IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 21, Issue 5, Ser. I (Sep - Oct 2019), PP 67-72
www.iosrjournals.org

Predicting Grade 10 Students performance in Mathematics using Machine Learning

Fatma Mohammed Haider AL-Zadjali, Dr. Syed Zakir Ali
Student of MSc-IT, Department of Computing, Faculty, Department of Computing
Middle East College, Muscat, Sultanate of Oman

Abstract: *The analysis and prediction of students' performance have become an essential topic in the educational sector nowadays. Data mining and machine learning techniques used to investigate data from education institutions. However, assessing the students' academic performance is not an easy task since the students' performance depends on different factors.*

This research proposed to predict students' second-semester mark from the begging of the second semester by taking the first-semester mark as an input variable. A sample of 1300 students with the low performance in Mathematics, was taken from 13 schools especially government schools in the Muscat region for the school year 2018-2019 to develop prediction models which can efficiently predict grade 10 students' marks in Mathematics. Since the second semester mark is unknown; K-means clustering algorithm implemented to find out the mean, minimum and maximum mark for each group and assigned each student the nearest mark to their first semester, once the data is labeled by defining the nearest mark supervised learning regression algorithms; ***linear Regression, Decision tree, and Neural Network were*** *applied to predict the second-semester mark. After receiving students' second-semester mark comparison done between results obtained by algorithms and student Actual mark; it's observed that linear regression and neural network performed as same with the accuracy of 90%. This research will help the teachers to know the students' performance in advance and take suitable action at the right time.*

Keywords: *Neural Network, Decision Tree, Linear Regression, Clustering, Regression*

Date of Submission: 16-09-2019 Date of acceptance: 01-10-2019

I. Introduction

The past several years have observed a rapid development in the application in the field of Artificial Intelligence in the educational area, supported by the evidence that it helps educational institutions to learn useful and new knowledge about students. Educational data mining is an emerging discipline involved with developing approaches for examining the unique types of data that come from educational institutions. One of the important areas of data mining application is the development of student models that would predict student performance to help educators to enhance the structure of their course and early identify those students who need special attention (Wakelamet al.2015).

Education is considered as an essential part of human resource progress as it is the first step for every human movement and ensures attainment of skills and knowledge that allow individuals to enhance their quality of life and increase their efficiency and effectiveness. Increase of social effectiveness will lead to producing of new sources, which improve the economic development of a country (Farooq *et al*.2011).

Students' mathematical achievement is very significant at the national level as shown by the intense interest of Sultanate of Oman on participating in *Trends in International Mathematics and Science Study* (TIMSS). TIMSS is an international assessment of the mathematics and Science achievement of the fourth and eighth-grade students around the world, conducted every four years in the United States (NCES 2019).

As per the rating system in the Ministry of Education the results measures on five benchmarks are Excellent (90-100), Very Good (80-89), Good (65-79), low performance (50-64) and fail (0-49). According to the data given by the Ministry of Education, Sultanate of Oman, there are a total of 8342 students in the Muscat region who are enrolled for grade-10 in the academic year 2018-19~~.~~ After analyzing the data, it is found that the total number of students with low performance are 2233 and the total number of students who failed in first semester are 2186.

These statistics proved that students of grade 10 have a weakness in mathematics hence; predicting weak students' performance before the final exam is very significant for the teacher because by identifying students with a low performance the teacher will be able to inform the students during their study and provide them additional support to improve their performance.

This research attempts to predict students' second semester mark from the begging of the second semester by taking first semester mark as input variable with the use of clustering and regression techniques .

Research Aim

This research aims to predict the students' second-semester mark by taking the first semester mark as an input variable. This done by using supervised and unsupervised machine learning techniques.

Related Work

There is no research conducted on predicting students' marks, most of the study has been done on factors affecting students performance and based on factors the classification done either pass or fail, good, very good and so on.

Machine Learning

Machine learning is a subdivision of artificial intelligence that provides computer application the capability to learn and enhance from experience without being expressly programmed and focuses on developing computer programs that will be able to access data and learn for themselves (Smola and Viswanathan, 2008). In this study the following Techniques and algorithms are used.

Clustering

Clustering is type of unsupervised machine learning method and it is a process of determining cluster or natural grouping within multidimensional data based on some similarity measure, therefor similarity measures are principle components in clustering algorithms. A distance measure is the most popular way to evaluate a similarity measure. The objective of clustering is to determine patterns and structures in high dimensional data and group similar data together. Clustering algorithms are based on two types known as partitional and hierarchical clustering. Algorithms in hierarchical clustering generate a cluster tree by using merging techniques or heuristic splitting .However partitional clustering divide the dataset into number of clusters in order to minimize some criteria for example square error function and be treated as optimization problems. K-means and Fuzzy clustering are types of partitional clustering.

Regression

Regression is type of supervised machine learning techniques used to predict quantitative value and evaluate the relationships between variables. There are many types of regression algorithms such as linear regression, support vector, random forest and decision tree (Dave, 2018).

K-means

k-means algorithm is most widely used it is a method of grouping observations into a particular number of disjoint clusters, each cluster is associated with a centroid and each point is assigned to the cluster with the closest centroid . Number of cluster must be specified by the user, the means in k-mean indicate to averaging of the data. The goal of k-means clustering is to minimize the squared error function and total cluster variance (Omran*et al*.2007).

Linear Regression

Linear regression is one of the most common type of regression method, used to predict the response variable Y depending on the input variable X .The objective is to create a linear relationship between predictor and target variable, therefore we can use this formula to predict the value of Y when X values are available. There are two methods in liner regression as follows

- Simple linear regression :when there is single input variable
- Multiple linear regressions: when there are multiple input variables

Liner regression formula is based on $y= \beta 0 + \beta 1.x$ where y is output variable, $\beta 0$ is intercept, $\beta 1$ is coefficient of x and x is input training data(Dave,2018).

Decision Tree

Decision tree is supervised machine learning algorithm in a tree shaped diagram used in order to identify a course of action each branch of the tree represents a possible decision or reaction. Decision tree can solve both regression and classification problems. **regression tree** is used when target variables is continuous or numerical in nature, regression model fit to the target variables by using each of the independent variables and each split is made based the sum of squared error (Soofi and Awan , 2017).

Neural networks

Neural networks have occurred as an essential tool for classification and regression. The latest massive research activities in classification have proven that neural networks are appropriate to several conventional classification methods. Back propagation is Multi-Layer Perceptions feed forward Artificial Neural network models that map sets of input data to output and involve of a number of neurons characterized into multiple layers and it considered as non-linear models used to solve predication problems. It works by resembling the nonlinear relationship between the input and the output by modifying the weight values (Jebaseeli and Kirubakaran, 2013).

II. Analysis and Discussion of Results

Clustering

The reason behind groping the students into clusters is to predict student second semester mark from the begging of second semester by taking first semester mark as an input variable as mentioned before , to do so k-means clustering applied on the dataset of students' first semester result. The dataset has been trained with 2,3,4,5,6 clusters .The total variances in the dataset that is explained by the clusters as shown in table 1 .We can observed from the below table 1 that cluster 5 has the highest total variances of 95.0%, for this reason 5 clusters selected for the existing dataset which indicated good fit.

Clusters	Total variances
2	72.1 %
3	85.6%
4	90.5%
5	**95.0%**
6	92.6 %

Table1: Total variance Explained by Clusters

After grouping the students into clusters k-means clustering determine the mean marks of each group. Second step, the mean, min, and max mark for each group were identified as well as number of students in each group as shown in Table 2.

Group	Mean Mark	Max Mark	Min Mark	Number of students
1	42.59387	47	39	261
2	52.20286	56	48	350
3	60.32806	64	57	253
4	33.76562	38	29	256
5	22.63889	28	5	180

Table2: Clusters Details

In this study sample of 1300 students with low performance were taken for the school year 2018/2019 the same procedure was followed as mention before ,the students were grouped into 5 clusters; then the student was assigned to the nearest mark to their first semester mark. For example if the student first semester mark is 45 and the mean mark of group is 42.59387, minimum is 39 and max is 47, so the student will be given 47 as his or her mark is nearest to 47, as represented in table3 . After defining the nearest mark for each student regression algorithms will apply to make the prediction. The nearest mark will be the response value or the output variable and the first-semester mark is the input variable.

Student No	First semester mark	Group	Mean Mark	Max Mark	Min Mark	Nearest Mark
1	45	1	42.59387	47	39	47
2	55	2	52.20286	56	48	56
3	64	3	60.32806	64	57	64
4	30	4	33.76562	38	29	33.76562
5	18	5	22.63889	28	5	22.63889

Table3: The procedure of assigning students the nearest mark

Linear Regression

Linear regression has been implemented to predict students' second semester mark after defining the nearest mark as mentioned before. The data divided into two sets training 70% and testing 30%.Summary of the model shows that F statistic and P value tests the null hypothesis that all the model coefficients are 0. Residual standard error is 0.72 that indicate how far observed Y values are from the predicted value, the intercept is 0.202 166 which estimated mean Y value when all X are 0. After receiving students second semester mark comparison was done between student actual mark and predicted mark; the overall differences between the actual and predicted mark is 10%, the correlation between the predicted mark and students' actual mark is 0.9062984 which indicate strong correlation. Table 5 shows sample of comparison between student actual mark and predicted mark.

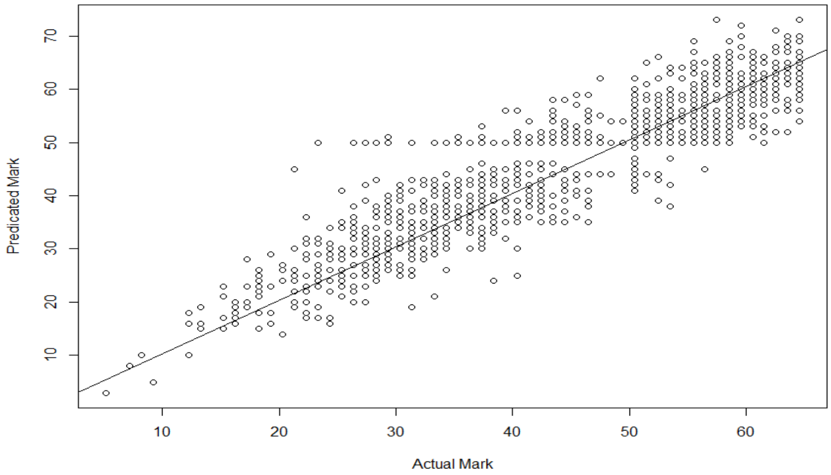

Figure1: The Linear correlation between student Actual mark and predicted mark.

Decision Tree Regression

Decision tree implemented for regression purpose to predict students second semester mark same procedure followed, the data divided into two sets training 70% and testing 30% .After receiving students' second semester mark comparison was done between student actual mark and predicted mark as shown in table 5, the correlation between the actual and predicted mark was done it is about 0.890085 which indicate strong correlation.

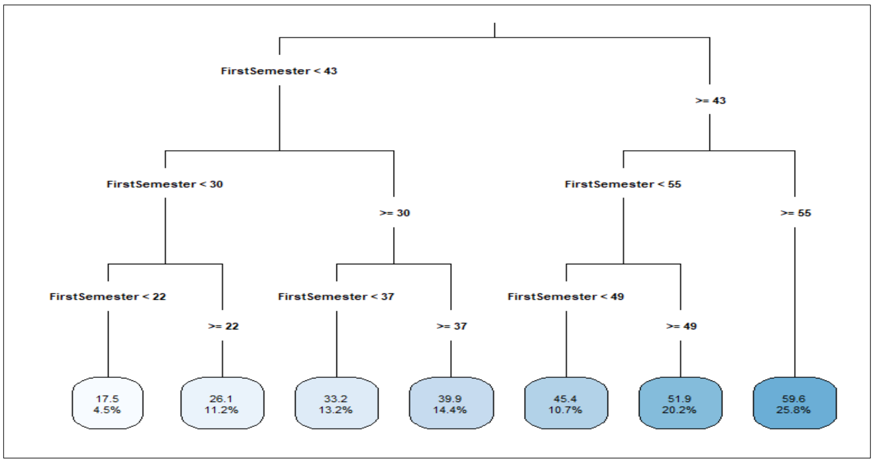

Figure2: Decision Tree rules for predicting students' second semester mark

Neural Network

Neural network has been applied also to predict students' second semester mark and the accuracy of the model was checked by checking the correlation between the students' actual mark and predicted mark and it is 0.906287 which indict strong correlation. Comparison was done between the predicted mark and actual mark as shown in table5.

III. Discussion of Results

For predicting students second semester mark two techniques implemented clustering and regression. Students were grouped into 5 groups as mentioned before and three regression algorithms implemented, the accuracy of algorithms as shown in table4 .We can observed from below table linear regression and neural network have the highest accuracy then decision tree. As it is seen from the results linear regression and neural network models can be successfully used to predict expected mark of grade 10 student's in Mathematics at Sultanate of Oman

Regression Algorithms	Accuracy
linear Regression	0.9062984
Decision Tree	0.890085
Neural Network	0.906287

Table4: Performance of Algorithms

Student ID	Actual Mark	Predicated Mark by linear Regression	Predicated Mark by Decision Tree	Predicated Mark by Neural Network
1	26	30.31502432	33.18939	29.99340936
2	53	52.42455728	51.85514	52.62864148
3	43	45.42698765	45.43263	45.73108986
4	32	27.34902033	26.08904	27.1700946
5	60	62.51628696	59.56583	62.45027923
6	55	50.41304633	51.85514	50.75074545

Table5: Comparison between the Predicted mark and Actual mark

IV. Conclusion

This research has proposed to predict grade 10 students' second-semester mark by taking the first-semester mark as an input variable in Mathematics. This has been done in order to help the teachers to know the students' performance in advance and take suitable action on students such as individual counseling, appropriate advising etc., at the right time which can minimize the failure rate.

This was achieved by implementing machine learning algorithms and techniques. Clustering and regression were used .Results indicated that linear regression and neural network models can be successfully used to predict expected mark of a low-performing student's in Mathematics with accuracy of 90%.

References

[1]. Wakelam ,E.Jefferies ,A, Davey ,N and Sun ,Y(2015)'The Potential for Using Artificial Intelligence Techniques to Improve e-Learning Systems ' [Online] Available from https://pdfs.semanticscholar.org/641e/9fe2856abaa7dbf59c3712aa2ad470d49242.pdf[14 Apr 2019]

[2]. Farooq,M.Chaudhry ,A.Shafiq ,M .and Berhanu ,G.(2011)'FACTORS AFFECTING STUDENTS' QUALITY OF ACADEMIC PERFORMANCE: A CASE OF SECONDARY SCHOOL LEVEL ' Journal of Quality and Technology Management ' [Online]II(II) ,01 - 14Available from https://www.researchgate.net/profile/Muhammad_Farooq108/publication/284150574_Factors_affecting_students'_quality_of_academic_performance_A_case_of_secondary_school_level/links/578c9fbb08ae254b1de84371/Factors-affecting-students-quality-of-academic-performance-A-case-of-secondary-school-level.pdf?_sg%5B0%5D=0P-bV8xN20jraXU3DXpbVHA_YiYxayipg4gbr0_wcc4t_bvuLxc14wrmWoK4uLwDFG06lMBrlYQOW1Cp9S2b0g.f8obC93LH8Mv8dnDmHUsOo-PnfALTqc8f7p0xlHkWkoi2KJgOp9jKmSP724wEcbTXD6IBkEjTVQXuVV2U48NpQ&_sg%5B1%5D=NcD3ZoNFnFCjOXioGkOCugQkuBbfEDglANlFcuohblFayKtkOQRMrhgnrafPfb-3KZKNpdEeIlG2Bae5eJwc6ogYrm6It_rLEosSw5iL8wpa.f8obC93LH8Mv8dnDmHUsOo-PnfALTqc8f7p0xlHkWkoi2KJgOp9jKmSP724wEcbTXD6IBkEjTVQXuVV2U48NpQ&_iepl [18 Apr 2019]

[3]. National Center For Education Statistics(2019)Trends in International Mathematics and Science Study (TIMSS) [Online]Available from https://nces.ed.gov/timss/ [17 Apr 2019]

[4]. Omran, M, Engelbrecht, A and Salman ,A(2007)An Overview of Clustering Methods[Online] Available fromhttps://www.researchgate.net/publication/220571682_An_overview_of_clustering_methods[28 Jul 2019].

[5]. Soofi,A and Awan ,A(2017) 'Classification Techniques in Machine Learning: Applications and Issues'Journal of Basic & Applied Sciences [Online] 13,459-465,Available fromhttps://www.researchgate.net/publication/319370844_Classification_Techniques_in_Machine_Learning_Applications_and_Issues[26 Jul 2019].

[6]. Vijayarekha,K.(2019)Back Propagation Neural Network[Online]Available fromhttps://nptel.ac.in/courses/117106100/Module%208/Lecture%204/LECTURE%204.pdf[30June 2019].

[7]. Smola, A and Vishwanathan,S. (2008) Introduction to Machine Learning [Online] Available fromhttps://alex.smola.org/drafts/thebook.pdf[22Jul 2019].

[8]. Sehra,C(2018)Decision Trees Explained Easily[Online] Available fromhttps://medium.com/@chiragsehra42/decision-trees-explained-easily-28f23241248[28 Jul 2019].

[9]. Brownlee, J. (2016)Supervised and Unsupervised Machine Learning Algorithms[Online],Available fromhttps://machinelearningmastery.com/supervised-and-unsupervised-machine-learning-algorithms/[20June 2019].

Fatma Mohammed Haider AL-Zadjali. "Predicting Grade 10 Students performance in Mathematics using Machine Learning" IOSR Journal of Computer Engineering (IOSR-JCE) 21.5 (2019): 67-72