International Journal of Computing and Digital Systems
ISSN (2210-142X)
Int. J. Com. Dig. Sys. 8, No.3 (May-2019)

http://dx.doi.org/10.12785/ijcds/080303

Mining Students Outcomes: An Empirical Study

Fawzi Albalooshi[1], Hadeel AlObaidy[2] and Amal Ghanim[3]

[1,2,3]Department of Computer Science, University of Bahrain, Kingdom of Bahrain

Received 2 Jan. 2019, Revised 10 Feb. 2019, Accepted 7 Mar. 2019, Published 1 May 2019

Abstract: The purpose of the research presented in this paper is to extract students' individualized learning achievements from course information and assessments results for student groups and predict expected performance in future courses based on existing achievements in a set of student outcomes at the program level. Data mining techniques are used to process course information and extract students' achievements in a set of student outcomes. Two prediction algorithms namely single linear regression and multiple linear regression are applied to determine students' expected performance in future courses. Specialized programs have been designed, implemented and processed in stages to achieve the results as described in this paper. The results show that by using data mining techniques individualized students' performance information can successfully be extracted to provide formative feedback. This information is further processed by applying prediction algorithms to determine expected students' performance in future courses depending on the available historical records for courses.

Keywords: Educational Data Mining, Prediction, Linear regression, Multi-valued regression, Association, Learners' achievements.

1. Introduction

There are many challenges faced by educational managers, faculty, and students in order to build strategies to improve learning, providing a competitive environment, and to predict the performance of students beforehand to make timely decisions [1]. In this paper, we implement Educational Data-Mining (EDM) techniques to extract useful information and deduce others to provide educationalist and students useful information that improves learning.

Our academic computer science department maintains electronic assessment records of students' achievements for each offered course. Each course has Course Intended Learning Outcomes (CILOs) that map to a set of Students' Outcomes (SOs) at the program level. Each course assessment (including quizzes, tests, assignments, projects and final examination questions and any others) maps to one or more CILO. Such mapping of assessment to CILOs and CILOs to SOs enable us to determine the student's performance as the extent he or she achieves the eleven SOs (referred at as a–k) set for the academic program in computer science. Appendix A shows the SOs and their description.

We currently maintain assessment records and derive achievements reports at the course level, but individual student's achievement of SOs is currently not required. The main aim of this paper is to use data mining techniques to capture students' assessment outcomes, course information, and program information to achieve the following objectives. Firstly, provide the learner with prompt individualized valuable informative feedback of his/her existing performance towards achieving the SOs. Secondly, to provide the learner with a predictive report on his/her possible achievement in future courses based on current performance. These two objectives will motivate the student to take the necessary steps to improve his/her performance in upcoming semesters to satisfy the graduation requirements and his/her expectations. To achieve these objectives the research question was set as "How to utilize available students course assessments data to extract individualised achievements and predict their performance in coming semesters?". The research methodology followed is to review the literature, design and implement a prototype system using data mining techniques, assess the results and set future research goals if necessary.

This paper is presented as follow: Section 2 presents the literature review on EDM. Section 3 describes the proposed EDM system including architecture, data preparation and pre-processing, mining hidden knowledge and predicting students' future achievements. A discussion on the system's achievements and results is presented in Section 4, and conclusions are presented in Section 5.

2. Educational Data Mining

"Data Mining is considered to be a new paradigm, but due to its significance in decision making, it has been successfully applied to a variety of domains including education," say Mueen, Zafar, and Manzoor [2]. A new

E-mail: falblooshi@uob.edu.bh, halobaidy@uob.edu.bh, aghanim@uob.edu.bh

field is introduced known as EDM, which is now trending and is essential for educational fields. It helps in determining the usefulness of learning systems [3], analysing learner academic performance [4], and developing an early warning system [5]. According to Linan and Perez [6] "EDM develops and adapts statistical, machine-learning and data-mining methods to study educational data generated basically by students and instructors". Prediction of student's performance is important and a challenging task. Academic progress involves numerous factors to be taken into consideration while assessing the performance of the students. These new methods of joining both Decision Support Systems (DSS) and Data Mining (DM) can turn the future of academics as well. For instance, if such systems are used for the prediction of the students' scores in a particular course would not only improve learning but would also help in enriching the teaching techniques [7]. Some examples of the use of EDM applications to enhance student learning are as follows.

A holistic model is developed by Ranjan and Malik [8] for educational purposes using data-mining techniques for exploring the effects of probable changes in processes related to admissions, course delivery and recruitments. Using the data-mining techniques the paper tries to help the students for the counselling by uncovering of the hidden trends and patterns to make accurate predictions through a higher level of analytical sophistication. This process helps in enhancing the research and academic decision making using a combination of the explicit knowledge base, sophisticated analytical skills and academic domain knowledge through DM techniques.

Mueen et al. [2] studied and compared the results of the three different DM classification algorithms (Naïve Bayes, Neural Network, and Decision Tree) to predict and analyse the performance of students. The performance was assessed based on their academic record and forum participation. The result of the study showed that the Naïve Bayes classifier outperforms the other two algorithms by achieving over 86% of prediction accuracy.

In another study carried out by Ahmad, Ismail, and Aziz [9] using classification techniques such as Naïve Bayes, Decision Tree and Rule-Based. They proposed a framework for predicting the results of students from the first year of the bachelor taking computer science courses. The dataset included the student's previous academic records, demographics and family background information. The rule-based technique showed the best result out of the three.

In an experiment carried out by Huang and Fang [10], the predictive performance of students was calculated using multivariate regression models. Input such as cumulative grade point average (GPA) in addition to grades the students earned in some pre-requisite courses was used to obtain the final exam marks in some courses. Their model used multiple criteria such as R-square, shrinkage, the average prediction accuracy and the percentage of good predictions. The results help the instructors to take some proactive measures in accordance with the results to assure the good performance of students.

A study by Abu [11] finds a qualitative model that best classifies the performance of a student based on related personal and social factors. The study explores multiple factors that can influence the performance of a student. Another interesting study put forward by Althaf, Basha, Ramesh Kumar, Govardhan and Ahmed [12] presents the techniques they used to predict the future result of the students. The study used temporal association rule mining to discover the hidden relations between the sequence and the subsequences of the events. The study tries to find the prediction for the students based on the social group categories (urban and rural) they belong to such as government and private sector colleges and for different courses. They successfully derived a prediction mechanism for the success of students' course, social status and grade wise.

Angeline [13] analysed the students' performance using an association-rule technique with the Apriori algorithm. Where the students are classified based on numerous factors such as involvement in doing assignments, internal assessment tests, attendance, etc., which helped to predict the students' performance and identify to which category they belong (average or below average) so the necessary steps can be taken to improve their results. This would not only help with knowing the category to which the student belongs, but also to help to match the organization's requirements with students profiling. This standardization is extended to account for minimum support and confidence thresholds.

The study by Elbadrawy, Polyzou, Ren, Sweeney, Karypis, and Rangwala [14] take prediction to a new level where the grade of the student can be identified for a given activity/course and also the risk of failing the course can be known. They present the class of linear multi-regression models that are designed in order to create models that are personalized to each student. These models take into consideration a number of features, which consists of students' past performance, course characteristics and also their engagement and efforts. Personalization is achieved by using the estimate regression models that are shared with different students along with the student-specified linear combinations. They conducted a performance study on a large set of students, courses and activities where they observed that these models increased the accuracy of the performance prediction by over 20%.

Al-Saleem, Al-Kathiry, Al-Osimi, and Badr [15] introduce a performance prediction model where the future results are predicted using their previous records using different classification techniques. According to them, this will help the students in knowing what courses to take in future. They suggest that this model can be integrated into a recommender system that would help students in deciding the courses they would like to do based on their predicted performances which have been calculated using their previous grades. This can be used while the student had to choose his future career where he would be successful. This idea is somehow related to the research presented here as it also predicts the results for the future courses that would be registered by the students.

In an experiment by Majeed and Junejo [1], they used machine learning and DM techniques to predict the grade of the student before the final examination of the student. The results were highly positive and showed that the predictions made were accurate for 96% of the students. Therefore, prediction provides an important opportunity for both the faculty and students to take timely and informed decisions.

Koutina and Kermanidis [16] tried to find the most efficient machine learning technique that would predict the final grades of the Ionian University Informatics Postgraduate students. They chose five academic courses with an individual dataset and experimented with six well-known classification algorithms. The dataset was enriched with demographics, in-term performance, and in-class behaviour. The classification algorithms Naïve Bayes and 1-NN achieved the best results.

An empirical study carried out by Alsaffar [17] shows that using educational data mining the results of the students can be predicted by adding some synthetic attributes to the classification algorithm. In his work, he also identifies the relevant attributes that play an influential role in the prediction of the results. Using his work, it can be seen what attributes play an important role so that they can be really worked on and improved to enhance the performances of the students. Therefore, it will be helpful to use those algorithms with two synthetic attributes in order to obtain better results.

The EDM studies mentioned above are helpful in their own way and they do help the instructors to carry out some form of prediction of students' performance that would be useful to enhance their studies and get better results rather than fail the course. The gap the research published in this paper fills as explained in section 3 is to use data mining techniques to extract student's achievements and predict his/her expected achievements in future courses as part of the academic program. A student's achievements in the eleven SOs is a reliable indicator of his/her abilities that can be used to predict performance in future courses. Their use as input variables for the prediction algorithms is unique and have never been used before so is the prediction approach in which the most accurate predicted mark for a course from two different processing methods is selected.

3. PROPOSED EDUCATIONAL DATA MINING SYSTEM

A. *Existing System for Course Assessment and the Achievement of SOs Reports*

The Department of Computer Science and all departments in the College of Information Technology requires course instructors to prepare the CILO-SO assessment report for each taught course. This report is currently developed as a series of linked excel sheets maintained by course instructors. The course worksheet has three input sheets. In the first, the instructor inputs the course details including the number of students and the mapping of the course CILOs to the program SOs in a tabular form. In the second, the instructor maps the course assessment activities such as quizzes, tests, assignments, projects and final examination questions to the CILOs. The third sheet holds the list of registered students and the marks they score in each of the assessment activities carried out during the course. The CILOs for the courses map via assessment activities to the appropriate program SOs they contribute to, and courses contribute to different SOs with different weights. The eleven SOs are referred at as "a to k" in the college as discussed in the introduction. For example, the course ITCS341 maps and contributes to the SOs a, b, c, d, i, j and k only. This mapping is established through the different course assessments activities resulting to the course contributing to each SO with a different weight. For example, ITCS341 contributes to SO 'a' with 22.730 points out of 352.551 overall points and to 'd' with 91.186 out of 352.551. The CILO report for the course is automatically generated by the excel worksheet (using hidden formulas and links between the various sheets) showing the students' (as a group belonging to a course) achievements towards each of the CILOs as a percentage. Similarly, the SO report is automatically generated by the excel program for the course presenting the students' achievements towards each of the SOs as a percentage. The sole purpose of this report is to present the extent to which the students have achieved the CILOs and SOs for the course.

B. *Proposed EDM System Architecture*

The architecture of the proposed system shown in Fig. 1 is based on the data mining model proposed by Fayyad, Piatetsky-Shapiro, and Smyth [18]. The first processing step is data cleaning and extraction; the clean data then undergoes a prepossessing step involving data selection, mapping, and classification and filtering to produce the transformed data; the system then mines hidden data to determine a particular student's achievement; and then prediction rules are applied to determine his/her expected future achievement. The three major processing stages are described in more details in the following sections 3.C through to 3.E.

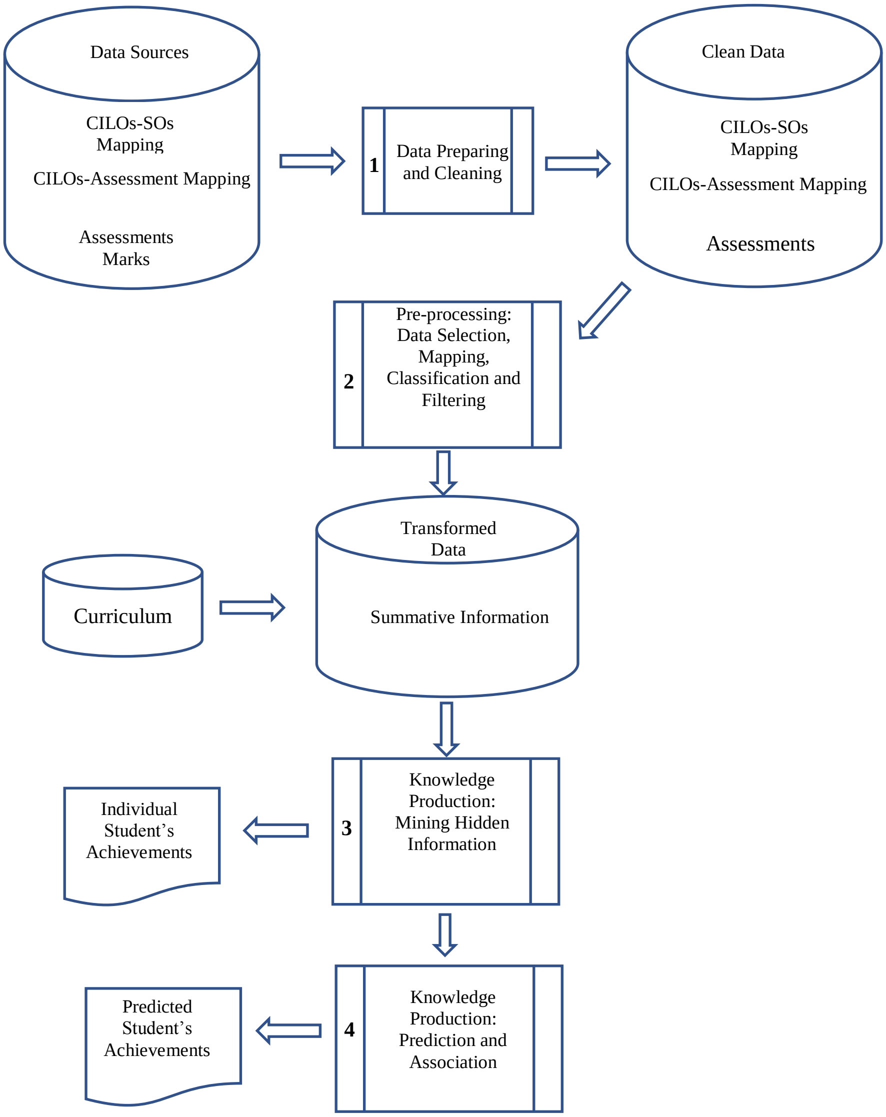
Data Sources
CILOs-SOs Mapping
CILOs-Assessment Mapping
Assessments Marks
1
Data Preparing and Cleaning
Clean Data
CILOs-SOs Mapping
CILOs-Assessment Mapping
Assessments
2
Pre-processing: Data Selection, Mapping, Classification and Filtering
Transformed Data
Summative Information
Curriculum
3
Knowledge Production: Mining Hidden Information
Individual Student's Achievements
4
Knowledge Production: Prediction and Association
Predicted Student's Achievements

Figure 1. Proposed system architecture.

C. Data Preparation and Pre-processing

The core data for each taught course is available in three different files. The first is "CILOs-SOs Mapping" holds general information about the course such as title, instructor, academic year and semester, and section number; and more importantly it holds course CILOS mapped to the SOs. The second is "CILOs-Assessment Mapping" which holds the details of mapping each course assessment to the relevant CILOs. The third is "Assessments Marks" which maintains the details of all assessment results including overall mark and grade for all students attended and completed the course in a given semester. Following the DM processing steps, data cleaning had to be performed first to ensure that input data is uniform and clean from unnecessary control characters and noise. This processing stage is shown in Fig. 3 as "1. Data Preparing and Cleaning". The second processing step was to write a program that extracts and processes the students' assessment details from the three input files and record them in a new file holding "summative information", in which there is an entry for each completed course by a student. The degree program has 43 courses for which 1,563 students are enrolled in and a total of 4,247 completed course records. The abstract algorithm shown in Fig. 2 outlines this process and is shown in Fig. 1 as "2. Pre-processing: Data Selection, Mapping, Classification, and Filtering".

❖ For each course:

- Course details are selected such as code, title, section, semester and year;
- Mapping, of course, CILOs to SOs are read and stored in a table;
- For each student:
    - ○ Mapping of assessments to CILOs are read and stored in a table;
    - ○ Student's assessments marks are read;
    - ○ Student's achievements of SOs are calculated using the mapping of assessments to CILOs and CILOs to SOs;
    - ○ Student's total SO achievements and highest value of each SO are calculated and the percentage achievements for each SO is calculated;
    - ○ Results are written to the summative information data file.

Figure 2. Processing: data selection, mapping, classification, and filtering.

D. *Mining Hidden Knowledge: Student's Achievements Report*

As discussed in section 3.C the summative achievements of all students in each course they completed are available in the summative information file. For a particular student, the program reads all completed courses and keeps an overall cumulative record of the student's achievements in each of the eleven SOs by adding up achievements from individual courses for each of the SOs. A line of entry is written for each course to the output file followed by summative results of SOs achievements. It presents the student with his/her overall achievements so far in each of the SOs as a percentage. The algorithm shown in Fig. 3 outlines this process which is shown as "3. Knowledge Production: Mining Hidden Information" in Fig. 1. The bar chart shown in Fig. 4 presents a sample student's SOs achievements in all courses s/he completed.

❖ For a given student:

- For each completed course:
    - ○ The student's course achievements are read;
    - ○ The course grade and each SO achievement is recorded/reported as a percentage of the overall for the SO;
    - ○ The total SO for the course is also recorded/reported.
- The cumulative SO achievements are recorded/reported.

Figure 3. Knowledge production: mining hidden information.

E. *Predicting Student's Future Achievements*

The students are assessed to the extent they achieve in the SOs based upon which a course grade is given. Therefore, a strong correlation between the two exists. This relationship is used to predict students' future grades in courses based on their latest achievements in the SO's. The data available to us are of two types. The first is their overall achievements in the SOs taken as a mean, and the second is their achievements in each of the SOs. To determine the significance of our hypothesis and data available to us Minitab [19] was used. At first, the single regression model was analysed with sample data from a few courses. The predictor (X) being the average SO achieved by the student so far and Y the resultant mark out of 100. Fig. 5 shows the regression analysis for the model using sample data from a course. As shown in the table the P value for the F-test is 0 suggesting that the model provides a better fit than the intercept-only model and is statistically significant. R-sq is high and R-sq(pred) is close to it, the P-value for the T-Value is significantly low and VIF for X is 1. All these indicators suggest that the model is highly reliable to predict new observations.

Since there are individual contributors (i.e. SO achievements) to the students' overall mark for a course the data was analysed for the possibility of a multiple linear regression model. The predictors would be the students' SO achievements required for a course to determine his resultant mark. Fig. 6 shows the regression analysis for the model. The overall P value for the F-test is 0 suggesting that the model provides a better fit than the intercept-only model. R-sq, R-sq(adj) and R-sq(pred) are high and the latter two measures are close to the first. There were concerns with the high value of VIF (more the 10) for a number of predictors as shown in the table suggesting high multicollinearity. A well-known solution to the multicollinearity problem is to use least squares estimation which was implemented to carry out multi-variant regression to obtain a prediction for the multiple variants.

Based on the curriculum plan, the grade for each of the unregistered courses is determined by the prediction programs. The two supervised learning techniques discussed in the last two paragraphs were implemented using the algorithms published by Sedgewick and Wayne [20]. Out of the two predictions the best was chosen based on the standard error (S) calculated during the process. The assessments in every course contribute to several SOs, it was therefore, more appropriate to choose multiple linear regression as the first choice for prediction. The formula that sets the relationship between the course SOs and the overall mark is as follows "mark = β0 + β1SOa + ... + β11SOk". The number of SOs a course contributes to, differ from one to another and the prediction program had to generate the number of beta (β) values suitable for each course. The student's cumulative SOs achievements in all completed courses are fed to the multiple linear regression program. The second choice was single linear regression using the student's overall SOs mean as an input calculated as "mark = β0 + β1μ". At first, the code for the course to be predicted is read and if the prerequisite for the course has been completed the code is passed to the prediction program. All available entries in the summative information file for students who completed the same course in previous semesters are read to be used as input values set for the two regression programs to generate the appropriate regression equations. The mark range for each letter grade is also established from historical records for the same course. At the same time, the standard deviation (S) of the errors of predictions is calculated as shown in (1) where σest is the standard error of the estimate, Y is the actual score for a course, Y' is the predicted scores for a course and N is the number of scores. This value helped us to decide on the most accurate prediction method for the course, and that is the method with the lowest S value. Fig. 7 shows the algorithm to predict outcomes of future courses.

Table I shows the prediction table for ITCS452. The results of the two prediction methods are shown and compared with the actual course mark and S is calculated. In this example, single linear regression had the lowest S value and thus the most accurate prediction. Fig. 8 shows a scatter chart with the actual marks (in blue) and the single linear regression predicted marks (in red). Table I also shows the grades for each of the marks. The grades for the predicted marks are determined based on the marks' ranges for the grades from previously completed ITCS452 courses, i.e. in a previous semester a student scoring 65.5 would have earned D+ and a student who scored 87.8 would have earned 'A–'. Table II shows the actual and predicted marks and grades for ITCS341. Multiple linear regression had the least S and therefore the most reliable prediction. Similarly, to ITCS452 grades for ITCS341 are set based on previously used ranges, i.e. a student scoring 89.7 would have earned 'A–' in earlier semesters and 'A' grade started at a mark of 90 and above. Fig. 9 shows the scatter chart for the actual marks (in blue) and the multiple linear regression marks (in red).

$$\sigma_{est} = \sqrt{\frac{\sum (Y - Y')^2}{N}} \qquad (1)$$

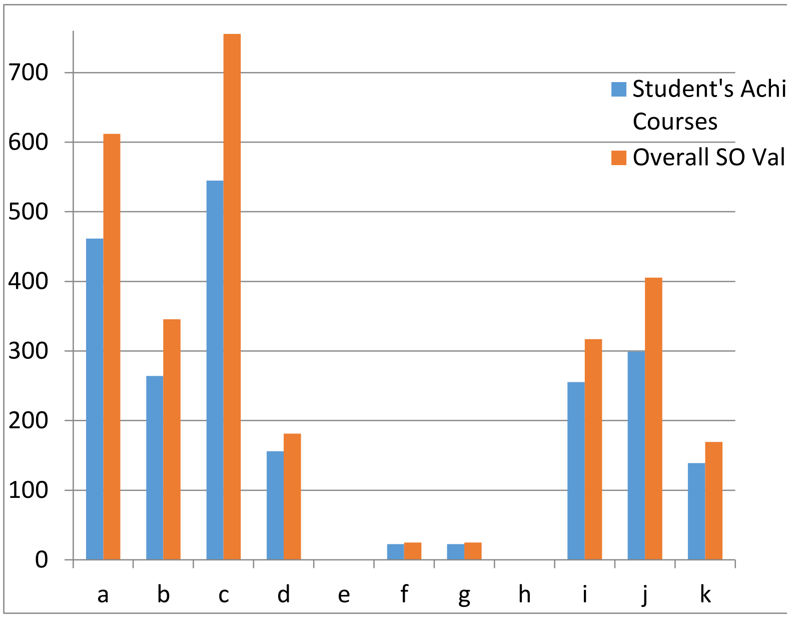

Figure 4. Bar chart showing a sample of a student's SOs achievements in completed courses.

```
Analysis of Variance

Source       DF   Adj SS    Adj MS   F-Value   P-Value
Regression    1   3934.5   3934.47   1071.91     0.000
  X           1   3934.5   3934.47   1071.91     0.000
Error        52    190.9      3.67
Total        53   4125.3

Model Summary

      S     R-sq  R-sq(adj)  R-sq(pred)
1.91586   95.37%     95.28%      93.73%

Coefficients

Term       Coef  SE Coef  T-Value  P-Value   VIF
Constant   4.91     2.40     2.05    0.045
X         95.76     2.92    32.74    0.000  1.00

Regression Equation

Y   =   4.91 + 95.76 X

Fits and Diagnostics for Unusual Observations

Obs       Y     Fit   Resid  Std Resid
  2  58.000  50.328   7.672       4.75  R  X
  3  65.000  69.212  -4.212      -2.27  R
  8  75.000  79.386  -4.386      -2.31  R
 40  68.000  73.886  -5.886      -3.13  R
R  Large residual
X  Unusual X
```

Figure 5. Regression analysis: Y versus X.

```
Analysis of Variance

Source       DF   Adj SS    Adj MS   F-Value   P-Value
Regression    6  4001.47   666.912    253.06     0.000
  X1          1   134.10   134.097     50.88     0.000
  X2          1     0.16     0.158      0.06     0.807
  X3          1    29.66    29.663     11.26     0.002
  X4          1     1.85     1.849      0.70     0.407
  X5          1     2.55     2.550      0.97     0.330
  X6          1     2.95     2.946      1.12     0.296
Error        47   123.86     2.635
Total        53  4125.33
Model Summary

      S     R-sq  R-sq(adj)  R-sq(pred)
1.62339   97.00%     96.61%      94.15%
```

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	5.91	4.36	1.35	0.182	
X1	1.021	0.143	7.13	0.000	3.76
X2	-0.054	0.219	-0.25	0.807	12.60
X3	0.814	0.243	3.35	0.002	53.03
X4	0.209	0.250	0.84	0.407	83.57
X5	0.530	0.538	0.98	0.330	1.75
X6	-0.283	0.268	-1.06	0.296	10.97

Regression Equation

Y = 5.91 + 1.021 X1 - 0.054 X2 + 0.814 X3 + 0.209 X4 + 0.530 X5 - 0.283 X6

Fits and Diagnostics for Unusual Observations

Obs	Y	Fit	Resid	Std Resid	
2	58.000	50.514	7.486	5.72	R
3	65.000	68.048	-3.048	-2.31	R
49	66.000	68.995	-2.995	-2.25	R

R Large residual

Figure 6. Regression analysis: Y versus X1, X2, X3, X4, X5, X6.

- For a given student:
  - The curriculum plan is read
  - For each uncompleted course with a completed pre-requisite
    - The prediction algorithms single- and multi-valued regression are trained using existing students' achievements and scores who completed the same course
    - The lower mark for each awarded letter grade is also recorded
    - Prediction is made using the multi-valued method
    - Prediction is made using the single-valued method
    - Based on applying the two methods on the actual data the most accurate method is determined using standard deviation
    - The regression method to determine the expected mark for the course is applied
    - The grade is determined based on completed courses allocation of grades to marks

Figure 7. Knowledge production: prediction and association.

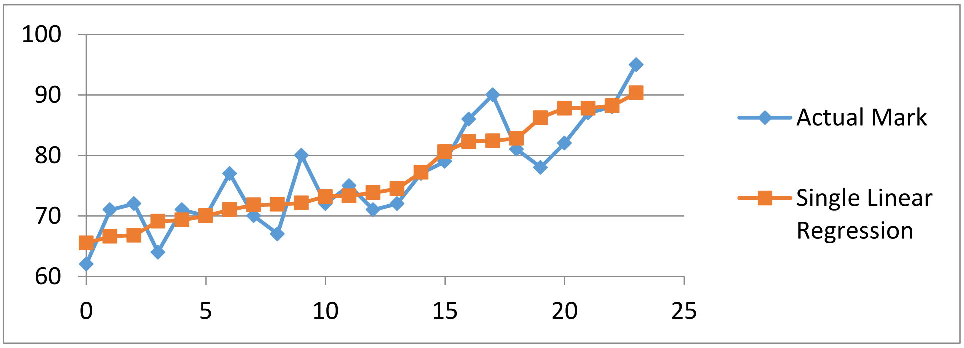

Figure 8. Scattered chart showing single linear prediction results compared with actual results.

TABLE I. PREDICTIONS COMPARED TO ACTUAL MARK FOR ITCS452.

SR No.	C-code	Grade	Actual Mark (Y)	Multiple Linear Regression (YM')	Diff^2 (Y-YM')^2	Grade	Single Linear Regression (YS')	Diff^2 (Y-YS')^2	Grade
0	ITCS452	D	62	56.5	30.25	F	65.5	12.25	D+
1	ITCS452	C	71	65.7	28.09	D+	66.6	19.36	D+
2	ITCS452	C	72	64	64	D+	66.8	27.04	D+
3	ITCS452	D+	64	63.7	0.09	D	69.1	26.01	C-
4	ITCS452	C	71	67	16	C-	69.3	2.89	C-
5	ITCS452	C	70	66.9	9.61	D+	70	0	C
6	ITCS452	B-	77	69.5	56.25	C-	71	36	C
7	ITCS452	C	70	72.7	7.29	C+	71.8	3.24	C
8	ITCS452	C-	67	69.4	5.76	C-	71.9	24.01	C
9	ITCS452	B	80	72.5	56.25	C+	72.1	62.41	C+
10	ITCS452	C	72	69.3	7.29	C-	73.2	1.44	C+
11	ITCS452	C+	75	63.7	127.69	D	73.3	2.89	C+
12	ITCS452	C	71	71.5	0.25	C	73.8	7.84	C+
13	ITCS452	C+	72	73.6	2.56	C+	74.5	6.25	C+
14	ITCS452	B-	77	79.3	5.29	B-	77.2	0.04	B-
15	ITCS452	B-	79	78.4	0.36	B-	80.6	2.56	B
16	ITCS452	B+	86	72.4	184.96	C+	82.3	13.69	B
17	ITCS452	A	90	85.8	17.64	B	82.4	57.76	B
18	ITCS452	B	81	84.7	13.69	B	82.8	3.24	B
19	ITCS452	B-	78	79.9	3.61	B-	86.2	67.24	B+
20	ITCS452	B	82	82.7	0.49	B	87.8	33.64	A-
21	ITCS452	A-	87	85.7	1.69	B	87.8	0.64	A-
22	ITCS452	A-	88	88.4	0.16	A-	88.2	0.04	A-
23	ITCS452	A	95	93.2	3.24	A	90.3	22.09	A
S					5.174094			4.245439	

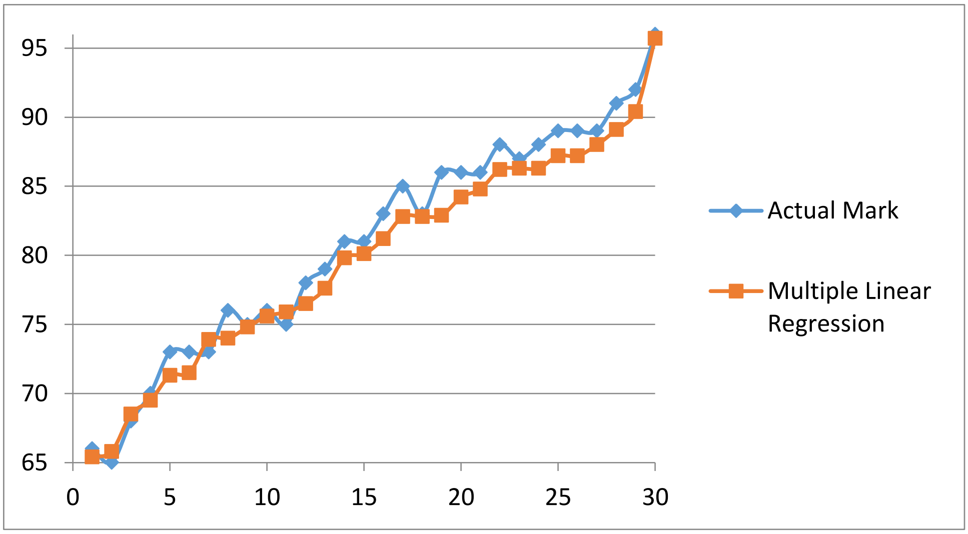

Figure 9. Scattered chart showing multiple linear prediction results compared with actual results.

TABLE II. PREDICTIONS COMPARED TO ACTUAL MARK FOR ITCS341.

SR No.	C-code	Grade	Actual Mark (Y)	Multiple Linear Regression (YM′)	Diff^2 (Y-YM′)^2	Grade	Single Linear Regression (YS′)	Diff^2 (Y-YS′)^2	Grade
1	ITCS341	D+	66	65.4	0.36	D+	68.6	6.76	C-
2	ITCS341	D+	65	65.8	0.64	D+	69.2	17.64	C-
3	ITCS341	C-	68	68.5	0.25	C-	73.9	34.81	C
4	ITCS341	C	70	69.5	0.25	C-	70.9	0.81	C
5	ITCS341	C	73	71.3	2.89	C	72.5	0.25	C
6	ITCS341	C	73	71.5	2.25	C	72.1	0.81	C
7	ITCS341	C	73	73.9	0.81	C	76.2	10.24	C+
8	ITCS341	C+	76	74	4	C+	74.9	1.21	C+
9	ITCS341	C+	75	74.8	0.04	C+	75.7	0.49	C+
10	ITCS341	C+	76	75.6	0.16	C+	76.8	0.64	C+
11	ITCS341	C+	75	75.9	0.81	C+	79.4	19.36	B-
12	ITCS341	B-	78	76.5	2.25	C+	78.4	0.16	B-
13	ITCS341	B-	79	77.6	1.96	B-	79.2	0.04	B-
14	ITCS341	B	81	79.8	1.44	B-	80.6	0.16	B
15	ITCS341	B	81	80.1	0.81	B	81.8	0.64	B
16	ITCS341	B	83	81.2	3.24	B	82	1	B
17	ITCS341	B+	85	82.8	4.84	B	83.8	1.44	B+
18	ITCS341	B+	83	82.8	0.04	B	85.4	5.76	B+
19	ITCS341	B+	86	82.9	9.61	B	83.7	5.29	B+
20	ITCS341	B+	86	84.2	3.24	B+	85.3	0.49	B+
21	ITCS341	B+	86	84.8	1.44	B+	85.8	0.04	B+
22	ITCS341	A-	88	86.2	3.24	B+	87.1	0.81	A-
23	ITCS341	A-	87	86.3	0.49	B+	86.4	0.36	B+
24	ITCS341	A-	88	86.3	2.89	B+	87.4	0.36	A-
25	ITCS341	A-	89	87.2	3.24	A-	87.4	2.56	A-
26	ITCS341	A-	89	87.2	3.24	A-	87.6	1.96	A-
27	ITCS341	A-	89	88	1	A-	89.1	0.01	A-
28	ITCS341	A	91	89.1	3.61	A-	89.7	1.69	A-
29	ITCS341	A	92	90.4	2.56	A	90.7	1.69	A
30	ITCS341	A	96	95.7	0.09	A	96.7	0.49	A
S					1.433992			1.983011	

4. Discussion

The input data files discussed in section 3.C are primarily maintained to assess students' performance as a group registered in a course. Using data mining techniques, the necessary information to produce individualized students' reports of two types are extracted. The first, 'descriptive' presenting the student's current performance so far in the enrolled program in terms of the level of achievement in each of the SOs s/he was assessed in. The second, 'predictive' primarily based on the student's current level of achievement as determined in the first part, and at the same time, it is compared to the performance of other students who completed courses not yet registered by this student. Using single linear and multivalued linear regression techniques, we were able to determine the students expected performance and grades in courses s/he will register for in the coming semesters. The standard error obtained while applying the prediction methods on actual data from past records is used to determine the accuracy of results. The results show that there is no single best method for prediction of the student's future achievements, and accuracy varied from one course to another based on historical data for the course.

Existing approaches to student performance analysis and prediction presented in section 2 varied on the data upon which the analysis and prediction were based. The data included past academic records, classroom participation, demographics, social background, performance in assignments, attendance, and other data that is general and cannot be directly linked to the nature of topics taught in the courses. Unlike other work, our approach in students' performance analysis and prediction is based on the set of SOs for the whole program as discussed in section 1 and presented in Appendix A. All courses offered as part of the B.Sc. in computer science program map to one or more of the same set of SOs (which are used to measure the students' acquired knowledge, abilities and skills) and therefore can be more accurate indicators/predictors of a student's abilities based upon which achievements in future courses can be predicted. Furthermore, two techniques are used to predict a future course grade and the most accurate based on the input prediction variables and historical data is selected. A detailed analysis of prediction models (single and multiple linear) using Minitab is presented in subsection 3.E. Hence, the approach followed in this paper is unique compared to other approaches presented in section 2.

5. Conclusion

In this paper, data mining techniques were presented as tools to process courses' assessment data to generate useful information that can be used to improve learning abilities and students' performance. A review of the literature for background information on EDM was carried out as presented in section 2. The proposed system that prepares students' assessment data (primarily based on ABET-Accreditation Board for Engineering and Technology accreditation requirements) available to us for mining is presented in section 3. Students learning outcomes from completed courses and their mapping to SOs are mined and individualised students' achievements are extracted, presented and future achievements are predicted, such useful information has a major positive impact on learning abilities and students' progress. It represents important feedback to students from which they can locate their areas of weaknesses and strengths. This feedback would help students and their advisors to prepare personalised plans to improve one or more specific personal skill(s) and maintain others of strengths. The findings clearly demonstrate that the classification of students learning outcomes for courses into broad knowledge and skill areas (SOs) and mapping assessment outcomes to them enable accurate prediction of expected future achievements using DM techniques. A major challenge that is unique in the work presented here compared to other similar research work is the use of single and multi-valued regression techniques based on students' SOs values to predict future performance. The work showed that students' achievements in the eleven SOs is a very reliable indicator of their abilities based upon which future achievements can be predicted. The accuracy of each prediction of the two methods used in this work is determined for each course and the most accurate is chosen to decide on the expected student grade for a course. As discussed in section 4 the work presented here is unique compared to similar work presented in section 2 and presents a contribution to the field.

References

[1] Majeed, A. E. and Junejo, K. N. (2016). "Grade Prediction Using Supervised Machine Learning Techniques". In: 4th Global Summit on Education GSE 2016.

[2] Mueen, A., Zafar, B. and Manzoor, U. (2016). "Modeling and Predicting Students' Academic Performance Using Data Mining Techniques". International Journal of Modern Education and Computer Science, 8(11), pp. 36–42.

[3] Kotsiantis, S. B. (2012). "Use of machine learning techniques for educational purposes: A decision support system for forecasting students grades," Artificial Intelligence Review, 37(4), pp. 331–344.

[4] Charu, C. A. (2014). "An Introduction to Data Classification", Data Classification, Chapman and Hall/CRC, pp 1–36, 2014.

[5] Hongbo, D., Yizhou, S., Yi, C., and Jiawei, H. (2014) "Probabilistic Models for Classification," Data Classification, Chapman and Hall/CRC, (pp. 65-86), 2014.

[6] Linan, L., C. and Perez, A., A., J. (2015). "Educational Data Mining and Learning Analytics: differences, similarities, an time evolution". International Journal of Educational Technology in Higher Education, 12, Springer

[7] Arnold, D. and Sangra, A (2018). "Dawn or dusk of the 5th age of research in educational technology? A literature review on (e-) leadership for technology-enhanced learning in higher education (2013–2017)". International Journal of Educational Technology in Higher Education, 2018, 15:24.

[8] Ranjan. J. and Malik. K., (2007) "Effective educational process: a data‐mining approach", VINE, 37(4), pp. 502‐515, https://doi.org/10.1108/03055720710838551.

[9] Ahmad, F., Ismail, N., and Aziz, A. (2015). "The prediction of students' academic performance using classification data mining techniques". Applied Mathematical Sciences, 9, pp. 6415–6426.

[10] Huang, S. and Fang, N. (2013). "Predicting student academic performance in an engineering dynamics course: A comparison of four types of predictive mathematical models". Computers & Education, 61, pp. 133–145.

[11] Abu, A. (2016). 'Educational Data Mining & Students' Performance Prediction'. International Journal of Advanced Computer Science and Applications, 7(5).

[12] Althaf, S.K., Basha, H., Ramesh Kumar, Y.R., Govardhan, A. and Ahmed, M. Z., (2012). "Predicting Student Academic Performance Using Temporal Association Mining". International Journal of Information Science and Education. 2(1), pp. 21–41.

[13] Angeline. D (2013). "Association Rule Generation for Student Performance Analysis using Apriori Algorithm". The SIJ Transactions on Computer Science Engineering & its Applications (CSEA). 1(1).

[14] Elbadrawy, A., Polyzou, A., Ren, Z., Sweeney, M., Karypis, G., and Rangwala, H. (2016) "Predicting Student Performance Using Personalized Analytics," in Computer, vol. 49, no. 4, pp. 61–69, Apr. 2016.

[15] Al-Saleem, M., Al-Kathiry, N., Al-Osimi, S. and Badr, G. (2015). "Mining Educational Data to Predict Students' Academic Performance". Machine Learning and Data Mining in Pattern Recognition, pp. 403–414.

[16] Koutina, M. and Kermanidis, K. (2011). "Predicting Postgraduate Students' Performance Using Machine Learning Techniques". IFIP Advances in Information and Communication Technology, pp. 159–168.

[17] Alsaffar, A. H. (2017) "Empirical study on the effect of using synthetic attributes on classification algorithms", International Journal of Intelligent Computing and Cybernetics,10(2), pp. 111–129, https://doi.org/10.1108/IJICC-08-2016-0029.

[18] Fayyad, U., Piatetsky-Shapiro, G., Smyth, U. (1996). "From Data Mining to Knowledge Discovery: An Overview", In Fayyad, Piatetsky-Shapiro, G., Smyth, U. Advances in Knowledge Discovery and Data Mining, AAAI Press / The MIT Press, Menlo Park, CA, pp. 1–34.

[19] Minitab. Available at http://www.minitab.com [Accessed 1st November 2018].

[20] Sedgewick Robert and Wayne Kevin (2007). Introduction to Programming in Java: An Interdisciplinary Approach. Published by Addison-Wesley Publishing Company, USA 2007. Available at: https://introcs.cs.princeton.edu/java/home/ [Accessed 3rd July 2018].

Dr Fawzi Albalooshi is currently a faculty member in the department of computer science at the college of IT in the University of Bahrain since 1996. Has published many research articles in international journals and conferences. He has authored and edited books in Computer Science. Earned his PhD in 1996 from the University of Wales in the field of Software Engineering. His research interests include software engineering, programming languages and algorithms, intelligent decision support systems and data mining. Main contact email is falblooshi@uob.edu.bh.

Dr Hadeel Alobaidy is an Assistant Professor at the University of Bahrain, College of Information Technology, Department of Computer Science, specializing in Software Engineering and web engineering. Her current research interests Human-Computer interaction, Software Engineering, Web engineering, Ontology, Semantic Web and Data Mining.

Dr Amal Saleh Ghanem received her BSc Degree in Computer Science from the University of Bahrain, the Kingdom of Bahrain in 2001; MSc Degree in Computer Science from Manchester University, the UK in 2004; and PhD Degree in Computer Science from Curtin University of Technology, Australia in 2010. Her research interests include data mining, machine learning, and database management systems. Currently, she is an Assistant Professor in the Department of Computer Science at the University of Bahrain. Amal can be contacted at aghanim@uob.edu.bh.

Appendix A - Student Outcomes for Computer Science Program

a. An ability to apply knowledge of computing and mathematics appropriate to the program’s student outcomes and to the discipline.
b. An ability to analyse a problem and identify and define the computing requirements appropriate to its solution.
c. An ability to design, implement, and evaluate a computer-based system, process, component, or program to meet desired needs.
d. An ability to function effectively on teams to accomplish a common goal.
e. An understanding of professional, ethical, legal, security and social issues and responsibilities.
f. An ability to communicate effectively with a range of audiences.
g. An ability to analyse the local and global impact of computing on individuals, organizations, and society.
h. Recognition of the need for and an ability to engage in continuing professional development.
i. An ability to use current techniques, skills, and tools necessary for computing practice.
j. An ability to apply mathematical foundations, algorithmic principles, and computer science theory in the modelling and design of computer-based systems in a way that demonstrates comprehension of the trade-offs involved in design choices.
k. An ability to apply design and development principles in the construction of software systems of varying complexity.