# *Context Driven Data Mining to Classify Students of Higher Educational Institutions*

Subhashini Bhaskaran Sailesh (Author1)
Phd Student, Brunel University, London and Ahlia University, Bahrain
Subhashini.Bhaskaran@brunel.ac.uk

Dr. Kevin J Lu (Author2)
Lecturer, Business analytics, Brunel University, London
Prof. Mansoor Al Aali (Author3)
President, Ahlia University, Bahrain

**Abstract— Literature shows that knowledge about contextual factors associated with student time to degree and CGPA could play an important role in enabling HEIs to make more accurate and informed decisions that enhance student learning. It is also seen that such knowledge could be discovered using data mining process hidden in past data of students and used for prediction of student performance as part of the decision making process. In line with this argument in this study time to degree (total number of semesters taken to graduate) and CGPA of students were studied taking into account course difficulty and semester as contextual factors. CRISP-DM process was employed to mine student data. Results showed that classification could be used as the model for understanding about student course taking pattern, CGPA, course difficulty and semester and optimize the student time to degree in terms of the course taking pattern, course difficulty and semester to achieve best CGPA. The student data pertaining to a single programme of a single university were mined. Possible decisions in terms of student categorization based on course taking pattern, course categorization based on course difficulty, student advising and provision of learning support could be taken by using the outcomes of this research.**

**Keywords—HEIs; Data Mining; KDDM; Time to Degree; Student Performance; Context-Awareness**

## I. INTRODUCTION

Many organizations including higher educational institutions (HEIs) have been striving to improve their decision making process (Pheng & Arain, 2006). Through that decisions making process HEIs want to enhance learning experience of students, their satisfaction and retention. However student satisfaction is linked with student performance (Lyons, 1999; Roszkowski & Ricci, 2005; Schreiner, 2009), hence it is vital for HEIs to find ways to improve their decision making process to improve student performance. Enhancement of student performance in terms of optimization of time to degree, course taking pattern and GPA are found to be major areas of concern of HEIs (Knight,2000). Moreover, certain factors including course taking patterns, time to degree, and knowledge about contextual factors are found to be influencing decision making processes in HEIs by researchers (Lotkowski et al., 2004) although such findings have not been found to be generalisable or established in different contexts. In addition, literature shows that the current level of understanding on how these factors could be used to make useful decisions to enhance student performance is not very clear, an aspect that suggests that further examination of these factors and their relationship to student performance is needed (Bowen et al., 2009).

A major source that could be used to extract knowledge regarding factors that influence decision making process in HEIs is past student data (Phang, 2013). Extant literature shows that some knowledge required to make decisions can be extracted from data resident in data warehouses in 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 available as a part of student data and warehoused using computer systems. Such knowledge is easily observable, understandable and usable for making decisions such as grouping of students and student advising. But there could be unobservable knowledge hidden in student data set which when extracted and applied to the decision making process could produce more accurate decisions leading to greater enhancement in the performance of students. For instance literature shows that there can be hidden knowledge in student dataset about patterns of courses (Kovacic,2010;).That is to say if knowledge about course patterns is extracted from the student dataset, it can support decision makers in HEIs to understand how improve student performance in many ways, for instance grouping courses and students based on the pattern and also identify associated factors like the time of their registration (e.g. which semester or year) or course features (e.g. course difficulty) to determine the most optimum time to degree. Obviously if knowledge hidden is not brought to the fore it cannot be used and it is possible HEIs have ignored such knowledge unknowingly which may have profound effect on the accuracy and usefulness of the decisions made. For instance when students are advised and grouped in sections, such advising and grouping if supported by knowledge about a set of courses or pattern of courses that may have influence on CGPA or student time to degree, then there is a greater chance that students could be provided with a more precise advise or grouped based on that knowledge leading to better CGPA and more optimized time to degree. On the other hand, if the advising or grouping lacks such knowledge then it is possible HEIs have overlooked vital factors that could influence the student performance in optimizing their time to degree or enhancing their CGPA or both.

Extraction of hidden knowledge from student data, usually considered as big data, is no ordinary task as it may require special techniques like data mining (DM) to reveal that knowledge. Review of relevant literature shows that there is scope to examine how data mining techniques could be used to discover hidden knowledge (usually called as the knowledge discovery and data mining (KDDM) process) and use that knowledge in decision making in HEIs to improve student performance, particularly in regard to course taking pattern and time to degree (Kovacic,2010). Thus the focus of this research is to identify a data mining technique that could be used to determine the course taking pattern of students and relate them to student time to degree using the extracted knowledge to assist in decision making. In addition this research has examined certain other factors (e.g. contextual factors like course difficulty, time of course registration like which semester or year) not addressed in the literature and not currently used in the decision making process in the HEIs but could affect the student course taking pattern and time to degree and hence the decision making process. In particular this research focused on the contextual factors and their impact on the mined data that has influence on the decision making process, an area of study not well researched in the literature (Vert et.al, 2010). The contextual factors under study in this research are course difficulty, specific semester (time). The end result of this exercise is the classification of courses determined using specific factors like course difficulty, grouping of courses using course difficulty, and evaluation of the pattern of courses using the CGPA and course difficulty pattern. Classification is a pattern and is considered as hidden knowledge that needs to be extracted through data mining. Thus in order to know how this can be achieved using KDDM process, the next section discusses about the KDDM process and how it could be used for this research.

### About knowledge discovery and data mining

Raw data cannot provide knowledge unless it is interpreted to derive information. In turn information becomes knowledge if it is put to use. These arguments point towards the need to define data, information and knowledge. Although the knowledge used by HEIs in decision making is extracted from the data that resides in data warehouses created by the HEIs, the extraction processes largely use simple computer algorithms or query or manual ways. However such knowledge extracted using those processes seem to lack depth and might have excluded essential knowledge that is hidden in the data that could not be extracted by those processes mentioned above (Fayyad, 1996). The result is the use of incomplete knowledge in decision making processes in HEIs. For instance course taking pattern of students of HEIs characterized by contextual information associated with such patterns could not be extracted from student dataset of HEIs, using simple computer algorithms or query or manual ways. As mentioned earlier any decision made in HEIs with regard to student performance that does not involve contextual knowledge could be incomplete.

To overcome this problem researchers have developed sophisticated processes for knowledge extraction and discovery of hidden knowledge in the recent past, leading amongst which is the data mining process, also known as knowledge discovery data mining process (KDDM process). Even though significant advances have been made in developing data mining processes, literature points out that there is hardly any evidence of the use of data mining processes in decision making in HEIs. One of the reasons for this state of affairs could be the lack of consensus amongst the researching community with regard to identifying a single standardized DM process that could be used in all contexts. While data mining has been argued to be a very useful process in knowledge discovery that is hidden in large data sets as well as extraction of discovered knowledge, the availability of a wide variety of data mining processes having significant differences among them can cause difficulty for the users in determining which one is the best amongst them. Besides, each process has its own advantages, disadvantages and limitations and there is no one DM process that fits 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 that could clearly guide the HEIs in choosing an appropriate data mining process for implementation and use the outcomes from those processes for decision making. Therefore the objective of this research is to improve decision making process in HEIs related to student time to degree by classifying courses and students using such contextual information as course difficulty, semester and CGPA using a randomly chosen KDDM process for demonstrating the relationship between course taking pattern and CGPA evaluated against the course difficulty in a particular semesters.

Thus this paper has contributed to the understanding of the importance of contextual factors in the decision making process of HEIs using DM at the methodological level. At the practical level this paper contributes to improving the quality (e.g. accuracy and adequacy of decision making leading to better learning experience of students) of decision making at HEIs related to classification of courses, classification of students, optimization of time to degree and enhancement in the performance of students (CGPA) using student data and discovery of course registration pattern hidden in the student dataset.

## II. RELATED LITERATURE

### *A. Critical factors that affect decision process in HEIs and the importance of DM in that process*

Some of the factors that have been found to affect the decision making processes in HEIS in the literature are 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) and course-taking patterns (Ronco, 1996; Bahr, 2010;Kovacic, 2010; Vialardi et al., 2011). Although these factors have been found to be critical to decision making in HEIs, literature is silent on how course-taking pattern as a factor could affect student performance (CGPA) in terms of time to degree. In addition, literature has highlighted that data collected by HEIs on these three factors (viz. course-taking patterns, student performance (GPA) and time to degree) as part of the student data, can have hidden knowledge that could be used in decision making in HEIs implying that data mining as a process could be used by HEIs as a supporting mechanism in their decision making process. For instance hidden knowledge could be the optimum time to degree determined based on the set of courses registered in by students and generated as a pattern and classified based on course difficulty in a particular semester. But research outcomes produced in this area have hardly suggested conclusive DM ways by which course taking pattern could be used as a factor to predict student performance and the time to degree as part of the decision making process in HEIs taking into account the course difficulty as a contextual factor an important gap. This implies that there is a need to understand how course-taking pattern affects student performance and their time to degree and whether data mining process could be used to discover knowledge hidden in student data including contextual factors. This research addresses this issue by studying a decision making process in HEIs pertaining to enhancement of student performance in terms of time to degree and cumulative grade point average (CGPA) as a business case. Using extant literature (see section 2 A) it is argued that in order to enhance the performance of students, the researcher could hypothesize that there is a pattern of courses taken by students in semesters that acts as the independent factor, knowledge about which if extracted from student data, could enable the students and institutions to determine the most optimum time to degree and achieve the best possible CGPA. It is found from the literature that this aspect has not been well understood especially using the methodology of data mining. Thus the critical factors that were considered in this research were: course taking pattern of students, time to degree and CGPA. However as explained earlier these critical factors need to be studied along with contextual factors for better decision making and hence this research has chosen course difficulty, and semester as supporting factors for study. These factors have been described in Section 2 A. In general it is seen that the effect of contextual factors on the decision making process in HEIs alongside critical factors, is an area not addressed in the literature (Vert et al. 2010).

Furthermore, student registration pattern and its linkage to student performance and time to degree has been identified as a complex phenomenon in literature and usual processes such as

Query processing tools, OLAP, artificial intelligence and machine learning have not been found to be useful to determine such patterns (Vialardi et al. 2011). In such a situation some (e.g. Fayyad et al., 1996) suggest the use of KDDM process to discover hidden knowledge in terms of patterns from large databases for decision making. Such a discovery could enable HEIs to predict the course taking pattern of students and making decisions related to optimization of time to degree, grouping students as well as advise students on the courses to be registered. For instance it may be possible for some student advisers in HEIs to decide on the set of courses the student should register in a semester based on the knowledge gained from the pattern of courses identified using data mining process. Such a decision could enable advisers to guide the student to choose those courses for registration in particular semesters which in turn could make the student graduate within an optimum time to degree and with the best possible CGPA.

### *B. About contextual factors*

The different types of context factors that affect HEIs in decision making related to student performance are defined in the following Table 1:

| Type of context | Example of contexts used in HEIs | Includes details about |
|---|---|---|
| Domain Context describes domain specific context | Student Context | The student who is subject of the business problem or understanding; his/her background; previous education; family details; and income level. |
| | Course Context | The course like course description; credits; weightage; course type; course importance; timings; class size; and class location. |
| | Faculty and Teaching Context | The faculty; background; experience; education; and different kinds of teaching methods or techniques. |
| | Student Transcript Context | The student transcript like student grades of the courses taken in semesters; semester GPA; and passed credits in a semester |
| | Student Graduation Context | The student graduation; final GPA (CGPA); time to degree; and destination. |
| Data Mining Context defines the characteristics related to the data-mining task | Data Context | Concerns related to the dataset to be used for mining process |
| | Attribute Context | Attributes using which the prediction has to be made. |
| | Performance Context | The time consumed for data mining process. |

Table 1: Types of context

At this point it is necessary to note that identification of contextual attributes that could have bearing on the pattern of courses that the students opt to register in, the student's time to degree and CGPA. This is a major challenge as student data has seldom been studied to define contextual attributes. Therefore it is seen that on one hand student data needs to be mined to discover hidden knowledge containing contextual information whereas on the other a data mining process model needs to be chosen that can enable the discovery of such knowledge with regard to HEIs.

### *C. Overview of KDDM Process to mine contextualised dataset*

The (KDDM) process is a multiphase process that includes: business understanding (also sometimes referred to as domain understanding), data preparation, modeling, evaluation and deployment or implementation phases (see Figure 1). KDDM processes developed are many. Some of the widely used KDD processes found in the literature include process centric view of KDD proposed by (Fayyad et al., 1996); practical view of KDD with a human centred approach proposed by (Brachman and Anand, 1996); process model which emphasised the cooperation of a data mining analyst and domain expert proposed by (John, 1997); Cross Industry Standard Process Model (CRISP-DM) developed by (Chapman et al., 2000); and SEMMA process was developed by the SAS Institute (SAS, 2008).

Literature shows that KDDM process models have limitations and a single process cannot be used to address every business activity (Fettke et al., 2012). In addition pros and cons of using any particular model need to be known prior to the selection and implementation of a KDD process. One of the most significant limitations of the KDD process is the lack of contextual information that can be associated with patterns discovered. Thus there is a need to introduce the concept of context at some point in the KDD process which is discussed in the following paragraph.

Context driven processing is motivated by the surroundings and semantics of meaning describing an event.

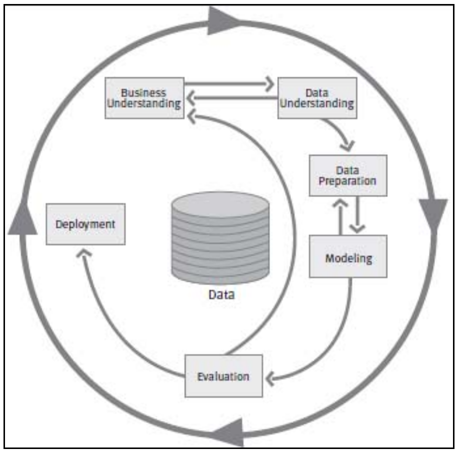


Fig 1: CRISP-DM model described in the literature (Adapted from Chapman et al., 2000)

An example of an event could be explained using spatiotemporal dataset which is expected to explain about a spatial or temporal happening that might take place at a particular time t and location x (Rao et al., 2012). For instance disasters like floods or earthquakes can be considered as spatiotemporal events and datasets containing information on these events can be considered to describe the time and location at which the event has taken place. Applying similar arguments to the case of student performance in HEIs (time to degree and CGPA are temporal factors and HEI is spatial factor), this research aims to mine student dataset and discover knowledge using the pattern of courses that the students opt to register in a semester in order to achieve the optimum time-to-degree and optimum CGPA in a specific context. Such knowledge could enable HEIs to make high quality decisions which are expected to be better than those made using existing processes.

## III. IMPLEMENTATION OF CRISP-DM TO DISCOVER STUDENT COURSE TAKING PATTERN CHARACTERISED BY CONTEXTUAL FACTORS

Based on the above arguments the concept of context processing is introduced in a chosen knowledge discovery process namely CRISP-DM process. The choice of CRISP-DM as the DM process is based on the fact that this data mining process has been acknowledged by researchers as a widely used process in the industry only and sparsely used in academia despite the advantages it offers.If CRISP-DM process can be applied to academic area, it may be possible to gain an understanding about the utility of this process to the academia. The CRISP-DM model provided in Figure 1 is proposed for application in this research.

In order to implement CRISP-DM methodology, IBM SPSS Modeler 17.0 was used in a limited manner. Important aspects that enabled the choice of IBM SPSS Modeler include the availability of a project tool (CRISP-DM) which helps to organise streams, output and annotations depending on the phases of data mining project, possibility to produce reports at any time during the project depending on the notes provided by the user for streams and CRISP-DM phases, and availability of guidance in terms of task lists for each step. In addition, Weka Explorer was used wherever needed in the modelling phase because of lack of facilities in IBM SPSS Modeler 17.0. The stages involved in the DM process (see Figure 1) were used. Data extracted from the Student information system of a university in Bahrain was used with permission. Data pertained to graduated students of Bachelors of Science in Accounting and Finance. Initial data set included 337 student records. The dataset contained information about student grades, registered courses and course details for the period 2003 to 2014. Some fields were readily available, while some were computed, for instance CGPA and time to degree. In addiction the following specific information was used from the dataset

- Students table – this database had all information about current and graduated students. As soon as a prospective applicant becomes a student the details are stored in the student database. No record from this data base was deleted.
- Course table – this is a table that had all information about the current and old courses that were offered to students.
- Semester table – this is a table that had all information about the semesters in which the students studied.
- Course registrations table – this is a table that had all information about the courses registered by students in various semesters (course taking pattern)
- Programme table – this is a table that had all information about the programmes that were and are currently offered in the university
- Programme plan – this is a table that had all the information about the curriculum of the programme.
- Transcript table – this is a table that had all the courses, grades scored by student in each course, semester GPA, semester completed credits and CGPA of all students.

Contextual factors (also called attributes) needed to support the critical factors were extracted from the dataset. In doing so the researcher relied upon the research outcomes produced by (Vert et al. 2010). The contextual factors extracted from the data set included contextual dimensions (eg. time, span, impact and similarity), information critical factors (time period, criticality, impact, spatial) and quality factors (how recent is the data, ambiguity and contradiction). However only specific and relevant factors were studied.

Each one of the contextual factors chosen for study has been described below.

***Contextual dimensions***

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

***Information criticality factors (ICF).***

- time period of information collection (last 5 years graduate data from 2009-2014 was taken by deleting the old graduates).
- impact (e.g. semester GPA, passed credits in a semester)
- ancillary damage of miss classification (e.g. CGPA and time to degree).
- spatial extent data set coverage (e.g. College of business and finance).

***Quality of the data***

- currency (data collected over the period 2009-2014), how recently was the data collected, is the data stale and smells bad (last 5 years graduate data was taken removing the old graduates).
- ambiguity (e.g. course difficulty, course weightage, course length, course discipline difficulty and course level)
- contradiction, what does it really mean when conflicting information comes in different sources (Student potential)

The resulting set of contextual variables extracted from the student set studied included course difficulty, course weightage, course level, semester GPA (SGPA) and semester.

Amongst these variables only course difficulty and semester were used for making decisions using mined data. Once the contextual factors were identified, the DM process was initiated. The outcome from each one of the steps involved in the CRISP-DM process is explained next using the definition of the steps provided by Chapman et al. (2000).

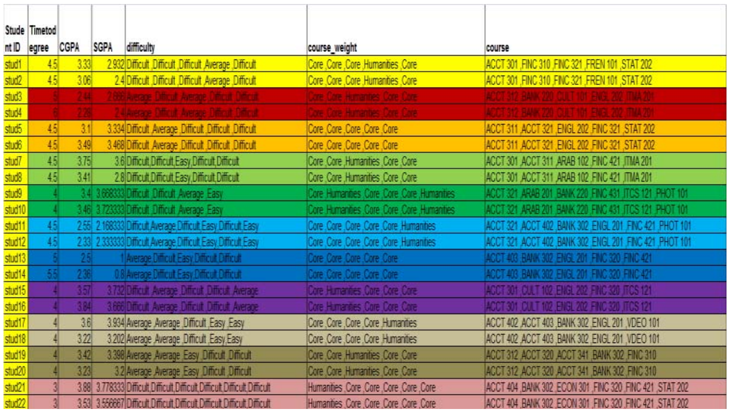
Business Understanding: To predict the optimum time to degree taking into account course taking patterns, CGPA and contextual factors namely course difficulty, Semester GPA to classify courses and students using discovered knowledge.

Data Understanding: This involved collection of data from student registration system of a university in Bahrain. This was achieved using SQL queries. The result was a dataset that could be mined to address the business problem. An examination of the dataset revealed possible relationship between CGPA and the student course registration data by semester. The dataset was assessed for quality problems including incomplete values, extreme values.

Data Preparation: This step involved the selection of student records from the dataset, cleaning the data, constructing, integrating and formatting the data. The resultant dataset contained limited information related to courses registered, semester details, time to degree, course difficulty and SGPA. Course difficulty was computed using the formula recommended by Zainudin(2012). The dataset contained 50.

Modelling: This step involved the use of modeling algorithm that enabled selection of modeling technique, generation of test design, building a model and assessment of the model. The modeling technique chosen was classification and the chosen algorithm is Genetic Algorithm which is a heuristic search algorithm that uses cross over, mutation and fitness function to generate the pattern.As part of the test design a training, test and validation data sets were generated. The model was built based on the following parameters: CGPA $\leq 4$, time to degree $\geq 3.5$ and $\leq 6$, programme under consideration – bachelors in accounting, number of courses registered in $\geq 4$, semester = 3 and course difficulty $< 1$. The final model is a table with student records and fields (CGPA, time to degree, semester and course difficulty) (see figure 2) which has columns studentid, time to degree, CGPA, SGPA, difficulty, course weight and course taking patterns.

The course taking patterns were assessed using course difficulty as a measure to evaluate. The course difficulty was measured using a 5 point scale i.e. very difficult, difficult, average, easy and very easy. The assessment showed that the model built is accurate as the measurement of course difficulty with respect to each course in the pattern of individual students matched for every course and was consistent across the courses in which the students registered.



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

Fig 2: Classification of pattern of courses achieved using Genetic Algorithm

Evaluation: Evaluation involved the summarization of assessed results in terms of achieving business goals. In this research the business goals to be achieved were to predict the optimum time to degree taking into account course taking patterns, CGPA and contextual factors namely course difficulty and semester to classify courses and students using discovered knowledge. It can be seen from previous step that students can be classified into 3 major time to degree categories that is 3, 4 and 4.5. The number of students who could be accounted under these categories is 18 out of 22. Amongst this 2 students achieved a time to degree of 3 whereas 8 students apiece achieved a time to degree of 4 and 4.5 respectively. The remaining outcomes pertain to students who have achieved time to degree exceeding 4.5 with 1 student achieving 5, 2 students 5.5 and 1 student 6. The last 4 have not been discussed in this research paper with regard to the business goals as more interesting patterns of courses have emerged with 18 students whose time to degree fell between 3 and 4.5 years. An analysis of the 18 students' pattern showed that the highest GPA of 3.88 has been scored by a student with code stud21 who has registered in 6 courses in semester 3 with all the 6 courses having course difficulty measured as difficult and achieved a time to degree of 3. In another student case with student code stud21 the same set of courses as the one taken by student stud22 has been found in the course taking pattern and has achieved a time to degree 3 and CGPA of 3.53. Results of stud21 and stud22 indicate the best results using which decisions can be taken. In addition other inferences that can be made are:

*1)* Even with the maximum of 6 courses registered in a semester, with all courses classified as difficult, 1 student has scored the highest GPA in the list of 22 with shortest time to degree. This implies that the course taking pattern comprising 6 courses could lead to the shortest time to degree with set of courses measured as difficult. Course difficulty has no impact on CGPA and time to degree.

*2)* While the first inference could be considered exceptional because only 2 students seem to fall in that category a set of 8 students were found to have achieved a time to degree of 4 years but in all these cases the number of courses taken as a pattern was only 5. This result is slightly counter intuitive for even with 5 courses in a semester the student could have completed the degree within a time to degree that is less than 4 years. However that is not the case. Additionally even with less number of courses (5) found in the pattern the students have not scored a GPA exceeding 3.88 that was achieved by stud21, which is also an anomalous situation. Lesser number of courses in a pattern could provide better opportunity for a student to score a higher GPA as the course load is lower than those students who have achieved a time to degree of 3 who had registered for six courses. Besides, an important aspect that emerges from the inspection of the model with regard to the 8 students is that the course pattern shows that the course difficulty measured for the courses found in the pattern were within only three points namely difficult, average and easy. This is not the case with the 2 students who achieved lower time to degree of 3 and higher GPA of 3.88 and 3.53, in whose case the course difficulty measured was related to only one point in the course difficulty scale namely "difficult". Thus lower number of courses, having a pattern of courses whose difficulty measure is a mixture of difficult, average and easy does not imply lower time to degree or higher GPA which is in contrary to the common belief in academia. Generally it is believed that if a student registers in less courses in a semester then higher GPA could be scored. This is not the case. One more point that needs to brought out is that when the course taking pattern is exactly the same with regard to two or more students the CGPA is not same even though the time to degree and course difficulties are the same. This implies that even if the course taking pattern is the same for any two students, while achieving the same time to degree the students need not achieve the same CGPA. Therefore it is not possible to conclude that a particular set of courses only will contribute to a particular time to degree or CGPA. Similar arguments could be made with regard to students who have achieved a time to degree of 4.5.

*3)* The most significant point that emerges is that students when categorized based on the course taking pattern do not perform in the same way as it can be expected. This implies that time to degree can be 3 years or 6 years regardless of the course taking pattern as found in the case one set of two students (stud21 and stud22). Alternatively it is also possible to categorise students who have registered in five courses considered as optimum but whose course taking pattern contains common courses and course difficulty of the five courses measured as difficult and average as indicated in Table 1. This implies that course difficulty is an important factor that affects time to degree and CGPA. Such students could be achieving time to degree either 4 or less than 4. Infact HEIs could motivate such students to complete their programme within a time to degree 4. For instance students with codes stud15 and stud16 who have registered for atleast 5 courses in semester 3 and whose course taking pattern shows that they have registered for courses whose difficulty fell in the range between difficult and average have achieved a time to degree 4 and a CGPA more than 3.57 and 3.84 respectively could be encouraged to approach the performance of students whose time to degree is 3 years (stud21 and stud22). Encouragement could include advising students stud15 and stud16 to register in courses similar to the ones of students and stud21 and stud22 with same course difficulty. Special attention could be given to those students so that their performance could be improved. HEIs could take a decision in categorizing such students and courses so that more number of students could achieve best CGPA a time to degree ≤ 4.

## CONCLUSION AND LIMITATIONS

The outcome of this research shows that HEIs can determine the pattern of courses students could register in each semester that could enable them to graduate within a time to degree that is less than 4 years. While analyzing the course difficulty of the courses found in the student course taking pattern (which should approach the measure difficult) and number of courses found in the pattern either being equal to 5 or 6, less than 4 years appears to be the most optimum time to degree that could be achieved by students with best possible CGPA regardless. Where students are falling in other categories where the course taking patterns differ from those students achieving the most optimum time to degree, further categorization of students could be done by HEIs alongside the categorization of courses based on the measure of difficulty. Similarly HEIs could make decisions on grouping students who have very similar course patterns with a minimum of 5 courses in each semester and provide additional support in learning to ensure that they achieve a better CGPA and optimum time to degree. Another significant finding is that none of the course taking patterns discovered and classified, contained a course whose measure of difficulty approached either the ‘very difficult’ level or very easy’ level, indicating that generally students who have achieved a time to degree of $\leqslant 4.5$ do not find the courses very difficult. The results of this research clearly enable prediction of time to degree of students which could be $\leqslant 4$ such using different student categorization scenarios in terms of their course taking pattern and the difficulty of courses they register in to achieve the most optimum CGPA In addition students can be advised suitably to choose courses that could be determined based on the course difficulty so that they score a higher CGPA than what they normally do. Finally to achieve lower time to degree the HEIs could decide on how to provide learning support to students depending on the categorization. Limitations of this research include that the dataset pertaining to only one semester and 50 students was analysed and only two contextual factors were considered for study. Future studies should include more semesters, more number of students and other potential contextual factors such as student potential, course weightage and course difficulty pattern in their study to determine course taking patterns for achieving the optimum time to degree and the best CGPA.

## ACKNOWLEDGMENT

The authors acknowledge the support rendered by Mr.S.Gowrishankar, Ahlia University, Bahrain in the process of developing the paper.

## REFERENCES

[1] 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.

[2] Astin, A., 1971. Predicting academic performance in college: Selectivity data for 2300 American colleges.. New York: The Free Press.

[3] Athiyaman, A., 1997. Linking student satisfaction and service quality perceptions: The case of university education. European Journal of Marketing, 31(7), pp. 528-540.

[4] 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.

[5] Bolchini, C. et al., 2007. A dataoriented survey of context models. SIGMOD Rec.(ACM) doi:10.1145/1361348.1361353. ISSN, 36(4), pp. 19-26.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] Davenport, T. . H., 2010. The New World of "Business Analytics", s.l.: International Institute for Analytics.

[11] 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.

[12] 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.

[13] Fayyad, U., Piatetsky-Shapiro, G., Smyth, P. & Uthurusamy, R., 1996. Advances in knowledge discovery and data mining.. MIT Press..

[14] 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.

[15] Helgesen, O. & Nesset, E., 2007. What accounts for students’ loyalty? Some field study evidence.. International Journal of Educational Management, 21(2), pp. 126-143.

[16] Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design science in information systems research. MIS Quarterly, 28(1), 75–105.

[17] John, G. H., 1997. Enhancements to the data mining process. s.l.:PhD thesis, Stanford University.

[18] 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.

[19] Kovacic, Z. J., 2010. Early prediction of student success: Mining student enrollment data. s.l., Proceedings of Informing Science & IT Education Conference .

[20] Kurgan, L. & Musilek, P., 2006. A survey of knowledge discovery and data mining process models.. Knowledge Engineering Review, 21(1), pp. 1-24.

[21] 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).

[22] 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.

[23] 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.

[24] Minaei-Bigdoli, B., Kashy, D. A., Kortemeyer, G. & Punch, W. F., 2003. 33rd ASEE/IEEE Frontiers in Education Conference. Boulder,CO, IEEE.

[25] Pheng, L. S. & Arain, F. M., 2006. A KNOWLEDGE-BASED SYSTEM AS A DECISION MAKING TOOL FOR EFFECTIVE MANAGEMENT OF VARIATIONS AND DESIGN IMPROVEMENT: LEVERAGING ON INFORMATION TECHNOLOGY APPLICATIONS. ITcon, Volume 11.

[26] 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.

[27] 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.

[28] Saraiva, P., Lourenço, L. and Louro, A.I.C.P., 2015, August. Quality Assessment in Higher Education Institutions. In *Toulon-Verona Conference" Excellence in Services"*.

[29] SAS, I., 2008. SEMMA data mining methodology. [Online] Available at: http://www.sas.com

[30] Schilit, B., Adams, N. & Want, R., 1994. Context-Aware Computing Applications. s.l., First International Workshop on Mobile Computing Systems and Applications.

[31] 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.

[32] Tinto, V., 1975. Dropouts from higher education: A theoretical synthesis of recent literature. A Review of Educational Research, Volume 45, pp. 89-125.

[33] 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.

[34] 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.