Design and Implementation of an Intelligent System to Predict the Student Graduation AGPA

Sameh Ismail
Sultan Qaboos University, Oman
Shubair Abdulla
Sultan Qaboos University, Oman

Since Accumulated Grad-Point Average (AGPA) is crucial in the professional life of students, it is an interesting and challenging problem to create profiles for those students who are likely to graduate with low AGPA. Identifying this kind of students accurately will enable the university staff to help them improve their ability by providing them with special academic guidance and tutoring. In this paper, using a large and feature rich dataset of marks of high secondary school subjects, we developed a data-mining model to classify the newly-enrolled students into two groups; "weak students" (i.e. students who are likely to graduate with low AGPA) and "normal students" (i.e. students who are likely to graduate with high AGPA). We investigated the suitability of evolving fuzzy clustering methods to predict the ability of students graduating in five disciplines at Sultan Qaboos University in the Sultanate of Oman. A solid test has been conducted to determine the model quality and validity. The experimental results showed a high level of accuracy, ranging from 71%-84%. This accuracy revealed the suitability of evolving fuzzy clustering methods for predicting the students' AGPA.

Keywords: *Educational data mining, evolving fuzzy clustering, prediction, AGPA*

Introduction

The size and complexity of educational data warehouses have increased the challenges associated with managing and analysing relevant data. Recent decades have witnessed the birth of a research field in educational data mining (EDM) (Peña-Ayala, 2014) with the purpose to discover knowledge from such a huge amount of data. Simply put, EDM is a computer-based information system devoted to scanning huge amounts of educational data, generating information, and discovering knowledge (Anjewierden, 2011). EDM has emerged as a new branch of Data Mining (DM), which is a process of analysing data and extracting useful knowledge which is previously unknown (Witten, Frank, & Hall, 2011). Hence, any EMD approach will inherit the characteristics of DM approaches. Examples of these characteristics include: the theoretical model that grounded the baseline of the approach (such as:

machine learning and probability), the nature of the model to be designed (supervised and unsupervised) (T.-M. Huang, Kecman, & Kopriva, 2006), the task to be performed, the techniques used to build the approach, and the frameworks followed to deploy the approach on computers and internet. EDM has recently been applied in many areas, i.e. predicting student performance, students modelling, providing recommendations for students, detecting undesirable student behaviours, and planning and scheduling. Among these educational areas, the accurate prediction of student performance is useful for universities in many different ways.

Having profiles for the newly enrolled students who are likely to graduate with low AGPA is an interesting and challenging problem. Therefore, predicting the graduation AGPA for the students who are currently in their initial stages has always attracted the attention of academic advisors. The DM models could help to achieve these objectives. This research has been inspired from our observational work conducted at Sultan Qaboos University (SQU). Despite the fact that the SQU only admits outstanding students from high secondary schools, we noticed a small percentage of students who achieved low AGPA upon graduation. The goal of this research is to identify those students who will graduate with a low AGPA. To achieve this goal, we employ evolving fuzzy clustering data mining models (Lughofer, 2011) to predict the AGPA graduation figure for newly-enrolled students. This research also seeks to answer the following questions:

1. To what extent can evolving fuzzy clustering methods predict the AGPA graduation figure based on the students' marks in secondary school?
2. How effective is the prediction process in identifying the students who are "weak" and the students who are "normal"?

The significance of this research can be summarized in two ways. First, it helps the management of the colleges to predict the level of graduating students before they actually graduate. It gives indicators as to what needs to be done in terms of academic guidance, future study plans, and development of educational applications in different disciplines, especially the educational technology. Second, the research will assist the management of student counselling centres at universities in providing academic counselling services, educational services, and psychological services. All these services are necessary to create opportunities to raise the level of weak students, and consequently achieve the desired success in meeting the needs of society and the labour market.

This paper is organized with a review of the literature, a research methodology where the CRISP-DM framework is explained, followed by the implementation of our data collection and mining process following this framework. The research results are presented and discussed, and lastly, conclusions from the research and direction of future developments.

Related Work

This section discusses some related research. First, we review two categories in which most papers have been published. Second, we describe the research that belongs to our research category. DM has spread throughout the field of business to predict future trends and behaviours of consumers. Its potential has been exploited in the educational field to (1) assess students' learning performance, (2) provide feedback and learning recommendations based on students' learning behaviours, (3) evaluate learning material and web-based courses, and (4) detect a typical students' learning behaviour (He, 2011).

Although EDM still in its early phase (Peña-Ayala, 2014), there are many problems in the educational environment that have been addressed by EDM researchers. Most of the research published is categorized as "Predicting Students Performance". The main goal of the research in this category is to estimate how well learners accomplish a given task or how they will reach a specific learning level. The first example of this category in our review is the work published by (Kabakchieva, 2011). They seek patterns to predict students' performance at university level based on their personal and pre-university characters. Several classification algorithms are selected and applied, including the k-Nearest Neighbours (kNN), the common decision tree algorithm C4.5, two Bayesian classifiers (NaiveBayes and BayesNet), and two rule learners (OneR and JRip). The work presented by (Zimmermann, 2011) analysed the statistical relationship between B.Sc. and M.Sc. achievements. A random-forest algorithm was used to estimate decision trees for regression on random partitions of the dataset that is not subjected to an admission-induce selection bias. The authors in (Campagni, 2012) analysed the path of how learners performed in exams over the degree e-Learning time. The clustering DM task is fulfilled using K-means method in relation to the dataset to measure the performance of the students in terms of graduation time and final grade.

The second category, which will be reviewed, is known as "Modelling Student's Behaviour". This category seeks to predict learners' behaviours in order to adapt the system according to their tendencies. For example, the authors in (Toscher, 2010) scanned sample of data to predict students' abilities to answer questions. They employed kNN algorithm to search for a number of students who had the most similar results and to predict outcomes using a weighted mean of the results. The research in (Köck & Paramythis, 2011) has analysed learner's behaviour along known learning dimensions to discover learning dimensions and problem solving. The Discrete Markov Models have been employed to detect the styles of problem solving. The work presented in (Bayer, 2012) is another instance of modelling the students' behaviour. They developed a method for mining educational data in order to learn a classifier to predict student success and student dropouts. Seven different machine-learning methods have been employed for prediction. The main characteristics used to shape the EDM approaches are the task, which need to be performed, and the technique to mechanize the proposal. The task represents the main goal of the model

and the technique is the approach chosen to accomplish this goal. Different DM techniques have been employed to accomplish the DM tasks. Table 1 shows examples of data mining tasks and techniques.

In weighing up the pros and cons of the publication reviewed, we have noted the accuracy and significance about the findings and conclusions. However, the main critique is upon the methodologies applied. We believe that some EDM problems have a fuzzy nature and therefore, there is a need to employ fuzzy clustering methodologies to learn and adapt the knowledge of the EDM systems.

Among the different categories of EDM works, this research belongs to the category of "Predicting Students Performance" (Romero & Ventura, 2010). The objective of this category is to predict an unknown variable (grades or scores) that describes the students learning behaviours. Most of the student performance prediction studies have focused on a student's previous data such as his or her grades in a specific course (Qasem, 2006) and a list of grades in some subject (Kotsiantis & Pintelas, 2005). A student's performance at secondary school plays a crucial role in drawing clear picture of his or her final AGPA. According to our knowledge, this research is the first attempt to predict the student's level at graduation time based on their performance in secondary school subjects, which qualified them to join their current university department in the first place. Also, this research uses a fuzzy clustering method, namely the kNN-based Evolving Fuzzy Clustering Method (kEFCM) (A. Shubair & Al-Nassiri, 2015). kEFCM is selected as the main modelling technique to identify the students who are likely to graduate with low AGPA. The kEFCM prediction method belongs to a new kind of algorithm used for cluster analysis: Evolving Fuzzy Clustering, which is popular in recently published data mining literature. In addition, kEFCM needs few efforts to be tuned, it has few parameters, and it is very easy to implement. kEFCM is an enhancement of the traditional kNN algorithm, which is an instance-based clustering algorithm. kEFCM overcomes the problems of computational cost and clustering complexity diagnosed in kNN clustering method. The great merit of kEFCM lies in the fact that it is a dynamic evolving clustering method (A Shubair, Ramadass, & Altyeb, 2014).

Table 1 Examples of DM tasks and techniques

Paper	**DM Task**	**DM Technique**
(Hsia, Shie, & Chen, 2008)	classification	1-decision trees 2-link analysis
(C.-T. Huang, Lin, Wang, & Wang, 2009)	classification	1-decision trees 2- neural network
(Ranjan, 2008)	estimation	1-decision trees

		2-Bayesian models
(Hilbert, 2007)	estimation	1-association-rule
(Yao-Te, Yu-Hsin, Ting-Cheng, & Jen, 2008)	classification	1-association-rule 2-genetic algorithm

Methodology

Two guides have been defined in the context of efforts used to establish industrial standards that represent the implementation steps of data mining applications. These are known as: (i) Cross Industry Standard Process for Data Mining (CRISP-DM) (Shearer, 2000) and (ii) Sample, Explore, Modify, Model and Assess (SEMMA) (Mosaddar & Shojaie, 2013). In this research, we have selected the most popular data mining process model, CRISP-DM. The reason behind the selection is that the CRISP-DM is more complete than SEMMA (MF, 2008), and also this model has been used regularly to describe the approaches used to tackle DM problems (Şen, Uçar, & Delen, 2012) (Mariscal, Marbán, & Fernández, 2010). Moreover, it provides us with an organized and scientific way for conducting our research and increases the possibility of reaching an acceptable level of reliability and accuracy. It is also worth mentioning that we have adapted the CRISP-DM in line with the educational nature of the research, especially in terms of the terminology. For instance, we classified the objectives into two levels. At the root level, we named the business objectives as project goals that can be analysed at the branch level into DM objectives. The rationale for the adaptation process is that we found that CRISP-DM was not updated for a long time, which led to some difficulties with regards designing software application for new DM topics as in EDM.

CRISP-DM Process Model

The CRISP-DM model (Chapman et al., 2013) breaks down the whole DM process into 6 phases: Project Understanding, Data Understanding, Data Preparation, Modelling, Evaluation, and Deployment. The output of executing one phase initiates input for the subsequent phase. The sequence of phases is flexible, moving forth or back is allowed.

1. **Project Understanding:** Usually, the most important phase of any data mining research is the phase of project understanding that focuses on: (i) understanding the research goals, (ii) assess the situation (iii) converting the goals into a data mining problem definition, and (iii) developing a preliminary plan to achieve the objectives. In the step of determining project goals, understanding the goal of the project is critical to ensure that the research will produce right answers to right, properly addressed questions. The second step involves listing the resources available to the research, the requirements and the constraints. The step of data mining definition problem is totally based on the step of determining project goals. For example, if the project goal is: to raise the academic level of the student, the

data mining problem could be: predicting the students who will perform at low academic level. If the project goals cannot be effectively converted into a data-mining problem, the objectives should be redefined. Finally, this phase should develop a research plan to describe the intended plan for tackling the data-mining problem defined. The research plan may include outlining specific research steps and a proposed timeline.

2. **Data Understanding:** In general, this phase involves three steps: (i) collecting data, (ii) describing data, (iii) exploring data, and (iv) verifying the quality of the data. In the first step, the necessary data should be acquired and integrated. To avoid any potential delay, knowing in advance whether the data should be collected from several different sources is very critical. During the step of data description, the properties of acquired data should be examined to produce a report on the examination results. The examination process concerns about data format, data quantity, and records and fields in each table. The key question that will be responded in this step is: does the data acquired satisfy the relevant requirements or not? The data quality verification step examines the acquired data quality in terms of data completion, missing values, and data confliction with common sense (e.g., graduated students are teenagers). Possible common items to check include: blank field, missing attributes, and spelling of values.

3. **Data Preparation:** The data preparation phase aims at constructing the input for the modelling tool(s) in phase 4. By using the initial raw data, this phase constructs the final dataset that will be fed into the modelling tool. The main steps of this phase are: (i) data selection, (ii) data cleansing, (iii) data construction, (iv) data-integration, (iiv) data format. In data selection step, a decision will be taken against the data used with an explanation why certain data was included or excluded. At the data-cleansing step, only the clean subsets of data should be used to avoid putting the data mining analysis in a question. For each quality problem reported in the "verify data quality" step in phase 2, there will be an outline explaining how the problem is addressed. After the data is cleaned, the data construction step starts. The task may involve producing derived attributes which are new attributes constructed from existing attributes. For example, transforming the grades of students to numeric values. These transformations are often required for operating with the modelling tools or algorithms. Data integration step involves merging information from multiple tables to create new table(s). Another task might be performed in this step, the data aggregation that refers to summarising information from multiple records or tables.

4. **Modelling:** In this phase, a modelling technique is selected and applied on the data prepared. Typically, this phase includes four steps: (i) modelling technique selection, (ii) model testing, (iii) model building, and (iv) model assessment. In the first step, an appropriate modelling technique, i.e. decision tree and neural networks, is selected to build a model. The main function of the model testing is

to determine the strength of the model. Setting up the modelling technique parameters to optimal values and choosing quality measurements should be carried out in this step. Model building, closely related to the testing step, aims at running the modelling tool on the prepared dataset to create model. Model assessment, the final step, and judges the success of the model according to the data mining process criteria and the desired test design taking into account the project objectives.

5. **Evaluation:** Before proceeding to the final deployment phase, it is important to review the construction of the model in order to assure a proper achievement of the project objectives. The steps of this phase are: (i) evaluation of the results, (ii) process review, and (iii) determine the next step. In the previous phase, the tests conducted dealt with factors such as the accuracy reliability of the model, while the step of results evaluation assesses how successfully the model meets the project objectives. One of the options applied is to test the model on real-world data and observe the compatibility between the output and the project objectives. The process review step aims at discovering if any important factor or task has been overlooked and deciding to move on to deployment or steeping back to cover a missing thing.

6. **Deployment:** Creation of the model is not the end of the research in some cases. The model must be presented in a way that can be used in a professional environment. This phase is often involves applying the model created within an organization's decision-making processes. Depending on the requirements, this phase can be as simple as a report generation or as complex as a repeatable data mining process implementation. Mainly, this phase involves four steps: (i) plan deployment, (ii) plan monitoring, (iii) produce final report, and (iv) review project.

In the subsequent sections we will explain our research phases in light of the CRISP-DM guide.

Project Understanding

According to the above illustration of the project-understanding phase, our research goal is to identify the students who will graduate with a low AGPA. Identifying such students helps the advisors and management of the colleges to providing a capacitor academic guidance for them. The step taken in the assessment of our research situation revealed the research requirements list, as illustrated in Table 2.

Table 2 Research requirement list

Ser	Requirement	Description
1	The data	The student marks in the secondary school certificate & The AGPA of graduate students

2	The software	MATLAB & Data mining tool (fuzzy clustering tool)
3	The college of education laws	list of disciplines in the college & grading system applied

The situation steps taken in the assessment also revealed some technological constraints. Initially, we found that the size of students' data collected was insufficient to be used for modelling. There were two reasons for this insufficient size: (1) too few students in some departments of the college of education, especially the department of Arabic Language Studies and the department of Islamic Studies, (2) no secondary school subjects related to the students acceptance in some discipline. As part of the solution, we collected data from five cohorts to increase the data size.

In light of our research goal, the requirements, and the constraints discussed, we set the objectives of the data mining modelling as follows:

1. To use evolving fuzzy clustering to predict the graduation AGPA for newly enrolled students.
2. To classify the students as "low students" and "normal students".

The final step taken in this phase was the development of preliminary plan to achieve the research goals and objectives as illustrated in Figure 1.

Data Understanding

The data collection task has been performed by gathering the data received from the Deanship of Admission and Registration at Sultan Qaboos University (SQU). In the process of data description, initially, we can describe the data sample obtained as unclassified. It was saved in many MS Excel sheets. The sheets contain marks of secondary school subjects as well as the graduation AGPA of 940 students from the cohorts 2007-2012. During the categorization, the data sample was categorized into five disciplines from scientific and pedagogy fields: 150 Science (SCI), 191 Instructional & Learning Technologies (ILT), 69 Arabic (ARB), 95 Islamic Education (ISL), 435 English (ENG) from the academic years 2007 to 2012. Each piece of discipline student data is stored in a separate table. The science and ILT disciplines represent the scientific field while the remaining three disciplines represent the pedagogy field. The reason behind choosing the disciplines is that the students' acceptance in these five disciplines depends to a large degree on their marks in some related secondary school subjects. These related subjects may have a significant impact on the students' graduation AGPA. Some other disciplines in the College such as physical and art education have been excluded from the research as the condition of accepting students depends not only on marks in some secondary school subjects but also there is an aptitude test.

After the sample categorization was completed, we tackled the data-mining question: "what are the input attributes and the output attributes of the prediction task" by

exploring the field names of each data table separately. The findings of the data exploration step revealed that the prediction task output is the AGPA and the input is the students' marks in 3 subjects, which will be described in the data preparation phase. In terms of verifying the data, we identified some course data was missing and that other data was invalid, i.e. some marks were entered with values higher than the maximum mark of the course.

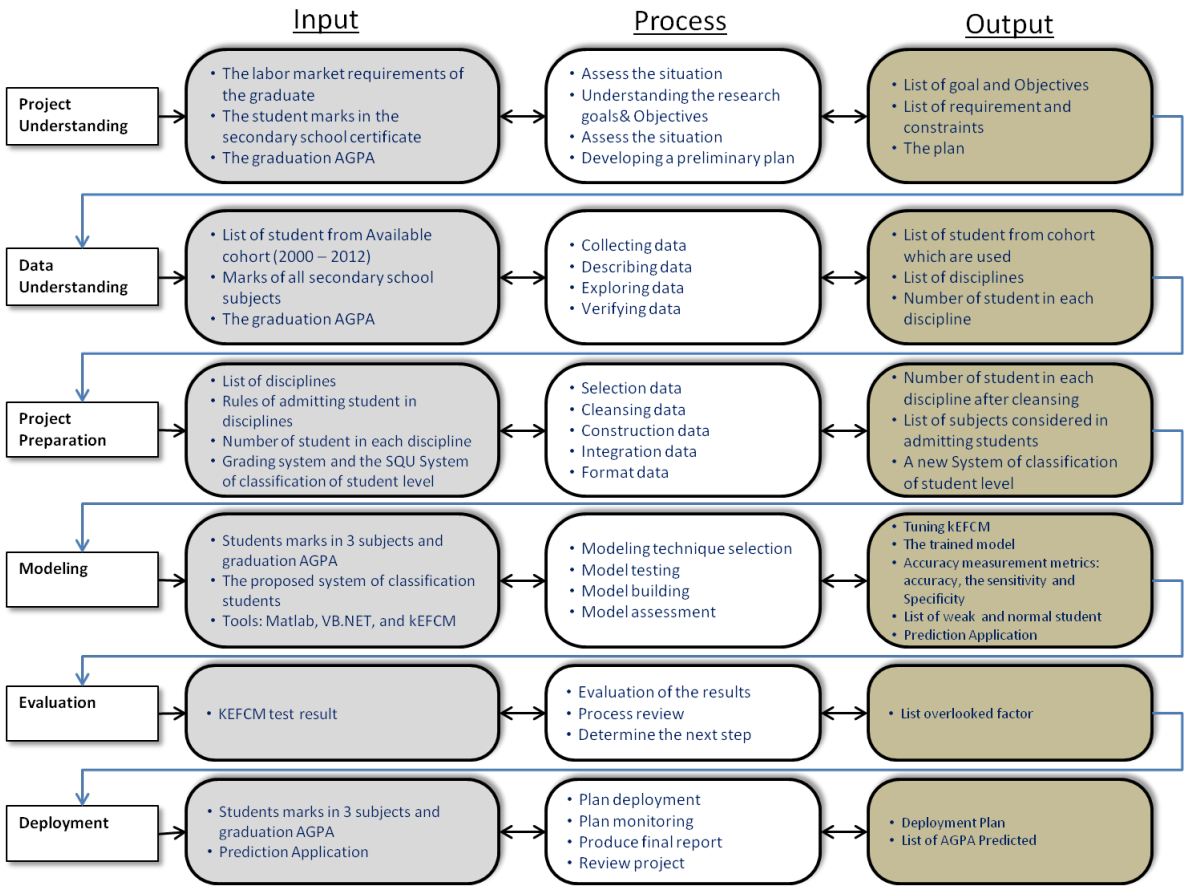

Figure 1 The research plan

Data Preparation

Regarding the data selection, Table 3 shows the secondary school subjects that have been selected for each discipline as the input for the prediction task. The selection of these subjects was according to the student guide published by the College of Education at SQU (SQU, 2010-2011).

Table 3 The secondary school subjects adopted for each discipline

Disciplines	**Subjects**		
	Subject1	**Subject -2**	**Subject -3**
ENG	English	Arabic	Islamic
ARB	Arabic	English	Islamic

ISL	Islamic			Arabic	English
ILT	Math			Arabic	English
SCI	Physic	Chemistry	Biology	Math	English

As aforementioned in the data verification step in the previous phase, the data cleansing actions are carried out by excluding records with invalid and missing values. We solved this data quality problem by excluding the invalid data entries from the sample. Although the number of records has been reduced from 940 to 923, our sample is still within reasonable limits, as showed in Table 4.

Table 4 Sample of the Research

No	**Field**	**Disciplines**	**2007/ 2008**	**2008/ 2009**	**2009/ 2010**	**2010/ 2011**	**2011/ 2012**	**SUM**
1	Pedagogy	ENG	92	91	57	94	89	423
2		ARB	14	19	9	11	14	67
3		ISL	14	16	21	23	19	93
4	Scientific	ILT	40	38	35	39	26	198
5		SCI	49	44	-	-	49	142

In the step of data construction, two actions have been taken. The first action concerned the SCI discipline. As stated in the SQU student guide, 5 subjects are considered when new students are accepted onto this discipline, while only 3 subjects are considered for other disciplines. To unify the data mining modelling tool input, we combined the marks of three subjects, which are similar in nature, namely Physics, Chemistry, and Biology, into one input mark through calculating the average. The second action regarded classifying the students. Figure 2 shows the students classification method, which was followed in this research. According to the student guide published by the College of Education at SQU, a student who achieves a semester and cumulative GPA, which is below 2.00, will be placed on probation. Also, a student who achieves GPA greater than or equal to 2.00 and below 3.15 shall be considered as medium student. In order to provide the medium and probation students with a special follow-up that helps them advancing to the excellent level, we combined the medium and weak levels into a "weak student" class, and classified the excellent students who did not need such a follow-up as "normal students". Initially, the DM modelling technique classified all the students into A, A-, B+, B, B-, C+, C, C-, D+, D, and F classes based on their predicted AGPA. Then, any student who belonged to classes from F up to B- was classified as "weak students"; otherwise, they were classified as "normal students".

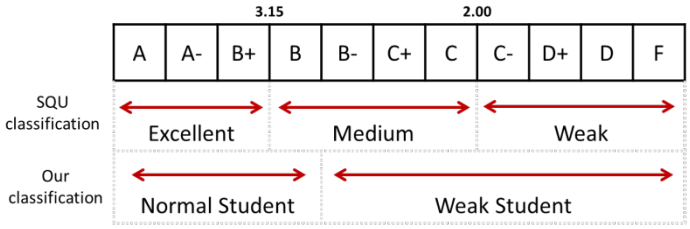

Figure 2 SQU & our students' classification

The main reason behind combining the students of medium and weak levels into one level, "weak student" level, is to increase their recruitment opportunities. Several interviews have been made with officials of private companies and institutions reporting that the excellent students (AGPA between B – A) have a priority in recruitment. In regards to performing the data integration step, five tables have been created, one for each discipline. The separation of disciplines in tables was to facilitate the test of the modelling tool.

The data is ready to be fed into the modelling tool now. The final form of the data is: (subject1, subject2, and subject3). We added one more field, the AGPA as a benchmarking. This small change, which represents the data format step, was needed to make our data suitable for testing the model, especially in comparing the predicted AGPA with the actual.

Modelling Technique

In the first step of this phase, we present our selection of the modelling technique, kEFCM. For a given set of training data, kEFCM orders all the points in such a way that the points in the same cluster are more similar to each other than to those in other clusters. Geometrically, the clusters are circles. The least-squares method is used in determining the circle centre and radius. The membership degree of a point in a cluster is identified by the Euclidean distance between the point and the cluster centre. For a new unseen before point, if it lies in a cluster then the point takes the cluster class. Otherwise, kEFCM compares it to the most similar students to in order to make a classification decision. After setting a classification decision against the point, kEFCM evolves its knowledgebase in order to decrease the false classification rate.

Figure 3 shows a clustering system case resulted by the kEFCM for the Physics marks and the Chemistry marks for a sample of 100 students from the Science department. The colour of points and clusters refer to a particular AGPA grade. The red colour indicates AGPA=A grade; the yellow colour indicates AGPA=B grade and so on.

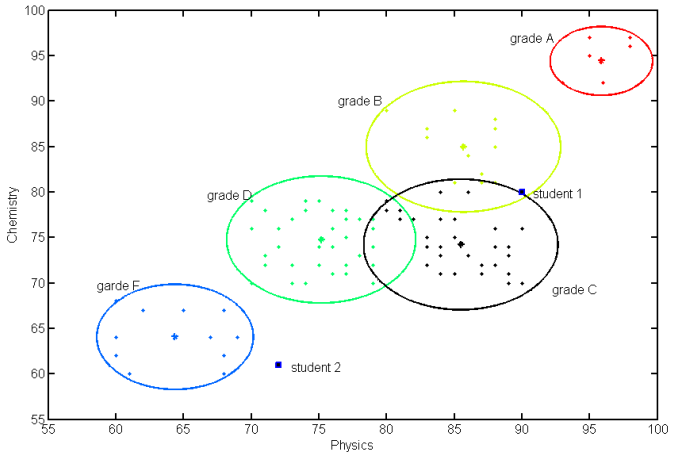

Figure 3 A case of 100 students AGPA cluster based on physics and chemistry marks

Now suppose that there were two new students (indicated as black squares) without classification and that we would like to classify them as either "weak students" or "normal students" based on other students with similar Physics and Chemistry marks. The new student 1 earned 90 and 80 in Physics and Chemistry respectively. Since the student's marks place him into cluster "grade B" class (yellow colour), we would thereby classify him as "normal student" "grade B" easily. Regarding the new student 2 who earned 72 and 61 in Physics and Chemistry respectively, he would be classified as a "weak students", "grade F" since that the closest centre on the scatter plot belongs to "weak students" (the blue colour).

After running the kEFCM to build the model, we start testing the model's quality and validity, to determine the strengths and weaknesses of the model. Typically, a modelling technique operates in two phases: a training phase and a testing phase. During the training phase, the modelling technique is supplied with a set of training examples, each with a class label. This phase conceptually generates a trained model. The testing phase, which is totally based on the trained model, is performed to predict unseen elements. Figure 4 illustrates a graphical representation of the model test.

In order to eliminate the deviation associated with the random sampling of the data elements in establishing the predictive accuracy of a data mining model, a special type of holdout process, known as N-fold cross-validation, is followed throughout the testing process. Generally, the N-fold cross-validation means setting aside some parts of data elements for training and the rest for testing. In this research, the data is split into 10 folds, with each fold containing 92 students. The classification model is trained and tested 10 times; each time is trained on all but one fold (831 students) and tested on the remaining one fold (92 students). A single training/testing process involved input 831 data records for training and 92 data record for testing.

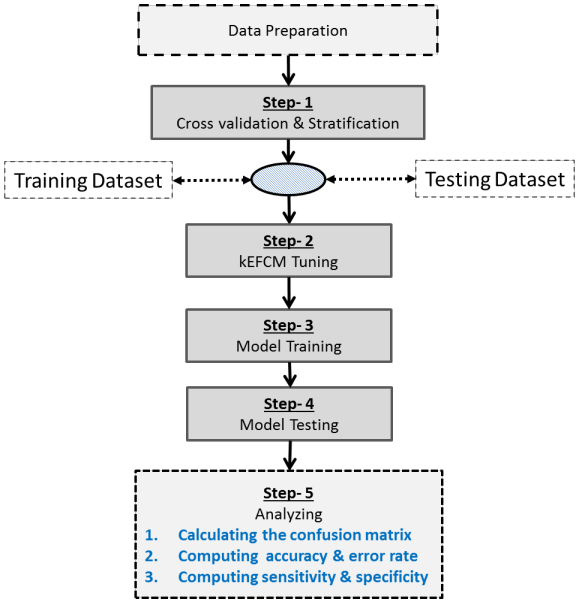

Figure 4 Graphical illustration of model testing step

Figure 5 illustrates graphically the model training and testing processes.

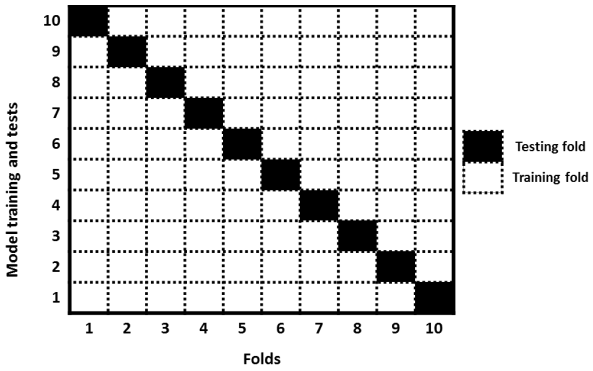

Figure 5 Graphical illustrations of model training and testing

To ensure that the same proportion is used for "weak students" and "normal students" in each fold, a further step has been taken which is commonly used and is known as stratification (Witten & Frank, 2005). It has been carried out by breaking up the students' data into two subsets; the first one contains all "weak students" and the second one contains the "normal students". Each fold was constructed by randomly selecting the fixed number of each subset so that each type of students (weak and normal) is represented evenly over the folds.

Before approaching the experiments, a tuning process is carried out. kEFCM has only one parameter *k*, the number of the most similar students to be considered in making classification decisions. The process of building classifiers highly depends on *k* value. Therefore, the value of k must be set carefully; a small value may maximize the

probability of misclassification, and a large value may make the *k* nearest objects distant from the right class. The obvious best solution is to employ a cross-validation technique (Witten & Frank, 2005) to determine the appropriate value for *k*. The 10-fold cross-validation technique is used on randomly selected 92 students (10% of the sample) from the disciplines. Five values have been tested: 3, 5, 7, 9, and 11. Figure 6 shows the results obtained for each value over the 10-fold. The *k*=5 is selected to configure kEFCM as it gives the highest accuracy related to our sample size.

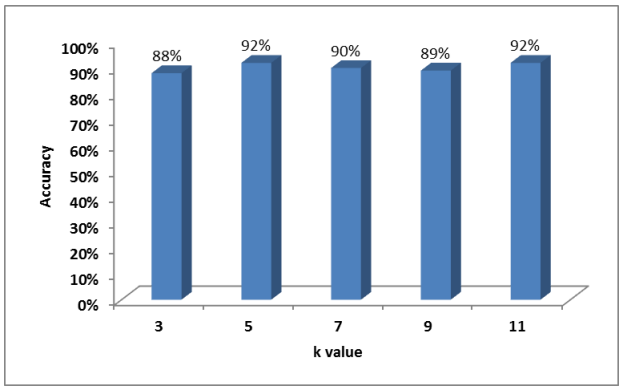

Figure 6 Cross-validation technique results for tuning kEFCM

The test outcomes of all folds are compiled into a confusion matrix, which is a device used to illustrate how a model is performing in terms of false positives and false negatives. Figure 7 depicts the confusion matrix complied upon finishing each test.

kEFCM Results		
Weak Student	Normal Student	
TP	FN	Weak Student
FP	TN	Normal Student

Actual Results

Figure 7 Confusion matrix

Assuming that the "weak student" state is "positive" and the "normal student" state is "negative", we compared the kEFCM results against the actual results by registering the *TP* (true positive), *FN* (false negative), *TN* (true negative), and *FP* (false positive) variables:

1. *TP*: "weak student" and kEFCM classified it as "weak student".
2. *FN*: "normal student" but the kEFCM classified it as "weak student".
3. *TN*: "normal student" and the kEFCM classified it as "normal student".
4. *FP*: "weak student" but the kEFCM classified it as "normal student".

Based on these metrics, we computed the kEFCM accuracy and error rate on each test fold. The accuracy, also referred to as the prediction rate, is the percentage of students who are correctly classified. The opposite of accuracy is the error rate, which is the percentage of students who are incorrectly classified. The specific formulas for kEFCM accuracy (*a*) and kEFCM error rate (*e*) are given in the following equations:

$$a = \frac{(TP + TN)}{(TP + FN + TN + FP)} \dots\dots\dots\dots (1)$$

$$e = \frac{(FP + FN)}{(TP + FN + TN + FP)} \dots\dots\dots\dots (2)$$

For any classification problems and in most cases, we can derive the Sensitivity (*sn*) and Specificity (*sp*) (Tung, Quek, & Guan, 2012), which are commonly used for analysis of any binary classifiers.

The sensitivity (*sn*) measures the proportion of students who are "weak student" (*TP*) out of all the students who are actually weak (*TP+FN*). With a higher sensitivity, fewer normal students will be predicted as "weak students":

$$sn = \frac{TP}{(TP + FN)} \dots\dots\dots\dots (3)$$

On the other hand, specificity (*sp*) measures the proportion of students that are "normal students" (*TN*) out of all students that are actually normal (*TN+FP*). With a higher specificity, fewer weak students are predicted as "normal students":

$$sp = \frac{TN}{(TN + FP)} \dots\dots\dots\dots (4)$$

The fundamental idea behind sensitivity and specificity analysis is to measure how well kEFCM is able to identify "weak" students" and "normal students".

The model-building step involved integrating the kEFCM tool into an application. Our application, shown in Figure 8, is developed using MATLAB and VB.NET. The big challenge, which we faced, was the execution of the MATLAB complex code within a VB.NET environment. We overcame this problem by creating a DLL file inside the MATLAB environment and then registering and calling it within VB.NET. Initially, the application loads training data from a semicolon delimited text file that contains five fields, course1, course2, course3, class, and APGA. To predict the AGPA of a student, firstly, three marks should be fed into the text boxes, and then the predicted AGPA will be shown after clicking on the predict button.

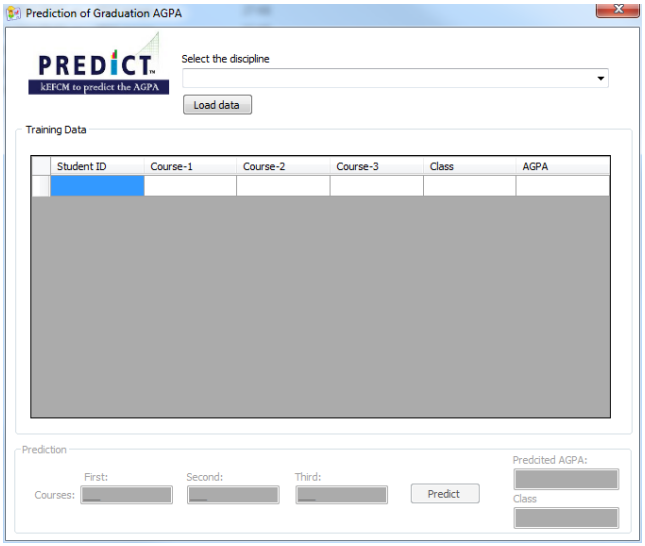

Figure 8 AGPA prediction program

In the final step of the modelling phase, we hereby present the testing results. The prediction results of the kEFCM modelling method for each discipline are presented in Table 5. The results presented are the 10-fold cross-validation results. The leftmost column shows the test number. The columns 2-5 illustrate the number of TP, FN, FP, and TN and the columns 6-9 illustrate the accuracy (*a*), the error rate (*e*), the sensitivity (*sn*), and the specificity (*sp*) for each test. Figure 9 compares the averages of *a*, *e*, *sp*, and *sn* for all disciplines. The results indicate that kEFCM performed reasonably well overall across the disciplines. Among them, the best prediction accuracy is produced for ENG with overall accuracy of 84% and the lowest overall accuracy produced for ARB, with an overall accuracy of 71%.

The calculation of sensitivity and specificity reflected a very good level of accuracy in terms of measuring how effective kEFCM classifies "weak students" and "normal students". The overall sensitivity of all disciplines ranged from 76% to 86%, and the overall specificity ranged from 18% to 86%.

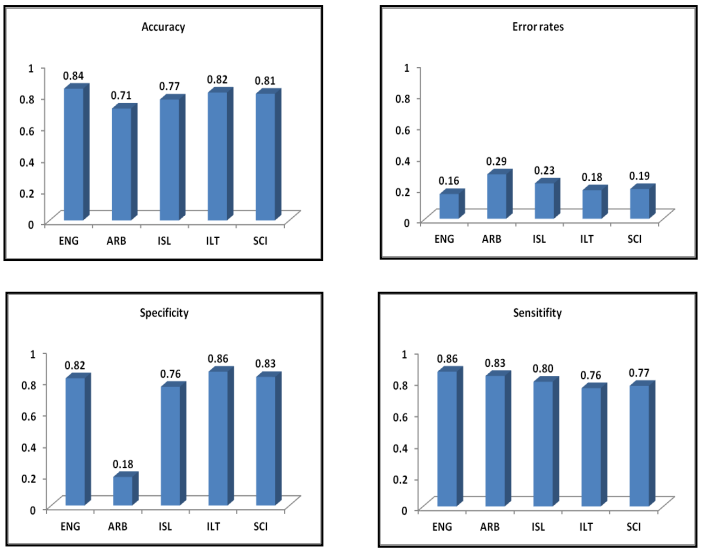

Figure 9 Averages of *a*, *e*, *sp*, and *sn* for all disciplines

Table 5 Cross-validation results for all disciplines

Test#	TP#	FN#	FP#	TN#	*a*	*E*	*sn*	*sp*
English discipline (ENG)								
1	24	4	3	11	83.33%	16.67%	85.71%	78.57%
2	18	5	5	14	76.19%	23.81%	78.26%	73.68%
3	19	5	2	16	83.33%	16.67%	79.17%	88.89%
4	18	2	5	16	82.93%	17.07%	90.00%	76.19%
5	20	2	3	17	88.10%	11.90%	90.91%	85.00%
6	23	1	2	16	92.86%	7.14%	95.83%	88.89%
7	24	4	3	11	83.33%	16.67%	85.71%	78.57%
8	20	2	5	15	83.33%	16.67%	90.91%	75.00%
9	15	5	5	17	76.19%	23.81%	75.00%	77.27%
10	16	2	1	17	91.67%	8.33%	88.89%	94.44%
Arabic discipline (ARB)								
1	5	1	1	0	71.43%	28.57%	83.33%	0.00%
2	3	1	1	2	71.43%	28.57%	75.00%	66.67%
3	4	0	2	1	71.43%	28.57%	100.00%	33.33%
4	5	1	1	0	71.43%	28.57%	83.33%	0.00%
5	5	1	1	0	71.43%	28.57%	83.33%	0.00%
6	3	1	2	1	57.14%	42.86%	75.00%	33.33%
7	4	2	1	0	57.14%	42.86%	66.67%	0.00%
8	4	2	1	0	57.14%	42.86%	66.67%	0.00%
9	5	0	1	1	85.71%	14.29%	100.00%	50.00%
10	4	0	0	0	100.00%	0.00%	100.00%	0.00%
Islamic Education discipline (ISL)								
1	3	0	0	9	100.00%	0.00%	100.00%	100.00%
2	3	1	2	3	66.67%	33.33%	75.00%	60.00%
3	2	2	1	4	66.67%	33.33%	50.00%	80.00%
4	1	2	1	5	66.67%	33.33%	33.33%	83.33%
5	3	0	2	4	77.78%	22.22%	100.00%	66.67%
6	3	0	2	4	77.78%	22.22%	100.00%	66.67%
7	4	0	1	4	88.89%	11.11%	100.00%	80.00%
8	3	1	3	2	55.56%	44.44%	75.00%	40.00%
9	5	1	0	3	88.89%	11.11%	83.33%	100.00%
10	4	1	1	6	83.33%	16.67%	80.00%	85.71%

Instructional & Learning Technologies (ILT)								
1	5	2	1	10	83.33%	16.67%	71.43%	90.91%
2	8	2	2	6	77.78%	22.22%	80.00%	75.00%
3	5	3	1	9	77.78%	22.22%	62.50%	90.00%
4	9	2	1	6	83.33%	16.67%	81.82%	85.71%
5	5	3	2	8	72.22%	27.78%	62.50%	80.00%
6	7	1	1	9	88.89%	11.11%	87.50%	90.00%
7	5	1	1	11	88.89%	11.11%	83.33%	91.67%
8	8	2	2	6	77.78%	22.22%	80.00%	75.00%
9	6	2	1	10	84.21%	15.79%	75.00%	90.91%
10	5	2	1	9	82.35%	17.65%	71.43%	90.00%
Science Discipline (SCI)								
1	6	1	3	4	71.43%	28.57%	85.71%	57.14%
2	5	3	2	4	64.29%	35.71%	62.50%	66.67%
3	4	2	1	7	78.57%	21.43%	66.67%	87.50%
4	5	2	1	6	78.57%	21.43%	71.43%	85.71%
5	3	1	1	9	85.71%	14.29%	75.00%	90.00%
6	5	1	2	6	78.57%	21.43%	83.33%	75.00%
7	6	0	1	7	92.86%	7.14%	100.00%	87.50%
8	7	1	1	5	85.71%	14.29%	87.50%	83.33%
9	5	2	0	7	85.71%	14.29%	71.43%	100.00%
10	2	1	1	12	87.50%	12.50%	66.67%	92.31%

3.1.5 Evaluation Phase

The first step of this phase evaluates the kEFCM testing results. Two aspects which were noted:

1. The low overall accuracy produced for ARB, 71%: since the sample size does affect the overall accuracy, this low accuracy is not a surprise. For the data used, there is a proportional relationship between the sample size and the accuracy of kEFCM. In Figure 10, the proportional relationship is drawn. The test on the biggest sample (ENG, 423 students) produced the highest overall accuracy, 84%, followed by ILT 198 students with overall accuracy of 82%, followed by SCI 142 students with overall accuracy of 81%, followed by ISL 93 students with overall accuracy of 77%, followed by ARB 67 students with overall accuracy of 71%, the lowest accuracy.
2. The overall specificity, 18% produced for ARB discipline: this low specificity is justified by the low number of "normal student" class entered to kEFCM during the cross-validation. As in Table 5, the number of "normal student" class equals 1 in some tests conducted on the students of ARB discipline as in tests 1 & 3 for instance, and hence, the specificity rate is deemed to be low.

With reference to the second step, namely the process review, although the model appeared satisfactory in terms of the testing results, we conducted a more thorough review of the implementation of CRISP-DM phases. The review has highlighted two factors as follows:

1. The insufficient number of students from the Arabic department in the data sample.
2. There is some sort of task overlapping between the step of data verifying, phase 2 and the step of data cleansing, phase 3.

Despite these factors, we decided to finish the project. The decision was based on the results of the assessment that has shown clearly that these factors were not having a big impact on the quality of the model.

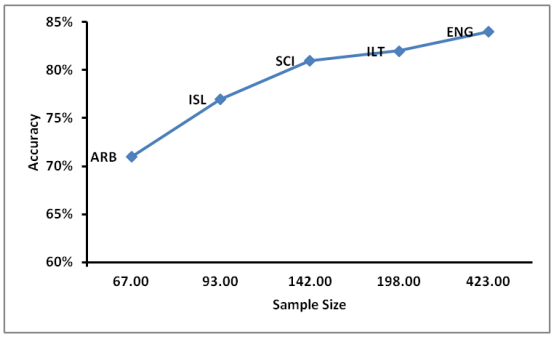

Figure 10 The proportional relationship between sample size and accuracy

Deployment Phase

This phase involves four steps as described in CRISP-DM Process Model, the first step is to produce the deployment plan, which summarizes the strategy for the software application deployment. Our deployment plan includes the necessary tasks and explains how to perform them, as follows:

1. Collect the current new enrolled students' marks in the secondary school certificate.
2. Extract the marks based on the disciplines as described in Table 4.
3. Predict the graduation AGPA for each student.
4. Classify the students into "weak students" and "normal students"
5. Suggest intensive curriculums of academic advice for the students who are classified as "weak students".
6. Compare the predicted graduation AGPA with the AGPA achieved by the students after the intensive curriculums of academic advice.
7. Measure the improvement of the student level.
8. Repeat steps (6) and (7) at the end of each academic year.

The application in the deployment phase was limited to the first four elements of this plan. The main reason for this limitation is that since our plan targets the newly enrolled students, step 5 onwards needs at least 3-4 years to confirm the student's AGPA improvement. Whenever the student progresses in his study and moves to a

new stage, we need to record the AGPA improvement in accordance with the predicted AGPA, and this is a cooperative process, which needs involving different departments from different specialties. At this point our research may stop since the ultimate goal of our research is to predict the graduation AGPA.

4. Results and Discussion

The results and discussion section is organized to address the research questions.

To what extent can evolving fuzzy clustering methods predict the graduation AGPA based on the students' marks in the secondary school?

To predict the graduation AGPA using the evolving clustering methods, we selected the kEFCM-modelling tool. The prediction was done by relating the unknown AGPA to a set of five known AGPAs according to distance function and letting the majority vote predict the AGPA. Number five was chosen as a result of a process called kEFCM tuning. After constructing the kEFCM model, we started measuring the accuracy of the prediction, which has been determined by applying the kEFCM to a sample of 923 students from five different disciplines in the College of Education at SQU. The results reported high prediction accuracy, between 71% and 84%. Randomly selected 10 examples of AGPA prediction results along with actual AGPA of a set of students are presented in Table 6.

Table 6 Examples of Prediction results

Student ID	Student AGPA	Predicted AGPA
87612	B-	B-
87612	B+	B+
86630	A-	A
87129	B-	B-
85773	C	C+
86630	B+	B+
85773	B-	B-
85773	D+	D+
87612	B-	B-
87129	B-	B-

With reference to the kEFCM test results and as illustrated in this table, the evolving fuzzy clustering systems could be successfully employed to predict the graduation AGPA for newly enrolled students.

How effective the prediction process in identifying the students who are "weak" and the students who are "normal"?

SQU only admits the outstanding students and seeks to provide them with what the need in order to fill labour market needs. Therefore, it is necessary to maintain a high level of students throughout the period of study in the university. Having profiles for the students who are likely to graduate with low AGPA (weak students) helps the management to plan for the process of allocating resources that will improve their level. SQU classifies the graduate students into three groups, from grade "F" – "C-": weak students; from grade "C+" to "B": medium students; and from grade "B+" to "A": excellent students. We combined the medium and weak students into one class, namely the "weak students", and set the excellent students as normal students, as shown in Figure 2. This adjustment has been done to ensure providing the students who are medium and weak with a consistent, special follow-up to advance to the excellent level. A software application has been developed to accept a student's set of marks and classify him or her as either weak or normal after predicting his or her graduation AGPA. Figure 12 depicts the classification results of the training data as well as the classification results of the predicted AGPA. Where the program predicts AGPA as B, this refers to "normal" for a student who earned marks 98, 95, and 92 in the secondary school. The accuracy of prediction achieved the software, which was between 71%-84% demonstrated the effectiveness of the prediction process in in identifying the students who are "weak" and the students who are "normal".

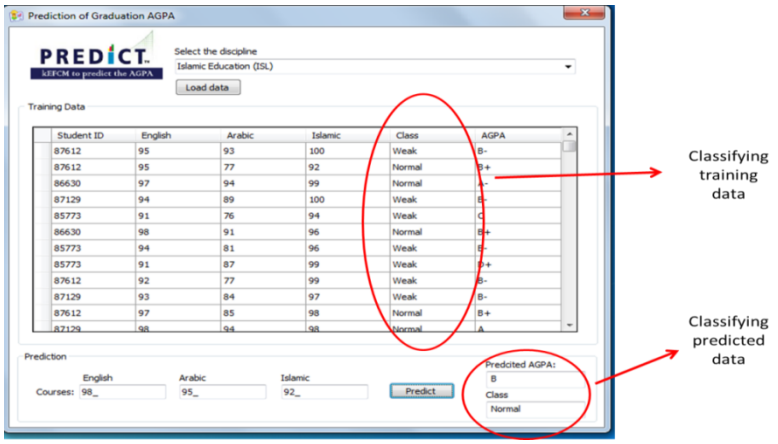

Figure 12 Classification the training data and predicted results

Conclusion and Future Research

EDM is a very useful tool for a wide variety of educational problems where large amounts of data are collected and analysed. As the awareness of the capabilities of EDM increases, researchers are increasingly identifying its usefulness in solving seemingly unsolvable education-related problems. Some of the notable emerging DM applications in the field of education include: predicting student performance, students modelling, recommendations for students, detecting undesirable student's

behaviour, and planning and scheduling. Our research, which focuses on predicting students' graduation AGPA, is another instance, which confirms this claim.

Generally, the objective of predicting students' performance DM systems is to predict unknown variables, such as grades, that describe the learning behaviours of students. These systems are useful for helping the management and academic advisors to plan for the process of allocating resources that improve the weak student's academic level. To accomplish the objective of prediction, we followed the CRISP-DM, which is a popular guide that provides systematic and structured way of conducting DM research. CRISP-DM application process consists of six phases. In the last phase, the deployment phase, we developed a plan of deployment. The plan is divided mainly into two parts, predicting the student's graduation AGPA and observing the AGPA improvement as the student progresses in his or her study. The application of this plan was limited to predicting the graduation AGPA step since the second part needs at least 3-4 years to be implemented. We noticed that the accuracy of an EDM approach depends on the dataset; the larger the sample size, the greater the accuracy of the results. The overall accuracy, 71%, produced for the ARB department was relatively low. This is because among the dataset of 923 students, there were only 67 students from the ARB department.

As the kEFCM experiments illustrated, EDM techniques can accurately predict the student graduation AGPA, and hence allow for identification of the weak students who need a capacitor academic guidance. Such prediction would potentially help in raising the level of weak students at SQU and achieve the desired success in meeting the needs of society and the labour market.

Though our system demonstrated an encourage results, still there are some data quality issues to be resolved. Issues faced by the system are data completeness, reliability, and accuracy. As thousands of data records are usually collected for such kind of EDM systems, if duplicate records, missing values, or presence of unneeded data happens in data, the all steps in the prediction process would be badly affected.

Taking the research results into account, we recommend using kEFCM for developing EDM approaches that could serve in many educational areas, such as discovering the reasons behind the decline of the academic level of students at the university stage of their education.

References

Anjewierden, A., Gijlers, H., Saab, N., & De-Hoog, R. (2011). *Brick: mining pedagogically interesting sequential patterns*. Paper presented at the the 4th international conference on educational data mining.

Bayer, J., Bydzovská, H., Géryk, J., Obsıvac, T., & Popelínsky´, L. (2012). *Predicting drop-out from social behaviour of students*. Paper presented at the 5th international conference on educational data mining, Chania, Greece.

Campagni, R., Merlini, D., & Sprugnoli, R. (2012). *Analyzing paths in a student database*. Paper presented at the 5th international conference on educational data mining, Chania, Greece.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2013). CRISP-DM 1.0 Step-by-step data mining guide. *SPSS Inc*.

He, W. (2011). Using text mining to uncover students' technology - related problems in live video streaming. *British Journal of Educational Technology, 42*(1), 40-49.

Hilbert, K. Sch¨onbrunn and A. (2007). *Data mining in higher education*. Paper presented at the Annual Conference Gesellschaft f¨ur Klassifikation e.V., Freie Universit¨at Berlin, Berlin, Germany.

Hsia, T., Shie, A., & Chen, L. (2008). Course planning of extension education to meet market demand by using data mining techniques – an example of Chinkuo technology university in Taiwan. *Expert Systems with Applications, 34*(1), 596-602. doi: http://dx.doi.org/10.1016/j.eswa.2006.09.025

Huang, C., Lin, W., Wang, S., & Wang, W. (2009). Planning of educational training courses by data mining: Using China Motor Corporation as an example. *Expert Systems with Applications, 36*(3, Part 2), 7199-7209. doi: http://dx.doi.org/10.1016/j.eswa.2008.09.009

Huang, T., Kecman, V., & Kopriva, I. (2006). *Kernel based algorithms for mining huge data sets: Supervised, semi-supervised, and unsupervised learning.* Springer, Berlin, Heidelberg.

Kabakchieva, D., Stefanova, K., & Kisimov, V. (2011). *Analyzing university data for determining student profiles and predicting performance*. Paper presented at the 4th international conference on educational data mining, Eindhoven, Holland.

Köck, M., & Paramythis, A. (2011). Activity sequence modelling and dynamic clustering for personalized e-learning. *User Modeling and User-Adapted Interaction, 21*(1-2), 51-97.

Kotsiantis, S., & Pintelas, P. (2005). *Predicting students marks in hellenic open university.* Paper presented at the Advanced Learning Technologies, 2005. ICALT 2005. Fifth IEEE International Conference on Advanced Learning Technologies.

Lughofer, E. (2011). *Evolving fuzzy systems-Methodologies, advanced concepts and applications*. Springer, Berlin, Heidelberg.

Mariscal, G., Marbán, Ó., & Fernández, C. (2010). A survey of data mining and knowledge discovery process models and methodologies. *Knowledge Engineering Review, 25*(2), 137.

Azevedo, A., & Santos, M. (2008). *KDD, SEMMA and CRISP-DM: a parallel overview*. Paper presented at the IADIS European Confence on Data Mining, Holland - Amsterdam.

Mosaddar, D., & Shojaie, A. (2013). A data mining model to identify inefficient maintenance activities. *International Journal of System Assurance Engineering and Management, 4*(2), 182-192.

Peña-Ayala, A. (2014). Educational data mining: A survey and a data mining-based analysis of recent works. *Expert Systems with Applications, 41*(4), 1432-1462.

Al-Radaideh, Q., Al-Shawakfa, E., K Al-Najjar, M. (2006). *Mining Student Data Using Decision Trees.* Paper presented at the 2006 International Arab Conference on Information Technology.

Ranjan, J., & Khalil, S. (2008). Conceptual Framework of Data Mining Process in Management Education in India: An Institutional Perspective. *Information Technology Journal, 7*(1), 16.

Romero, C., & Ventura, S. (2010). Educational Data Mining: A Review of the State of the Art. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on, 40*(6), 601-618. doi: 10.1109/TSMCC.2010.2053532

Şen, B., Uçar, E., & Delen, D. (2012). Predicting and analyzing secondary education placement-test scores: A data mining approach. *Expert Systems with Applications, 39*(10), 9468-9476.

Shearer, C. (2000). The CRISP-DM model: The new blueprint for data mining *Journal of Data Warehousing, 5*, 13-22.

Shubair, A., Ramadass, S., & Altaher, A. (2014). kENFIS: kNN-based evolving neuro-fuzzy inference system for computer worms detection. *Journal of Intelligent and Fuzzy Systems*.

Shubair, A., & Al-Nassiri, A. (2015). kEFCM: kNN-Based Dynamic Evolving Fuzzy Clustering Method. *International Journal of Advanced Computer Science and Applications(IJACSA), 6*(2), 5-13.

SQU, College of Education. (2010). *Student Guide.* Retrieved from http://www.squ.edu.om/Portals/43/College_of_Education/guides/Attachment%20A-Student%20Guide.pdf

Toscher, A., & Jahrer, M. (2010). *Collaborative filtering applied to educational data mining*. Paper presented at the KDD 2010 cup 2010 workshop: Knowledge discovery in educational data.

Tung, S., Quek, C., & Guan, C. (2012). SoHyFIS-Yager: A Self-organizing Yager based Hybrid neural Fuzzy Inference System. *Expert Systems with Applications*.

Witten, I., & Eibe, F. (2005). *Data Mining: Practical machine learning tools and techniques*: Morgan Kaufmann.

Witten, I., Eibe, F., & Hall, M. (2011). *Data Mining: Practical Machine Learning Tools and Techniques: Practical Machine Learning Tools and Techniques*: Elsevier.

Wang, Y., Cheng, Y., T., Chang, & Jen, S. (2008). *On the application of data mining technique and genetic algorithm to an automatic course scheduling system.* Paper presented at the 2008 IEEE Conference on Cybernetics and Intelligent Systems.

Zimmermann, J., Brodersen, K., Pellet, J., August, E., & Buhmann, J. (2011). *Predicting graduate-level performance from undergraduate achievements*. Paper presented at the 4th international conference on educational data mining, Eindhoven, Holland.