated after multiple iterations 4. New Additional Data Preparation (Merge)- in this step the 2 datasets namely	- Calibration and application of DM methods to the prepared data	-Evaluation of the generated knowledge from the business Perspective	-Presentation of the discovered knowledge in a customer-oriented way. Performing deployment, monitoring, maintenance, and writing final report	There is lack of guideline or tool on how to approach this step and complete it without challenge (Step1). Lack of clarity on the list of all data quality checks to be performed and on the ways to address or overcome data quality issues (Step2). The prepared data might not be suitable for the DM method as the data is not formatted to suit it (Step3). Feedback loop is not discussed in detail in the documentation (Step4). Lack of documentation on guidelines of evaluation (Step5). Lack of feedback loop from deployment to business understanding (Step6).

From the comparison it can be seen that modified CRISP-DM model has been developed to overcome one of the limitations of the CRISP-DM model by incorporating the contextual data understanding, contextual data preparation, merging of general and contextual data, generating contextualized course taking pattern, feedback loops between the modelling stage and contextual data preparation stage and feedback loop between the evaluation stage and business understanding stage. Process theory was used to change the original CRISP-DM process. The resulting artefact was used to generate course taking pattern, course difficulty level pattern, optimum CGPA and optimum time-to-degree. The test results are given Section 6.4. It can be seen that a new relationship has been generated through the modified artefact as depicted in equation 6.1 reproduced below.

(Optimum CGPA, Optimum Time-to-degree) = function of (course taking pattern, course difficulty, semester number) → (6.1)

Thus it can be seen that the problem of addressing research questions RQ1 – Part 1 and RQ1 – Part 2 partially due to lack of generating optimum CGPA and time-to-degree has been fully overcome in equation 6.1. Thus while literature shows that there is hardly any KDDM process that could generate patterns associated with contextual information, the modified CRISP-DM process has overcome this limitation in the literature. From this it can be concluded that RQ1 – Part 3 is fully addressed. Thus this research has contributed to the growing body of knowledge related to KDDM by developing a new artefact namely modified CRISP-DM process that can mine and generate course taking pattern of students contextualized by course difficulty level pattern.

7.2.4 RQ2:

Is it possible to predict the optimum CGPA and optimum time-to-degree of undergraduate students in terms of the course taking pattern and course difficulty level pattern extracted from the educational dataset of a higher education institution using a KDDM process?

This question was addressed in Section 6.4. The discovered course taking patterns of a select set of students was linked to the course difficulty pattern of the same set of students and the students were arranged in the descending order with respect to the CGPA (see Table 6.11). This table showed the shortest time-to-degree (3 years) and highest CGPA (3.88) scored by a student with code Stud21. This student's performance in terms of CGPA and time-to-degree showed the optimum figures. 3 years was the lowest time-to-degree that could be achieved by any student. Incidentally 3.88 happened to be the highest CGPA scored by any student.

This record of Stud21 became the reference to compare and predict any other student's performance and suggest ways to improve by matching the course taking pattern of Stud21 with other students. The modified CRISP-DM process was run to check whether by simulation the performance of any student (e.g. Stud22) who has scored is lower CGPA than Stud21 and any other student (Stud7) who has taken longer time-to-degree than Stud21 could be theoretically improved. The experiment described in Section 5.4.8 showed that it is possible to enhance the performance of Stud22 and Stud7 by reorganizing their course taking pattern and course difficulty level pattern in line with that of Stud 21. The results showed that there is a clear possibility that the combination of course taking patterns and course difficulty level patterns can be used to predict the optimum CGPA and time-to-degree. This finding is unique as no similar contribution to knowledge by any other researcher could be found in the extant literature. But it must be stated that while the simulation shows positive results, the assumption here is that the students would perform the same way they were assumed to perform although such an assumption was made based on the past performance of those students. The developed artefact was used to generate course taking pattern, course difficulty level pattern, optimum CGPA and optimum time-to-degree. The findings of this research is supported by similar outcomes obtained by other researchers for instance Vajirkar et al. (2003) who have have developed a similar KDDM process model artefact including contextual factors.

From the above discussion it can be seen that equation 5.1 could be realized to predict student performance in terms of achieving optimum CGPA and optimum time-to-degree using course taking pattern and course difficulty pattern. Thus it can be said that RQ2 has been fully addressed.

7.2.5 Summary

From the above discussions it can be seen that the research questions set for this research have been addressed and the gap existing in the literature has been partially filled In addition the discussions show that the main assumptions depicted through equations 3.9 and 3.10 had to be re-specified and more accurate relationship has been derived in equation 5.1. From this equation it can be seen that optimum CGPA and time-to-degree could be predicted by course taking pattern and course difficulty level pattern.

7.3 Conclusion

The outcome of the research derived and presented in chapters 2, 3,4,5,6 clearly demonstrate that this research contributes to the body of KDDM process knowledge and HEI performance management field. In addition the outcomes have demonstrated the contributions to knowledge, theory, method and practice.

7.3.1 Contribution to knowledge

This research contributes primarily to the body of KDDM knowledge. In addition the outcome of this research contributes to the decision making process in HEIs to enhance student performance.

7.3.1.1 Contribution to KDDM knowledge

Foremost this research has developed a KDDM artefact by modifying the CRISP-DM process model that is useful to mine educational data and extract course taking pattern of students contextualized by course difficulty level pattern and predict optimum CGPA and time-to-degree. Literature shows that KDDM process models including CRISP-DM process model do not generate patterns characterized by contextual information (Vert et al. 2010). This research has overcome this difficulty. This contribution is a novelty. The nearest research outcome that could be compared with this contribution is the one produced by Vajirkar et al. (2003). While the work of Vajirkar et al. (2003) tested a new KDD process to find if context based pattern can be generated and found that it could be, however the results obtained by Vajirkar et al. (2003) could not be generalized due to limitations. Particularly when compared with the findings of this research which dealt with a large and complex dataset of the students comprising 44 courses per student, several semesters, CGPA for each course registered in by a student and the course difficulty pattern of each student's course taking pattern and tested using different datamining techniques, the findings of Vajirkar et al. (2003) are seen to suffer from lack of application of different datamining techniques and use of a very small dataset. Thus the findings of this research become more valuable when compared to that of Vajirkar et al. (2003) as it has the ability to deal with large datasets that have complex variables. This is a major contribution to knowledge. Particularly the research has been able to demonstrate how a very popular KDDM process like CRISP-DM could be modified to deal with complex contextual factors with unique patterns. In addition this research has developed new computer programmes and algorithm to perform different functions as given in Table 7.2.

Table 7.2, Computer programmes and algorithm to perform different functions

Technical work	Remarks

SQL Query	1. To extract data from the Student information system to create the input dataset for clustering, association rules and classification (genetic algorithm). 2. To extract data from the Student information system to create the input dataset for classification technique (Genetic algorithm) after contextual factor extraction steps were embedded in CRISP-DM (genetic algorithm).
Genetic algorithm	1.Genetic algorithm was modified to extract course taking patterns in CRISP-DM process 2.Genetic algorithm was modified to extract course taking patterns and course difficulty patterns in modified CRISP-DM process
Visual Studio and C#	1. Pseudocode proposed by Vert et al. 2010 for checking if the dataset have contextual factors (course difficulty). 2. Simulation of course taking pattern is done by writing a program to simulate with that of student with optimum time-to-degree.

From Table 7.2 it can be seen that new programmes were developed including GA which provide a practical method to mine educational data to predict optimum CGPA and time-to-degree using course taking patterns of students and course difficulty level pattern. In the literature knowledge is available to develop computer programmes to realize some functions using which new computer programmes were developed to make the modified CRISP-DM process function to extract hidden patterns of courses taken by students as well as the course difficulty level pattern.

A significant contribution in this direction is the development of a GA that is able to extract hidden contextual factor pattern from educational data along with course taking pattern useful to predict optimum CGPA and time-to-degree. This GA can generate models in any HEI that is similar to the one where this research was conducted. This is an important contribution to knowledge as previously no artefact that could predict the optimum CGPA and time-to-degree using patterns of courses taken by students as well as the course difficulty level pattern. Although CRISP-DM process is a cross industry standard process, it has not been applied to HEIs. This research has demonstrated the usefulness of CRISP-DM process model to support HEIs in enhancing student learning experience in terms of achieving optimum CGPA and time-to-degree as well as decision making. In this study genetic algorithm was modified to generate course taking patterns to predict optimum CGPA and time-to-degree an outcome that has not been dealt with in the literature in regards to the context HEIs although the use of Genetic Algorithm and its modification for application to different contexts is found in the literature (e.g. Ramjeet et al, 2011; Papagelis and Kalles, 2001; Lakshmi et al, 2013). In this research, genetic algorithm was modified to generate course taking patterns to predict optimum CGPA and time-to-degree.

This research demonstrated that common data mining algorithms cannot be used to extract hidden knowledge in the educational dataset and there is a need to use KDDM process under certain circumstances where existing algorithms are unable to extract hidden knowledge, for instance course difficulty level pattern.

7.3.1.2 Contribution to knowledge related to performance management of HEIs

This research provides new relationships hitherto not thought of in the domain of HEIs in the process of improving their performance. Time-to-degree is a major factor that affects students, HEIs and other stakeholders. That this factor could be linked to the concept of course taking pattern and determine the optimum time-to-degree and CGPA is a novel way to understand the performance of the students. While in the literature there are sporadic examples giving hint of this relationship, hardly any research has been conducted to study the educational data to extract this relationship hidden in the dataset over the entire tenure of the student from enrollment to the programme till graduation. Studies like Kovacic (2010); Bahr (2010) have shown that course taking patterns are linked to time-to-degree. Studies like Belcheir, 2000; Volkwein & Lorang, 1996; DesJardins et al., 2002 show that GPA can be associated with time-to-degree. But there is hardly any study that shows the relationship course taking pattern, CGPA and time-to-degree. Similar contribution in the extant literature is hard to find. This research contributes to knowledge in developing those relationships and establishing them through the method of KDDM. Thus achieving equation 6.1 and the corresponding hypothesis outlined in Section 6.4 is a contribution.

In addition the usefulness of the contextual factors in association with course taking pattern to predict accurately the optimum time-to-degree and CGPA is another new relationship this research has brought out. This contribution is new as no similar contributions could be found in the relevant literature. Significantly these relationships were found to be hidden in the educational dataset and have been extracted from the educational dataset. Using this relationship it is possible for HEIs to predict and improve student performance with regard to optimum CGPA and time-to-degree thus helping students to have a better learning experience. Even though Baker (2008) has emphasised that issues of time, sequence, and context show significant roles in the study of educational data there is not much work done on extracting contextual information. Further it has been observed in the literature that existence of huge data of students from similar learning experiences but in very different contexts gives leverage for studying the influence of contextual factors on learning and learners (Knoblauch & Hoy, 2008).

In spite of enough emphasis on contextual knowledge to be generated in the education domain not much work has been at the process (KDDM) level to generate contextual factors. This research has found the usefulness of the contextual factors in association with course taking pattern to predict accurately the optimum time-to-degree and CGPA by developing a new artefact.

7.3.2 Contribution to method

This research has developed a new method to test the equation 6.1. Using this method it is possible discover both contextual factor pattern and course taking pattern which is unique. For instance literature shows that if a dataset is to be mined to discover knowledge through a KDDM process only one dataset can be fed at a time (Sharma et al. 2012). This research has developed modified CRISP-DM process model that could handle two datasets, one the contextual dataset and the other the general dataset. Similar research effort is difficult find in the literature (Chapman et al. 2000).

A new GA has been developed that can handle two different data sets and generate two different patterns namely course taking pattern and course difficulty level pattern. No similar research effort that has handled two datasets simultaneously has been found in the literature (Chapman et al. 2000; Fayyad et al. 1996a; Anand & Buchner, 1998; Cabena et al. 1998; Cios et al. 2000) related to HEIs.

7.3.2.1 The CRISP-DM model guidelines were successfully used to test the modified CRISP-DM process model.

7.3.2.2 Simulation was carried out to test the modified model to predict any future occurrence of the event that is predicting the performance of future students and enhance it. In the absence of a facility to deploy the modified CRISP-DM process the simulation method provides a way to understand the performance of the model which is in line the literature related to design science (Hevner et al. 2004).

7.3.2.3 Complex data understanding and preparation functions have been dealt with simple computer programmes that are reliable and valid. These are in line with similar efforts found in the literature (Pyle, 1999; Witten et al. 2016) which confirm that the contribution although new to the field HEIs are in line with the contribution of others.

7.3.2.4 Course difficulty can be measured on five point nominal scale with points ‘very difficult’, ‘difficult’, ‘average’, ‘easy’ and ‘very easy’. This scale provides a simple way to measure and classify course difficulty level of courses for each student. A search through the KDDM literature did not yield similar nominal scale to be found to have been developed to test the course difficulty level pattern.

7.3.3 Contribution to practice

The modified CRISP-DM process model can be deployed in HEIs as this model has been practically tested on an actual educational data set. The model has potential benefits to enhance student performance in terms of completing their programme within an optimum time-to-degree scoring optimum CGPA. Using the outcome of the modified CRISP-DM process it is possible for HEIs to develop policies and procedures to profile students, to advice students, to guide students during enrollment, to plan teaching methods, to change or design a curriculum, to enhance teaching facilities to support students who cannot cope up, to monitor student performance and recruit new students. Further HEIs can make decisions with regard to support mechanisms to address the problems of underperforming students using the course taking pattern and course difficulty pattern. In addition, HEIs can optimise on offering courses in each semester by studying the course taking pattern of students.

7.4 Limitations

The main limitation of this research lies in the fact that the results are unlikely to be generalizable in the absence of test results conducted in other HEIs. In addition, the outcome of this result could have been different if other contextual factors had been included in the investigation (e.g. course complexity and student potential, refer Appendix 25 for additional contextual factors). Further the model used to address the research question could have been any other KDDM process other than CRISP-DM. Knowledge about the performance of other KDDM processes (e.g. SEMMA, KDD). Each process has unique characteristics that need to be understood before processing the data.

7.5 Recommendations for future research

To begin with it is recommended that this model could be applied to any other HEI to generalise the outcomes of this research. In addition more contextual factors (refer Appendix 25) like student potential, course complexity factors, student background factors, faculty attributes, classroom attributes could be included to know their effect on the relationship between optimum

CGPA and time-to-degree, course taking pattern and course difficulty pattern. The study of contextual factors identified in this section assume significance as it is possible that future research outcomes can produce a more accurate result pertaining to prediction of time-to-degree and CGPA. From the literature (see Appendix 25) it can be seen that the contextual factors that have been recommended by other researchers have the potential to contribute to the body of the knowledge in predicting a more accurate time-to-degree and CGPA in terms of course taking patterns as associated contextual factors. It is possible that future researchers can handle the contextual factors individually or groups of factors. This are complex concepts and hence there is a need to understand the influence of these factors on time-to-degree and CGPA and hence student performance. It is therefore recommended that student performance could be studied more elaborately using KDDM processes as the one developed in this research by identifying and including specific contextual factors in this study which may benefit student and HEIs. Besides the performance of students could be linked to the employment requirements of the students and generate new patterns to determine the optimum CGPA and time-to-degree. The relationship developed in this research regard to the optimum CGPA and time-to-degree, course taking pattern and course difficulty pattern could be tested using a different KDDM to know whether the concepts developed in this research could work.

References

Adelman, C. (1999). Answers in the toolbox: Academic Intensity,attendance patterns, and Bachelor's Degree attainment. Washington DC, Department of Education, Office of Educational Research and Improvement.

Alazmi, A.R. and Alazmi, A.R., 2012. Data mining and visualization of large databases. International Journal of Computer Science and Security, 6(5), pp.295-314.

Aggarwal, CC. (2015). Data Mining: The Textbook. Berlin, Germany: Springer.

Agrawal, R. and Srikant, R. (1995). Mining sequential patterns. In Eleventh International Conference on Data Engineering.

Aggarwal, Charu C. and Reddy, Chandan K. (2013). Data clustering: algorithms and applications. CRC Press.

Al-Dubaee, S. A. and Ahmad, N. (2010). Multilingual Lossy Text Compression Using Wavelet Transform. First International Conference on Integrated Intelligent Computing. (ICIIC). pp. 39-44.

Aljohani, N. R. and Davis, H. C. (2012). Learning analytics in mobile and ubiquitous learning environments. Proceedings of the 11th World Conference on Mobile and Contextual Learning: mLearn '12, Helsinki, Finland: Centre of Learning Sciences and Technologies.

Aslam, S. and Ashraf, I. (2014). Data Mining Algorithms and their applications in Education Data Mining, International Journal of Advance Research in Computer Science and Management Studies, Vol 2, Issue 7.

Astin, A., L. Tsui, and J. Avalos. (1996). Degree attainment rates at American colleges and universities: Effects of race, gender, and institutional type. Los Angeles, Calif.: Higher Education Research Institute, Graduate School of Education, University of California,Los Angeles.

Astin, A.W. (1993). Four critical years. San Francisco, CA: Jossey-Bass.

Astin, A.W. (1993). What matters in college?: Four critical years revisited. San Francisco: Jossey-Bass.

Astin, A. W., & Oseguera, L. (2004). The declining "equity" of American higher education. Review of Higher Education, 27(3), 321-341.

Astin, A.W. & Oseguera, L. (2005). Degree Attainment Rates at American Colleges and Universities. Revised Edition. Los Angeles: Higher Education Research.

Athiyaman, A. (1997). Linking student satisfaction and service quality perceptions: The case of university education. European Journal of Marketing, 31(7), pp. 528-540.

Bahr, P. R., (2010). The bird's eye view of community colleges: A behavioral typology of first-time students based on cluster analytic classification. Research in Higher Education, 51(8), pp. 724-749.

Bailey, T.C. and Gatrell, A.C.(1995). Interactive spatial data analysis. Longman Scientific & Technical Essex.

Bajpai, P. and Kumar, M., (2010). Genetic algorithm–an approach to solve global optimization problems. Indian Journal of computer science and engineering, 1(3), pp.199-206.

Baker, R.S.J.d., Barnes, T., Beck, J.E. (Eds.) (2008). Proceedings of the 1st International Conference on Educational Data Mining.

Baker, R.S.J.D., (2010). Data mining for education. International encyclopedia of education, 7(3), pp.112-118.

Baker, R. S. J. D., & Yacef, K. (2009). The state of educational data mining in 2009: A review and future vision. Journal of Educational Data Mining, 1(1), 1–15.

Bandaru, S., Ng, A.H. and Deb, K., (2017). Data mining methods for knowledge discovery in multi-objective optimization: Part B-New developments and applications. Expert Systems with Applications, 70, pp.119-138.

Bawden, D., (2013). Data Mining and Decision Support: Integration and Collaboration. Journal of Documentation.

Belcheir, M. J. (2000). Predicting the probability of graduating after four, six, and ten years. research report. No. BSU-RR-2000-01.

Bess, J.L. and Dee, J.R., (2008). Understanding college and university organization: Dynamics of the system (Vol. 2). Stylus Publishing, LLC.

BIS (2014). Improving the Student Learning Experience – a national assessment (BIS Research Paper no 169). London: Department for Business, Innovation and Skills.

Bienkowski, M., Brecht, J. and Klo, J., (2012). The learning registry: building a foundation for learning resource analytics. Proceedings of the 2nd international conference on learning analytics and knowledge (pp. 208-211). ACM.

Bowen, W.G., Chingos, M.M. and McPherson, M.S. (2009). Crossing the finish line: Completing college at America's public universities. Princeton University Press.

Burkom, H., Murphy, S., and Shmueli, G. (2007). Automated time series forecasting for biosurveillance. Statistics on Medicine. 26(22). pp. 4202–4218.

Brachman, R. & Anand, T., (1996). The process of knowledge discovery in databases:A human-centered approach.In U. Fayyad, G. Paitestsky-Shapiro, P. Smith, & R. Uthuruswamy (Eds.), Advances in knowledge discovery and data mining. AAAI Press, pp. 36-57.

Brinckmann, J., Grichnik, D. and Kapsa, D., (2010). Should entrepreneurs plan or just storm the castle? A meta-analysis on contextual factors impacting the business planning–performance relationship in small firms. Journal of Business Venturing, 25(1), pp.24-40.

Cabena, P., Hadjinian, P., Stadler, R., & Verhees, J. (1998). Discovering data mining: From concepts to implementation. Prentice Hall.

Cabrera, A.F., Burkum, K.R. and La Nasa, S.M., (2005). Pathways to a four-year degree. College student retention: Formula for student success, pp.155-214.

Cai C.Z., Han L. Y., Ji Z. L., Chen X., and Chen Y. Z. (2003) SVM-Prot: Web-based support vector machine software for functional classification of a protein from its primary sequence. Nucleic Acids Research, vol. 31, pp. 3692–3697.

Catley, C., Smith, K., McGregor, C. & Tracy, M., (2009). Extending CRISP-DM to Incorporate Temporal Data Mining of Multi dimensional Medical Data Streams: A Neonatal Intensive Care Unit Case Study. s.l., Computer-Based Medical Systems, 2009. CBMS 2009. 22nd IEEE International Symposium.

Chapman, P., Clinton, J., Kerber, R. & Khabaza, T., (2000). CRISPDM 1.0 step-by-step data mining guide. Technical report, CRISP-DM , s.l.: CRISP-DM.

Campbell, J. W., & Blakey, L. S. (1996). Assessing the impact of early remediation in the persistence and performance of underprepared community college students. Paper presented at AIR Conference, Albuquerque, NM.

Cao, L. and Tay, F. (2009). Feature selection for support vector machines in financial time series forecasting. In Intelligent Data Engineering and Automated Learning. Lecture Notes in Computer Science, vol. 1983. Springer, 41–65.

Carr, J., (2014). An introduction to genetic algorithms. Senior Project, pp.1-40.

Carlos Ordonez (2004). Programming the K-means Clustering Alogrithm in SQL. ACM Int. Conf. on Knowledge Discovery and Data Mining, 2004.

Charu C. Aggarwal , Jiawei Han , Jianyong Wang , Philip S. Yu,(2003). A framework for clustering evolving data streams, Proceedings of the 29th international conference on Very large data bases, p.81-92, September 09-12, 2003, Berlin, Germany.

Chen-Huei, C. ,Atish, S. and Huimin, Z.,(2008) “A text mining approach to Internet abuse detection,” Information Systems and e-Business Management, vol. 6, no. 4, pp. 419-439, 2008.

Cheqing Jin et al. Dynamically Maintaining Frequent Items over a Data Stream, in the proceedings of CIKM USA, 2003.

Cios, K., Teresinska, A., Konieczna, S., Potocka, J., & Sharma, S. (2000). Diagnosing myocardial perfusion from PECTbull’s-eye maps – A knowledge discovery approach. IEEE Engineering in Medicine and Biology Magazine, Special Issue onMedical Data Mining and Knowledge Discovery, 19(4), 17–25.

CRISP-DM (2003). CRoss industry standard process for data mining 1.0: Step by step data mining guide. <http://www.crisp-dm.org/> Retrieved 01.10.10

Complete College America. (2014). Four-year myth. Indianapolis, IN: Author. Retrieved from http:// completecollege.org/wp-content/uploads/2014/11/4- Year-Myth.pdf

Conrad, C., Gasman, M., Lundberg, T., Nguyen, T.H., Commodore, F. and Samayoa, A.C., 2013. Using educational data

to increase learning, retention, and degree attainment at minority serving institutions (MSIs). A Research Report of Penn

Graduate School of Education, GSE.

Cormode, G., Muthukrishnan, S., and Zhuang, W. 2007. Conquering the divide: Continuous clustering of distributed data

streams. In Proceedings of the IEEE 23rd International Conference on Data Engineering. 1036–1045.

Cortez, P., Silva, A.(2008),' Using data mining to predict secondary school student performance',Proceedings of 5th Future Business Technology Conference, Oporto, Portugal.

Crawford, Kate, Kate Miltner and Mary L. Gray.(2014). "Critiquing Big Data: Politics, Ethics, Epistemology." International Journal of Communication 8, 1663–1672. http://ijoc.org/ index.php/ijoc/article/view/2167/1164.

Cullinane, J. and Lincove, J.A., 2014, May. The effects of institutional inputs on time to degree for traditional and nontraditional college students. In annual Association of Education Finance and Policy conference, San Antonio, TX.

Cullinane, J.P., 2014. The path to timely completion: supply-and demand-side analyses of time to bachelor's degree completion (Doctoral dissertation).

Cunha, J. M. and Miller, T. (2012). Measuring value-added in higher education. Presented as part of Context for Success. Available at: www.hcmstrategists.com/contextforsuccess/papers/CUNHA_MILLER_PAPER.pdf.

Darlington, R.B. and Hayes, A.F.,(2016). Regression analysis and linear models: Concepts, applications, and implementation. Guilford Publications.

Daempfle, P. A., (2003). An Analysis of the High Attrition Rates Among First Year College Science,Math,and Engineering Majors. Journal of College Student Retention, 5(1), pp. 37-52.

Daniel, B., (2015). Big data and analytics in higher education: Opportunities and challenges. British journal of educational technology, 46(5), pp.904-920.

Davenport, T. . H., (2010). The New World of "Business Analytics", s.l.: International Institute for Analytics.

Dekker, G., Pechenizkiy, M., Vleeshouwers, J(2009), 'Predicting students drop out: a case study',Proceedings of the 2nd International Conference on Educational Data Mining (EDM'09), Cordoba, Spain (2009).

Delavari, N. , Phon-Amnuaisuk, S., Beikzadeh, M. R.(2007), "Data mining application in higher learning institutions", *Inform. Edu.-Int. J.*, vol. 7, no. 1, pp. 31-54.

DeShields, O. W., Kara, A. & Kaynak, E., (2005). Determinants of business student satisfaction and retention in higher education: Applying Herzberg's two-factor theory. International Journal of Educational Management, 19(2), pp. 128-139.

De Francisci Morales, G., Bifet, A., Khan, L., Gama, J. and Fan, W., (2016). Iot big data stream mining. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 2119-2120). ACM.

Dey, A. K.(2001). Understanding and Using Context. Personal Ubiquitous Comput., 5(1), pp. 4-7.

Derya Birant & Alp Kut, (2007) "ST-DBSCAN: An algorithm for clustering spatio-temporal data", Data & Knowledge Engineering, Volume 60, Issue 1, January, Pages 208-221. [17].

DesJardins, S. L., Ahlburg, D. A., & McCall, B. P. (2002). A temporal investigation of factors related to timely degree completion. Journal of Higher Education, 73(5), 555-581.

Dianhui Wang, Guang-Bin Huang (2005) Protein Sequence Classification Using Extreme Learning Machine. Proceedings of International Joint Conference on Neural Networks (IJCNN2005), Montreal, Canada.

Douglass, J.A., Edelstein, R. and Haoreau, C., 2013. Seeking smart growth: The Idea of a California global higher education Hub.

Dora Cai, Y. , David Clutter , Greg Pape , Jiawei Han , Michael Welge , Loretta Auvil,(2004),MAIDS: mining alarming incidents from data streams, Proceedings of the 2004 ACM SIGMOD international conference on Management of data, June 13-18, 2004, Paris, France [doi>10.1145/1007568.1007695].

Dwivedi, T. and Singh, D., (2016). Analyzing Educational Data through EDM Process: A Survey. International Journal of Computer Applications, 136(5), pp.13-15.

Du, W., Lin, H., Sun, J., Yu, B. and Yang, H.,(2016), August. Combining Statistical Information and Distance Computation for K-Means Initialization. In Semantics, Knowledge and Grids (SKG), 2016 12th International Conference on (pp. 97-102). IEEE.

Džeroski, S., (2003). Multi-relational data mining: an introduction. ACM SIGKDD Explorations Newsletter, 5(1), pp.1-16.

Dzeroski, S., (2007). Inductive logic programming in a nutshell. Introduction to Statistical Relational Learning [16].

Edelstein, R., 2014. Globalization and Student Learning: A Literature Review and Call for Greater Conceptual Rigor and Cross-Institutional Studies. Research & Occasional Paper Series: CSHE. 6.14. Center for Studies in Higher Education.

Elliott, K. M. & Healy, M. A., (2001). Key factors influencing student satisfaction related to recruitment retention. Journal of Marketing for Higher Education, 10(4), pp. 1-11.

Evgeniy Gabrilovich , ShaulMarkovitch , 2004, “Text Categorization with Many Redundant Features: Using Aggressive Feature Selection to Make SVMs Competitive with C4.5”, ICML.

Fayyad, U., Piatetsky-Shapiro, G., Smyth, P. & Uthurusamy, R.,(1996). Advances in knowledge discovery and data mining. MIT Press.

Federico Michele Facca and Pier Luca Lanzi.(2003) “Recent Developments in Web Usage Mining Research” in proceedings of Data Warehousing and Knowledge Discovery Volume 2737, Pg 140-150.

Fettke, P., Vella, A. L. & Loos, P., (2012). From Measuring the Quality of Labels in Process Models to a Discourse on Process Model Quality: A Case Study. Maui, HI , IEEE.

Fisher, W.D., Camp, T.K. and Krzhizhanovskaya, V.V.,(2017). Anomaly detection in earth dam and levee passive seismic data using support vector machines and automatic feature selection. Journal of Computational Science, 20, pp.143-153.

Florian Verhein & Sanjay Chawla, (2005) Mining Spatio-Temporal Association Ruls, Sources, Sinks, Stationary Regions and Thoroughfares in Object Mobility Databases, Technical Report Number 574, The University of Sydney.

Gamarra, C., Guerrero, J.M. and Montero, E., (2016). A knowledge discovery in databases approach for industrial microgrid planning. Renewable and Sustainable Energy Reviews, 60, pp.615-630.

García, S., Ramírez-Gallego, S., Luengo, J., Benítez, J.M. and Herrera, F., (2016). Big data preprocessing: methods and prospects. Big Data Analytics, 1(1), p.9.

Goyal, M. and Vohra, R., (2012). Applications of data mining in higher education. International journal of computer science, 9(2), p.113.

Gillmore, G.M. and Hoffman, P.H., (1997). The graduation efficiency index: Validity and use as an accountability and research measure. Research in Higher Education, 38(6), pp.677-697.

Greenhow, C., Robelia, B., & Hughes, J. (2009). Learning, teaching, and scholarship in a digital age: Web 2.0 and classroom research: What path should we take now? Educational Researcher, 38, 246–259.

Han, J., Pei, J. and Kamber, M., (2011). Data mining: concepts and techniques. Elsevier.

Holland, J. H. (1975). Adaptation in Natural and Artificial Systems. University of Michigan Press. (Second edition: MIT Press, 1992.).

Holland, J.H. and Reitman, J.S., (1977). Cognitive systems based on adaptive algorithms. Acm Sigart Bulletin, (63), pp.49-49.

Han, J., Pei, J. and Kamber, M., (2011). Data mining: concepts and techniques. Elsevier.

Hancer, E. and Karaboga, D., (2017). A comprehensive survey of traditional, merge-split and evolutionary approaches proposed for determination of cluster number. Swarm and Evolutionary Computation, 32, pp.49-67.

Hanna, M. (2004). Data mining in the e-learning domain. In Campus-Wide Information Systems, Volume 21, Number 1, 29-34.

Helgesen, O. & Nesset, E., (2007). What accounts for students' loyalty? Some field study evidence. International Journal of Educational Management, 21(2), pp. 126-143.

Herrouz, A., Khentout, C., Djoudi, M. (2012) Overview of Visualization Tools for Web Browser History Data, IJCSI International Journal of Computer Science Issues, Vol.9, Issue 6,No3, November 2012, pp. 92-98.

Herzog, S., (2006). Estimating student retention and degree completion time: Decision trees and neural networks via regression. New directions for institutional research, 2006(131), pp.17-33.

Hiremath, R. and Patil, P., (2016). A Study-Knowledge Discovery ApproachesAnd Its Impact With Reference To Cognitive Internet Of Things (CIOT). International Journal of Information, 6(1/2).

Hollands, F.M. and Escueta, M., (2017). EdTech Decision-making in Higher Education.

Hua, L.T., 2011, July. Sustainable competitive advantage for market leadership amongst the private higher education institutes in Malaysia. In 2nd International Conference on Business and Economic Research (2nd ICBER 2011) Proceeding.

Hunt, E.B., Marin, J., Stone, P.J.(1996), Experiments in Induction (Academic Press, New York, 1966.

Jason T. L. Wang, Qic Heng Ma, Dennis Shasha, Cathy H Wu (2000) Application of Neural Networks to Biological Data Mining: A case study in Protein Sequence Classification. KDD, Boston, MA, USA, pp: 305-309.

Ishitani, T. T. (2003). A longitudinal approach to assessing attrition behavior among first generation students: Time-varying effects of pre-college characteristics. Research in Higher Education, 44(4), 433-449.

Ishitani, T. T. (2005, May). Studying educational attainment among first-generation students in the United States. Paper

presented at AIR Conference, San Diego, CA.

Ishitani, T. T., & Snider, K. G. (2006). Longitudinal effects of college preparation programs on college retention. IR applications. Volume 9., Association for Institutional Research; 10p.

Lane, J.E. ed., 2014. Building a smarter university: Big data, innovation, and analytics. SUNY Press.

Larkin, H. and Watchorn, V., 2012. Changes and challenges in higher education: what is the impact on fieldwork education?. Australian occupational therapy journal, 59(6), pp.463-466.

Li, W., Han, J. and Pei, J., 2001. CMAR: Accurate and efficient classification based on multiple class-association rules. In Data Mining, 2001. ICDM 2001, Proceedings IEEE International Conference on (pp. 369-376). IEEE.

Luan, J., 2002. Data Mining and Knowledge Management in Higher Education-Potential Applications.

Jeffrey W. Seifert, ―Data Mining: An Overview‖ CRS Report for Congress, 2004.

Jewell, D., Barros, R.D., Diederichs, S., Duijvestijn, L.M., Hammersley, M., Hazra, A., Holban, C., Li, Y., Osaigbovo, O., Plach, A. and Portilla, I., 2014. Performance and capacity implications for big data. IBM Redbooks.

John, G. H., 1997. Enhancements to the data mining process. s.l.:PhD thesis, Stanford University.

Kamarainen JK, Kyrki V, Ilonen J, Kälviäinen H. Improving similarity measures of histograms using smoothing projections. Pattern Recognition Letters. 2003;24(12):2009–2019. doi: 10.1016/S0167-8655(03)00039-4.

Kaushik Sinha, Xuan Zhang, Ruoming Jin, and Gagan Agrawal (2005). Using Data Mining Techniques to Learn Layouts of Flat-File Biological Datasets . In Proceedings of IEEE Symposium on Bioinformatics and Bioengineering (BIBE).

Khan, K., (2013). Spatio-temporal variation in educational status and level of socio-economic development in Aligarh district, UP, India. International Journal of Management, IT and Engineering, 3(12), p.410.

Kharade B and Wagh K.,(2016). Data Analytics in Educational Management System. IJCA Proceedings on National Conference on Advances in Computing, Communication and Networking ACCNET 2016(5):22-25, June 2016. Full text available.

Khoo, S., Khoo, S., Ha, H., Ha, H., McGregor, S.L. and McGregor, S.L., (2017). Service quality and student/customer satisfaction in the private tertiary education sector in Singapore. International Journal of Educational Management, 31(4), pp.430-444.

Kirchdoerfer, T. and Ortiz, M., (2017). Data Driven Computing with Noisy Material Data Sets. arXiv preprint arXiv:1702.01574.

Klopfenstein, K. (2000). The effect of AP participation on time to college graduation: Technical report. A Critical Examination of the Advanced Placement Program. Cambridge, MA: Harvard Press.

Khobragade, M.V.B., Patil, M.L.H. and Patel, M.U., (2015). Image retrieval by information fusion of multimedia resources. IMAGE, 4(5).

Knight, W. E., (1994). Why the five-year (or longer) bachelors degree ?:An exploratory study of time-to-degree attainment. New Orleans, LA, Association for Institutional Research forum.

Knight, C., (2000). Engaging the student in the field instruction relationship: BSW and MSW students' views. Journal of Teaching in Social Work, 20(3-4), pp.173-201.

Knight, W. E. (2004). Time to bachelor's degree attainment: An application of descriptive, bivariate, and multiple regression techniques. IR applications, volume 2, September 8, 2004., Association for Institutional Research; 16p.

Knight, W. E., & Arnold, W. (2000). Towards a comprehensive predictive model of time to bachelor's degree attainment. Presented at AIR Conference, Cincinnati, OH.

Kumpošt, M., 2009. Context Information and user profiling (Doctoral dissertation, Masarykova univerzita, Fakulta informatiky).

Knoblauch, D. and Hoy, A.W., 2008. “Maybe I can teach those kids.” The influence of contextual factors on student teachers’ efficacy beliefs. Teaching and Teacher Education, 24(1), pp.166-179.

Koopmans, M., 2011. Time Series in Education: The Analysis of Daily Attendance in Two High Schools. Online Submission.

Kovacic, Z. J., 2010. Early prediction of student success: Mining student enrollment data. s.l., Proceedings of Informing Science & IT Education Conference .

Kaur, P., Singh, M. and Josan, G.S., 2015. Classification and prediction based data mining algorithms to predict slow learners in education sector. Procedia Computer Science, 57, pp.500-508.

Kumar, D. and Bhardwaj, D., 2011. Rise of data mining: Current and future application areas. IJCSI International Journal of Computer Science Issues, 8(5).

Kumar, M., Singh, A.J. and Handa, D., 2017. Literature Survey on Educational Dropout Prediction.

Kumar, A., Tyagi, A.K. and Tyagi, S.K., 2014. data Mining: Various Issues and Challenges for Future A Short discussion on Data Mining issues for future work. International Journal of Emerging Technology and Advanced Engineering, 4(1), pp.1-8.

Kurgan, L. & Musilek, P., 2006. A survey of knowledge discovery and data mining process models. Knowledge Engineering Review, 21(1), pp. 1-24.

Lin, T., Kaminski, N., and Bar-joseph, Z. 2008. Alignment and classification of time series gene expression in clinical studies. Bioinf. 24, 13, 147–155.

Lin, S.H. and Ho, J.M., 2002, July. Discovering informative content blocks from Web documents. In Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 588-593). ACM.

Liebowitz, J., 2017. Thoughts on Recent Trends and Future Research Perspectives in Big Data and Analytics in Higher Education. In Big Data and Learning Analytics in Higher Education (pp. 7-17). Springer International Publishing.

Levien, R. E., & Maron, M. E. (1967). A computer system for inference execution and data retrieval. Communications of the ACM, 10(11) (November 1967), 715–721.

Lotkowski, V. A., Robbins, S. B. & Noeth, R. J., 2004. The Role of Academic and Non-Academic Factors in Improving College Retention, Iowa City, IA: ACT Policy Report.

Li, J., Yang, B. & Song, W., 2009. A New Data Mining Process Model for Aluminum Electrolysis. Qingdao, P. R. China, Proceedings of the International Symposium on Intelligent Information Systems and Applications (IISA'09).

Lotkowski, V. A., Robbins, S. B. & Noeth, R. J., 2004. The Role of Academic and Non-Academic Factors in Improving College Retention, Iowa City, IA: ACT Policy Report.

Letkiewicz, J., Lim, H., Heckman, S., Bartholomae, S., Fox, J.J. and Montalto, C.P., 2014. The path to graduation: Factors predicting on-time graduation rates. Journal of College Student Retention: Research, Theory & Practice, 16(3), pp.351-371.

Macfayden, L. P., & Dawson, S. (2010). Mining LMS data to develop an ‘‘early warning’’ system for educators: A proof of concept. Computers and Education, 54(2), 588–599.

Maki, P.L., 2017. Real-time Student Assessment: Meeting the Imperative for Improved Time to Degree, Closing the Opportunity Gap, and Assuring Student Competencies for 21st Century Needs. Stylus Publishing, LLC.

Manso J A, Times V C, Oliveira G, Alvares L & Bogorny V, (2010) "DB-SMoT: A Direction based Spatio-Temporal Clustering Method", Fifth IEEE International Conference on intelligent Systems IEEE IS 2010.

Marbán, Ó., Mariscal, G. & Segovia, J., 2009. A Data Mining & Knowledge Discovery Process Model. I-Tech, Vienna, Austria: Data Mining and Knowledge Discovery in Real Life Applications, Julio Ponce and Adem Karahoca (Ed.),ISBN: 978-3-902613-53-0.

Maroko, A., Maantay J.A. and Grady K. (2011). Using geovisualization and geospatial analysis to explore respiratory disease and environmental health justice in New York City. In Geospatial Analysis of Environmental Health (pp. 39-66). Springer Netherlands.

Meenakumari, J. and Kudari, J.M.,(2015) Learning Analytics and its challenges in Education Sector a Survey.

Menon, M.E., Terkla, D.G. and Gibbs, P. eds., 2014. Using data to improve higher education: Research, policy and practice. Springer.

Michelle Davis, "Data Tools Aim to Predict Student Performance," Education Week Digital Directions, February 8, 2012.

Milkman, K.L., Chugh, D. and Bazerman, M.H., 2009. How can decision making be improved?. Perspectives on psychological science, 4(4), pp.379-383.

Miguel Gomes da Costa Júnior and Zhingo Gong. "Web Structure Mining: An Introduction" in proceedings of the 2005 IEEE International Conference on Information Acquisition June 27-July 3, 2005, Hong Kong and Macau, China.

Minaei-Bigdoli, B., Kashy, D. A., Kortemeyer, G. & Punch, W. F., 2003. 33rd ASEE/IEEE Frontiers in Education Conference. Boulder,CO, IEEE.

Mitchell, T. M. (1997). Machine Learning. McGraw-Hill, Inc., New York, NY, USA, 1 edition.

More, S. and Mishra, D.K., 2012. Multimedia Data Mining: A Survey. Pratibha: International Journal of science, spirituality, business and technology (ijssbt), 1(1).

Nan Jiang and Le Grunewald. Research Issues in Data Stream Association Rule Mining, SIGMOD Record, 2006: 35(1).

Nandeshwar, A., Menzies, T. and Nelson, A., 2011. Learning patterns of university student retention. Expert Systems with Applications, 38(12), pp.14984-14996.

Nanopoulos, A., Alcock, R., AND Manolopoulos, Y. 2001. Feature-Based classification of time-series data. In Information Processing and Technology. 49–61.

Nesbit, J. C., Xu, Y., Winne, P. H., Zhou, M., (2008), "Sequential pattern analysis software for educational event data", in Proc. Int. Conf. Methods Tech. Behav. Res. Netherlands,pp.1-5.

Neagu, B., Grigoraş, G., Scarlatache, F., Schreiner, C. and Ciobanu, R., 2017, April. Patterns discovery of load curves characteristics using clustering based data mining. In Compatibility, Power Electronics and Power Engineering (CPE-POWERENG), 2017 11th IEEE International Conference on (pp. 83-87). IEEE.

Nikolovski, V., Stojanov, R., Mishkovski, I., Chorbev, I. and Madjarov, G., Educational Data Mining: Case Study for Predicting Student Dropout in Higher Education.

Nithya, D.P., Umamaheswari, B. and Umadevi, A., 2016. A survey on educational data mining in field of education. International Journal of Advanced Research in Computer Engineering & Technology (IJARCET), 5(1), pp.69-78.

Oklahoma Council on Student Affairs.(OCSA) (1996). Time-to-degree completion. A system-wide survey of Oklahoma college and university students. Oklahoma City: Author.

Oladokun, V.O., Adebanjo, A.T. and Charles-Owaba, O.E., 2008. Predicting students' academic performance using artificial neural network: A case study of an engineering course. The Pacific Journal of Science and Technology, 9(1), pp.72-79.

Ouyang, R., Ren, L., Cheng, W., AND Zhou, C. 2010. Similarity search and pattern discovery in hydrological time series data mining. Hydrol. Process. 24, 9, 1198–1210.

Osmanbegović E., Suljić M : Data Mining Approach for Predicting student performance, Economic Review – Journal of Economics and Business, Vol. X, Issue 1, May 2012.

Oladipupo, O.O. and Oyelade, O.J., 2010. Knowledge discovery from students' result repository: association rule mining approach. International Journal of Computer Science and Security, 4(2), pp.199-207.

O'Hagan, A. and White, A., 2015. Improved model-based clustering performance using Bayesian initialization averaging. arXiv preprint arXiv:1504.06870.

Pan, D., 2010. An Integrative Framework for Continuous Knowledge Discovery. Journal of Convergence Information Technology, 5(3).

Paul Bradley, Usama Fayyad, Cory Reina: Scaling Clustering Algorithms to Large Databases. KDD-98, pages 9-15.

Pesaran, M., Pettenuzzo, D., and Timmermann, A. 2006. Forecasting time series subject to multiple structural breaks. Rev. Econ. Studies 73, 4, 1057–1084.

Pheng, L. S. & Arain, F. M., 2006. A KNOWLEDGE-BASED SYSTEM AS A DECISION MAKING TOOL FOREFFECTIVE MANAGEMENT OF VARIATIONS AND DESIGN IMPROVEMENT: LEVERAGING ON INFORMATION TECHNOLOGY APPLICATIONS. ITcon, Volume 11.

Picciano, A.G., 2012. The Evolution of Big Data and Learning Analytics in American Higher Education. Journal of Asynchronous Learning Networks, 16(3), pp.9-20.

Pitter, G. W., LeMon, R.E., & Lanham, C.H. (1996). Hours to graduation: A national survey of credit hours required for baccalaureate degrees.

Popivanov, I. and Miller, R. 2002. Similarity search over time-series data using wavelets. In Proceedings of the International Conference on Data Engineering. 212–224.

Povinelli, R., Johnson, M., Lindgren, A., and Ye, J. 2004. Time series classification using Gaussian mixture models of reconstructed phase spaces. IEEE Trans. Knowl. Data Engin. 16, 6, 779–783.

Prineas, M., & Cini, M. (2011, October). Assessing learning in online education: The role of technology in improving student outcomes (NILOA Occasional Paper No.12). Urbana, IL: University of Illinois and Indiana University, National Institute for Learning Outcomes Assessment.

Pullman M., McGuire, K., and Cleveland, C., "Let me count the words: Quantifying open-ended interactions with guests," Cornell Hotel and Restaurant Administration Quarterly, vol. 46, no. 3, pp. 323-343, 2005.

Pyle, D., 1999. Data preparation for data mining (Vol. 1). morgan kaufmann.

Rao, K., Govardhan, A. & Rao, K., 2012. SPATIOTEMPORAL DATA MINING: ISSUES, TASKS AND APPLICATIONS. International Journal of Computer Science & Engineering Survey (IJCSES), 3(1).

Ratcliff, J. L., Jones, E. A., and Hoffman, S. (1992). *Handbook on Linking Assessment and General Education.* University Park, PA: National Center on Postsecondary Teaching, Learning, and Assessment.

Redpath, R. & Srinivasan, B., 2004. A Model for Domain Centered Knowledge Discovery in Databases. Budapest, Hungary, Proceedings of the IEEE 4th International Conference On Intelligent Systems Design and Application August,(ISDA 2004), ISBN: 9637154302.

Rennolls, K. & Al-Shawabkeh, A., 2008. Formal structures for data mining, knowledge discovery and communication in a knowledge management environment. Intelligent Data Analysis, 12(2), p. 147–163.

Ronco, S. L., 1996. How Enrollment Ends: Analyzing the Correlates of Student Graduation, Transfer and Dropout with a Competing Risks Model, Tallahassee, Fla.: AIR Professional File, No. 61 Association for Institutional Research.

Rutkowski et al. Decision trees for mining data streams based on the McDiarmid's bound. IEEE Transactions on Knowledge and Data Engineering, 2013; 25(6).

Ruan, T. . L. D., 2007. An extended process model of knowledge discovery in database. Journal of Enterprise Information Management, 20(2), pp. 169 - 177.

Ramesh, G., Maniatty, W. and Zaki, M.J., 2002, May. Indexing and Data Access Methods for Database Mining. In DMKD.

Ramzan, M. and Ahmad, M., 2014, March. Evolution of data mining: An overview. In IT in Business, Industry and Government (CSIBIG), 2014 Conference on (pp. 1-4). IEEE.

RAO, C. R. (1948). "The utilization of multiple measurements in problems of biologica! classification." J. Roy. Stat. Soc., B 10, 159-193.

Richard A. Huebner: A Survey of educational data-mining Research, Research in Higher Education Journal (2013).

Roiger, Richard J. Data mining: a tutorial-based primer. CRC Press, 2017.

Romero, C., & Ventura, S. (2010). Educational data mining: a review of the state of the art. IEEE Transactions on systems, man, and cybernetics, part C: applications and reviews, 40(6), 601–618.

Shannon, C.E. ,Amathematical theory of communication.BELLSyst. Tech. J. 27(1), 379–423,625–56 (1948).

Sailesh, S.B., Lu, K.J. and Al Aali, M., 2016, August. Profiling students on their course taking patterns in higher educational institutions (HEIs). In Information Science (ICIS), International Conference on (pp. 160-167). IEEE. doi: 10.1109/INFOSCI.2016.7845319 URL: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7845319&isnumber=7845284

SAS, I., 2008. SEMMA data mining methodology. [Online] Available at: http://www.sas.com.

Singh, M., Nagar, H. and Sant, A., 2016. K-mean and EM Clustering algorithm using attendance performance improvement Primary school Student.

Sin, K. and Muthu, L., 2015. APPLICATION OF BIG DATA IN EDUCATION DATA MINING AND LEARNING ANALYTICS--A LITERATURE REVIEW. ICTACT journal on soft computing, 5(4).

Sowan, B. and Qattous, H., 2017. A Data Mining of Supervised learning Approach based on K-means Clustering. International Journal of Computer Science and Network Security (IJCSNS), 17(1), p.18.

Schilit, B., Adams, N. & Want, R., 1994. Context-Aware Computing Applications. s.l., First International Workshop on Mobile Computing Systems and Applications.

Schmidt, W.H., 1983. High School Course- Taking: its Relationship to Achievement. J. Curriculum Studies, 15(3), pp.311-332.

Schneider, M. (2010). Finishing the first lap: The cost of first-year student attrition in America's four-year colleges and universities. Washington, DC: American Institutes for Research. Retrieved from http://www.eric.ed.gov/ ERICWebPortal/contentdelivery/servlet/ERICServlet?accno=ED512253.

Sheikh, Y.A., 2017. Higher Education in India: Challenges and Opportunities. Journal of Education and Practice, 8(1), pp.39-42.

Sharma, S., Osei-Bryson, K.-M. & Kasper, G. M., 2012. Evaluation of an integrated Knowledge Discovery and Data Mining process model. Expert Systems with Applications, Volume 39, p. 11335–11348.

Sharma, B.R., Kaur, D. and Manju, A., 2013. A review on data mining: its challenges, issues and applications. International Journal of Current Engineering and Technology ISSN, pp.2277-4106.

Shacklock, X., 2016. From bricks to clicks: the potential of data and analytics in higher education. Higher Education Commission.

Shakir Mohamed, David Rubin and Tshilidzi Marwala (2006) Multi-class Protein Sequence Classification Using Fuzzy ARTMAP. IEEE Conference pp: 1676 – 1680.

Shekhar, S. and Chawla, S. Spatial Databases: A Tour. Pretice Hall (ISBN 0-7484-0064-6), 2003.

Singh, S. and Gupta, P., 2014. Comparative study ID3, cart and C4. 5 decision tree algorithm: a survey. Int J Adv Inf Sci Technol [Internet], 27, pp.97-103.

Sivarajah, U., Kamal, M.M., Irani, Z. and Weerakkody, V., 2017. Critical analysis of Big Data challenges and analytical methods. Journal of Business Research, 70, pp.263-286.

Skinner, E. (2011). Challenges of college transfer: Senate bill 1440: the student transfer reform act. iJournal: Insight into Student Services, (27). Retrieved from http://www.ijournalccc.com/articles/issue_27/content/senate-bill-1440- student-transfer-achievement-reform-act

Sofo, F., Colapinto, C., Sofo, M. & Ammirato, S., 2013. Adaptive Decision Making and Intellectual Styles. s.l.:Springer.

Sokol, L. and Chan, S., 2013. Context-based analytics in a big data world: better decisions. IBM RedBooks Point-of-View Publication.

Soman, K.P., Diwakar, S. and Ajay, V., 2006. Data Mining: Theory and Practice [WITH CD]. PHI Learning Pvt. Ltd.

Song, H. and Li, G. 2008. Tourism demand modelling and forecasting–A review of recent research. Tour. Manag. 29, 2, 203–220.

Stock, W.A., Siegfried, J.J. and Finegan, T.A., 2011. Completion rates and time-to-degree in economics PhD programs. The American Economic Review, 101(3), pp.176-193.

Stuart P. Lloyd. Least squares quantization in pcm. IEEE Transactions on Information Theory, 28(2):129–136, 1982.

Sudipta Guha, D.Gunopulos, N. Kaudas. Correlating synchronous and asynchronous data streams, in the proceedings of SIGKDD 2003 held from august 24th -27th, USA, 2003.

Sujath1, P., Thailambal, G. and Sheela Angalin Ruby R. "Study of Web Content Mining and Its Tools." International Journal of Engineering & Computer Science Vol.3,Issue 8 (August-2014)

Sunita Sarawagi , Shiby Thomas , Rakesh Agrawal, Integrating association rule mining with relational database systems: alternatives and implications, Proceedings of the 1998 ACM SIGMOD international conference on Management of data, p.343-354, June 01-04, 1998, Seattle, Washington, USA [doi>10.1145/276304.276335]

Surajit Chaudhuri, Usama M. Fayyad, Jeff Bernhardt: Scalable Classification over SQL Databases. ICDE 1999: 470-479.

Surajit Chaudhuri: Data Mining and Database Systems: Where is the Intersection? Data Engineering Bulletin 21(1) 1998.

Tinto, V., 1975. Dropouts from higher education: A theoretical synthesis of recent literature. A Review of Educational Research, Volume 45, pp. 89-125.

Torgo, L., 2016. Data mining with R: learning with case studies. CRC press.

Underwood, D.G. and Rieck, J.R., 1999. Setting graduation rate thresholds. Journal of College Student Retention: Research, Theory & Practice, 1(3), pp.255-266.

University of Oxford. International Strategy Office, 2015. International trends in higher education 2015.

Vajirkar, P., Singh, S. and Lee, Y., 2003, September. Context-aware data mining framework for wireless medical application. In International Conference on Database and Expert Systems Applications (pp. 381-391). Springer Berlin Heidelberg.

Valery A. Petrushin and Latifur Khan, ―Multimedia Data Mining and Knowledge Discovery‖, Springer, 2007 pp. 3- 17.

VandanaKorde, C NamrataMahender, March 2012, "Text Classification And Classifiers: A Survey", International Journal of Artificial Intelligence & Applications(IJAIA),Vol 3, No 2,.

Venkatadari M., Dr.Lokanataha C. Reddy, ―A Review on Data Mining from Past to Future‖, International Journal of Computer Applications, pp.19-22, vol. 15, No. 7, Feb 2011.

Volkwein, J. F. &. L. W. G., 1996. Characteristics of extenders: Full-time students who take light credit loads and graduate in more than four years. Research in Higher Education, 37(1), pp. 43-68.

Vlachos, M., Lin, J., Keogh, E., and Gunopulos, D. 2003. A wavelet-based anytime algorithm for k-means clustering of time series. In Proceedings of the Workshop on Clustering High Dimensionality Data and Its Applications. 23–30.

Vermunt, J.D., 2005. Relations between student learning patterns and personal and contextual factors and academic performance. Higher education, 49(3), pp.205-234.

Verma, H., Agrawal, R.K. and Sharan, A., 2016. An improved intuitionistic fuzzy c-means clustering algorithm incorporating local information for brain image segmentation. Applied Soft Computing, 46, pp.543-557.

Vert, G., Chennamaneni, A. & Iyengar, S. S., 2010. Potential Application of Contextual Information Processing To Data Mining. Las Vegas Nevada, USA., Proceedings of the 2010 International Conference on Information & Knowledge Engineering, IKE 2010, July 12-15, 2010.

Vialardi, . C., Chue, J., Peche, J. P. & Alvarado, G., 2011. A data mining approach to guide students through the enrollment process based on academic performance. User Modeling and User - Adaptation Interaction, Volume 21, pp. 217-248.

Vialardi-Sacin, C., Bravo-Agapito, J., Shafti, L., & Ortigosa, A. (2009). Recommendation in higher education using data mining techniques. In Proceedings of the 2nd international conference on educational data mining (pp. 190–199).

Voorhees, R.A. and Cooper, J.D., 2014. Opportunities and barriers to effective planning in higher education. In Using Data to Improve Higher Education (pp. 25-38). SensePublishers.

Wang, Y.,Tseng, M.,Liao, H.(2009), "Data mining for adaptive learning sequence in English language instruction", Expert Syst. Appl. J. Vol.36, pp.7681–7686.

Weise, M. and Christensen, C., 2014. Hire education. Christensen Institute for Disruptive Innovation. http://www. christenseninstitute. org/wpcontent/uploads/2014/07/Hire-Education. pdf.

Wellman, J. V. (2008, November/December). The higher education funding disconnect: Spending more, getting less. Change. Retrieved from http://www. changemag.org/A rchives/Back%20Issues/November-December%20 2008/full-funding-disconnect.html.

Wellman, J. V. (2010a). Connecting the dots between learning and resources (Occasional Paper No. 3). Champaign, IL: University of Illinois at UrbanaChampaign, National Institute for Learning Outcomes Assessment. Retrieved from http://www.learningoutcomesassessment.org/documents/ wellman.pdf.

Wellman, J. (2010b). Making it real: Incorporating cost management and productivity improvements into financing decisions. New England Journal of Higher Education, 24(3), 30–32.

Wellman, J. (2011). Financial characteristics of broad access public institutions (Paper prepared for The Changing Ecology of Higher Education Project). Stanford, CA: Center for Education Policy Analysis, Stanford University. Retrieved from https://cepa.stanford.edu/conference-papers/financialcharacteristics-broad-access-public-institutions.

Witten, I.H., Frank, E., Hall, M.A. and Pal, C.J., 2016. Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann.

Xuan Zhang and Gagan Agrawal. A Tool for Supporting Integration Across Multiple Flat-File Datasets. In Proceedings of the Conference on Bioinformatics and Bioengineering (BIBE). IEEE Computer Society, October 2006.

Xu, L., Jiang, C., Wang, J., Yuan, J. and Ren, Y., 2014. Information security in big data: privacy and data mining. IEEE Access, 2, pp.1149-1176.

Xu, D., Jaggars, S.S. and Fletcher, J., 2016. How and Why Does Two-Year College Entry Influence Baccalaureate Aspirants' Academic and Labor Market Outcomes? A CAPSEE Working Paper. Center for Analysis of Postsecondary Education and Employment.

Yang, J., Wang, W., and Yu, P. S. 2001. Infominer: mining surprising periodic patterns. In Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining. ACM Press, 395–400.

You, J. Liu, L. Li, and K.H. Cheung.(2002), On data mining and warehousing for multimedia information retrievalǁ, From Proceeding (365) Artificial and Computational Intelligence – 2002.

Yu, P. S. and Chen, Eds. IEEE Computer Society Press, Taipei, Taiwan, 3–14.

Zeidenberg, M. and Scott, M.,(2011). The content of their coursework: Understanding course taking patterns at community colleges by clustering studenttranscript.

Zhai, M. and Skerl, J., (2000). The impact of remedial English courses on student college-level coursework performance and persistence. In Bridges to the future: Building linkages for institutional research: Proceedings of the 27 th Annual Conference North East Association for Institutional Research (pp. 233-244).

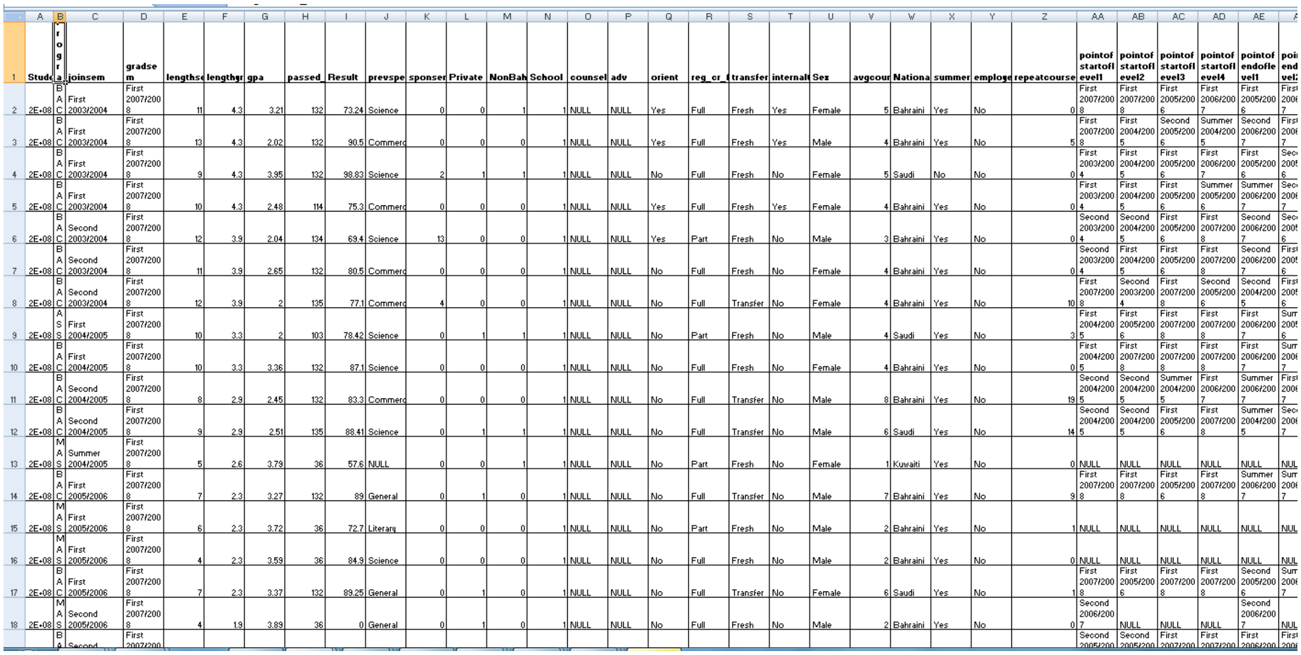
Zhang, L., and Liu, X. (2008), “Personalized instructing recommendation system based on web mining” in Proc. Int. Conf. Young Comput. Sci. Hunan, China, pp.2517-2521.

Zhang, C., Yu, P. . S. & Bell, D., (2010). Introduction to the Domain-Driven Data Mining Special Section. IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, 22(6).

Zhong, S., Khoshgoftaar, T., and Seliya, N. (2007). Clustering-Based network intrusion detection. Int. J. Reliab. Qual. Safety Engin. 14, 2, 169–187.

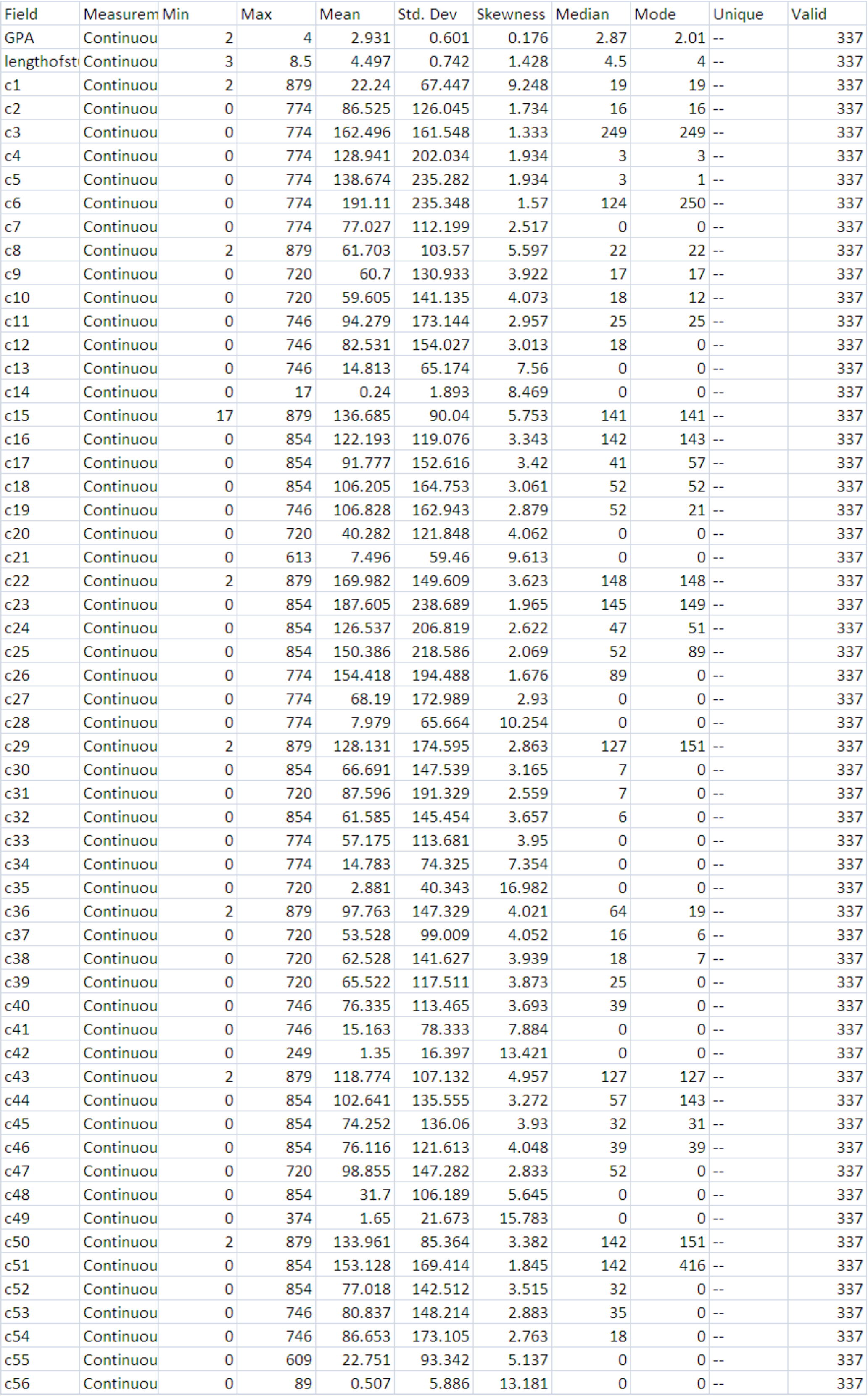
Appendix

Appendix 1 Data Collection Report



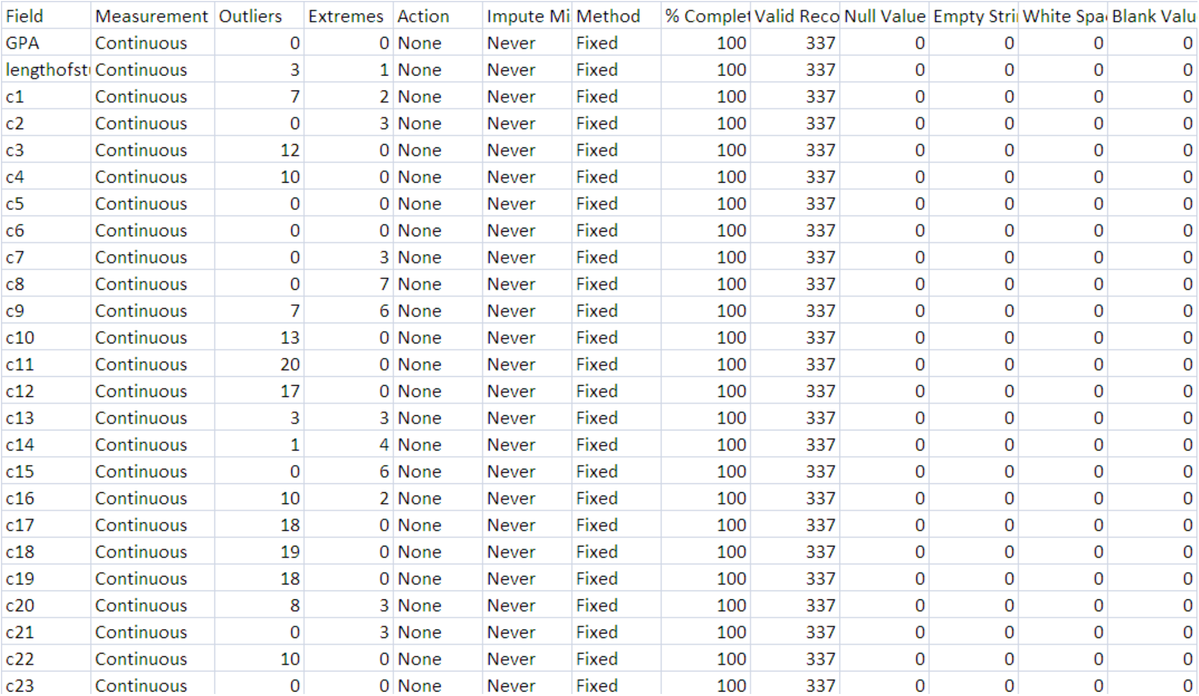
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD	AE	AF
1	**Stude**	**rogra**	**joinsem**	**gradsem**	**lengthse**	**lengthyr**	**gpa**	**passed_**	**Result**	**prevspe**	**sponser**	**Private**	**NonBah**	**School**	**counsel**	**adv**	**orient**	**reg_cr_**	**transfer**	**internalt**	**Sex**	**avgcour**	**Nationa**	**summer**	**employe**	**repeatcourse**	**pointofstartoflevel1**	**pointofstartoflevel2**	**pointofstartoflevel3**	**pointofstartoflevel4**	**pointofendoflevel1**	**poi end vel2**
2	2E+08	BAC	First 2003/2004	First 2007/2008	11	4.3	3.21	132	73.24	Science	0	0	1	1	NULL	NULL	Yes	Full	Fresh	Yes	Female	5	Bahraini	Yes	No	0	First 2007/2008	First 2007/2008	First 2005/2006	First 2006/2007	First 2005/2006	First 2006 7
3	2E+08	BAC	First 2003/2004	First 2007/2008	13	4.3	2.02	132	90.5	Commerc	0	0	0	1	NULL	NULL	Yes	Full	Fresh	Yes	Male	4	Bahraini	Yes	No	5	First 2007/2008	First 2004/2005	Second 2005/2006	Summer 2004/2005	Second 2006/2007	First 2006 7
4	2E+08	BAC	First 2003/2004	First 2007/2008	9	4.3	3.95	132	98.83	Science	2	1	1	1	NULL	NULL	No	Full	Fresh	No	Female	5	Saudi	No	No	0	First 2003/2004	First 2004/2005	First 2005/2006	First 2006/2007	First 2005/2006	Seco 2005 6
5	2E+08	BAC	First 2003/2004	First 2007/2008	10	4.3	2.48	114	75.3	Commerc	0	0	0	1	NULL	NULL	Yes	Full	Fresh	Yes	Female	4	Bahraini	Yes	No	0	First 2003/2004	First 2004/2005	First 2005/2006	Summer 2005/2006	Summer 2006/2007	Seco 2006 7
6	2E+08	BAC	Second 2003/2004	First 2007/2008	12	3.9	2.04	134	69.4	Science	13	0	0	1	NULL	NULL	Yes	Part	Fresh	No	Male	3	Bahraini	Yes	No	0	Second 2003/2004	Second 2004/2005	First 2005/2006	First 2007/2008	Second 2006/2007	Seco 2005 6
7	2E+08	BAC	Second 2003/2004	First 2007/2008	11	3.9	2.65	132	80.5	Commerc	0	0	0	1	NULL	NULL	No	Full	Fresh	No	Female	4	Bahraini	Yes	No	0	Second 2003/2004	First 2004/2005	First 2005/2006	First 2007/2008	Second 2006/2007	First 2005 6
8	2E+08	BAC	Second 2003/2004	First 2007/2008	12	3.9	2	135	77.1	Commerc	4	0	0	1	NULL	NULL	No	Full	Transfer	No	Female	4	Bahraini	Yes	No	10	First 2007/2008	Second 2003/2004	First 2007/2008	Second 2005/2006	Second 2004/2005	First 2005 6
9	2E+08	ASS	First 2004/2005	First 2007/2008	10	3.3	2	103	78.42	Science	0	1	1	1	NULL	NULL	No	Part	Fresh	No	Male	4	Saudi	Yes	No	3	First 2004/2005	First 2005/2006	First 2007/2008	First 2007/2008	First 2006/2007	Sum 2005 6
10	2E+08	BAC	First 2004/2005	First 2007/2008	10	3.3	3.36	132	87.1	Science	0	0	0	1	NULL	NULL	No	Full	Fresh	No	Female	4	Bahraini	Yes	No	0	First 2004/2005	First 2007/2008	First 2007/2008	First 2007/2008	First 2006/2007	Sum 2006 7
11	2E+08	BAC	Second 2004/2005	First 2007/2008	8	2.9	2.45	132	83.3	Commerc	0	0	0	1	NULL	NULL	No	Full	Transfer	No	Male	8	Bahraini	Yes	No	19	Second 2004/2005	Second 2004/2005	Summer 2004/2005	First 2006/2007	Summer 2006/2007	First 2006 7
12	2E+08	BAC	Second 2004/2005	First 2007/2008	9	2.9	2.51	135	88.41	Science	0	1	1	1	NULL	NULL	No	Full	Transfer	No	Male	6	Saudi	Yes	No	14	Second 2004/2005	Second 2004/2005	First 2005/2006	First 2007/2008	Summer 2004/2005	Seco 2006 7
13	2E+08	MAS	Summer 2004/2005	First 2007/2008	5	2.6	3.79	36	57.6	NULL	0	0	1	1	NULL	NULL	No	Part	Fresh	No	Female	1	Kuwaiti	Yes	No	0	NULL	NULL	NULL	NULL	NULL	NUL
14	2E+08	BAC	First 2005/2006	First 2007/2008	7	2.3	3.27	132	89	General	0	1	0	1	NULL	NULL	No	Full	Transfer	No	Male	7	Bahraini	Yes	No	9	First 2007/2008	First 2007/2008	First 2005/2006	First 2007/2008	Summer 2006/2007	Sum 2006 7
15	2E+08	MAS	First 2005/2006	First 2007/2008	6	2.3	3.72	36	72.7	Literary	0	0	0	1	NULL	NULL	No	Part	Fresh	No	Male	2	Bahraini	Yes	No	1	NULL	NULL	NULL	NULL	NULL	NUL
16	2E+08	MAS	First 2005/2006	First 2007/2008	4	2.3	3.59	36	84.9	Science	0	0	0	1	NULL	NULL	No	Full	Fresh	No	Male	2	Bahraini	Yes	No	0	NULL	NULL	NULL	NULL	NULL	NUL
17	2E+08	BAC	First 2005/2006	First 2007/2008	7	2.3	3.37	132	89.25	General	0	1	0	1	NULL	NULL	No	Full	Transfer	No	Female	6	Saudi	Yes	No	1	First 2007/2008	First 2005/2006	First 2007/2008	First 2007/2008	Second 2005/2006	Sum 2006 7
18	2E+08	MAS	Second 2005/2006	First 2007/2008	4	1.9	3.89	36	0	General	0	1	0	1	NULL	NULL	No	Full	Fresh	No	Male	2	Bahraini	Yes	No	0	Second 2006/2007	NULL	NULL	NULL	Second 2006/2007	NUL
		BA	Second	First 2007/200																							Second 2005/200	Second 2005/200	First 2007/200	First 2007/200	First 2006/200	First 200

Appendix 2 Data Description Report



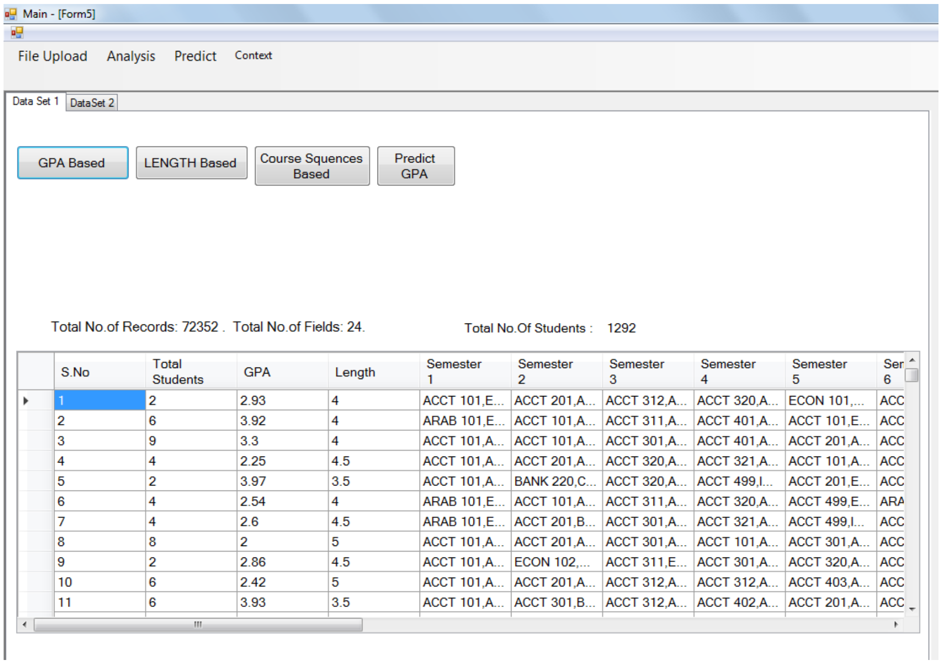
Field	Measurem	Min	Max	Mean	Std. Dev	Skewness	Median	Mode	Unique	Valid
GPA	Continuou	2	4	2.931	0.601	0.176	2.87	2.01	--	337
lengthofst	Continuou	3	8.5	4.497	0.742	1.428	4.5	4	--	337
c1	Continuou	2	879	22.24	67.447	9.248	19	19	--	337
c2	Continuou	0	774	86.525	126.045	1.734	16	16	--	337
c3	Continuou	0	774	162.496	161.548	1.333	249	249	--	337
c4	Continuou	0	774	128.941	202.034	1.934	3	3	--	337
c5	Continuou	0	774	138.674	235.282	1.934	3	1	--	337
c6	Continuou	0	774	191.11	235.348	1.57	124	250	--	337
c7	Continuou	0	774	77.027	112.199	2.517	0	0	--	337
c8	Continuou	2	879	61.703	103.57	5.597	22	22	--	337
c9	Continuou	0	720	60.7	130.933	3.922	17	17	--	337
c10	Continuou	0	720	59.605	141.135	4.073	18	12	--	337
c11	Continuou	0	746	94.279	173.144	2.957	25	25	--	337
c12	Continuou	0	746	82.531	154.027	3.013	18	0	--	337
c13	Continuou	0	746	14.813	65.174	7.56	0	0	--	337
c14	Continuou	0	17	0.24	1.893	8.469	0	0	--	337
c15	Continuou	17	879	136.685	90.04	5.753	141	141	--	337
c16	Continuou	0	854	122.193	119.076	3.343	142	143	--	337
c17	Continuou	0	854	91.777	152.616	3.42	41	57	--	337
c18	Continuou	0	854	106.205	164.753	3.061	52	52	--	337
c19	Continuou	0	746	106.828	162.943	2.879	52	21	--	337
c20	Continuou	0	720	40.282	121.848	4.062	0	0	--	337
c21	Continuou	0	613	7.496	59.46	9.613	0	0	--	337
c22	Continuou	2	879	169.982	149.609	3.623	148	148	--	337
c23	Continuou	0	854	187.605	238.689	1.965	145	149	--	337
c24	Continuou	0	854	126.537	206.819	2.622	47	51	--	337
c25	Continuou	0	854	150.386	218.586	2.069	52	89	--	337
c26	Continuou	0	774	154.418	194.488	1.676	89	0	--	337
c27	Continuou	0	774	68.19	172.989	2.93	0	0	--	337
c28	Continuou	0	774	7.979	65.664	10.254	0	0	--	337
c29	Continuou	2	879	128.131	174.595	2.863	127	151	--	337
c30	Continuou	0	854	66.691	147.539	3.165	7	0	--	337
c31	Continuou	0	720	87.596	191.329	2.559	7	0	--	337
c32	Continuou	0	854	61.585	145.454	3.657	6	0	--	337
c33	Continuou	0	774	57.175	113.681	3.95	0	0	--	337
c34	Continuou	0	774	14.783	74.325	7.354	0	0	--	337
c35	Continuou	0	720	2.881	40.343	16.982	0	0	--	337
c36	Continuou	2	879	97.763	147.329	4.021	64	19	--	337
c37	Continuou	0	720	53.528	99.009	4.052	16	6	--	337
c38	Continuou	0	720	62.528	141.627	3.939	18	7	--	337
c39	Continuou	0	720	65.522	117.511	3.873	25	0	--	337
c40	Continuou	0	746	76.335	113.465	3.693	39	0	--	337
c41	Continuou	0	746	15.163	78.333	7.884	0	0	--	337
c42	Continuou	0	249	1.35	16.397	13.421	0	0	--	337
c43	Continuou	2	879	118.774	107.132	4.957	127	127	--	337
c44	Continuou	0	854	102.641	135.555	3.272	57	143	--	337
c45	Continuou	0	854	74.252	136.06	3.93	32	31	--	337
c46	Continuou	0	854	76.116	121.613	4.048	39	39	--	337
c47	Continuou	0	720	98.855	147.282	2.833	52	0	--	337
c48	Continuou	0	854	31.7	106.189	5.645	0	0	--	337
c49	Continuou	0	374	1.65	21.673	15.783	0	0	--	337
c50	Continuou	2	879	133.961	85.364	3.382	142	151	--	337
c51	Continuou	0	854	153.128	169.414	1.845	142	416	--	337
c52	Continuou	0	854	77.018	142.512	3.515	32	0	--	337
c53	Continuou	0	746	80.837	148.214	2.883	35	0	--	337
c54	Continuou	0	746	86.653	173.105	2.763	18	0	--	337
c55	Continuou	0	609	22.751	93.342	5.137	0	0	--	337
c56	Continuou	0	89	0.507	5.886	13.181	0	0	--	337

Appendix 3 Data Quality Report



Field	Measurement	Outliers	Extremes	Action	Impute Mi	Method	% Complet	Valid Reco	Null Value	Empty Stri	White Spa	Blank Valu
GPA	Continuous	0	0	None	Never	Fixed	100	337	0	0	0	0
lengthofst	Continuous	3	1	None	Never	Fixed	100	337	0	0	0	0
c1	Continuous	7	2	None	Never	Fixed	100	337	0	0	0	0
c2	Continuous	0	3	None	Never	Fixed	100	337	0	0	0	0
c3	Continuous	12	0	None	Never	Fixed	100	337	0	0	0	0
c4	Continuous	10	0	None	Never	Fixed	100	337	0	0	0	0
c5	Continuous	0	0	None	Never	Fixed	100	337	0	0	0	0
c6	Continuous	0	0	None	Never	Fixed	100	337	0	0	0	0
c7	Continuous	0	3	None	Never	Fixed	100	337	0	0	0	0
c8	Continuous	0	7	None	Never	Fixed	100	337	0	0	0	0
c9	Continuous	7	6	None	Never	Fixed	100	337	0	0	0	0
c10	Continuous	13	0	None	Never	Fixed	100	337	0	0	0	0
c11	Continuous	20	0	None	Never	Fixed	100	337	0	0	0	0
c12	Continuous	17	0	None	Never	Fixed	100	337	0	0	0	0
c13	Continuous	3	3	None	Never	Fixed	100	337	0	0	0	0
c14	Continuous	1	4	None	Never	Fixed	100	337	0	0	0	0
c15	Continuous	0	6	None	Never	Fixed	100	337	0	0	0	0
c16	Continuous	10	2	None	Never	Fixed	100	337	0	0	0	0
c17	Continuous	18	0	None	Never	Fixed	100	337	0	0	0	0
c18	Continuous	19	0	None	Never	Fixed	100	337	0	0	0	0
c19	Continuous	18	0	None	Never	Fixed	100	337	0	0	0	0
c20	Continuous	8	3	None	Never	Fixed	100	337	0	0	0	0
c21	Continuous	0	3	None	Never	Fixed	100	337	0	0	0	0
c22	Continuous	10	0	None	Never	Fixed	100	337	0	0	0	0
c23	Continuous	0	0	None	Never	Fixed	100	337	0	0	0	0

Appendix 4 CGPA based course taking pattern

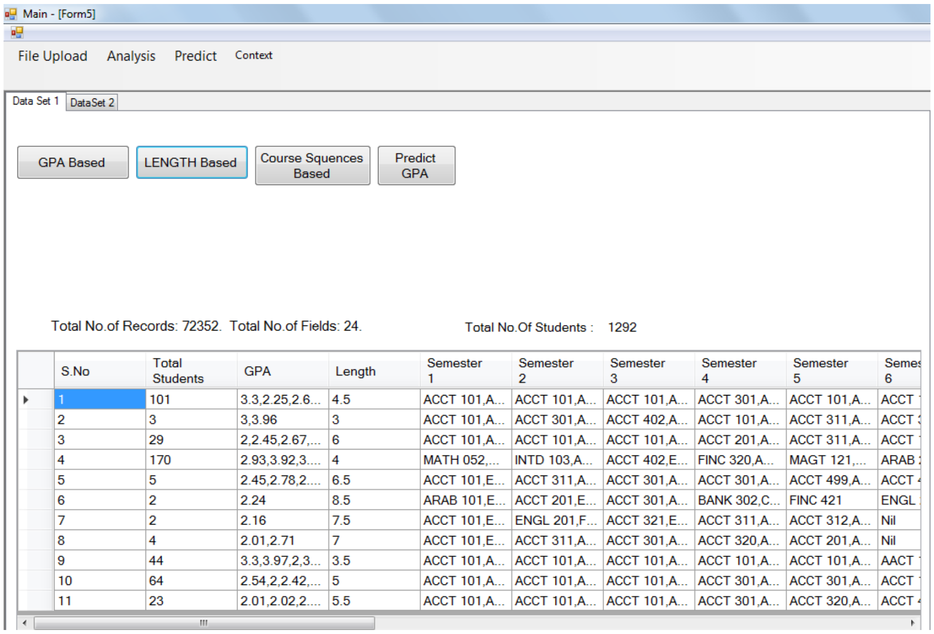


Total No.of Records: 72352 . Total No.of Fields: 24.

Total No.Of Students : 1292

S.No	Total Students	GPA	Length	Semester 1	Semester 2	Semester 3	Semester 4	Semester 5	Ser 6
1	2	2.93	4	ACCT 101,E...	ACCT 201,A...	ACCT 312,A...	ACCT 320,A...	ECON 101,...	ACC
2	6	3.92	4	ARAB 101,E...	ACCT 101,A...	ACCT 311,A...	ACCT 401,A...	ACCT 101,E...	ACC
3	9	3.3	4	ACCT 101,A...	ACCT 101,A...	ACCT 301,A...	ACCT 401,A...	ACCT 201,A...	ACC
4	4	2.25	4.5	ACCT 101,A...	ACCT 201,A...	ACCT 320,A...	ACCT 321,A...	ACCT 101,A...	ACC
5	2	3.97	3.5	ACCT 101,A...	BANK 220,C...	ACCT 320,A...	ACCT 499,I...	ACCT 201,E...	ACC
6	4	2.54	4	ARAB 101,E...	ACCT 101,A...	ACCT 311,A...	ACCT 320,A...	ACCT 499,E...	ARA
7	4	2.6	4.5	ARAB 101,E...	ACCT 201,B...	ACCT 301,A...	ACCT 321,A...	ACCT 499,I...	ACC
8	8	2	5	ACCT 101,A...	ACCT 201,A...	ACCT 301,A...	ACCT 101,A...	ACCT 301,A...	ACC
9	2	2.86	4.5	ACCT 101,A...	ECON 102,...	ACCT 311,E...	ACCT 301,A...	ACCT 320,A...	ACC
10	6	2.42	5	ACCT 101,A...	ACCT 201,A...	ACCT 312,A...	ACCT 312,A...	ACCT 403,A...	ACC
11	6	3.93	3.5	ACCT 101,A...	ACCT 301,B...	ACCT 312,A...	ACCT 402,A...	ACCT 201,A...	ACC

Appendix 5 Time-to-degree based course taking pattern

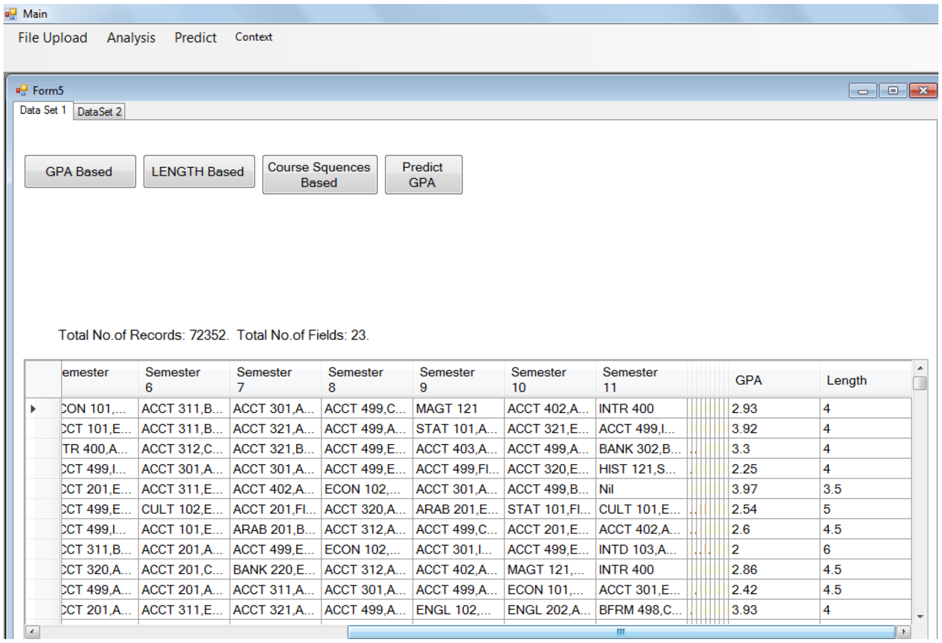


Total No.of Records: 72352. Total No.of Fields: 24.

Total No.Of Students : 1292

S.No	Total Students	GPA	Length	Semester 1	Semester 2	Semester 3	Semester 4	Semester 5	Semes 6
1	101	3.3,2.25,2.6...	4.5	ACCT 101,A...	ACCT 101,A...	ACCT 101,A...	ACCT 301,A...	ACCT 101,A...	ACCT
2	3	3,3.96	3	ACCT 101,A...	ACCT 301,A...	ACCT 402,A...	ACCT 101,A...	ACCT 311,A...	ACCT
3	29	2,2.45,2.67,...	6	ACCT 101,A...	ACCT 101,A...	ACCT 101,A...	ACCT 201,A...	ACCT 311,A...	ACCT
4	170	2.93,3.92,3....	4	MATH 052,...	INTD 103,A...	ACCT 402,E...	FINC 320,A...	MAGT 121,...	ARAB
5	5	2.45,2.78,2....	6.5	ACCT 101,E...	ACCT 311,A...	ACCT 301,A...	ACCT 301,A...	ACCT 499,A...	ACCT
6	2	2.24	8.5	ARAB 101,E...	ACCT 201,E...	ACCT 301,A...	BANK 302,C...	FINC 421	ENGL
7	2	2.16	7.5	ACCT 101,E...	ENGL 201,F...	ACCT 321,E...	ACCT 311,A...	ACCT 312,A...	Nil
8	4	2.01,2.71	7	ACCT 101,E...	ACCT 311,A...	ACCT 301,A...	ACCT 320,A...	ACCT 201,A...	Nil
9	44	3.3,3.97,2,3...	3.5	ACCT 101,A...	ACCT 101,A...	ACCT 101,A...	ACCT 101,A...	ACCT 101,A...	AACT
10	64	2.54,2,2.42,...	5	ACCT 101,A...	ACCT 101,A...	ACCT 101,A...	ACCT 301,A...	ACCT 301,A...	ACCT
11	23	2.01,2.02,2....	5.5	ACCT 101,A...	ACCT 101,A...	ACCT 101,A...	ACCT 301,A...	ACCT 320,A...	ACCT

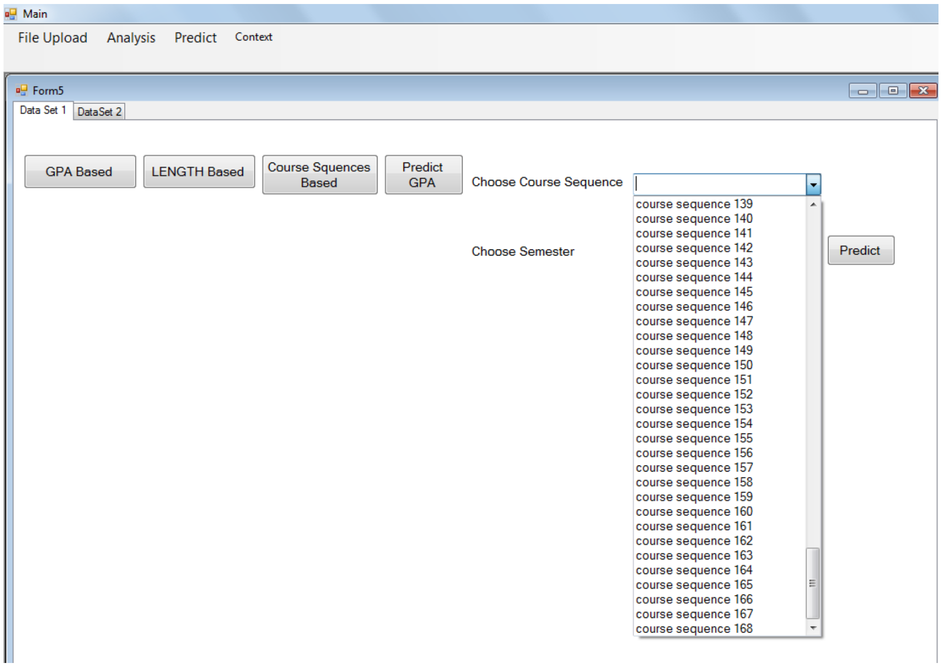
Appendix 6 Course sequences for various CGPA and time-to-degree

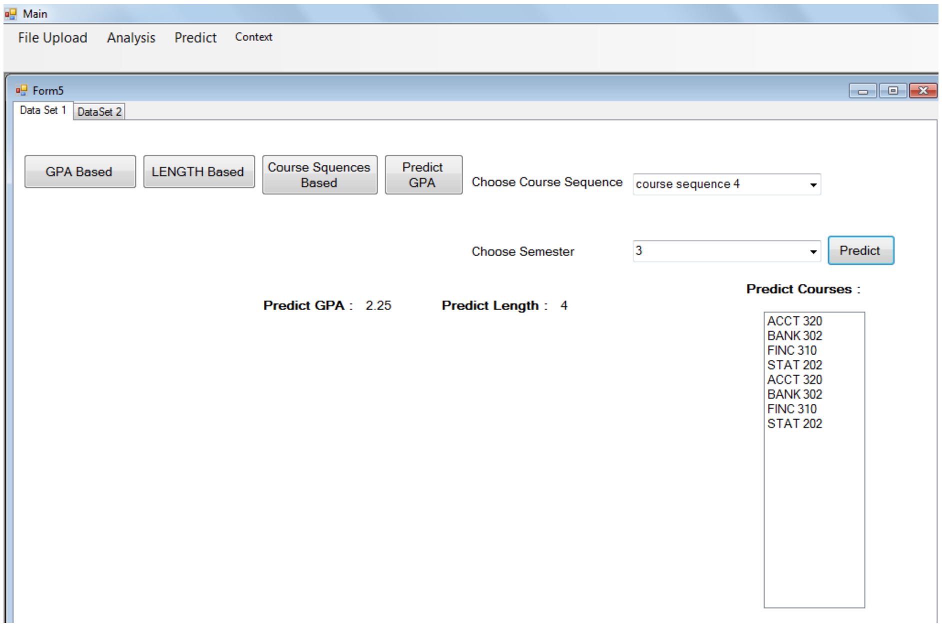


Total No.of Records: 72352. Total No.of Fields: 23.

emester	Semester 6	Semester 7	Semester 8	Semester 9	Semester 10	Semester 11	GPA	Length
:ON 101,...	ACCT 311,B...	ACCT 301,A...	ACCT 499,C...	MAGT 121	ACCT 402,A...	INTR 400	2.93	4
:CT 101,E...	ACCT 311,B...	ACCT 321,A...	ACCT 499,A...	STAT 101,A...	ACCT 321,E...	ACCT 499,I...	3.92	4
TR 400,A...	ACCT 312,C...	ACCT 321,B...	ACCT 499,E...	ACCT 403,A...	ACCT 499,A...	BANK 302,B...	3.3	4
:CT 499,I...	ACCT 301,A...	ACCT 301,A...	ACCT 499,E...	ACCT 499,FI...	ACCT 320,E...	HIST 121,S...	2.25	4
:CT 201,E...	ACCT 311,E...	ACCT 402,A...	ECON 102,...	ACCT 301,A...	ACCT 499,B...	Nil	3.97	3.5
:CT 499,E...	CULT 102,E...	ACCT 201,FI...	ACCT 320,A...	ARAB 201,E...	STAT 101,FI...	CULT 101,E...	2.54	5
:CT 499,I...	ACCT 101,E...	ARAB 201,B...	ACCT 312,A...	ACCT 499,C...	ACCT 201,E...	ACCT 402,A...	2.6	4.5
:CT 311,B...	ACCT 201,A...	ACCT 499,E...	ECON 102,...	ACCT 301,I...	ACCT 499,E...	INTD 103,A...	2	6
:CT 320,A...	ACCT 201,C...	BANK 220,E...	ACCT 312,A...	ACCT 402,A...	MAGT 121,...	INTR 400	2.86	4.5
:CT 499,A...	ACCT 201,A...	ACCT 311,A...	ACCT 301,A...	ACCT 499,A...	ECON 101,...	ACCT 301,E...	2.42	4.5
:CT 201,A...	ACCT 311,E...	ACCT 321,A...	ACCT 499,A...	ENGL 102,...	ENGL 202,A...	BFRM 498,C...	3.93	4

Appendix 7 Different course sequences generated from Genetic Algorithm





Choose Course Sequence course sequence 4

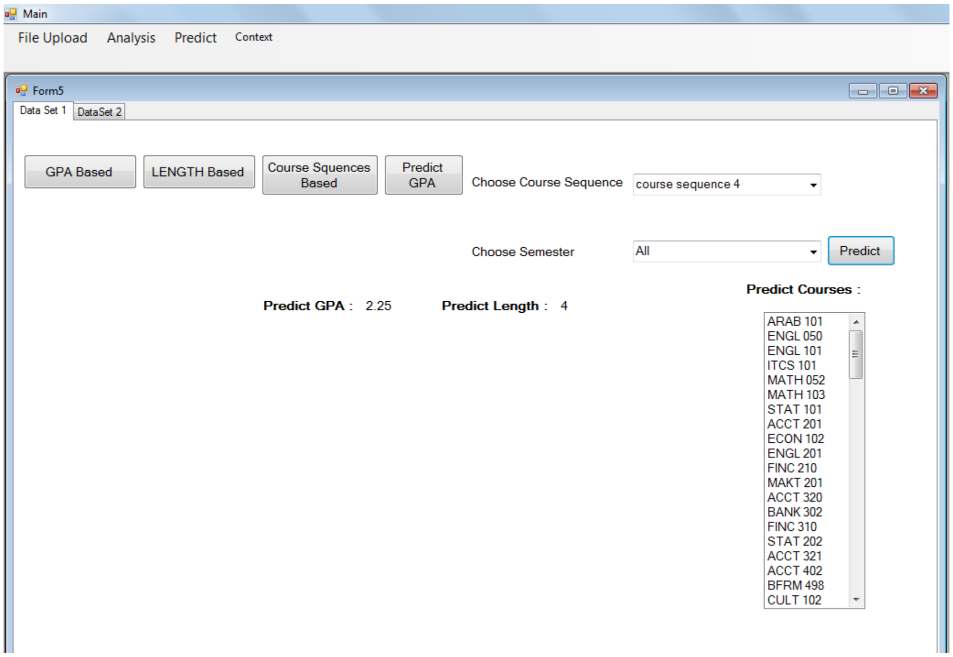
Choose Semester 3

Predict GPA : 2.25 **Predict Length :** 4

Predict Courses :

ACCT 320
BANK 302
FINC 310
STAT 202
ACCT 320
BANK 302
FINC 310
STAT 202

Appendix 8 Predict time-to-degree and CGPA for current students data based on their courses registered



Choose Course Sequence: course sequence 4

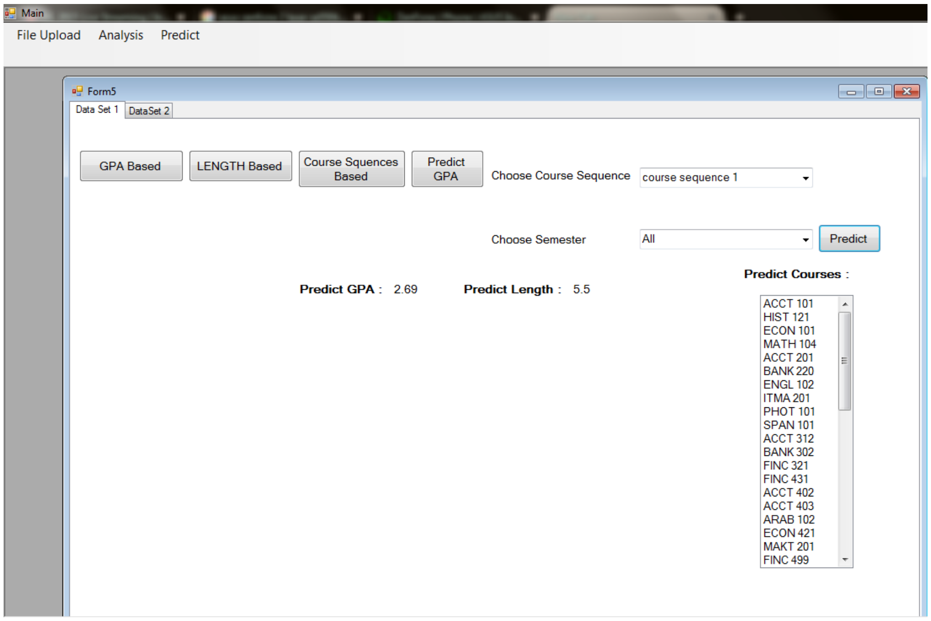
Choose Semester: All

Predict GPA : 2.25

Predict Length : 4

Predict Courses :

ARAB 101
ENGL 050
ENGL 101
ITCS 101
MATH 052
MATH 103
STAT 101
ACCT 201
ECON 102
ENGL 201
FINC 210
MAKT 201
ACCT 320
BANK 302
FINC 310
STAT 202
ACCT 321
ACCT 402
BFRM 498
CULT 102



Choose Course Sequence: course sequence 1

Choose Semester: All

Predict GPA : 2.69

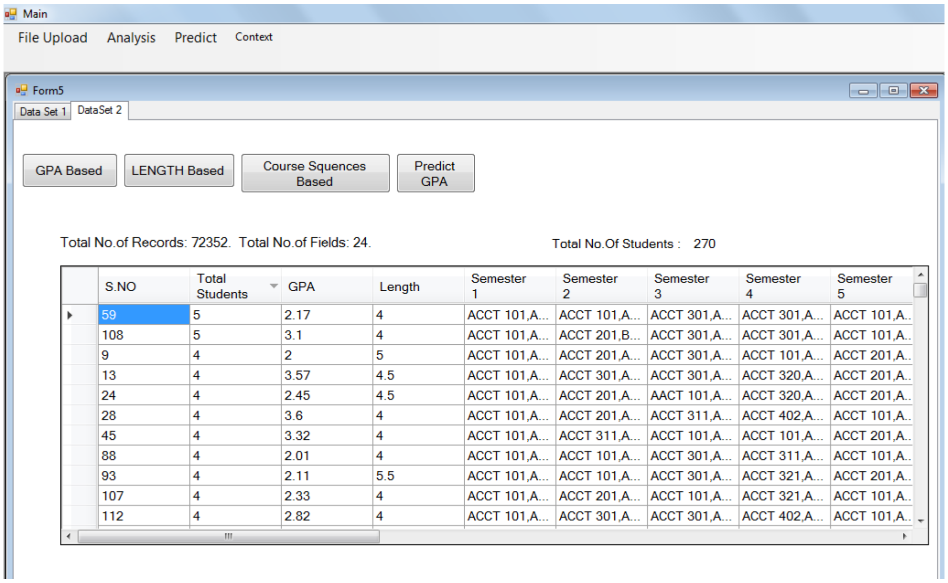
Predict Length : 5.5

Predict Courses :

ACCT 101
HIST 121
ECON 101
MATH 104
ACCT 201
BANK 220
ENGL 102
ITMA 201
PHOT 101
SPAN 101
ACCT 312
BANK 302
FINC 321
FINC 431
ACCT 402
ACCT 403
ARAB 102
ECON 421
MAKT 201
FINC 499

Appendix 9 Prediction of course taking patterns for various CGPA and time-to-degree

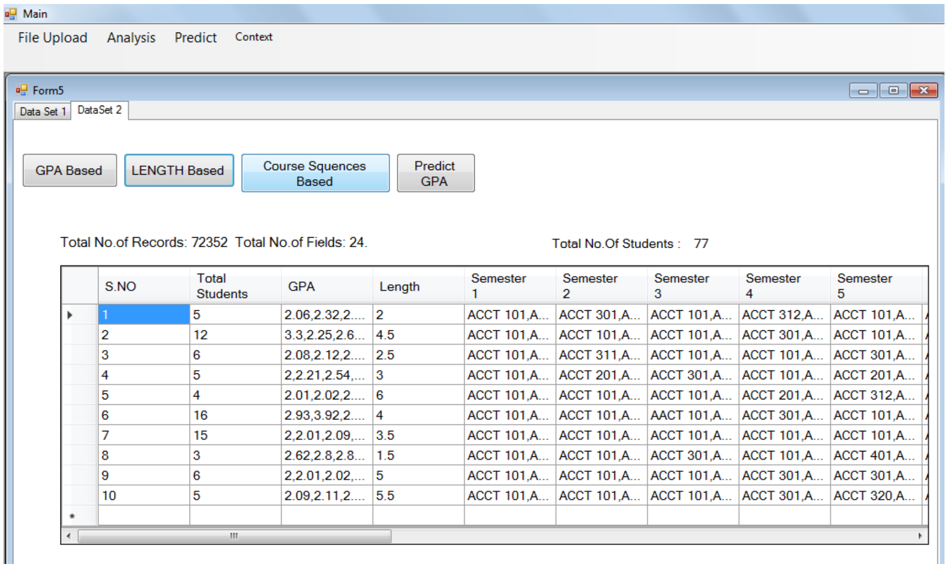
To predict the completion time and GPA of current students with their difficulty data and courses already registered by them.



Total No.of Records: 72352. Total No.of Fields: 24.

Total No.Of Students : 270

S.NO	Total Students	GPA	Length	Semester 1	Semester 2	Semester 3	Semester 4	Semester 5
59	5	2.17	4	ACCT 101,A...	ACCT 101,A...	ACCT 301,A...	ACCT 301,A...	ACCT 101,A..
108	5	3.1	4	ACCT 101,A...	ACCT 201,B...	ACCT 301,A...	ACCT 301,A...	ACCT 101,A..
9	4	2	5	ACCT 101,A...	ACCT 201,A...	ACCT 301,A...	ACCT 101,A...	ACCT 201,A..
13	4	3.57	4.5	ACCT 101,A...	ACCT 301,A...	ACCT 301,A...	ACCT 320,A...	ACCT 201,A..
24	4	2.45	4.5	ACCT 101,A...	ACCT 201,A...	AACT 101,A...	ACCT 320,A...	ACCT 201,A..
28	4	3.6	4	ACCT 101,A...	ACCT 201,A...	ACCT 311,A...	ACCT 402,A...	ACCT 101,A..
45	4	3.32	4	ACCT 101,A...	ACCT 311,A...	ACCT 101,A...	ACCT 101,A...	ACCT 201,A..
88	4	2.01	4	ACCT 101,A...	ACCT 101,A...	ACCT 301,A...	ACCT 311,A...	ACCT 101,A..
93	4	2.11	5.5	ACCT 101,A...	ACCT 101,A...	ACCT 301,A...	ACCT 321,A...	ACCT 201,A..
107	4	2.33	4	ACCT 101,A...	ACCT 201,A...	ACCT 101,A...	ACCT 321,A...	ACCT 101,A..
112	4	2.82	4	ACCT 101,A...	ACCT 301,A...	ACCT 301,A...	ACCT 402,A...	ACCT 101,A..

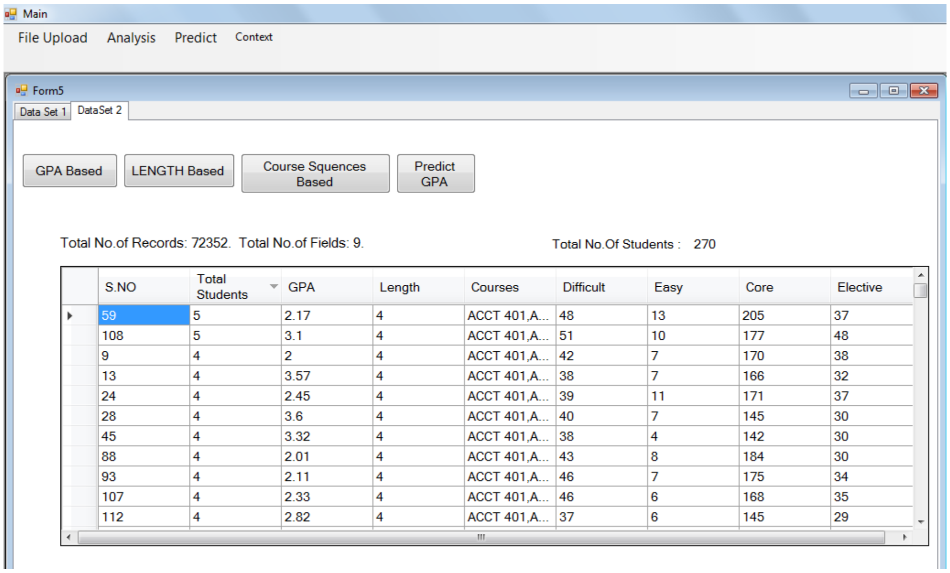


Total No.of Records: 72352 Total No.of Fields: 24.

Total No.Of Students : 77

S.NO	Total Students	GPA	Length	Semester 1	Semester 2	Semester 3	Semester 4	Semester 5
1	5	2.06,2.32,2....	2	ACCT 101,A...	ACCT 301,A...	ACCT 101,A...	ACCT 312,A...	ACCT 101,A...
2	12	3.3,2.25,2.6...	4.5	ACCT 101,A...	ACCT 101,A...	ACCT 101,A...	ACCT 301,A...	ACCT 101,A...
3	6	2.08,2.12,2....	2.5	ACCT 101,A...	ACCT 311,A...	ACCT 101,A...	ACCT 101,A...	ACCT 301,A...
4	5	2,2.21,2.54,...	3	ACCT 101,A...	ACCT 201,A...	ACCT 301,A...	ACCT 101,A...	ACCT 201,A...
5	4	2.01,2.02,2....	6	ACCT 101,A...	ACCT 101,A...	ACCT 101,A...	ACCT 201,A...	ACCT 312,A...
6	16	2.93,3.92,2....	4	ACCT 101,A...	ACCT 101,A...	AACT 101,A...	ACCT 301,A...	ACCT 101,A...
7	15	2,2.01,2.09,...	3.5	ACCT 101,A...	ACCT 101,A...	ACCT 101,A...	ACCT 101,A...	ACCT 101,A...
8	3	2.62,2.8,2.8...	1.5	ACCT 101,A...	ACCT 101,A...	ACCT 301,A...	ACCT 101,A...	ACCT 401,A...
9	6	2,2.01,2.02,...	5	ACCT 101,A...	ACCT 101,A...	ACCT 101,A...	ACCT 301,A...	ACCT 301,A...
10	5	2.09,2.11,2....	5.5	ACCT 101,A...	ACCT 101,A...	ACCT 101,A...	ACCT 301,A...	ACCT 320,A...

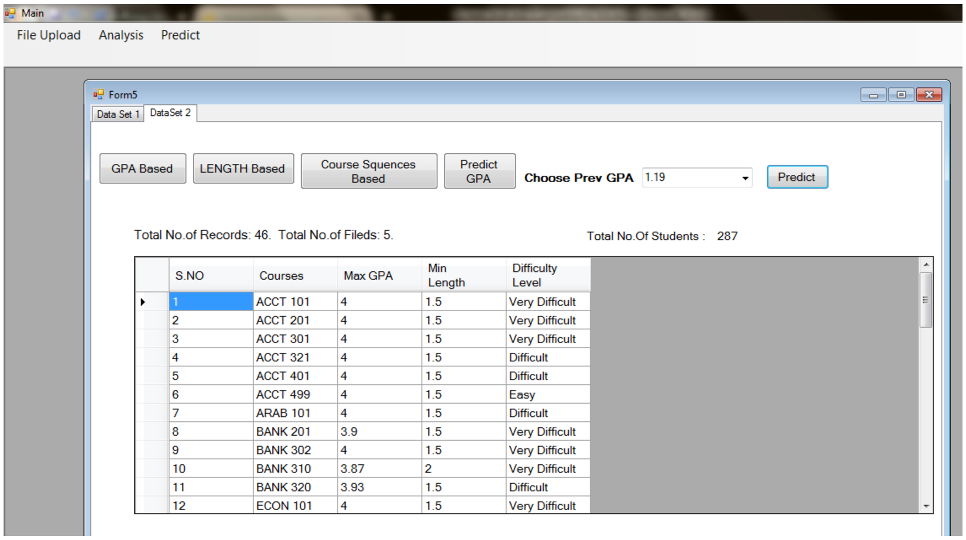
Appendix 10 Prediction of course taking patterns for various Course difficulty, CGPA and time-to-degree



Total No.of Records: 72352. Total No.of Fields: 9.

Total No.Of Students : 270

S.NO	Total Students	GPA	Length	Courses	Difficult	Easy	Core	Elective
59	5	2.17	4	ACCT 401,A...	48	13	205	37
108	5	3.1	4	ACCT 401,A...	51	10	177	48
9	4	2	4	ACCT 401,A...	42	7	170	38
13	4	3.57	4	ACCT 401,A...	38	7	166	32
24	4	2.45	4	ACCT 401,A...	39	11	171	37
28	4	3.6	4	ACCT 401,A...	40	7	145	30
45	4	3.32	4	ACCT 401,A...	38	4	142	30
88	4	2.01	4	ACCT 401,A...	43	8	184	30
93	4	2.11	4	ACCT 401,A...	46	7	175	34
107	4	2.33	4	ACCT 401,A...	46	6	168	35
112	4	2.82	4	ACCT 401,A...	37	6	145	29



Choose Prev GPA 1.19

Total No.of Records: 46. Total No.of Fileds: 5.

Total No.Of Students : 287

S.NO	Courses	Max GPA	Min Length	Difficulty Level
1	ACCT 101	4	1.5	Very Difficult
2	ACCT 201	4	1.5	Very Difficult
3	ACCT 301	4	1.5	Very Difficult
4	ACCT 321	4	1.5	Difficult
5	ACCT 401	4	1.5	Difficult
6	ACCT 499	4	1.5	Easy
7	ARAB 101	4	1.5	Difficult
8	BANK 201	3.9	1.5	Very Difficult
9	BANK 302	4	1.5	Very Difficult
10	BANK 310	3.87	2	Very Difficult
11	BANK 320	3.93	1.5	Difficult
12	ECON 101	4	1.5	Very Difficult

Appendix 11 ADREG System

A Student information system (SIS) is a Management Information System for education establishments to manage student data. Student Information Systems (often abbreviated as SIS systems) provide capabilities for registering students on courses, documenting grading, transcripts and results of student test and other assessment scores, build student schedules, track student attendance, and manage many other student-related data needs in a school. A SIS should not be confused with a Learning Management System (LMS) or Virtual Learning Environment (VLE) where course materials, assignments and assessment tests can be published electronically. The SIS can be considered as an Enterprise Resource Planning (ERP) system.

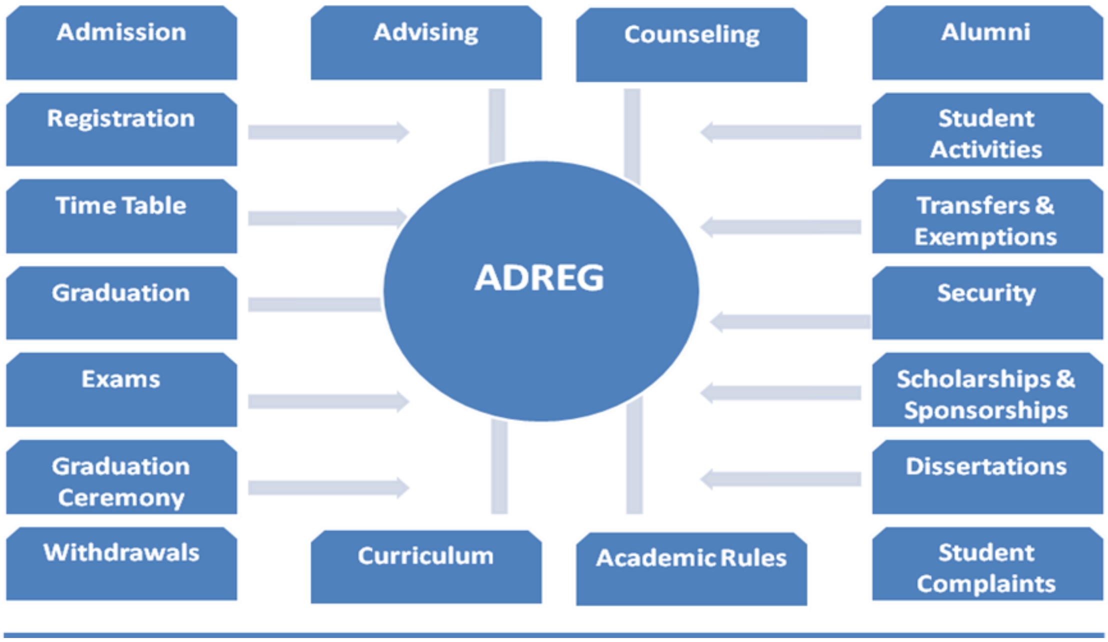
These systems vary in size, scope and capability, from packages that are implemented in relatively small organizations to cover student records alone, to enterprise-wide solutions that aim to cover most aspects of running large multi-campus organizations and their online schools with significant local responsibility. Many systems can be scaled to different levels of functionality by purchasing add-on "modules" and can typically be configured by their home institutions to meet local needs.

SIS automates or simplifies all the processes of a student's lifecycle from application and Financial Aid, to career services and online education.

ADREG (**Ad**mission and **Reg**istration) system is the SIS system of an anonymous University which automates the processes of admitting students, registering , time table of courses, master data like courses, staff, rooms, departments, colleges and course fees. It has modules to store information starting from application process till after students become part of alumni, the complete life cycle of the student in the university. The personal, background of the students, their course enrollments, grades, semester, curriculum and course timetables are recorded in the system. It contains student records from 2003 till 2017. Currently around 11,000 student records and around 200000 course enrollments are recorded.

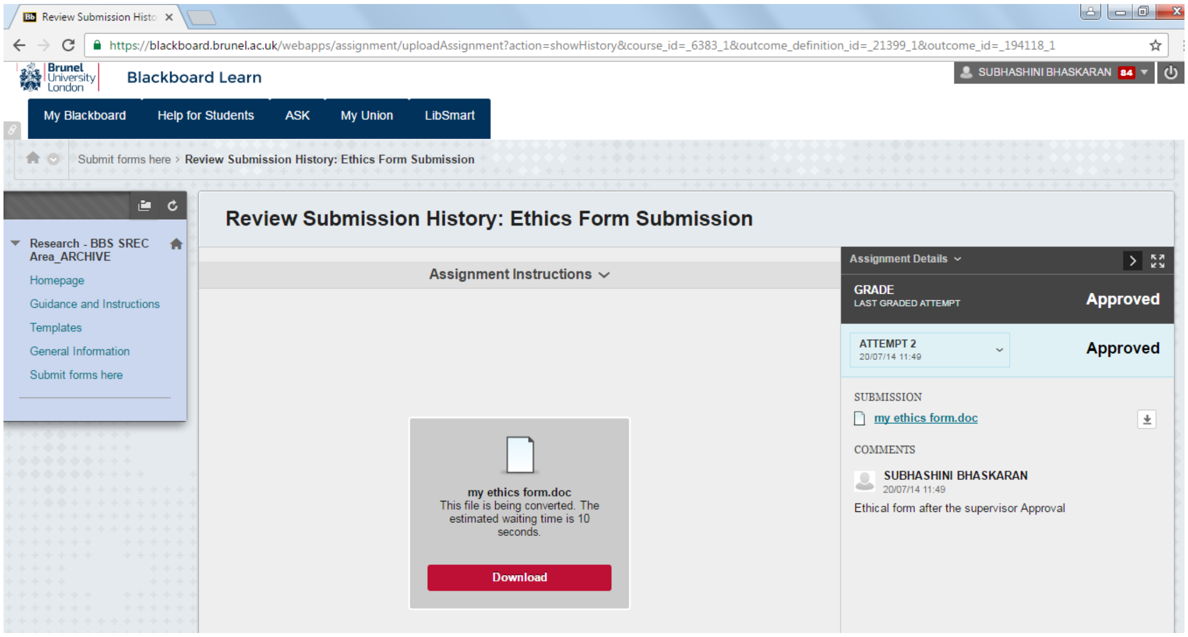
The modules of ADREG system are depicted in the figure below

ADREG System



Appendix 12 Ethical Approval

This research is guided by the university's code of ethics that provides a statement of principles and procedures for the conduct of the research work, highlighting what is and what is not considered ethical. In line with the requirements of the university's code of ethics, the screen shot of the ethical approval is given below.



https://blackboard.brunel.ac.uk/webapps/assignment/uploadAssignment?action=showHistory&course_id=_6383_1&outcome_definition_id=_21399_1&outcome_id=_194118_1

Submit forms here > Review Submission History: Ethics Form Submission

Review Submission History: Ethics Form Submission

my ethics form.doc
This file is being converted. The estimated waiting time is 10 seconds.

GRADE
LAST GRADED ATTEMPT Approved

ATTEMPT 2
20/07/14 11:49 Approved

SUBMISSION
my ethics form.doc

COMMENTS
SUBHASHINI BHASKARAN
20/07/14 11:49
Ethical form after the supervisor Approval

Appendix 13 Sample Student Attributes

Student Personal factors – student date of birth, name, nationality, previous education, previous education mark, previous education specialisation, student degree programme, financial aid, student type fresh or transfer.

Student course enrollment factors – course, lecturer taking the course, attendance.

Course factors- course, number of students who took the course (value of the course), study hours, weightage of the course (core, core elective, free and humanities), level of the course.

Appendix 14 Definitions of select terms

Semester	Every semester consists of around 15-18 weeks of academic work.
Academic Year	One academic year constitute two consecutive semesters.
Programme	An educational programme leading to award of a Degree, diploma or Certificate.
Course	Course is often stated to as subjects or papers, are a part of a programme and need not carry the same weightage. The courses should express the learning objectives and outcomes.
Credit	A unit by which the course work is measured. It determines the number of hours of teaching required per week. One credit is equivalent to one hour of teaching (lecture or tutorial) or two hours of practical work/field work per week.
Letter Grade	It is an index of the performance of students in a said course. Grades are denoted by letters O, A+, A, B+, B,B-,C+,C-, C,D,D+, P and F.
Grade Point	It is a numerical weight allotted to each letter grade on a 10-point scale.
Semester Grade Point Average (SGPA)	It is a measure of performance of work done in a semester. It is ratio of total credit points secured by a student in various courses registered in a semester and the total course credits taken during that semester. It shall be expressed up to two decimal places.
Credit Based Semester System (CBSS)	In the CBSS, the prerequisite for awarding a degree or diploma or certificate is approved in terms of number of credits to be completed by the students. . The credit based semester system provides flexibility in designing curriculum and assigning credits based on the course content and hours of teaching.
Choice Based Credit System (CBCS)	In this system students have the choice to select from the prescribed courses (core, elective or minor or soft skill courses). The choice based credit system facilitates the students to take courses of their choice, learn at their own pace, undergo additional courses and acquire more than the required credits, and adopt an interdisciplinary approach to learning.
Cumulative Grade Point Average (CGPA)	It is a measure of overall cumulative performance of a student over all semesters. The CGPA is the ratio of total credit points secured by a student in various courses in all semesters and the sum of the total credits of all courses in all the semesters. It is expressed up to two decimal places.
Transcript or Grade Card or Certificate	Based on the grades earned, a grade certificate shall be issued to all the registered students after every semester. The grade certificate will display the course details (code, title, number of credits, grade secured) along with SGPA of that semester and CGPA earned till that semester.

Appendix 15 US and UK Grading

US Grading	UK Grading
93-100%: A, Excellent	70% & above: Distinction
85-93%: B, Very good	60-69%: Merit
78-85%: C, Average	50-59%: Pass
70-77%: D, Below average	40-49%: Tolerated Fail
Below 69%: F, Failure	Below 40%: Fail

Appendix 16 Taxonomy of contextual aspects

Meaning of Context	Author and Paper	Field	Remarks	Experimentation or Analysis	Terms
Context, the cumulative history that is derived from data observations about entities (people, places, and things), is a critical component of analytic decision process. Without context, business conclusions might be flawed.	Context-Based Analytics in a Big Data World: Better Decisions, An IBM® Redbooks® Point-of-View publication Sokol, 2013	Big Data	By using context analytics with big data, organizations can derive trends, patterns and relationships from unstructured data and related structured data. These insights can help an organization to make fact-based decisions to anticipate and shape business outcomes. Entities are defined as people, places, things, locations, organizations, and events. Entities are an important focus of big data analytics. Context is defined as a better understanding of how entities relate. Cumulative context is the memory of how entities relate over time.	Nil	Context Analytics, Big data Analytics, Real-time analytics, Deep reflection analytics, predictive analytics
A general approach for context-aware adaptive mining of data streams that aims to dynamically and autonomously adjust data stream mining parameters according to changes in context and situations	Context-Aware Adaptive Data Stream Mining, Haghighia et al. 2009	Mobile devices/PDA, Datamining	The researchers proposed an overall method for context-aware adaptive data mining that includes context-awareness into universal data stream mining and allows real-time examination of data on board mobile devices in a clever and cost-effective manner. They achieved Context-awareness through Fuzzy Situation Inference (FSI) that assimilates fuzzy logic in the CS model, an official context modeling and cognitive approach for assisting pervasive computing environments.	A prototype was tested on the Nokia N95 mobile phone to represent a real-world scenario in the area of mobile healthcare for monitoring patients suffering from blood pressure fluctuations	Context, fuzzy logic, data mining
Context – background information Contextual item set mining	Contextual Item set Mining in DBpedia,	Data mining	The authors exhibit the capacity of contextual item set mining. Contextual item set mining excerpts frequent associations between items bearing in mind the background information. Each resultant item set is specific to a	Algorithm has been proposed and experimented	Contextual frequent pattern (CFP),Linked Open Data,

extracts frequent associations among items considering background information.	Rabatel et al, 2014		particular context and contexts can be related to each other's following the ontological structure. They used contextual mining on DBpedia data and proved the use of contextual information can refine the item sets obtained by the knowledge discovery process.		
Context refers to the background in which the consumer review was given in opinion mining	Mining Context Information from Consumer's Reviews Silvana Aciar	Data mining	The authors have tried to address the most critical issue on opinion mining which is how to extract information that can be understood and utilized by computers from written text by users/consumers in natural language. They employed classification text mining techniques to identify review's sentences containing contextual information to be then processing and incorporated in a recommender system.	Use of text mining tools to obtain classification rules to identify contextual sentences containing contextual information into a review	Contextual Information, Opinion Mining, Text Mining, sentiment analysis
Processes are performed in a particular context, but this context is often neglected during analysis in process mining which is used to discover and analyse business processes based on raw event data	Process Mining Put Into Context Wil M.P. van der Aalst and Schahram Dustdar	Data mining	The authors argue that the contexts in which the events take place should be considered when the processes are analysed. They distinguish four types of contexts like instance context, process context, social context and external context.	Nil	Process mining
Context could consist of any circumstantial factors of the user and domain that may affect the data mining process.	Context-Aware Data Mining Framework for Wireless Medical Application Vajirkar, 2003	Datamining	The authors proposed a context aware framework which considers context factors during datamining. In their framework, Context-aware Process Component deals with context factors for Data mining and Query processor components. In this component, Context factors such as Application Context, User Context, Domain Context, Data Context are defined. In order to identify context factors from the implicit information of the user's	Nil	Context-aware data mining

			query, processes including checking the attributes of the user query; learning about the entities of the query context are performed. They showed 2 scenarios as a proof of their concept.		
Notion of a context with that of the idea that information could be used to characterize a situation and thus could be responded to.	Potential Application of Contextual Information Processing To Data Mining, Vert et al. IKE 2010		The authors have proposed contextual information processing to be as a part of data mining process. No experiments or statistical proofs are given for the same.	Pseudocode and functions have been proposed.	Data Mining

Appendix 17 Big Data

Big data has been varyingly described in the literature. For instance, Manyika et al. (2011) explain big data as that data which is basically too big and moves too quickly leading to processing problems as conventional database systems are not having the capacity to process big data. On the other hand De Mauro et al. (2016, p.122) define big data as "*Big Data is the information asset characterized by such a high volume, velocity and variety to require specific technology and analytical methods for its transformation into value*". Ramirez et al. (2016, p.1) define big data as: "*The term "big data" refers to a confluence of factors, including the nearly ubiquitous collection of consumer data from a variety of sources, the plummeting cost of data storage, and powerful new capabilities to analyse data to draw connections and make inferences and predictions*". There are a host of other definitions found in the literature (Dutcher, 2014). While definitions vary one of the most common ways of describing big data is through its characterization as '5 Vs': Volume, Variety, Veracity, Velocity and Value (Rasi, 2016). Volume indicates the massive datasets of very high volume and variety indicates the breadth of data related to large numbers of individuals and depth of data on each individual. Veracity indicates the uncertainty of the quality of data and its strength or robustness when extracted from different sources. Velocity signifies the speed with which data is accumulated, transmitted and analysed in real time and value of big data is realized through the generation of information and knowledge that enables the discovery of hidden knowledge, gaining new insights, understanding of new associations and increase in efficiencies (Rasi, 2016). Examples of big data are given in Table 2.7.

Table 2.7, Examples of big data

Example of an entity or area where big data is generated	**Data size**	**Authors**	**Remarks**
Individual human genomes	1 x 10^{19} bytes	Gillings et al. (2016)	Due to growth in digital technology at 30-40 % per year leading to doubling of information on Earth during next century!
Digital information	5 x 10^{21} bytes		
All DNA nucleotides	5 x 10^{37} bytes		
Big data drawn from interactions on the web, online commercial transactions, e-government records, social media, mobile phone records, mobile apps, and sensors in objects linked to the to the Internet of Things.	Estimated to reach 44 zeta bytes in 2020	Davies (2016)	The amount of data produced worldwide is doubling every two years.

Table 2.7 shows the magnitude of the problem faced in regards to data collection that is taking place across the world. The challenges of handling such huge data also increase accordingly and literature shows that current methods used for data analysis are not useful to analyse such data (Österreich, 2016). For instance for analysing the US Vaccine Adverse Event Reporting System (VAERS) traditional data analysing techniques may not be useful (Habl et al. 2016) for instance SQL, RDBMS, OLAP and data warehousing. In place of these organisations may have to use modern techniques such as Hadoop, NoSQL, machine learning and cloud (Campbell et al. 2013).

Further, examples of the various sectors where big data is generated and analysed and the expected use are given in Table 2.8.

Table 2.8, Examples of application of big data (*Source*: OECD, 2013)

	Example of sectors where big data is generated	**Use of big data**	**Remarks**
1.	Insurance	Predictive analytics software to identify likely cases of public health insurance fraud before claims are paid	To save costs: Example in one case costs were save to the extent of three times when compared to its first year.
2.	Retail sector	To access huge data flows through its consumer loyalty cards	To provide tailored services to consumers.
3.	Utilities	Data analytics to identify overall consumption patterns	To forecast future demand and adjust prices and production capacities.
4.	Genetic data	To analyse genes of people in healthcare	In one case genetic data from 35,000 people were able to identify a genetic variant related to schizophrenia that would have been difficult or impossible to pick out of smaller samples
5.	Official statistics	Analysis of official statistics	• Application in road traffic data can provide rapid indicators of the level of economic activity. • Internet site price scan provide up-to-date information on inflation. • Social media chatter can serve to estimate levels of consumer confidence on a weekly basis.

Table 2.8 shows that big data is finding application in every sphere of activity a few of which have been highlighted above. This argument is also applicable to education data which is the focus of this research.

While big data has attracted considerable attention during the last decade due to its purported use it is worthwhile to consider its advantages as well as the challenges that exist to its exploitation.

Table 2.9 lists out the advantages and while Table 2.10 lists out challenges facing big data.

Table 2.9, Advantages that could be derived using big data (*Source*: Campbell et al. 2013)

	Advantages
1.	Optimising campaigns by analysing higher volumes of granular data like clickstream and weather data.
2.	Product & Pricing optimisation by analysing large volumes of data to assess likely impact of changes.
3.	Online retailers use Hadoop to recommend products and services based on user profile analysis and behavioural analytics
4.	Big Data technologies optimise the entire customer experience based on insights from analysing data across a variety of channels
5.	Machine Learning algorithms are used to identify and rank individuals with most influence for a particular topic
6.	Advanced text Analytics tools analyse unstructured data from sites such as Twitter and Facebook to determine sentiment
7.	Businesses are analysing terabytes of data from forums associated with hackers to predict and detect financial fraud and identity theft
8.	Financial institutions analyse large volumes of transaction data to determine exposure of financial assets and score potential customers for risk
9.	Companies are analysing the wealth of information collected by business systems and external sources to optimise business processes
10.	Process modelling and simulation allows companies to assess the impact of complex operational changes before they are implemented

Table 2.10, Challenges in using big data (*Source*: Mukherjee & Shaw, 2016; Pusala et al. 2016)

	Challenges
1.	Capture, storage, search, sharing, transfer, analysis, and visualization.
2.	Lack of knowledge on what is enough about big data
3.	Lack of understanding on the benefits of big data
4.	Lack of an appropriate reason to adopt big data
5.	Lack of business support
6.	Poor data quality in current systems
7.	Lack of executive commitment
8.	Cost/financial resources
9.	Transfer-prohibitive volume of Big data leads to accessing data with low latency.
10.	Machine failure is inevitable. Both transient (i.e., fail-recovery) and permanent (i.e., fail-stop) failures can occur during the execution of such applications.
11.	Big data applications either do not replicate the data or do it automatically through a distributed file system (DFS). Without replication, the failure of a server storing the data causes the re-execution of the affected tasks. Although the replication approach provides more fault-tolerance, it is not efficient due to network overhead and increase in the execution time of the job.
12.	Increasing concern over the confidentiality of the Big data in cloud environments.
13.	How to integrate all the data from different sources to maximize the value of data (data heterogeneity) is a challenge.

While Table 2.9 provides the advantages of using big data Table 2.10 provides the challenges being faced by the users and researchers in using big data. Challenges appear to be very daunting.

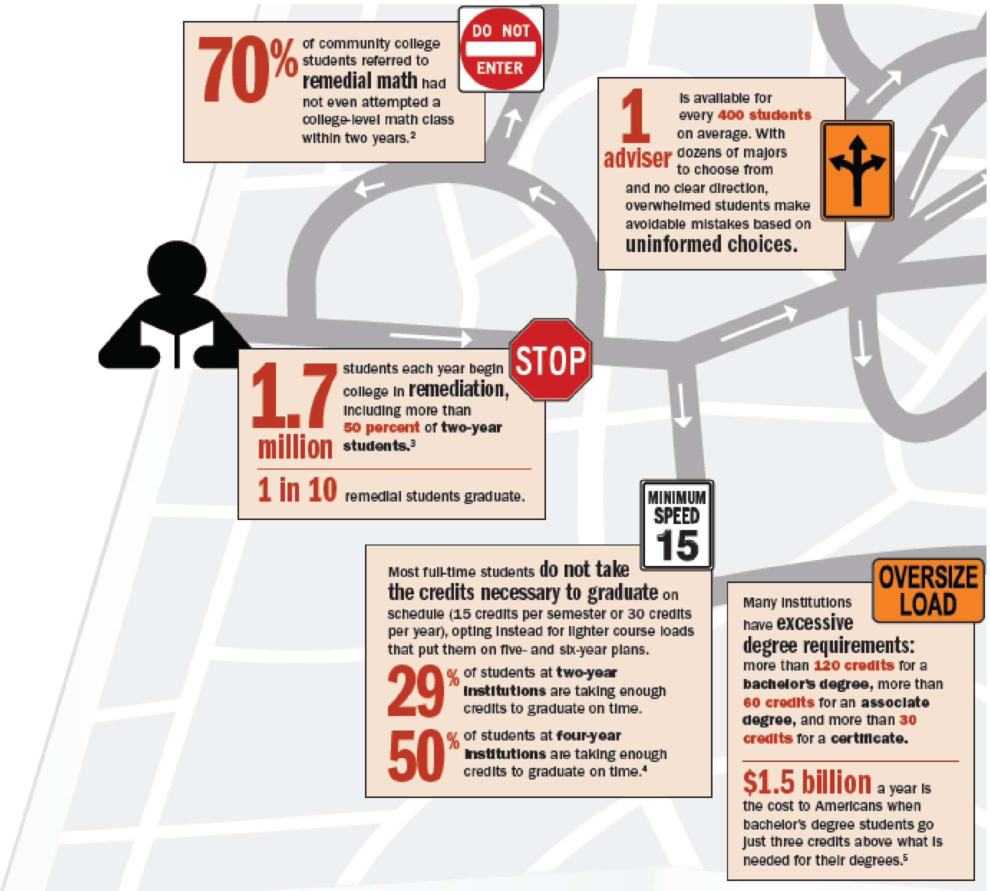
In light of this situation it is necessary to minimize problems and maximize on the advantages while using big data. How to achieve this? This is a major question that needs to be addressed.

Appendix 18 Theoretical aspects concerning big data

Alguliyev et al. (2017, p. 28) argue that data analysis is a bottleneck in various applications created due to lack of scalability of the underlying algorithms and due to the complexity of the data that needs to be analysed. It is further posited that exposition of the outcome of analysed data and extracting actionable knowledge from it through proper interpretation by non-technical experts is vital. These aspects have been described as major challenges as opportunities offered by big data and it is suggested in the literature that there is a need to have rethinking on these aspects considered as part of data management (Alguliyev et al. 2017, p. 28). Some argue that big data and investments in it have the potential to advance scientific knowledge and lay the base for the next stage of advancement in science, medicine and business (Gopalkrishnan et al. 2012; Madden, 2012). In continuation it is argued that this is a major challenge as big data has been associated with sophisticated analytical bottlenecks that cannot be addressed with currently available tools and practices found in the industry.

One of the ways by which this can be taken care of is by building scientific approaches and theoretical models (Alguliyev et al. 2017, p. 28). Theories gain importance. Similar sentiments are espoused by Bradlow et al. (2017, p. 85) who argue that theory is needed to guide where to look in the data and develop precise hypotheses that can be verified against the data. Although predictive algorithms that link input with output using past data, and machine learning are cited as examples of approaches that are commonly used big data analysis Bradlow et al. (2017, p. 85) argue that those approaches are not adequate to predict outcomes related to significant policy changes and hence theory is needed to guide managers to understand and predict outcomes. Theories related to data mining that explain how to analyse granular data, generating many correlations at different aggregation levels, visualizing, and leveraging the power of random sampling are identified as one of the ways by which managers can be guided to predict outcomes. However there is a thought that shows these theories may not be adequate as a theory working in context may not work in other contexts (Bradlow et al. 2017). For instance it is argued that predictions that could be made purely based on patterns generated by data mining approaches have limitations and do not go far beyond the learnings from a training set. For instance the much touted Google flu detector algorithm was severely criticized for over-predicting actual cases new or as yet unrecorded flu cases in the USA. In fact some went to the extent of saying that theory is no more needed in data mining (Bradlow et al. 2017; Anderson, 2008) which was falsified by the example of Google flu detector algorithm, considered theory free (Lazer et al. 2014). It can be implied from the above that supporting theories and concepts are needed for instance contextual theory to interpret and predict outcomes.

Appendix 19 Reasons for delayed Graduation (Source: Complete College America, 2014)



70% of community college students referred to **remedial math** had not even attempted a college-level math class within two years.[2]

1 adviser is available for every **400 students** on average. With dozens of majors to choose from and no clear direction, overwhelmed students make avoidable mistakes based on **uninformed choices.**

1.7 million students each year begin college in **remediation,** including more than **50 percent** of **two-year students.**[3]

1 in 10 remedial students graduate.

Most full-time students **do not take the credits necessary to graduate** on schedule (15 credits per semester or 30 credits per year), opting instead for lighter course loads that put them on five- and six-year plans.

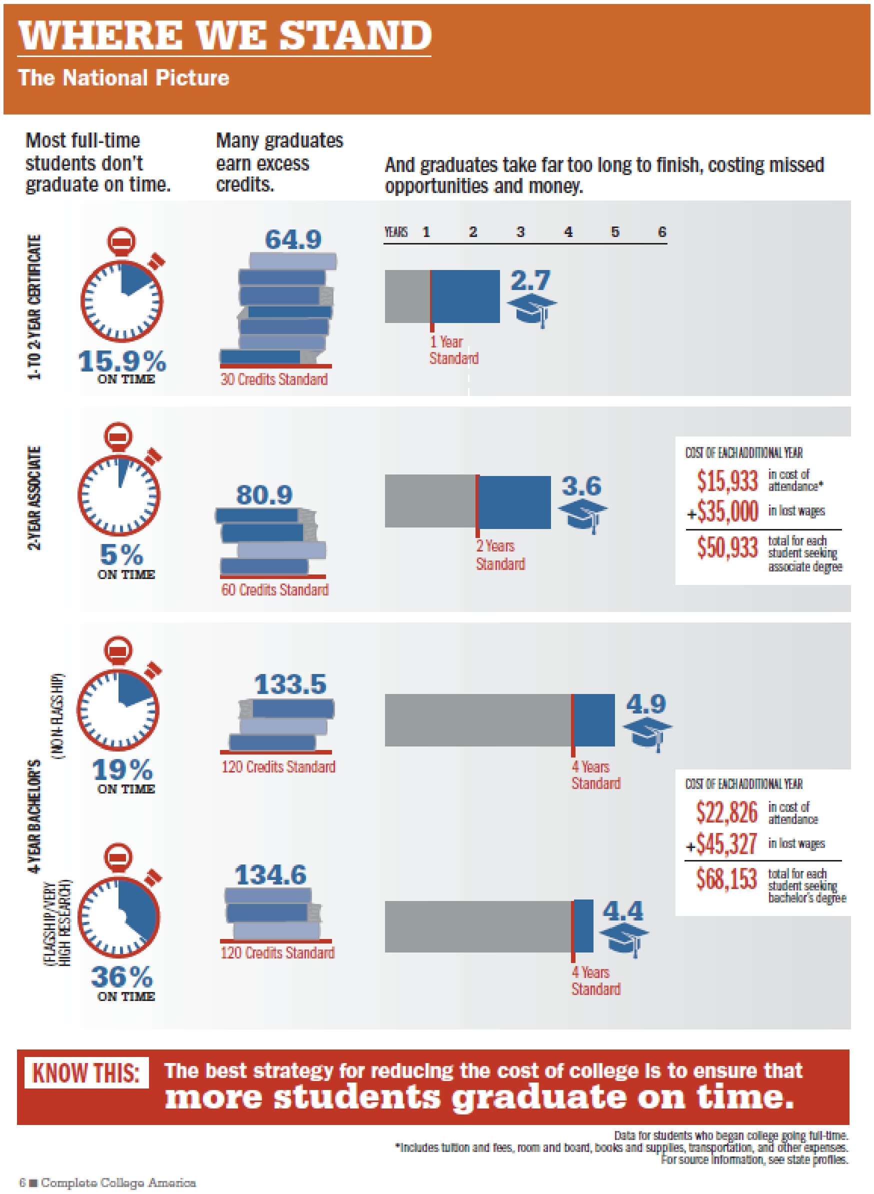
29% of students at **two-year institutions** are taking enough credits to graduate on time.

50% of students at **four-year institutions** are taking enough credits to graduate on time.[4]

Many institutions have **excessive degree requirements:** more than **120 credits** for a **bachelor's degree,** more than **60 credits** for an **associate degree,** and more than **30 credits** for a **certificate.**

$1.5 billion a year is the cost to Americans when bachelor's degree students go just three credits above what is needed for their degrees.[5]

Appendix 20 Percentage of On-time completion and additional costs on every extra year (Source: Complete College America, 2014)



WHERE WE STAND

The National Picture

Most full-time students don't graduate on time.

Many graduates earn excess credits.

And graduates take far too long to finish, costing missed opportunities and money.

COST OF EACH ADDITIONAL YEAR	
$15,933	in cost of attendance*
+$35,000	in lost wages
$50,933	total for each student seeking associate degree

COST OF EACH ADDITIONAL YEAR	
$22,826	in cost of attendance
+$45,327	in lost wages
$68,153	total for each student seeking bachelor's degree

KNOW THIS: The best strategy for reducing the cost of college is to ensure that **more students graduate on time.**

Data for students who began college going full-time.
*Includes tuition and fees, room and board, books and supplies, transportation, and other expenses.
For source information, see state profiles.

Appendix 21 Related literature about data and data analysis

All the organisations across the world are witnessing a dramatic increase in data and information. The facilities offered by the fields of computing and information technology enable the storage and analysis of large volumes of data. In HEIs volume of data is increasing exponentially with the number of students growing steadily. For instance one report shows that globally the number of internationally mobile students who moved to other countries was estimated to be 5 million in 2014 which is an increase of more than 100% when compared to the figure of 2.1 million in 2000 (University of Oxford International Strategy Office, 2015). In addition the landscape of a modern university is gradually changing. For instance a modern university is characterized by multi-culture environment with students coming to study from different countries supported by faculty belonging to different nationalities due to globalisation (Edelstein, 2014). Similarly curricula are being modified more frequently than before necessitated by job-market requirements and newer programmes are being offered based on the demand from students (Weise, 2014). In addition, HEIs are involved in implementing new strategies in order to forge ahead of other competing institutions. For instance, some universities focus on technological and engineering streams whereas some others want to excel in healthcare. These aspects invariably lead to an environment in HEIs that is challenging and complex making it important for the institutions to identify ways by which complex and challenging environment could be tackled. Amongst the different strategies that are used by HEIs, analysing data both past and current and use the results of the analysed data for making decisions and improving performance including those related to student learning, student experience and teaching. Thus educational data becomes very important for modern HEIs (Conrad et al. 2013).

While data and its analysis are becoming important in HEIs, dealing with data and its analysis is creating other challenges. Different types of data, collection of data, ever increasing volume of data, method of analysing data, interpreting the results of analysing data, choice of decisions from among alternatives, implementing decisions, monitoring the progress of implemented decisions, assessing the performance and ensuring enhancement in performance based on sound processes are posing serious challenges to HEIs. For instance, from Table 3.1 which provides a list of various types of data collected in HEIs it can be seen that different data types use different data analysing techniques. This aspect requires expertise on the part of HEIs to understand the data types and analysing techniques thoroughly before choosing the most appropriate analysing technique for analysing a particular type of data. Further the table shows that a variety of data types have been identified in the literature that are collected, stored in electronic form and used in various organisations, which is in itself a challenge. Additionally, Table 3.1 shows that some types of data are not used in the HEIs yet (e.g. Multimedia, Data streams and sensor data databases) probably due to lack of expertise on the part of HEIs to deal with particular types of data although recent developments indicate that HEIs have started using those data types. Even if expertise is provided to handle different types of data to the HEIs, does the literature support the HEIs with research outcomes that could guide the HEIs in overcoming the challenges that are faced by the HEIs is a moot question that requires examination. Ironically some of the challenges faced by HEIs are faced by the researching communities themselves solutions for which are eluding the researchers, evidence for which could be found from the interpretation of the information provided Table 3.1 which is discussed next.

Table 7.3, Data types and analysing methods

No.	Kinds of data	Characteristics	Businesses where this kind of data is found	Authors	Algorithms in which the data type can be used	Remarks with regard to education sector (Whether used or not)
1.	Flat files (semi structured)	Simple data files in text or binary format. The data in these files are transactions and occupy less space	Biological data, Educational data	Sinha et al. (2005); Zhang and Agrawal (2006), Al-Dubaee et al. (2010)	Almost all algorithms especially Association Rule Mining (ARM) (Ramesh et al. 2002)	Most of the educational data mining has been done with flat files (Aslam & Ashraf, 2014)
2.	Relational Databases and Data warehouses	Consists of set of tables Repository of data from multiple sources	Databases of most of the businesses	Carlos (2004); Sarawagi (1998); Chaudhuri et al. (1999); Chaudhuri (1998);Bradley et al. (1998)	Relational Data Mining(RDM) algorithms – Relational Decision trees, Relational Regression (Džeroski,2003;2007)	Most of the educational data mining have been done with data warehouses (Aslam & Ashraf, 2014)
3	Time series data or temporal data	Time related data stored as series of data points indexed (or listed or graphed) in time order	Stock market, Logged activities of any business	Economic forecasting (Song and Li, 2008), intrusion detection (Zhong et al. 2007), gene expression analysis (Lin et al. 2008), medical surveillance (Burkom et al. 2007), and hydrology (Ouyang et al. 2010)	k−means (Vlachos et al. 2003), k-center clustering (Cormode et al.2007); neural networks (Nanopoulos et al.2001) or Bayesian classification (Povinelli et al. 2004), Bayesian prediction (Pesaran et al. 2006),support vector machines (Cao and Tay, 2009)	Very few studies in education. Time series has been used to analyse attendance data (Koopmans, 2011).
4.	Sequence data	Data related to sequences of ordered events	Web click streams, Online learning system activities, Medical, DNA Sequences	Agrawal and Srikant 1995; Yang et al.2001	Neural Networks (Wang et al.2000);Support Vector Machines (Cai et al. 2003); Fuzzy ARTMAP (Mohamed et al. 2006)	To study learner behaviour with the help of sequential pattern mining (Wang et al.2009; Zhang & Liu,2008; Nesbit et al. 2008)
5.	Text data	Contains text documents in the form of long sentences or paragraphs or product specifications, reports, summary reports, error reports,	Web logs, online education systems	Pullman et al.(2005);Chen-Huei,2008	Support Vector Machines SVM (Evgeniy Gabrilovich, 2004), Vandana Korde et al, 2012.	Abdous and He (2011), Elrahman et al. (2010)
6.	Multimedia databases	Contains images, audio and video data	Voice mail systems, video on demand systems, speech based interfaces, the World Wide Web	Petrushin and Khan, 2007; Liu et al.2002	Multimedia data mining (MDM) - summarization, association, classification, clustering, trend analysis and deviation analysis(More & Mishra,2012)	Nil
7.	Spatial data	Contains	Spatial- Geographical	Spatial – Shekhar and Chawla,	Spatiotemporal association	Very few studies like (Khan, 2014)

	and spatio temporal data	spatial(location and geometry) related information which can be represented in the form of raster or vector data. Raster data consists of n-dimensional bit maps or pixel maps and vector data are represented by points, lines, polygons. The spatiotemporal object usually contains spatial, temporal and thematic or non-spatial attributes	databases, medical and satellite. Spatio temporal - event types are earth quake, hurricanes, road traffic jam and road accidents	2003; Bailey and Gatrell, 1995; Maroko, et al.2011. Spatio temporal - Bogorny and Shekhar,2010; Han,2003.	rules (Verhein & Chawla, 2005);Spatio temporal data clustering (Manso et al. ,2010; Birant & Kut, 2007)	
8.	Data streams and sensor data	Infinite volume , dynamically changing, flowing in and out in a fixed order, very quick response time	Scientific, engineering data, network traffic, stock exchange, telecommunication, weather or environment monitoring	Aggarwal et al. 2003; Cai et al.2004	Stream clustering, Frequent items (Cheqing Jin et al. 2003); Data Stream association rule mining (Jiang and Grunewald, 2006), Decision trees (Rutkowski et al.2013)	Nil
9.	World wide web	Worldwide online information services data where data objects are linked together to facilitate interactive services	Yahoo, America online	Herrouz et al.2012; Cooley,1997	Content mining (Sujath1 ,2014), structure mining (Junior and Gong,2005) and usage mining (Facca & Lanzi, 2003)	E-learning has a lot of world wide web data mining (Greenhow, Robelia and Hughes,2009; Hanna,2004; Macfayden & Dawson,2010)

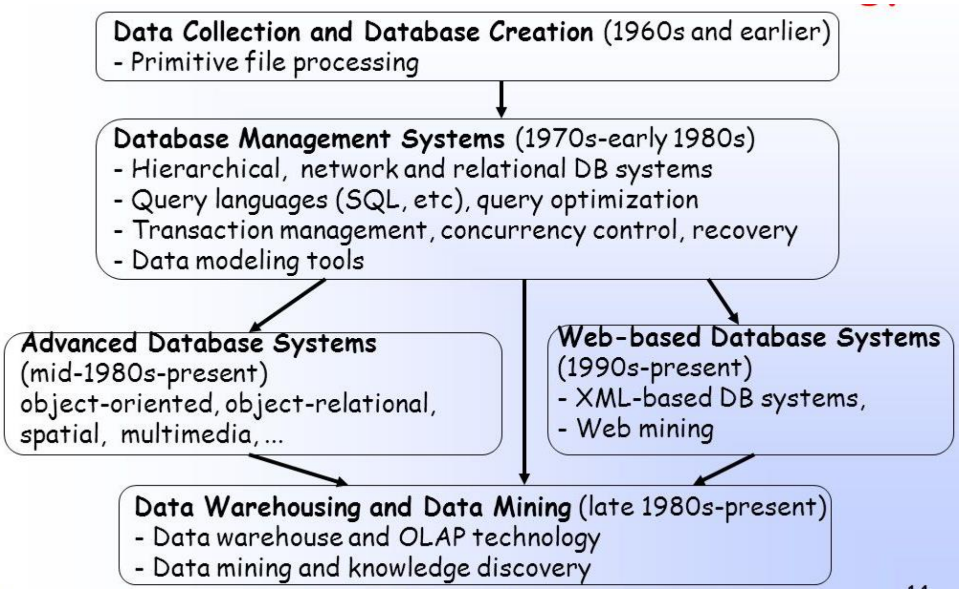
An important inference that could be derived from Table 3.1 is that multiple data analysing techniques have been developed to deal with each type of data which indicates that there is no unique way by which a particular type of data could be analysed. For instance neural networks, support vector machines and fuzzy ARTMAP are some algorithms that are used to analyse sequence data, a data type used in HEIs. Each one of these techniques has different characteristics yet could be used to analyse sequence data necessitating users to have accurate knowledge to determine which one of the analysing methods rather algorithms should be used to analyse sequence data. In the event users do not have such knowledge, then the choice of the method could be made based on arbitrary ways. Similar arguments could be put forth with regard to other data analysing algorithms or data analysing methods used to analyse the different data types.

These arguments show that there is a lack of consistency in regards to the standard using which data analysing methods were developed to analyse data due to which it might not have been possible for the users to choose the most appropriate method for data analysis. In addition those algorithms and methods are found to have limitations. For instance C4.5 and ID3 algorithms have been criticized to be suffering from the limitation that it can have a target attribute that has discrete values only which forces the users to search for another algorithm like regression trees or CART that could enable them to have target data with continuous values (Singh & Gupta, 2014).

Appendix 22 Evolution of Data mining

Data mining is an evolving field (Liebowitz, 2017). It is finding application in many sectors including healthcare, education and others. With data growth estimated at around 35% per year in fields like education, data mining was finding application in analysing large data. Table 3.2 gives an idea about how data mining has evolved.

Table 7.4, Evolution of Data Mining



Data Collection and Database Creation (1960s and earlier)
- Primitive file processing

Database Management Systems (1970s-early 1980s)
- Hierarchical, network and relational DB systems
- Query languages (SQL, etc), query optimization
- Transaction management, concurrency control, recovery
- Data modeling tools

Advanced Database Systems
(mid-1980s-present)
object-oriented, object-relational, spatial, multimedia, ...

Web-based Database Systems
(1990s-present)
- XML-based DB systems,
- Web mining

Data Warehousing and Data Mining (late 1980s-present)
- Data warehouse and OLAP technology
- Data mining and knowledge discovery

The term "Data mining" was introduced in the 1990s, but the field has a long history (Ramzan & Ahmad, 2014). The roots of data mining can be tracked back to three disciplines namely classical statistics, artificial intelligence, and machine learning. Statistics is the foundation on which data mining is established. For instance discriminate analysis, regression analysis, standard variance, standard distribution, standard deviation, cluster analysis, and confidence intervals are used to study data and data relationships (Kumar & Bharadwaj, 2011). Artificial intelligence or AI uses heuristics (technique designed to solve a problem more quickly when classic methods are too slow or fail to find solution) and applies human-thought-like processing to statistical problems. Examples of AI applications include the Apple's Siri, Google's self-driving cars and Facebook's image recognition software. Machine learning is the union of statistics and AI (Soman et al. 2006). Machine learning (Mitchell, 1997) is an established and well-recognized research area of computer science, mainly concerned with the discovery of models, patterns, and other regularities in data. Machine learning requires no prior assumptions about the underlying relationships between the variables about which data has been collected. AI blends AI heuristics with advanced statistical analysis. Using machine learning computer programs are written to learn about the data under study in such a way that programs enable users to make various decisions based on the quality of the data studied, use of statistics for fundamental concepts and addition of more advanced AI heuristics and algorithms to achieve the goals set.

While data mining was evolving it was aided by the rapid advances taking place in the computer hardware industry, which was essential for data mining. The steady progress of computer hardware technology in the past three decades led to large supplies of powerful and affordable computers, data collection equipment, and storage media. The advances taking place in the hardware side supported the progression that was taking place in the database and information industry. Thus a high number of databases and information repositories were made available for transaction management, information retrieval, and data analysis (Han & Kember, 2011). Data warehouse is an example of a repository of multiple heterogeneous data sources, organized under a unified schema at a single site that facilitates management decision making. Data warehouse technology includes data cleansing, data integration, and On-Line Analytical Processing (OLAP), an analysis technique featuring summarization, consolidation, and aggregation, as well as viewing information from different angles. OLAP can be used to test hypothesis about relationships and patterns in data. Online analytical processing (OLAP) was introduced to enable inexpensive data access and gain insights in to those data (Bawden, 2013). However, it should be noted that the typical operations used in data warehouses are similar to the ones used in the traditional OLTP databases and the users can issue a query to obtain a data table as a result. OLTP is a group of information systems that aids in and controls transaction-oriented applications, usually for data entry and retrieval transaction processing on a database management system. OLAP tools support multidimensional analysis and decision making but only with the support of additional data analysis tools that can support in-depth analysis, such as data classification, clustering, and the characterization of data changes over time. To overcome this difficulty database Query Languages were found useful as query languages could help the users in reporting facts and figures that are already stored in database. The discussions provided above have been summarised in Table 3.3.

Table 7.5, Evolution of data mining process

Evolutionary Step	Business Question	Enabling Technologies
Data Collection(1960s)	"What was my total revenue in the last five years?"	Computers, tapes, disks
Data Access	"What were unit sales in New	Relational databases

(1980s)	England last March?"	(RDBMS), Structured Query Language (SQL), ODBC
Data Warehousing & Decision Support (1990s)	"What were unit sales in New England last March? Drill down to Boston."	On-line analytic processing (OLAP), multidimensional databases, data warehouses
Data Mining (Emerging Today)	"What's likely to happen to Boston unit sales next month? Why?"	Advanced algorithms, multiprocessor computers, massive databases

Appendix 23 Knowledge Discovery and Data Mining (KDDM)

KDD is described as a process starting from use of prior knowledge, to construction of target data set, its cleaning and pre-processing, use of data mining algorithms, identify interesting patterns, evaluate patterns and consolidate the discovered knowledge (Fayyad et al. 1996a) (e.g. Figure 4.1). In literature it can be seen that specifically data mining as a step has been identified as part of the overall KDD process in which the user could apply particular data mining algorithms to obtain patterns and discover knowledge. However Fayyad et al. (1996a) and others (Reinartz 2002; Han & Kamber 2006; Kurgan & Musilek 2006) acknowledge that today data mining and KDD have come to be used interchangeably in the literature.

.

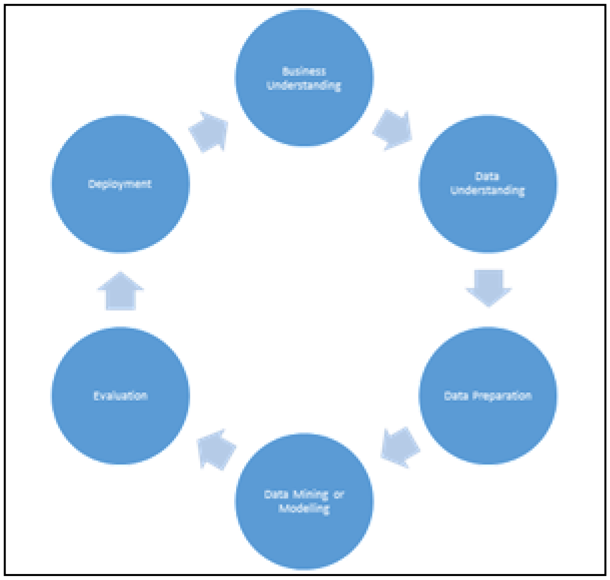


Figure 7.1, Representation of KDDM (*Source:* Fayyad et al. 1996)

The term data mining (DM) was coined by the MIS researchers while the term KDD was used by researchers involved in machine learning to describe the knowledge discovery process (Fayyad et al. 1996a). Researchers concerned with DM and KDD in due course combined the two concepts by proposing the use of the term ‘knowledge discovery and data mining’ (KDDM). It was argued that KDDM addresses two equally critical aspects namely the overall knowledge discovery process which includes pre-processing and post-processing of data as well as interpretation of the discovered patterns as knowledge; and particular data mining methods and algorithms aimed at solely extracting patterns from data.

KDDM Process Models

In continuation to the discussions provided Section 3.2 of chapter 3, Figure 4.2 has been provided to gain an idea of the evolution of the KDD process. As can be seen from Figure 4.2 the root of the currently available KDDM process models lies in the KDD process developed by Fayyad et al. (1996) with the exception of 5 A's model developed by de Pisón (2003).

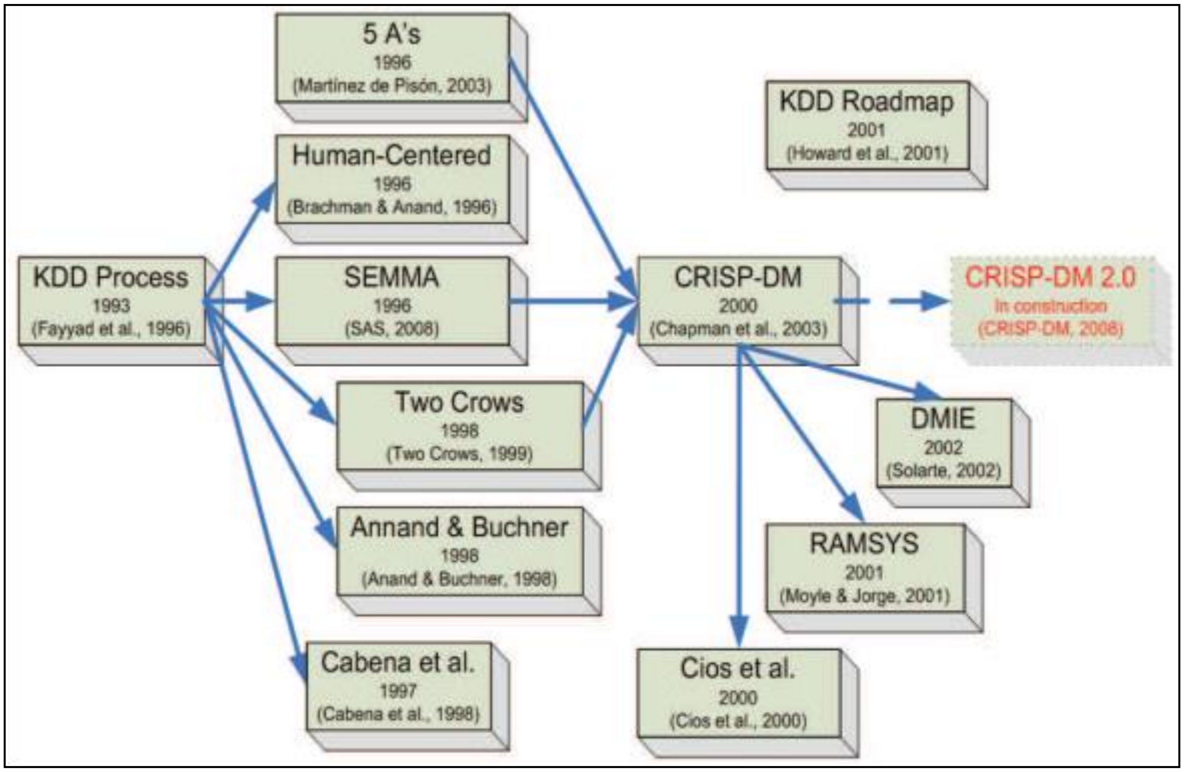


Figure 7.2, Evolution of DM & KD process models (*Source*: Marban et al. 2009)

Although the main scope of this research does not include a detailed review of all the models found in Figure 4.2, a discussion on some of the widely discussed models in the literature has been provided in Table 4.5. Prior to the discussion an explanation about the various steps commonly used in the KDDM processes was considered necessary in order to know the structure of a KDDM process. Knowledge about the steps is expected provide the foundation to analyse how the current KDDM processes mine data and the problems encountered by the users in using those KDDM process models.

KDDM as a process model requires some description in order to understand its capabilities and limitations. While capabilities provide knowledge about how KDDM process models could be used for data mining purposes, limitations are expected to reveal the areas that need enhancement. KDDM process models offer the up-to-date guidelines for the use of data mining. In the literature a number of KDDM process models have been discussed (Adriaans & Zantinge, 1996; Fayyad et al. 1996a; Anand & Buchner, 1998; Cabena et al. 1998; Cios et al. 2000; Chapman et al. 2000; Berry & Linoff, 1997; Edelstein, 1998; Han & Kamber, 2001; Klösgen & Zytkow, 2002; SAS Institute, 2003; Haglin et al. 2005). Fayyad et al. (1996a, 1996b) who proposed the widely recognized "KDD Process", defined the basic structure of KDDM process models which was used by researchers who developed other KDDM models. Kurgan and Musilek (2006) reviewed and evaluated of prominent KDDM process models and identified the essential set of steps that need part of a the data mining project. The steps mentioned in the KDDM process models begin with objective determination and end with the utilization of the discovered knowledge. It is significant to note that data mining is treated as one of the stages in the entire KDDM process. Although the various KDDM process models are alike still there are major differences that exist. Such differences while highlighting the inconsistency that exist in the KDDM process models developed so far, also indicate that no model can be considered as standard and applicable to every environment where data mining process can be deployed (Kurgan & Musilek, 2006). A discussion therefore on the various models could reveal the differences and inconsistencies in those models but it is beyond the scope of this research. However although the main scope of this research does not include a detailed review of all the models found in Figure 4.1, a taxonomy of some of the widely discussed models in the literature was thought to be necessary and has been provided in Table 4.5. The taxonomy includes information about the essential steps involved in those models which will provide an idea about the function to be executed in that step and the limitations surrounding those steps that need to be addressed.

Table 7.6, Taxonomy of KDDM Models

Model	**Step1**	**Step 2**	**Step 3**	**Step4**	**Step5**	**Step6**	**Limitations**
Generic	Step 1.Application Domain Understanding	STEP 2 Data Understanding	STEP 3 Data Preparation and Identification of DM Technology	Step 4: Data Mining	STEP 5 Evaluation	STEP 6 Knowledge Consolidation and Deployment	Lack of understanding of specific business problems (Step1). No mandatory step like in other models to understand them (Step1). Lack of clarity on the list of data quality checks to be performed and on the ways to address or overcome data quality issues (Step2). Detailed steps of the process is missing which can guide the data miner to carry out the tasks efficiently (Step3). Feedback loop is not discussed in detail in the documentation (Step4). Lack of documentation on guidelines of evaluation (Step5).

							Lack of monitoring of deployment (Step6). Lack of Contextual Processing.
Fayyad et al.	-Learning goals of the end-user and relevant prior knowledge	-Selection of a subset of variables and sampling of the data to be used in later steps	-Preprocessing of noise, outliers, missing values, etc, and accounting for time sequence information - Selection of useful attributes by dimension reduction and transformation, development of invariant data representation -Goals from Step 1 are matched with a particular DM method, i.e. classification, regression, etc. - Selection of particular data model(s), method(s), and method's parameters	-Generation of knowledge (patterns) from data, for example classification rules, regression model, etc	-Interpretation of the model(s) based on visualization of the model(s) and the data based on the model(s)	-Incorporation of the discovered knowledge into a final system, creation of documentation and reports, checking and resolving potential conflicts with previously held knowledge	There is a lack of understanding on specific business needs which need to be translated into DM goals as present in other models (Step1). There is a lack of data understanding stage where the data is explored for quality issues which are present in other models (Step2). The prepared data might not be suitable for the DM method as the data is not formatted to suit it (Step3). Feedback loop is not discussed in detail in the documentation (Step4). Lack of documentation on guidelines of evaluation (Step5).

							Lack of deployment and monitoring phase unlike other models (Step6). Lack of Contextual Processing
Cabena et al.	1.Understanding the business problem and defining business objectives, which are later redefined into DM goals.	-Identification of internal and external data sources, selection of subset of data relevant to a given DM task. It also includes verifying and improving data quality, such as noise and missing data. Determination of DM methods that will be used in the next step and transformation of the data into analytical model required by selectedDM methods	______	-Application of the selected DM methods to the prepared data.	-Interpretation and analysis of DM results; usually visualization technique(s) are used	-Presentation of the generated knowledge in a business-oriented way, formulation of how the knowledge can be exploited, and incorporation of the knowledge into organization’s systems	There is lack of guideline or tool on how to approach this step and complete it without challenge (Step1). There is lack of time to understand and explore the data as this stage includes preparation also (Step2). Lack of data preparation step which has a tedious sub processes unlike other models (Step3). Feedback loop is not discussed in detail in the documentation (Step4). Lack of documentation on guidelines of evaluation (Step5).

							Lack of deployment phase and monitoring unlike other models (Step6), Lack of Contextual Processing.
Anand & Buchner	-.Identification of human resources and their roles -Partitioning of the project into smaller tasks that can be solved using a particular DM method	-Analysis of accessibility and availability of data, selection of relevant attributes and a storage model -Elicitation of the project domain knowledge	-Selection of the most appropriate DM method, or a combination of DM methods -Pre-processing of the data, including removal of outliers, dealing with missing and noisy data, dimensionality reduction, data quantization, transformation and coding, and resolution of heterogeneity issues	-Automated pattern discovery from the pre-processed data	- Filtering out trivial and obsolete patterns, validation and visualization of the discovered knowledge	__________	Lack of understanding on specific business needs which need to be translated into DM goals (Step1). Lack of data understanding stage where the data is explored for quality issues (Step2). The data preparation is preceded by the DM stage where the method is chosen and there is no step to format the data according to the method (Step3). Feedback loop is not discussed in detail in the documentation (Step4). Lack of documentation on guidelines of evaluation (Step5).

							Lack of deployment phase unlike other models (Step6).
CRISP-DM	1.Understanding of business objectives and requirements,which are converted into a DM problem definition	2 Identification of data quality problems, data exploration, and selection of interesting data subsets	3 Preparation of the final dataset, which will be fed into DM tool(s), and includes data and attribute selection, cleaning, construction of new attributes, and data transformations	-Calibration and application of DM methods to the prepared data	-Evaluation of the generated knowledge from the business perspective	-Presentation of the discovered knowledge in a customer-oriented way. Performing deployment, monitoring, maintenance, and writing final report	There is lack of guideline or tool on how to approach this step and complete it without challenge (Step1). Lack of clarity on the list of all data quality checks to be performed and on the ways to address or overcome data quality issues. (Step2). The prepared data might not be suitable for the DM method as the data is not formatted to suit it (Step3). Feedback loop is not discussed in detail in the documentation (Step4). Lack of documentation on guidelines of evaluation (Step5).

							Lack of feedback loop from deployment to business understanding (Step6). Lack of Contextual Processing
Cios et al.	1.Defining project goals, identifying key people, learning current solutions and domain terminology, translation of project goals into DM goals, and selection of DM methods for Step 4	2.Collecting the data, verification of data completeness, redundancy, missing values, plausibility, and usefulness of the data with respect to the DM goals	3 Pre-processing via sampling, correlation and significance tests, cleaning, feature selection and extraction, derivation of new attributes, and data summarization. The end result is a data set that meets specific input requirements for the selected DM methods	-Application of the selected DM methods to the prepared data, and testing of the generated knowledge	- Interpretation of the results, assessing impact, novelty and interestingness of the discovered knowledge. Revisiting the process to identify which alternative actions could have been taken to improve the results	-Deployment of the discovered knowledge. Creation of a plan to monitor the implementation of the discovered knowledge, documenting the project, extending the application area from the current to other possible domains	Selection of DM method at this stage could bring out problems that the method would not suit the data to be understood in step 2(Step1) Other data quality checks like noise, data staleness, ambiguity etc. are not handled which could lead to problems with modelling(Step2) Detailed documentation on the steps of the process is missing which can guide the data miner to carry out the tasks efficiently (Step3). Feedback loop is not discussed in detail in the documentation (Step4).

							Lack of documentation on guidelines of evaluation (Step5). Lack of feedback loop from deployment to business understanding (Step6). Lack of Contextual Processing

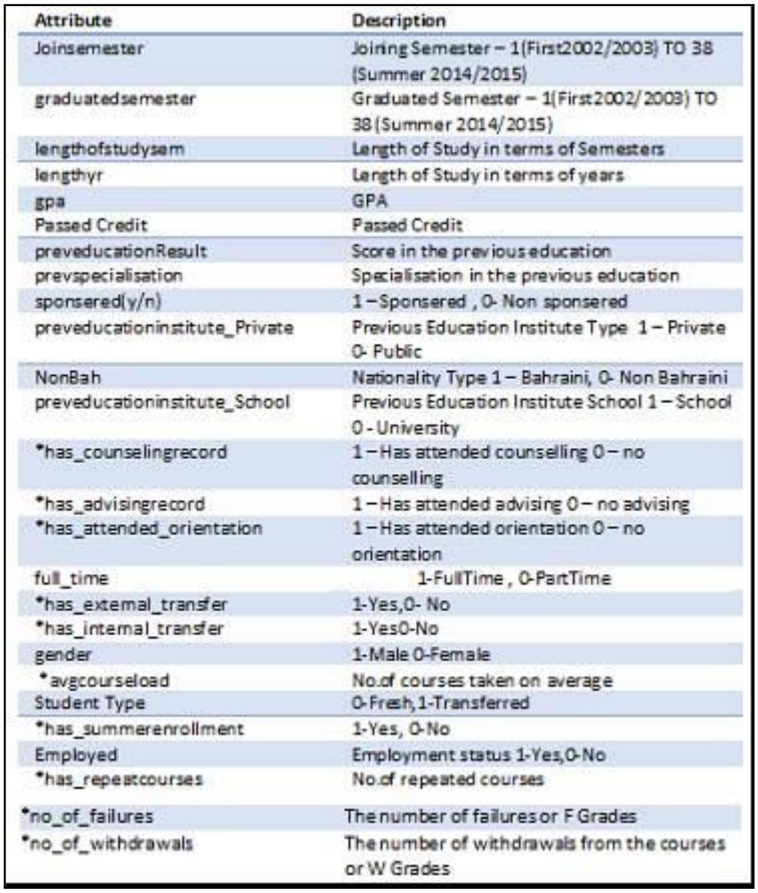
Appendix 24 Example of use of data mining as a method to answer questions:

Data mining involves application of algorithms to extract hidden pattern or knowledge from data. Data mining has six common tasks namely association rule mining, clustering, classification, regression, summarization and anomaly detection (see Section 3.4). One example was taken for investigation namely clustering. In addition data mining involved other processes such as data cleaning, data integration, data transformation, data mining, pattern evaluation and data presentation (Torgo, 2016). These aspects were also implemented.

As an example Equations 3.9 and 3.10 were verified by applying data mining technique called clustering technique to find if any hidden knowledge could be discovered in terms of course taking pattern and check the existence of any relationship in the dataset with regard to course taking pattern, CGPA and time-to-degree. The educational dataset from the anonymous university in Bahrain was used to find the above. Records of 1292 students were mined. Clustering technique was applied to find groups of students having similar course taking pattern and time-to-degree & CGPA. K-means clustering algorithm was applied to the dataset in Table 4.2. The dataset was extracted from the student information system of a university in Bahrain. The educational dataset mined consisted of the attributes that could be used to extract groups of students who have similar course taking patterns, time-to-degree and CGPA. The dataset was fed into data mining tool Weka version 3.7.11.

Prior to mining the educational dataset data was cleaned by identifying and correcting corrupt or erroneous records and then replace or delete dirty or uneven data. In addition missing data was checked and ensured that those records were either set right or deleted. No data integration was needed as all the attributes were already found in the sample educational dataset got from a single source. Further as part of data transformation step, the format of the data was changed to suit the algorithm (see Table 4.2).

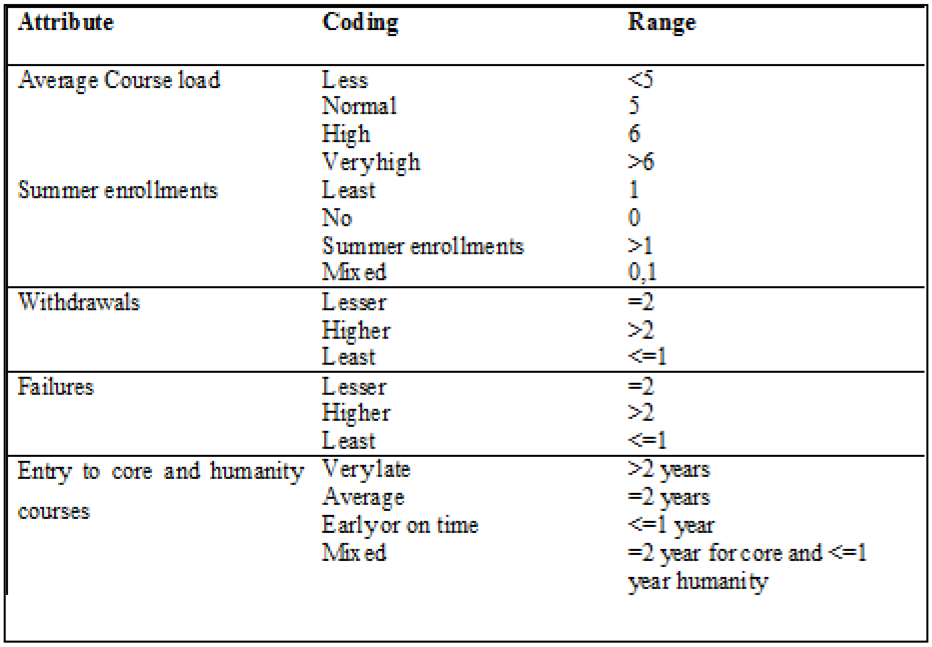
Table 7.7, Sample dataset used for clustering (data mining)



Attribute	Description
Joinsemester	Joining Semester – 1(First2002/2003) TO 38 (Summer 2014/2015)
graduatedsemester	Graduated Semester – 1(First2002/ 2003) TO 38 (Summer 2014/ 2015)
lengthofstudysem	Length of Study in terms of Semesters
lengthyr	Length of Study in terms of years
gpa	GPA
Passed Credit	Passed Credit
preveducationResult	Score in the previous education
prevspecialisation	Specialisation in the previous education
sponsered(y/n)	1 – Sponsered , 0- Non sponsered
preveducationinstitute_Private	Previous Education Institute Type 1 – Private 0- Public
NonBah	Nationality Type 1 – Bahraini, 0- Non Bahraini
preveducationinstitute_School	Previous Education Institute School 1 – School 0 - University
*has_counselingrecord	1 – Has attended counselling 0 – no counselling
*has_advisingrecord	1 – Has attended advising 0 – no advising
*has_attended_orientation	1 – Has attended orientation 0 – no orientation
full_time	1-FullTime , 0-PartTime
*has_external_transfer	1-Yes,0- No
*has_internal_transfer	1-Yes0-No
gender	1-Male 0-Female
*avgcourseload	No.of courses taken on average
Student Type	0-Fresh,1-Transferred
*has_summerenrollment	1-Yes, 0-No
Employed	Employment status 1-Yes,0-No
*has_repeatcourses	No.of repeated courses
*no_of_failures	The number of failures or F Grades
*no_of_withdrawals	The number of withdrawals from the courses or W Grades

It is brought out here that scales were developed to measure certain attributes on a nominal scale (see Table 4.3).

Table 7.8, Coding of output attributes from clustering

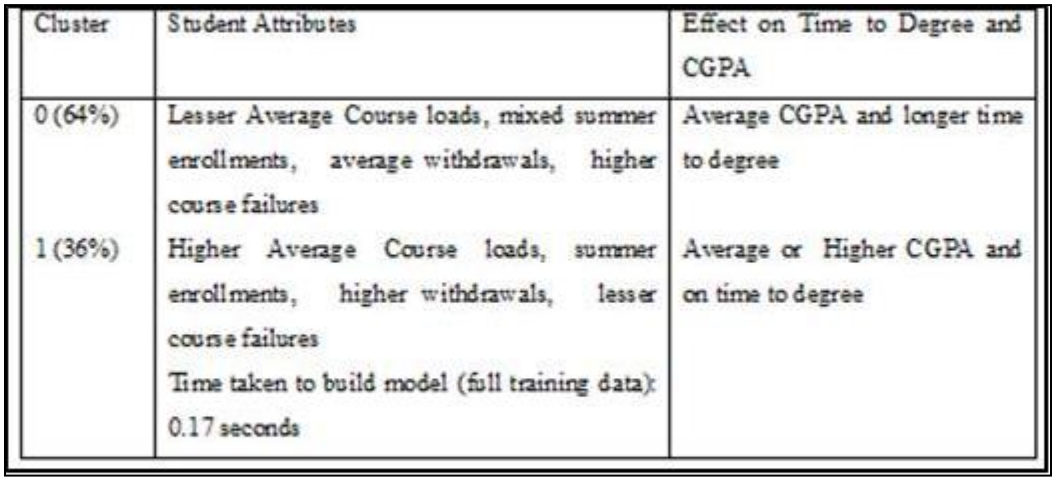


Attribute	Coding	Range
Average Course load	Less Normal High Veryhigh	<5 5 6 >6
Summer enrollments	Least No Summer enrollments Mixed	1 0 >1 0,1
Withdrawals	Lesser Higher Least	=2 >2 <=1
Failures	Lesser Higher Least	=2 >2 <=1
Entry to core and humanity courses	Verylate Average Earlyor on time Mixed	>2 years =2 years <=1 year =2 year for core and <=1 year humanity

K - means algorithm was used to extract groups of students as a pattern. Then pattern generated by the algorithm was evaluated. In this research the cluster quality was evaluated using sum of squared error (SSE). K-means by default generates two clusters and higher number of clusters could be generated by the algorithm by setting the number. For instance in Weka software it is possible to set the cluster number as k=3 or 4 or 5 or 6 so on and so forth. Here the default number of 2 clusters generated by the algorithm was evaluated and the SSE was checked and found to be 26.34. Then the cluster numbers were incremented by one using Weka software version 3.7.11 until 10 and the clusters generated were evaluated by using SSE values produced for k= 3 clusters, 4 clusters and 5 clusters and continued until 10 clusters were tested. The SSE values were found to be the least in 2 clusters. Thus after the clustering technique was run with the sample educational dataset fed in with k=2, 2 clusters were created which was accepted as it had the lowest SSE.

The data mining as report generated is provided in Table 4.4.

Table 7.9, Output clusters from clustering



Cluster	Student Attributes	Effect on Time to Degree and CGPA
0 (64%)	Lesser Average Course loads, mixed summer enrollments, average withdrawals, higher course failures	Average CGPA and longer time to degree
1 (36%)	Higher Average Course loads, summer enrollments, higher withdrawals, lesser course failures Time taken to build model (full training data): 0.17 seconds	Average or Higher CGPA and on time to degree

From Table 4.4 it can be seen that out of the 1292 students mined only four attributes are found to be related to average CGPA and longer time-to-degree in ‘cluster 0’ while four attributes are found to be related to average or higher CGPA and on time-to-degree in ‘cluster 1’. However from this information it is clearly difficult to predict the optimum CGPA and time-to-degree in terms of course taking pattern as a set of courses. In addition the outcome could be interpreted by the data miner in ways not related to the business goals of the HEIs. While clustering could help in generating groups of students with common attribute like students who have the ‘*higher average load’, ‘summer enrollment’, ‘higher withdrawals’ and ‘lesser course failures*’. Similar arguments could be extended to the other data mining techniques for instance association rule mining and classification and hence have not been discussed here as examples of data mining experiment that could be used to predict optimum CGPA and time-to-degree in terms of course taking pattern and course difficulty level pattern. In fact all the three techniques have been used elaborately in experiments concerning CRISP-DM process in Chapter 5 where the full details of how the techniques function could be gauged. However in the example of cluster provided here and the other examples dealt with in Chapter 5, it is clear that data mining alone cannot help to predict the optimum CGPA and time-to-degree using the knowledge discovered in terms of course taking pattern and course difficulty level pattern due to the limitations surrounding data mining techniques.

If business goals are not understood then the mined data, the generated patterns and discovered knowledge could be misinterpreted leading to erroneous decision making. Thus data mining projects need to be preceded by proper understanding of the business and the business goals to be achieved. For instance, in the current instance it is necessary to enhance the student learning experience of HEIs. One way to do this is to enable the students to achieve optimum CGPA and time-to-degree by predicting these two factors. Although there could be many ways that could be used to achieve this, one area that is not well understood is the concept of course taking pattern of students which has been found to have the potential of predicting optimum CGPA and time-to-degree. This understanding of the business and the goals to be achieved is essential prior to mining. If not understood then the mining would be carried out using wrong techniques and the results may not be useful for implementation or could lead to erroneous decision making. Similarly it is possible data is not properly understood or prepared. In that case the outcome of the mining process may not help in discovering the true knowledge. This stage must precede the data mining stage. Again the outcome of the mined data may be patterns that may or may not be relevant and the information fed back to the input stage to correct if needed. For instance if the pattern generated is going to produce pattern of all students and grouped in a large number of clusters, then the mined outcome may not be easy to understand. An evaluation stage must follow mining to know whether the pattern generated using data mining is meaningful or not and whether large clusters could be reduced to a number that could be meaningfully dealt with. In other times event the data mining algorithms will require special formatting of data (e.g. association rules cannot handle numerical data). Thus there must be a stage prior to the mining stage to understand the data and prepare data for mining.

The limitations mentioned above and the ones described in Section 3.4 point out that data mining as a concept will not be adequate to predict optimum CGPA and time-to-degree in terms of course taking pattern and course difficulty level patterns. Literature suggests in situations knowledge discovery and data mining (KDDM) could be used (Kovacic, 2010).

Appendix 25: Contextual factors

Contextual factor	Author
Student potential represents the competence of a student for a given course based on the grades he has obtained in all the courses (prerequisities or not) the student has taken up to the moment. Potential is calculated as a weighted average of those grades divided by their corresponding difficulties. The potential is represented by: $$Potential_{s,c,d} = \frac{\sum_{t \in SPC_{c,d}} \left(\frac{\sum_{v=1}^{H_t} G_{s,t,v} * W_t}{D_t} \right)}{\sum_{t \in SPC_{c,d}} W_t * H_t}$$ where *s*, student; *c*, current target course; *d*, distance for the potential calculation; *t*, course; *SPCc,d* , set of course taken so far *c* at distance *d*; *Ht* , number of times student *s* was enrolled in course *c*; *Gs,t,v* , grade from student *s* in the course *t* at attempt *v*; *Wt* number of credits from course *t*; *Dt* difficulty from the course *t*.	Vialardi et al.2010 has used the student potential in his recommender system to predict the enrolment pattern of students.
Courselength - Duration of course(if 50 minutes or 90 minutes or 150 minutes)	Glocker, 2009; Ewer, 2002; Austin and Gustafson, 2006
Class size of the course	Hoxby,2000; Krueger, 2003
Disciplinegrade - Average course difficulty of all courses in the discipline.	Vialardi et al.2010
Student attendance	Devadoss,1996
Prior learning variables	Cortez andSilva,2008
Student background	Astin 1982;Woodman,2001

Gender	Knight, 1994; Adelman, 1999; Boero, Laureti & Naylor ,2005
Race	OCSA, 1996