2018 International Conference on Advances in Computing and Communication Engineering (ICACCE-2018)
Paris, France 22-23 June 2018

Predicting Academic Performance of Students in UAE Using Data Mining Techniques

Ayesha Anzer, Hadeel A. Tabaza, Jauhar Ali
College of Engineering
Abu Dhabi University
Abu Dhabi, UAE
1063572@students.adu.ac.ae, 1007353@students.adu.ac.ae, jauhar.ali@adu.ac.ae

***Abstract*—In this paper, we implemented an approach to predict final exam scores from early course assessments of the students during the semester. We used a linear regression model to check which part of the evaluation of the course assessment affects final exam score the most. In addition, we explained the origins of data mining and data mining in education. After preprocessing and preparing data for the task in hand, we implemented the linear regression model. The results of our work show that quizzes are most accurate predictors of final exam scores compared to other kinds of assessments.**

Keywords— Prediction; Educational data mining; Linear regression; Academic performance

I. INTRODUCTION

The main objective of academic institutes is to offer quality education to their students. One of the important measure of quality education is students' academic performance. Identifying factors that affect the academic performance of students is of utmost priority for any academic organization.

Data mining is an analytical process that explores large data sets to establish relationships and identify patterns to find solutions for problems using data analysis. The tools of data mining allow enterprises to have predictions for future trends in the market. Predictive analytics is a common type of data mining which is widely used by organizations to make better business decisions [1]. The data mining process is made up of several steps as follows: data collection, pre-processing data, applying data mining, interpretation, evaluation, and deploying the results.

Data mining tools can be used in many applications. We decided to use data mining tools for Educational Data Mining (*EDM*). EDM is a developing discipline involved in establishing methods and approaches to have a better understanding of students and the learning setting they are in. It is used to explore the distinctive large-scale data which is growing rapidly. The data usually has a meaningful hierarchy, which needs to be decided by the data properties. Matters like sequence, time, and context have a significant role in educational data study [2].

The objective of this study is to predict students' performance in the final exam using their previous grades during the semester and see how accurate the model's predicted results are compared to the real results. To achieve this objective, we used students' data at Abu Dhabi University. To make predictions we opted to use linear regression as this is a commonly used data mining technique to predict continuous values.

We noticed some studies about the prediction of academic performance of students in multiple universities using different analysis tools, so we wanted to conduct our study based on the tools we see are suitable to give better results. The study is very much important, as the previous studies have provided with limited knowledge about the prediction of the academic performances. The academic performance is prioritized in this research for the development of knowledge about the students in UAE.

The rest of the paper is organized as follows. Section II discusses the previous work done by other researchers in the same field. Section III discusses the steps of data mining process used to carry out the research such as data collection, pre-processing, modeling, and evaluating. Section IV discusses the results and limitation of our work. Finally, Section V concludes our work and the results.

II. LITERATURE REVIEW

Yassein et al. [3] studied data of 150 students of Najran University. The software used was SPSS and clementine to identify which known factors can provide a quick indicator of anticipated performance. The techniques used in their research are feature reduction and classification. The result of their study reveals a relationship between assignments and practical work of the course and conclude that these two attributes result in a higher success rate. Also, they indicated that assignments have negative effect on academic performance. Their study also reveals that students' attendance plays an essential role in their academic performance. It is evident that the more the students are attentive in class the more they will understand the lecture and employ that during different assessments [3].

Ahmed et al. [4] apply data mining techniques to predict and analyze students' academic performance based on their forum participation and academic record. Students' data has been gathered from two different undergraduate courses. The models used are Naïve Bayes, Decision Tree (C4.5), and Neural Network (Multilayer Perception). After that, prediction performance of those classifiers was evaluated and compared. The result showed that Naïve Bayes was highly accurate by 86% than the other classifiers, which were 82.7% (C4.5) and 79.2% (Multilayer Perception) accurate, respectively [4].

Merchan et al. [5] analyzed the records of 932 students of systems engineering from El Bosque University to build a

978-1-5386-4485-0/18/$31.00 ©2018 IEEE

predictive model for students based on their academic performance. Data obtained is evaluated based on predicted input, data output, output depiction, theory, and model which are relevant for prediction accuracy. The results obtained depicts the performance of students through their learning process. Based on results, timely decision can be made to prevent academic risk [5].

It has been observed in Malaysia by Amirah et al. [6] that their existing system lacks to analyze and monitor student performance due to insufficient prediction method and lack of investigation on aspects which affect students' accomplishments. Consequently, a thorough literature review to predict student progress with different data mining techniques is proposed to enhance student achievements. Various data mining techniques used are Decision Tree, Neural Network, Naive Bayes, K-Nearest Neighbor, and Support Vector Machine. The main objective was to stipulate a synopsis of data mining techniques that are used by other researchers to forecast students' performance. Also, to focus on how prediction algorithms could be exploited to distinguish utmost vital attributes in a student's data. Students achievements and success can be enhanced more efficiently by using educational data mining techniques. It will benefit educational institute, educators, and students. As a result, it has been observed that Neural Network has the highest accuracy of 98%, Decision Tress has 91%, K-Nearest Neighbor has 83%, and Naive Bayes has the lowest accuracy of 76% [6].

Pooja et al. [7] provided a comprehensive survey of educations data mining done from the year 2002 to 2014 and its scope in the future. It has been noticed that extensive work is done for the usage of data mining techniques in education but still some areas are untouched and unified approach is not used. The previous work done in this area include student's course satisfaction, finding a set of weak students, faculty evaluations, predicting student drop out, course registration planning, etc. Other researchers came up with a significant relationship between employability and personality preferences such as work and life experience. This relationship is an important aspect of Employability Development Profile. Hence, employability is bound to dispositions and competencies instead of academic qualification [7].

Educational data mining is considered as developing a method to explore unique kinds of data related to educational context. In e-learning system, application of data mining is an iterative cycle. Romero et al. [8] provided a survey for some applications of data mining in learning management system and Moodle system, which is mostly used by universities worldwide. Their objective was to provide both theoretically and practically for all users related to such systems as e-learning administrators, online instructors, etc. This Moodle data mining tool will be used by online instructors by precluding the requirement for CMS administrators to pre-process data and apply data mining techniques. It will automatically pre-process Moodle data and simplify it to configure and execute data mining techniques through its data mining algorithms. The complete process of data mining for e-learning data is systematically described and how data mining techniques can be used for Moodle data like visualization, statistics, clustering, classification, and association rule. They used Weka and Keel systems which are free data mining tools to make it easier for any user to apply data mining without purchasing a commercial tool [8].

III. PROCESS OF DATA MINING IN ACADEMIC PERFORMANCE

This section includes steps used to carry out the data mining task in hand.

A. Collect Data

Data has been collected from Abu Dhabi University during the past four years from 2014 till 2017. The data gathered pertains to a computer programming course. This data has been provided based on years and gender on separate sheets as this course is taught separately to male and female students. These sheets comprise of different attributes like gender, quiz 1, quiz 2, quiz 3, quiz 4, assignment 1, assignment 2, assignment 3, midterm, and final exam grades. Each sheet comprised of a different number of records. All the sheets data was combined giving us a total of 182 examples including both genders.

B. Pre-Process The Data

After accessing the data, we preprocessed the data by modifying it to make all the quiz grades with the same weight, as well as the assignment. We aggregated the marks of all quizzes and all assignments to know the total quizzes and the total assignments grades. We also removed unneeded attributes and modified the missing values of the quizzes by adding the average of the other quizzes to the missing values and converted the final exam grade to a percentage. We also removed quiz 4 as in some semesters, quiz 4 was not been conducted. We applied data mining on 182 examples with 12 regular attributes. We decided to make the final exam grade as a special attribute/label, as this is the attribute we wanted to predict from the other attributes. We also wanted to know which of the other attributes contribute more in accurately predicting the final exam grade.

C. Apply Data Mining

Linear regression is used to predict an academic performance of a student in course assessments of a programming course. The purpose of using linear regression is that it is used to predict numeric and continuous values whereas other techniques like classification use discrete values. It is used to indicate the significant relationships between independent and dependent variables and their strength of impact. Whereas in our scenario, it is a relationship between attributes of course assessments and final exams.

We used the Rapid Miner tool to implement linear regression. The reason we chose to use this tool are [9]:

- Maximizes data productivity.
- Access data of any format.
- Robust statistical overview to explore and understand data quickly using graphical displays like scatter, histogram charts, etc.

- Provides data quality integration, and transformation tools.
- Provides different models like classification, regression, clustering etc. along with multiple operators.
- Offers different validation techniques and performance evaluation.

D. Interpret, Evaluate And Deploy The Results

Rapid Miner is used to import data from an excel sheet. The needed attributes are selected to perform the linear regression. The attribute of gender is converted to numerical to avoid error while applying regression. Final exam is selected as the label by using the Set Role operator. Linear regressions are applied inside the Cross-Validation operator as shown in Fig. 1 and Fig. 2.

The equation used to perform linear regression is:

$$Y = B_0 + B_1x_1 + B_2x_2 + \ldots + B_nx_n + \varepsilon$$

where $x_1, x_2, \ldots x_n$ are dependent variables, $B_1, B_2, \ldots B_n$ are coefficients of regression, B_0 is the y-intercept, ε is the residual (difference between actual and predicted values), and Y is the label value (the value to be predicted).

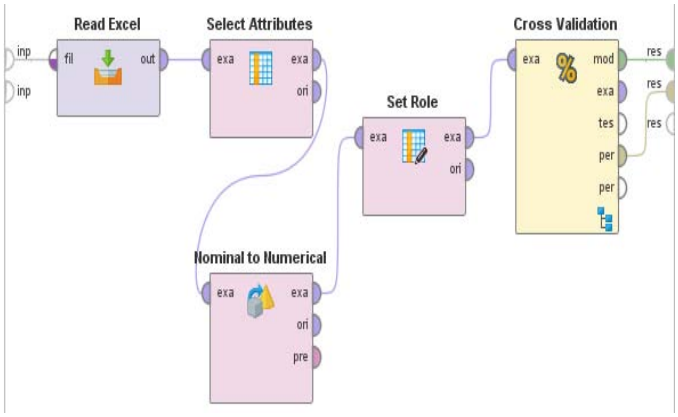

Fig. 1. Design of Model on Rapid Miner

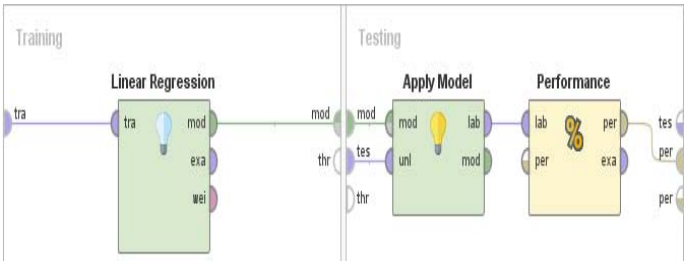

Fig. 2. Cross-validation of data set

In TABLE I, it is seen that there are multiple attributes, the coefficient represents the coefficient of the equation to get the results using each attribute. The standard error is the sampling error estimate, and it is to estimate the variation under the repeated sampling. As for the standardized coefficient, it is the estimate which results from the analysis of the regression, and it was standardized to make the variances of both independent and dependent variables become 1. For the tolerance, it means that it is the ability of the algorithm to learn when the data given is corrupted. The t-stat is the value coming from dividing the coefficient by the standard error. Therefore, the higher the absolute value of the attribute is, the better it is for the regression. The p-value is calculated from the t-stat, and it is the probability of having a result, which is the like the one we have in a random data collection. The code shows which attributes affect the regression the most. We see that we have different values to represent to know which attributes affect the value of the final exam. The attribute column shows which attribute in the excel sheet has more impact on the grade, and the one having more stars has more effect. It is seen that the Total quiz attribute has three stars, and total assignment attribute has two stars, which means that the total quiz has the most effect on the final exam. In the results, we can find t-stat, and p-value. These two values show how we pick the attributes with the most effect. As we explained earlier, the t-stat is the value coming from dividing the coefficient by the standard error. Therefore, the higher the absolute value of the attribute is, the better it is for the regression. The p-value is calculated from the t-stat, and it is the probability of having a result, which is like the one we have in a random data collection. It means that the lower the value of the p-value, the better it is, and the ones with the highest absolute value of t-stat, and the lowest value of the p-value, are the Total quiz, and the Total assignment value. According to the values, it can be noticed that the gender does not have any significance on the regression, so it is decided to run a test on the data after removing the attribute of the gender to show more realistic values. In this model, the squared correlation is 0.396, which is almost 0.4. Squared correlation is always between 0 and 1, and the closer it is to 1, the better it is. The values are not highly correlated according to this result.

TABLE I. LINEAR REGRESSION RESULTS

Attribute	Coefficient	Std. Err.	Std. Coeff.	Tolerance	t-Stat	p-Value	Code
Male	-2.11	4.26	-0.03	0.93	-0.50	0.62	
Female	2.11	4.26	0.03	0.93	0.50	0.62	
Quiz 1	0.57	1.39	0.05	0.42	0.44	0.66	
Quiz 2	0.85	1.39	0.070	0.32	0.61	0.54	
Quiz 3	-0.30	1.19	-0.03	0.40	-0.243	0.80	
Total assignment	2.76	1.14	0.17	0.85	2.43	0.02	**
Total quiz	3.43	1.26	0.45	0.16	2.72	0.01	***
Interept	-15.54	∞	?	?	0.00	1.00	

In TABLE II, values have been changed, and it's now showing that the Total assignment and the Total quiz attributes have the same effect on the final exam. The values of t-stat should be higher, and p-value should be lower. Therefore, it can be concluded that after running all the data we notice that the Total quiz and the Total assignment affect the marks of the final exam the most. In this model, a squared correlation is of 0.364, and this correlation is slightly lower than the previous one, and it gives a similar result that the values are not highly correlated.

TABLE II. RESULTS AFTER REMOVING THE GENDER ATTRIBUTE

Attribute	Coefficient	Std. Err.	Std. Coeff.	Tolerance	t-Stat	p-Value	Code
Quiz 2	0.49	1.43	0.04	0.30	0.35	0.73	
Quiz 3	-0.71	1.33	-0.06	0.37	-0.53	0.60	
Total assignment	2.83	1.04	0.18	0.85	2.72	0.01	***
Mid term	0.30	0.42	0.05	0.69	0.70	0.48	
Total quiz	1.96	1.28	0.51	0.21	3.10	0.003	***
Intercept	-19.57	9.84	?	?	-1.99	0.05	**

In TABLE III, attributes considered are, quiz 1, quiz 2, quiz 3, total assignment, mid-term, and final exam. It has been noticed that total assignment and quiz 2 has more impact on the final exam as it consists of 3 stars. Based on these attributes, we can compute the values of t-stat and p-value. The highest absolute value of t-stat are a total assignment and quiz two whereas lowest values of the p-value are a Total assignment and quiz 2. Hence, it has been concluded that total assignment and quiz 2 has more impact on a final exam, which means if a student did well in assignment there are more chances for them to get a good mark in their final. According to the values quiz 3 has no importance in regression, so this attribute can be excluded from the data to run a test. In this model, we got a squared correlation of 0.391. The values are not highly correlated according to this result.

TABLE III. RESULTS CONSIDERING QUIZZES AND TOTAL ASSIGNMENT ATTRIBUTES

Attribute	Coefficient	Std. Err.	Std. Coeff.	Tolerance	t-Stat	p-Value	Code
Quiz 1	2.60	1.00	0.20	0.61	2.60	0.01	**
Quiz 2	3.01	1.05	0.25	0.51	2.86	0.01	***
Quiz 3	1.58	0.97	0.14	0.54	1.62	0.11	
Total assignment	3.11	1.03	0.19	0.87	3.02	0.003	***
Intercept	-11.39	9.20	?	?	-1.237	0.29	

In TABLE IV, quiz is contributing the most to the final exam marks. Therefore, after removing the gender, the total assignment, and the total quiz, these results are achieved. According to the code attribute, the quizzes contributing the most to the final exam marks are quiz 1 and quiz 2 equally. Quiz 1, and quiz 2 have the highest t-stat values, which are 2.89, and 3.34 respectively. Also, quiz 1, and quiz 2 have the lowest p-value, which is 0.004, and 0.001 respectively. As for quiz 3, it has less contribution to the final exam marks. For this model, a squared correlation is of 0.352, which is the lowest correlation among all the models we created, but it is still in the range of 0.3.

TABLE IV. RESULTS CONSIDERING QUIZZES ATTRIBUTES

Attribute	Coefficient	Std. Err.	Std. Coeff.	Tolerance	t-Stat	p-Value	Code
Quiz 1	2.94	1.02	0.23	0.60	2.89	0.004	***
Quiz 2	3.54	1.06	0.29	0.50	3.34	0.001	***
Quiz 3	1.70	1.00	0.15	0.51	1.71	0.09	*
Intercept	8.36	6.63	?	?	1.26	0.21	

IV. DISCUSSION

In this paper, multiple linear regression method is used for the evaluation of earlier course assessments of how they affect the final examination score. Apart from that, this article has also provided knowledge about the process of using data mining in academic assessments. The use of linear regression model can also be witnessed in this article as it reveals the knowledge about the quizzes as the most important predictor of the final examination score. Apart from that, the results are accomplished after the eradication of gender criteria from the entire process of research. The results showed that Quiz 1 and Quiz 2 are equally important with highest rating of the t values whereas these quizzes are the lowest conceiver of the p values. The result of Quiz 3 has comparatively lower attributes of contribution of the final examination scores.

A limitation of the study is that it was conducted on the academic performance of the students in a programming course. The results may vary in non-programming courses.

Educational institutes can use the process of data mining to procure effective knowledge of the academic predictions about the examination results of the students. This will help in raising students retention as bad academic performance is a leading reason for many students who do not complete their planned course of studies. The future studies ought to incorporate adequate time to make extensive research using greater sample size to generate more general results for the analysis. In the future, we are considering conducting the study on both programming and non-programming-oriented courses to get better prediction for the results, and a better regression model for the relationships between the attributes used.

V. CONCLUSION

Academic data mining is considered as an emerging field related to developing different ways to explore the distinct type of data derived from educational context. Predicting performance of students can become difficult due to the large volume of educational data. In this paper, we elaborated how the application of data mining techniques can be useful to predict factors which will have more effect on marks of the final exam. Through this prediction, it will be easier for educators and students to improve their learning and teaching processes by knowing which attribute contribute more to achieve good grades. This paper reviewed historic data related to programming course to predict student performance. To carry out data mining, we used linear regression as our data was numeric. We experiment with selecting and avoiding some attributes of the same data set. Through results achieved, it can be concluded that total assignment and quizzes have more impact on the final exam. If a student does well in these factors, they can achieve good marks in the final exam.

REFERENCES

[1] A. Hughes, "What is data mining? - Definition from WhatIs.com", *SearchSQLServer*, 2008. [Online]. Available: http://searchsqlserver.techtarget.com/definition/data-mining. [Accessed: 30-Mar-2018].

[2] *Educationaldatamining.org*, 2018. [Online]. Available: http://educationaldatamining.org. [Accessed: 30-Mar-2018].

[3] Nawal Ali Yassein, Rasha Gaffer M Helali, and Somia B Mohomad, "Predicting Student Academic Performance in KSA using Data Mining Techniques," Journal of Information Technology & Software Engineering, vol. 7, no. 5, pp. 1-5, 2017.

[4] Ahmed Mueen, Bassam Zafar, and Umar Manzoor, "Modeling and Predicting Students' Academic Performance Using Data Mining Techniques," International Journal of Modern Education and Computer Science, vol. 11, pp. 36-42, November 2016.

[5] S. M. Merchan and J. A. Duarte, "Analysis of Data Mining Techniques for Constructing a Predictive Model for Academic Performance," IEEE Latin America Transactions, vol. 14, no. 6, pp. 2783-2788, JUNE 2016.

[6] Amirah Mohamed Shahiria, Wahidah Husaina, and Nuraini Abdul Rashida, "A Review on Predicting Student's Performance using Data Mining Techniques," Procedia Computer Science, vol. 72, no. 1, p. 414 – 422, 23 December 2015.

[7] Pooja Thakar, Anil Mehta, and Manisha, "Performance Analysis and Prediction in Educational Data Mining: A Research Travelogue," International Journal of Computer Applications, vol. 110, no. 15, pp. 60-68, January 2015.

[8] Cristobal Romero, Sebastian Ventura, and Enrique Garcıa, "Data mining in course management systems: Moodle case study and tutorial," Computers & Education, vol. 51, no. 1, p. 368–384, August 2008.

[9] RapidMiner, "RapidMiner Studio Visual Workflow Designer For Data Scientists,". [Online]. Available: https://rapidminer.com/products/studio/. [Accessed 30-Mar-2018].