Contextual Data Mining for Higher Educational Institutions

Subhashini Sailesh Bhaskaran, Mansoor Al Aali, Kevin Lu

Abstract: Context-awareness research has been carried out by many researchers. However, little has been done in building a context-aware data mining methodology which supports decision-making in HEIs. The usefulness of knowledge discovery data mining process (KDDM process) in HEIs were investigated to discover hidden knowledge that is contextualized, resident in student datasets and use them in decision making. It was experimented and found that not any of the KDDM processes include a contextual factor mining stage that is essential to take out hidden knowledge from datasets described by contextual factors. Therefore a new process was introduced in KDDM process that uncovered contextual data to be used to support business goal and produced a dataset at the preparation stage which generated data mining model that was contextual leading to the unearthing of course taking patterns that are contextualized. This discovery has enabled forecasting of optimum CGPA and time-to-degree.

Index Terms: HEIs, Data Mining, KDDM, Time to Degree, Student Performance, Context-Awareness.

I. INTRODUCTION

In the delivery of education in HEIs decision making processes assume significance. Those decision making processes are concerned with a number of activities in HEIs including teaching and student learning experience. For instance decisions could impact such aspects as curriculum design, categorization of students, timetabling improvements and student assessment that can be related to teaching and student learning experience (BIS, 2014). However accuracy and applicability of those decisions in HEIs are areas of major concern, because common decisions that are taken in HEIs at the institutional level do not always get implemented at the academic level (BIS, 2014). One reason for this could be that decisions usually made in HEIs use data that contain observable patterns or phenomena not entirely adequate to make accurate decisions.

Observable phenomena include data and information about student gender etc. However literature points out that there is a need to consider unobservable phenomena also that could have serious implications to the decision making process. Unobservable phenomena could include data and knowledge associated to student education (for instance, student potential, course difficulty). A major phenomenon that is found to be unobservable is the contextual factor that could contain knowledge helpful in HEIs decision making process. For instance when decisions are being made in HEIs regarding optimizing the student time to degree, knowledge related to course difficulty or student potential could help HEIs in determining the categorization of students and courses as well as providing additional support to students who need support. While literature review shows that contextual factors have the potential to enhance the quality of decision making, extracting such contextual factors from known resources and discover useful knowledge about those contextual factors has been a challenge (Vert et al., 2010). One resource that is very promising that could be used to extract contextual factors and knowledge about those contextual factors is student dataset. Recently researchers have started to focus on contextual factors that are important to HEIs that could be used in their decision making process although more needs to be done. One such area that is clearly providing evidence of this kind of interest and provides a large scope for investigation is Knowledge discovery. However serious questions are raised on the usefulness of existing KDDM processes to enable decision making involving contextual factors related to a particular business environment like HEI. For instance Vert et al (2010) argues that current KDDM processes need to be enhanced for their usefulness in supporting accurate decision making using contextual factors, an argument that implies that current KDDM processes are not capable of addressing unobservable knowledge that is contextualized. Some argue that further research is needed in this area to support users of KDDM processes (Singh, 2003; Vajirkar, 2003). Thus on the one hand there is a need to use KDDM processes to discover knowledge and on the other those KDDM processes must be enhanced to deal with contextual factors leading to the discovery of knowledge that is contextualized and useful to make more accurate decisions.

In the context of HEIs the current level of understanding of the decision makers is limited to visible and tangible factors available in the student data set (For e.g. Gender, age, nationality, grades). Most HEIs are yet to deploy any type of KDDM process to discover hidden knowledge in their decision making process. For instance, a question could be raised regarding predictability of a pattern of courses a student could register in a semester or semesters that would lead the student to achieve a lower time to degree (the time taken by a student to graduate). Such a question is unlikely to be answered straightaway using the data readily available in the dataset like student factors namely semester based number of courses enrolled or Grade Point Average (GPA) scores as the knowledge
related to pattern of courses that can be used to predict the
time to degree is not readily available in the dataset. Here
KDDM processes could be very useful although HEIs do not
appear to have deployed such processes yet in their decision
making process. The reason for this situation are that
manual processes or commonly used computing methods(e.g.
SQL queries) are not adequate enough to discover patterns
from hidden patterns from large datasets and HEIs lack
knowledge to extract those patterns from student dataset
using KDDM process. Moreover the use of extracted patterns
from student dataset using KDDMs in decision making can
suffer in terms of accuracy or achievement of business goals
as the extracted patterns lack any knowledge related to
textual factors. The main reason for this is that KDDM
processes that are currently used in the business domains or
discussed in the literature, that have the potential to be used
in HEIs, do not have steps to integrate contextual knowledge
concealed in the dataset, needed for making better decisions
to attain the business goals. This is a major gap in the
literature. Thus the objective of this research is to address this
gap to some extent by demonstrating a modified KDDM
process that extracts hidden knowledge from the student
dataset characterized by contextual factors that could be used
in making better decisions in HEIs to achieve specific
business goals. Towards this, in this research two contextual
factors namely course difficulty and semester have been
taken as examples to produce course taking patterns from
student dataset to forecast optimum time to degree and
Cumulative Grade Point Average (CGPA). This knowledge
about course taking patterns could be useful in taking many
decisions in HEIs that have bearing on teaching and
learning, for instance course categorization, student
categorization, forecast of optimal time to degree of students
and achievement of optimum CGPA.

II. RELATED LITERATURE

A. Review Stage

This section covers two aspects. One related to current
knowledge on contextual factors and their utility in KDDM
processes and the other related to KDDM
processes.Literature review shows that contextual factors
have been of interest researchers in various fields. TableI
provides a glimpse of current efforts of researchers taken in
different fields that utilize contextual factors in KDDM
processes to extract hidden knowledge from datasets that are
textualized.

<table>
<thead>
<tr>
<th>Meaning of Context</th>
<th>Author and Paper</th>
<th>Remarks</th>
</tr>
</thead>
</table>
| 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. | Contextualized Analytics in
Big Data World: Better
Decisions, An IBM® Redbooks®
Point-of-View publication | By using context analytics
organizations can derive
value from the data and
relationships from unstructured
data and related structured
data. These insights can help an
organization to make fast
decisions to anticipate and shape
business outcomes. |
| A general approach for context-aware adaptive
data mining of data streams that
aims to dynamically and autonomously adapt data
stream mining parameters according to changes
in context and situations. | ContextAware Adaptive
Data Stream Mining, Haghi, et al., 2009 | The researchers proposed
an overall method for context-aware adaptive
data mining that includes
textualization into universal data stream
mining and allows real time examination of data
on board mobile devices in a
coherent and effectiveness manner. They achieved
Contextualawareness through Fuzzy
Logic Inference System (FLS) that assimilates fuzzy
logic in the CS model, an
official context modeling
and cognitive approach for
assisting pervasive
computing environments. |
| Context as a concept has been identified as important
in the field of KDDM recently, due to the potential it has in
adding value to the discovered knowledge. Lack of
context in the discovered knowledge has been argued to be
a limitation in various KDDM processes by
researchers (Schilit et al., 1994; Dey, 2001; Bolchini
et al., 2007) due to the important role context could play in
decision making. (Vert et al., 2010)claim that a dataset
characterized by contextual attributes if fed to the
KDDM processes at a definite stage, then it is possible
that the knowledge discovered through the mining
processes will not only contain the normal attributes but
also the contextual attributes. These arguments could be
extended to the HEIs also. Thus based on these
arguments and as explained earlier there is a necessity to
investigate the usefulness of KDDM processes in HEIs
to discover hidden knowledge that is textualized,
resident in student datasets and use them in decision
making. Notably if a method to extract contextual
factors is identified and employed in KDDM processes
then it is probable to dig out hidden knowledge that is
textualized for better decision making. Furthermore,
amongst the various factors that have been discussed by
researchers in supporting decision making process in
HEIs, prediction of optimum time to degree and CGPA
by means of students’
course taking patterns in
association with contextual
factors namely course difficulty and semester do not find place. While there are arguments supporting a linkage between course taking patterns, CGPA and time to degree, there is hardly any research outcome that has substantiated this claim using data mining techniques particularly considering semester and course difficulty as contextual factors. This research aims to address this aspect.

B. KDDM Process

A number of KDDM processes have been used in many domains namely medicine, manufacturing, education and other businesses. Widely used KDDM processes are tabulated in Table 2 which indicates the steps involved in them. While literature shows the usefulness of KDDM processes in extracting hidden knowledge from datasets for decision making, those processes are also found to have limitations. Table 3 provides a good idea about the limitations affecting different KDDM processes.

Table 2, Evaluation of KDDM (adapted from Kurgan and Musilek, 2006)

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Area</td>
<td>Academic</td>
<td>Industrial</td>
<td>Academic</td>
<td>Industrial</td>
<td>Academic</td>
<td>N/A</td>
</tr>
<tr>
<td>No of step</td>
<td>5</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>4</td>
<td>4</td>
</tr>
</tbody>
</table>

Table 3, Limitations of KDDM

<table>
<thead>
<tr>
<th>Process Model</th>
<th>Limitation</th>
<th>Identified By</th>
<th>Modified step or stage</th>
<th>New Concept</th>
</tr>
</thead>
<tbody>
<tr>
<td>KDDM (Catley et al., 2012)</td>
<td>Lack of data collection step which is vital for the KDD techniques in some real applications such as information security and medical treatment</td>
<td>KDDM (2007)</td>
<td>Data collection step using previous mining results, it was added before data selections shown in Fig 2</td>
<td>Introduction of data collection step in data mining process to filter irrelevant data leading to better decision making. (Ji, 2007)</td>
</tr>
<tr>
<td>KDDM (Catley et al., 2012)</td>
<td>Lack of domain knowledge leading to decision making that maybe useful if such a knowledge is not part of the mined data.</td>
<td>Redpath &amp; Srinivasan (2004)</td>
<td>Proposed an architecture based on domain knowledge as shown in Fig 3</td>
<td>Not addressed in the literature</td>
</tr>
<tr>
<td>KDDM (Catley et al., 2012)</td>
<td>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.</td>
<td>Ruan (2009)</td>
<td>Not addressed in the literature</td>
<td>Not addressed in the literature</td>
</tr>
<tr>
<td>KDDM (Catley et al., 2012)</td>
<td>Lack of contextual information in the data mining process</td>
<td>Vert et al. (2010)</td>
<td>Not addressed in the literature</td>
<td>Not addressed in the literature</td>
</tr>
<tr>
<td>KDDM (Catley et al., 2012)</td>
<td>Insufficiency to handle multidimensional temporal data resulting in knowledge that cannot support decision making which is dependent on temporal issues.</td>
<td>Li et al. (2009)</td>
<td>Proposes mode had has 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 as shown in Fig 4</td>
<td>On-Line Analytical Processing (OLAP), six sigma, and domain knowledge (Li et al., 2009).</td>
</tr>
<tr>
<td>KDDM (Catley et al., 2012)</td>
<td>Lack of integrated process model.</td>
<td>Sharma et al. (2012)</td>
<td>Identification of task dependence between tasks of the same phase and different phases in the first row.</td>
<td>Building an integrated process model (Sharma et al., 2012).</td>
</tr>
</tbody>
</table>

The previous information reveals that not any of the KDDM processes include a contextual factor mining stage which is essential to unearth unseen knowledge from data identified by contextual factors. Even though there are few papers that have talked about context knowledge and context driven data mining, such knowledge do not propose how to link data mining to contextual factor to help in decision making in businesses including HEIs that lead to accomplishment of business goals. When such addition could be attained, the resultant KDDM model can have a higher foretelling power which is essential for making better decisions. This has been addressed in this research in a limited manner by adding the CRISP-DM process with a contextual factor mining stage. In this context, this research relies upon KDDM process developed by Chapman. CRISP-DM model is given in Figure 1.
Figure 1, Current CRISP-DM model described in the literature (Adapted from Chapman et al., 2000)

Figure 2, KDDM process model adapted from (Ruan, 2007)

Figure 3, KDDM process model adapted from (Redpath and Srinivasan, 2004)

Figure 4, KDDM process (Catley et al., 2009)
IV. TESTING AND EVALUATION

Use Refer figure 1. A broad summary of the different stages of CRISP-DM process has been provided in table 5 because the detailed discussion on the CRISP-DM process is beyond the scope of this paper. Further due to paucity of space each stage and its components have not been discussed in detail. Instead important components that are necessary to support the experiments conducted in this research are given below (for a detailed description of various components of CRISP-DM refer to Chapman et.al (2000)). In addition this section does not include discussion on any step related to contextual factor as the aim of this section is to fetch to see whether it can produce hidden pattern from the mined data including contextualised pattern.

Figure 6, Current CRISP-DM model steps described in the literature (Adapted from Chapman et al., 2000)

The essential steps of CRISP-DM are given below.

The detailed specification of the stages and components can be had from the authors.(i) Business Understanding: To forecast the optimal time to degree using course taking patterns and CGPA and to categorize courses and students with unearthed knowledge. For this purpose data from student registration system of a university in Bahrain was used. Weka was used as the data mining tool.

(ii) Data Understanding: SQL queries was used to retrieve the data (see table 3). The outcome was a dataset that can be mined to attend the business problem. An investigation of the dataset discovered probable association between semester wise student course registration data and CGPA. The dataset was evaluated for quality problems like extreme values, incomplete values.
Table 3, Data set without contextual factors extracted

(iii) Data Preparation: This stage concerned the choice of records of students from the dataset, clean-up, building, combining and arranges data. The resulting dataset enclosed partial information associated to enrolled student courses, semester information, CGPA and time to degree. The dataset contained 50 entries.

(iv) Modelling: This stage involved the use of modelling algorithm that enabled selection of modelling 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 produce the course taking pattern. As part of the test design training, test and validation data sets were generated. The model was created using the following parameter setting: CGPA ≤ 4, time to degree ≥ 3.5 and ≤ 6, programme under consideration – bachelors in accounting, number of courses registered in ≥ 4 and <= 6 and semester = 3. The final model is a table with student records and fields (CGPA, semester and time to degree) (see Table 5) having columns student ID (student_id), Cumulative CGPA (GPA), time to degree(len), and course taking patterns(course).

The time to degree were assessed using the course taking patterns and CGPA see table 5 which showed that time to degree can be associated courses in terms of registered number of courses and CGPA of students only.

(v) Evaluation: The results in Table 5 implied that the time to degree could be interpreted only in terms of the registered number of courses of students (without an observable pattern) for a particular CGPA, considered the highest amongst students belonging to the same programme (BSAF). However in the absence of an observable pattern the relationship between time to degree and the registered number of courses can only be interpreted using the number of courses registered in to achieve lower time to degree as in a semester as the attribute. However it can be seen that higher the number of courses, lower will be the time to degree which is perhaps obvious regardless of the CGPA which does not require data mining. Such a result will not be very useful to the HEIs to accurately classify the courses or students to allow them to attain optimal time to degree and CGPA because of lack of additional information that could be used to link course pattern to time to degree significantly.

Besides it lacks predictive power because of the obviousness associated with the outcome. Instead if a contextual factor was included in the dataset, for instance course difficulty, then it could act as missing information that could determine a enrolled course pattern of students. In this case the level of course difficulty associated with every course in the set of registered courses of students could be used to determine the optimum time to degree for a given CGPA. For instance student (stud21) has achieved the maximum CGPA of 3.88 possibly because all the six courses could be having a lower difficulty rating. However whether this actually is the case or not, cannot be confirmed due to lack of knowledge about course difficulty in the outcome produced by CRISP-DM process. This point towards the usefulness of contextual factors in discovering knowledge pertaining to course taking pattern. How this can be achieved is discussed in Section V.

(vi) Deployment: The model could not be deployed as evaluation of the outcome of CRISP-DM (see Table 5) showed that the discovered knowledge is not useful to accurately predict the course taking pattern and optimal CGPA and time to degree and classify either the students or courses which is the business goal.

At this point it can be summarized that if the business goal has to be achieved there is a need to discover hidden patterns in the dataset characterized by context which is only possible when the CRISP-DM process has steps to include contextualized dataset and mine it. Thus this research has developed an altered CRISP-DM process that could achieve the above as described next.

Table 5, Course taking patterns generated from Dataset devoid of Contextual Factors Extricated from Student Data

V. DEVELOPMENT OF MODIFIED CRISP-DM PROCESS (STAGE 2)

The modification of CRISP-DM process in line with the summarization provided in the previous paragraph is explained in a step-by-step fashion below taking into consideration the steps described in Section IV. The
modified process is depicted Figure 7 and steps A to F in this figure must be related to the discussions below. However in this section only the design aspects of the modification are discussed whereas the corresponding experimental details, numerical examples and interpretations are provided in Section VI. Hence every sub-section in this section could be directly cross referred to sub-sections (A1 to F1) in Section VI to understand how the design aspects have been translated to.

a) Business Understanding (Step A): This step does not significantly change and remain as explained in sub section (i) of section IV except that contextual factors namely course difficulty is also taken into account in prediction of time to degree.

b) Data Understanding (Step B): With respect to subsection (ii) of section IV this step has been divided into namely Contextual (Step B1) and General Data Understanding (Step B2). While step B1 enable the identification of course difficulty (contextual factor) in the dataset and extraction of hidden contextual data (course difficulty) from the dataset (here a data mining method needs to be introduced to discover contextual knowledge), the step B2 is the same as the one described in subsection (ii) of section IV. The outcome of step A can be seen to be linked to both the steps B1 and B2 in a reversible fashion which indicate a basic modification of the original CRISP-DM process.

c) Data Preparation (Step C): Like in Step B this step also have been divided into 2 subsections which step C1 and C2 with respect to subsection C of Section IV. While step C1 deals with contextual data namely course difficulty, in this step data preparation is conducted in similar line as described in subsection (iii) of section IV. Again step C2 is implemented in the same way as described in sub section (iii) of section IV. It can be seen in Figure 7 that the step B1 is linked to step C1 and step B2 is linked to step C2 in a reversible fashion which is another modification of the original CRISP-DM process.

d) Added Data Preparation (Step D): The outcomes of this step are a dataset that has extra information about course difficulty. This step is not originally part of the CRISP-DM process developed Chapman et. al, 2000. In this step the outcomes of steps C1 and C2 are merged into a single dataset that will be mined further. This is a major modification of the original CRISP-DM process.

e) Contextualized Modelling (Step E): This step although very similar to subsection (iv) of section IV differs from that step significantly as it has been modified to handle contextual data using the additional link provided to connect with step C1 and C2 in the reverse direction as well as using a modified Genetic Algorithm that uses course difficulty data. In addition the model was created The model was created using the following parameter setting: CGPA ≤ 4, time to degree ≥ 3 and ≤ 6, programme under consideration – bachelors in accounting and finance (BSAF), number of courses registered in ≥ 4 but <=6, semester = 3 and course difficulty which is measured using a 5 point scale (very difficult, difficult, average, easy and very easy). The final model expected to be produced is a table with student records and fields CGPA and time to degree (see Table 6) which has columns student identification (student_id), time to degree(len), Cumulative CGPA (GPA), course difficulty(difficulty) and course taking patterns. The time to degree were assessed using the course taking patterns and CGPA see table 5 which showed that time to degree can be associated courses in terms of registered number of courses and CGPA of students only.

f) Evaluation (Step F): Time to degree could be interpreted in terms of the registered number of courses of students, with an observable pattern that could be linked to course difficulty for a particular CGPA, considered the highest amongst students belonging to the same programme (BSAF). The use of course difficulty lies in enabling the interpretation of the most optimum course taking pattern of students that could lead to achieve the optimum time to degree and not by mere number of courses. For instance a student could register in six courses all of which could fall under the category of difficult and yet achieve a time to degree of 3 and the highest CGPA. On the other hand a student could register in six courses that could fall under the
category of course difficulty “easy” yet achieve a time to degree higher than 3 and lower CGPA. In this case it is possible to interpret that despite registering in the maximum number of courses in a semester, with the difficulty levels measured as “difficult”, a student can achieve a lower time to degree and highest CGPA. Such an interpretation could help in identifying and categorizing students who are capable of achieving lower time to degree regardless of the fact that they have registered in courses whose difficulty level could be higher and score high CGPA. The same argument could be used to categorize other students depending on the course taking pattern and the difficulty levels of those courses they have registered in but have taken longer time to degree and scored lower CGPA. This process could be continued to categorize students and the courses they have registered in to determine the level academic support that needs to be provided to enable them achieve optimum time to degree and high CGPA. Thus it can be seen that the business goals of categorising courses and students to predict optimum time to degree and highest CGPA could be achieved accurately using the contextual factors and a modified process which is demonstrated practically in Section VI.

g) Deployment (Step G): The deployment of the modified CRISP-DM process involves an analysis and understanding of the results produced by the process and at this stage a decision to deploy cannot be taken. Hence this section has been addressed in Section VI where results of the analysis can be used to determine the deployment.

In summary it can be seen that the newly designed modified CRISP-DM process has the potential to achieve the business goal of predicting reasonably accurately the optimum time to degree and CGPA in terms of course taking pattern of students and course difficulty leading to decision making about categorizing courses and students and provide better support to students. To verify this aspect the next section provides the details of the experiments conducted and analysis of the results.

Table 6. Course-taking patterns extracted from Dataset with Contextual Attributes mined from Student Data

The input to the KDDM process is the dataset characterised by contextual factor course difficulty. The business goals to be achieved are determining the optimum time to degree and CGPA, categorisation of courses that could predict optimum time to degree and CGPA using course registration patterns and course difficulty. Categorisation of students based on optimum time to degree, CGPA, course patterns and course difficulty.

Data preparation and data understanding stage covers the data quality factors of data. For detailed steps, readers have to refer to chapman et.al(2000). Genetic algorithm was used with knowledge of contextual factors course difficulty.

Model evaluation stage comprises of pattern extracted through the process with and without contextual detection stage.

The overall CRISP-DM process has not been discussed in detail and the discussion here is limited to demonstrating the difference between the knowledge discovered using patterns extracted through CRISP-DM process with and without the detection of course difficulty and semester as contextual factor. As far as the introduction of contextualised dataset was concerned initially the characteristics of course difficulty pertaining to every course under consideration in a particular semester 3 were utilised which included 'course GPA', 'weighted average of the course GPA', 'current course'; 'course equivalent to the current one', 'Set of equivalence courses for course', 'Total number of students in the course'. Using these characteristics pertaining to course difficulty a scale was developed to signify the level of difficulty of each course. The course difficulty was measured over 5 point scale very easy, easy, average, difficult and very difficult.
VI. EXPERIMENTS ON MODIFIED CRISP-DM PROCESS AND ANALYSIS OF RESULTS (STAGE 3)

This section describes experiments conducted on modified CRISP-DM process given in figure 7. The subsection (a1) to (g1) in this section has one to one correspondence to the subsection A to G in section V. Thus each subsection provides the experimental and analytical details conducted according to the design specification provided in the subsection A to G in section V.

a1) Business understanding (Step A): predicting reasonably accurately the optimum time to degree and CGPA in terms of course taking pattern of students and course difficulty leading to decision making about categorising courses and students and provide better support to students.

b1) Data Understanding (Step B): This step is divided into two: one addressing contextual data and the other addressing general data.

b1.1 Finding the existence or nonexistence course difficulty using a specially designed algorithm that was based on the formula devised by Zainuddin, 2012 to compute the course difficulty level (refer Zainuddin, 2012) and a pseudo code written based on the guideline given by Vert et al., 2010.

The outputs derived from the execution of the pseudocode is shown in the screen shot below.

b1.2 General Data

This section is same as previous sections as it deals with general data of students. The data extracted using the SQL queries is provided in Table 3.

c1) Data Preparation (Step C): Both the outputs of b1.1 and b1.2 were subjected to the same steps mentioned subsection iii of section IV.

d1) Added Data Preparation (Step D): The data prepared in step C1 and C2 see figure 7 are merged and the resulting dataset is shown in Table 4. This dataset is further fed into the modeling stage.

Table 4, Data set with contextual factors extracted

<table>
<thead>
<tr>
<th>student_id</th>
<th>pregressa</th>
<th>spa</th>
<th>semestres</th>
<th>sem_gradeports</th>
<th>course</th>
<th>difficulty</th>
</tr>
</thead>
<tbody>
<tr>
<td>stud 1</td>
<td>3.5</td>
<td>85</td>
<td>2.9</td>
<td>4.3</td>
<td>2.75</td>
<td>ACCT 101</td>
</tr>
<tr>
<td>stud 2</td>
<td>3.6</td>
<td>83</td>
<td>2.8</td>
<td>4.3</td>
<td>2.70</td>
<td>ARAB 101</td>
</tr>
<tr>
<td>stud 3</td>
<td>3.6</td>
<td>84</td>
<td>2.7</td>
<td>4.3</td>
<td>2.70</td>
<td>ACCT 101</td>
</tr>
<tr>
<td>stud 4</td>
<td>3.6</td>
<td>85</td>
<td>2.6</td>
<td>4.3</td>
<td>2.70</td>
<td>ARAB 101</td>
</tr>
<tr>
<td>stud 5</td>
<td>3.6</td>
<td>86</td>
<td>2.5</td>
<td>4.3</td>
<td>2.70</td>
<td>ACCT 101</td>
</tr>
<tr>
<td>stud 6</td>
<td>3.6</td>
<td>87</td>
<td>2.4</td>
<td>4.3</td>
<td>2.70</td>
<td>ARAB 101</td>
</tr>
<tr>
<td>stud 7</td>
<td>3.6</td>
<td>88</td>
<td>2.3</td>
<td>4.3</td>
<td>2.70</td>
<td>ACCT 101</td>
</tr>
<tr>
<td>stud 8</td>
<td>3.6</td>
<td>89</td>
<td>2.2</td>
<td>4.3</td>
<td>2.70</td>
<td>ARAB 101</td>
</tr>
<tr>
<td>stud 9</td>
<td>3.6</td>
<td>90</td>
<td>2.1</td>
<td>4.3</td>
<td>2.70</td>
<td>ARAB 101</td>
</tr>
<tr>
<td>stud 10</td>
<td>3.6</td>
<td>91</td>
<td>2.0</td>
<td>4.3</td>
<td>2.70</td>
<td>ARAB 101</td>
</tr>
</tbody>
</table>

Table 7, Course taking patterns extracted from Dataset with Contextual Factors Extracted from Student Data (KJ Lu, 2016).

f1) Evaluation (Step F): In line with the details of the design given in sub-section (f) of Section V the results obtained were evaluated. Evaluation included exploratory course taking pattern by allocating the course difficulty level for each enrolled course and links the course pattern to time to degree and assess the scored CGPA (Table 7). The course difficulty level was calculated using a five point scale (see Section V). The conjecture here is that more the number of enrolled courses of a student having difficulty level is greater than or equal to ‘difficult’ then more would be the time to degree and lesser will be the CGPA. For instance, take a student e.g. stud 6 in Table
7 who is shown to have registered in five courses of these the course difficulty is ‘Difficult’ with the left over 1 course as ‘Average’. The student CGPA is 3.49. It is sensible to anticipate that this student will take more time to graduate as he was able to finish the entire number of 44 courses in 4.5 years on an average of five courses in a semester and a CGPA around the same figure of 3.49. The course difficulty stage has possibly not allowed the student to enrol in more courses in a semester, say 6. Had ‘stud6’ enrolled in 6 courses/semester, then the student might have graduated in 4 years. Thus course taking pattern and the corresponding course difficulty measure of a course can be believed to have an effect on the time to degree. On the other hand if ‘stud6’ had registered in 6 courses whose course difficulty scores were a combination of ‘Easy’, ‘Average’ and ‘Difficult’ (for instance, a course taking pattern of 6 courses may be ARAB 102, ACCT 320, ENGL 201, BANK 302, STAT 202 with course difficulty measured as ’Easy’, ’Average’, ’Average’, ’Easy’, ’Difficult’, ’Difficult’ then there is a probability that ‘stud6’ might have scored the same CGPA of 3.49 or more and had taken lesser time to degree (4 years). Therefore it can be witnessed that course taking pattern if linked with the course difficulty then it is probable to forecast the optimal time to degree at an optimal CGPA. That is to say, it is possible to anticipate that when students enrol in courses where difficulty level is “difficult” then lesser would be the number of such courses a student could enrol owing to harder effort to complete with a higher CGPA. From this situation, it is probable to deduce that when students enrol in a set of courses in which difficulty levels are “difficult”, then the no. of courses in that set is liable to be lesser. This might be because, more the course difficulty level, difficult is the attempt needed to finish the course with higher CGPA. Therefore lower number of courses in the set provides greater opportunity to work hard leading to better chance of scoring higher CGPA at the cost of time to degree. That is:

f11) Course difficulty negatively influences the number of courses registered in a semester.

f12) Course taking pattern characterized by high number of difficult courses negatively influences CGPA and positively influences time to degree.

f13) Course taking pattern characterized by low number of difficult courses positively influences CGPA and negatively influences time to degree.

However this inference was not so obvious to predict and generalize when one considers the hidden knowledge in the dataset and results provided Table 7. Entirely opposite results have been obtained in the case two students namely ‘stud21’ and ‘stud22’. From Table 7 it can be seen that the statements in (f11), (f12) and (f13) are completely falsified. The evaluation of the results of the students ‘stud21’ and ‘stud22’ clearly shows:

f14) Course difficulty does not influence the number of courses registered in a semester. The students ‘stud21’ and ‘stud22’ have registered in a maximum of six courses all of them rated as ‘difficult’.

f15) Course taking pattern characterized by high number of difficult courses does not influence CGPA and negatively influences time to degree. It can be seen that the pattern of courses registered in by the student ‘stud21’ although comprises six courses whose level of difficulty is rated as ‘difficult’ has the scored the maximum CGPA and achieved the lowest time to degree of 3 years. Similarly for the student ‘stud22’ the course taking pattern although comprising of six courses whose level of difficulty is rated as ‘difficult’ has not scored the maximum CGPA but achieved the lowest time to degree of 3 years. This shows that the course taking pattern characterized by high number of difficult courses does not influence CGPA but negatively influences time to degree.

f16) Course taking pattern characterized by low number of difficult courses positively influences CGPA and negatively influences time to degree is falsified because high number of course rated as difficult can also positively influence CGPA and negatively influence time to degree as mentioned in (f15).

The previous findings and discussions evidently point out that the optimal time to degree and the maximum CGPA can be fixed by a set of six courses (highest authorized by a HEI in a semester, refer section III), with every course fixed as ‘Difficult’. When this finding is treated as optimal, in that case the following decision could be facilitated in the HEIs.

1. Categorizing students akin to ‘stud21’ and ‘stud22’ to allow them to graduate in 3 years.
2. Categorizing students akin to ‘stud6’ who have enrolled in just 5 courses and investigate 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’.
3. Categorizing 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’.
4. Assert the performance of students who have enrolled in 4 or 5 courses in a semester at an early stage in their academic career in the university using modified CRISP-DM to forecast their
optimal time to degree with higher CGPA based on past student results.

VII. CONCLUSION, LIMITATION AND FUTURE RESEARCH

The results obtained in this research as explained in Sections V and IV clearly point out to that the main objective of demonstrating a modified KDDM process that extracts hidden knowledge from the student dataset characterized by contextual factors that could be useful in making better decisions in HEIs to achieve specific business goals has been achieved. The modified process uses CRISP-DM as an example which has not been deployed in the context HEIs. Thus the outcome of this research also provides a demonstration of the use of CRISP-DM process in HEIs. Finally the modified CRISP-DM process provides a DM process that could uncover contextual data to be used to support business goal and produce a dataset at the preparation stage which is contextualized. This provides a new way of dealing with the model development at the DM stage in the CRISP-DM process. At this stage a modified genetic algorithm has been specifically developed to enable the CRISP-DM process to generate a model that is contextual leading to the discovery of course taking patterns that are contextualized. This discovery has enables prediction of optimum time to degree and CGPA, an aspect not covered in the extant literature. As far as limitations of the research are concerned, it must be pointed out that this research has used only one contextual factor namely course difficulty of the registered courses of just 25 students in only one semester. There is a need to cross-check this with more number of semesters and additional contextual factors for greater number of students. Future research could investigate into more number of semesters and use additional contextual factors including student potential, course weight age and course complexity which promise to enable decision makers to make more accurate decisions.

REFERENCES

AUTHORS PROFILE

Dr. Subhashini Bhaskaran Sailes ,
Assistant Professor, Ahlia University, Bahrain
Research interests include Data Mining, Data Science
Business Analytics, Data Analytics

Prof. Mansoor Al Aali
President, Ahlia university, Bahrain
Research interests include Data Mining, Data Science, Artificial Intelligence

Dr. Kevin Lu
Brunel University London

Contextual Data Mining for Higher Educational Institutions (HEIs)
Science & IT Education Conference .
42. Lu, K.J. and Sailesh, S.B., 2016. Context driven data mining to classify students of higher educational institutions.