Projection of Students' Exam Marks using Predictive Data Analytics

Ben George Ephrem[1], Balasupramanian.N[2], Huda Al-Shuaily[3]

Abstract— **One of the utmost responsibilities of the instructor is continuously tracking students' performance during the entire academic year, which plays an important role to understand their learning abilities, attitude and motivation. The prediction of the students' performance will help the students and learning process much easier. The field of Business Intelligence and Data analytics aims at finding out valuable insights from the raw data that can aid the education domain. Using school reports and questionnaires, the real-world data is collected to predict the performance of the students. The role of (FOSS) free and open source software is making data analytics motivating and effective in deriving insights in the area of interests. The growth of data analytics and role of Python is emerging in the field of predictive modelling that helps schools and universities in analysing the performance of the students' grades. This research paper makes use of the most commonly applied regression data analytics models to predict the students' final grade based on their performance in the internal assessments. A confidence estimate for the prediction accuracy, time for training and error rate are computed to find the performance of various regression models using the dataset obtained from the previous exam results. Based on the experiments in this research, the trained model foretells the student grade at the early stage of the course. This predictive model is implemented and tested using one of the most popular open-source data analytics tool python.**

***Index Terms*— Data analytics, Predictive model, Python, Datasets, Regression models, FOSS**

I. INTRODUCTION

Education is in a revolution phase and it is a prerequisite of modernization. It allows everyone to know the world beyond their own backgrounds and changes them to become rationalist and humanist. Acquiring knowledge as becomes easy to everyone through Massive open online courses, Wikipedia, Google, etc., [1]. Therefore, the traditional educational system is not of much use today. The students can learn and acquire knowledge easily because of the modernization of the traditional educational system. Even if the number of students in the class is large it gives teachers a facility to support each student individually [2]. Students' performance assessment is a continuous process and an essential part of the education process [3]. Grades are usually summarized as a single number or letter and how good a student was able to recognize and relate the knowledge communicated in a course. Using FOSS, a researcher can study, practise, modify, and redistribute the software without any conditions or with conditions. Free open source software (FOSS) offers the researchers the chance of creating their research on freely available data, publicly available data analysis algorithms, and software tools [4]. In this paper, we aim on predicting grades in traditional classroom-teaching where only the scores of students from past performance assessments are available.

II. RELATED WORK

Many researches have already scrutinized the value of standardized tests [5-7] admissions exams [8] and grade in the earlier programs [6] in predicting the intellectual success of students during their under graduation or schools. Actually, they decide on a positive correlation between the above mentioned predictors and success methods such as grade or degree accomplishment. Apart from the standardized tests, there are other relevant variables are used for predictions of the student's grade have been thoroughly scrutinized, resulting in the grade prediction from the prior education backgrounds and past grades obtained in certain courses or subjects [9,10] have a positive correlation.

The research work [9] studies that simple linear and more complex nonlinear models often lead to similar prediction correctness and determines that there is either no complicated nonlinear pattern to be found in the original data or the pattern cannot be predictable by their methodology. The research work [11] debates that the correctness of grade predictions often is middling due to diverse grading standards used in different classes and shows a complex rationality for grade predictions in single classes. Consequently, much research works aims in finding associations between a student's grade in a specific class and variables related to the student [12–21].

Significant elements were found to include the student's previous grade [12,13], [17–19], [21], performance in related subjects or courses [18,19,21], prior semester marks [15], performance in entrance exams [16], performance in the assignments of the class [22, 24], class presence [17], self-efficacy [20] and the student is repeating the same course in a class [19]. In these related research work, it is observed that the algorithms discussed has a lot of limitations and very difficult to apply in the field of education domains. Often related variables such as students' performance in class, grade, or self-efficacy are not available to the teacher since the data has not been gathered or is not available due to confidentiality reasons. However, the teacher has access to data collected from his/her own course. In one of the related

[1]Department of Information Technology, Higher College of Technology, P.O. 133, Al-Khuwair-33, Muscat, Sultanate of Oman (phone: 968-24143696; fax: 968-24473600; e-mail: ben.george@hct.edu.om)

[2]Department of Information Technology, Higher College of Technology, P.O. 133, Al-Khuwair-33, Muscat, Sultanate of Oman (phone: 968-24143696;fax:968-24473699;e-mail:balasupramanian.n@hct.edu.om)

[3]Department of Information Technology, Higher College of Technology, P.O. 133, Al-Khuwair-33, Muscat, Sultanate of Oman (phone: 968-24143696;fax:968-24473699;e-mail:582@hct.edu.om)

 ISSN: 1813-419X

research for predicting student's grade, the comments were gathered after the completion of each chapter in a course. These data are collected by asking students to fill three simple questionnaire items about their learning status. [25]. The three items so called, P(Previous), C (Current) and N (Next) comments. In another research related to the above one, only the C (Current) comments are mainly used for the analysis and the analysis uses the methodology, namely Japanese Morphological Analyzer Mecab, which extract words and part of speech [26]. Our research work mainly focuses on predicting the final grade based on the accessible data collected from each course teacher of the same course. The course we considered is Object-oriented programming-Java. Our research uses the best method, called the Pearson's correlation coefficient method to select the features from the existing list of features and use four different models to predict the target marks. This is little different in several aspects when compared to other relevant research work [22-24],[27-30].

In this research paper our aim is to focus on predicting students' grade using the following attributes such as quiz-1, quiz-2, class assignment, class activity, coursework(includes all assessments except mid exam), mid and final exam marks. Here we train the system with the collected data as mentioned above. Now we feed the dataset to many algorithms to predict the accuracy of grades by generating graphs, then these graphs and the error values are compared and we determine the algorithm that is best to predict the students' grade using the available datasets. Few hurdles were there in the dataset selection. First, the scores of the students in the quiz-I, class assignments might have little correlation with their scores in the remaining assessments. Second, even though the course material is same every year, but the assignments, exam complexity may change from year to year or semester to semester. Hence the prediction about the performance of the students' grade may change over the years. Third, students having a variety of different backgrounds are very difficult to predict their performance.

III. DATA ANALYTICS AND PREDICTIVE MODELING

Data analytics is the art of exploring raw data for the purpose of deriving valuable insights to end up with conclusions about that information. Almost many of the companies and organization have already started implementing data analytics to make improved decisions in their field of interests. Data mining and data analytics are different from each other by the scope, function and focus of the analysis. Data analytics basically aims on interpretation, the process by which conclusion is derived from what is previously known by the researcher [30]. Data analytics has several features and methods, about distinct techniques under a collection of names, in unique business and social science domains [31]. Predictive modeling is a procedure applied in data analytics to create a statistical model to derive the future insights. The main aim of predictive analytics is concerned with predicting possibilities and trends. In this modeling, data is gathered from the related predictors and then a numerical model is created, predictions are made and finally the model is revised as additional data are available [32]. The predictive modeling processes involve running one or more algorithms on the existing data sets from where predictions will be carried out. Therefore, this process itself is a repetitive process which usually requires training the model using several other models on the same data set and finally arriving on the finest model based on the business requirements. The main role of predictive algorithms is performing data mining and numerical analysis in order to influence trends and patterns insights from data [33].

IV. REGRESSION ANALYSIS

Regression analysis permits scientists to form numerical models that can be used to forecast the value of one variable from the information of another variable. There are a number of specific regression techniques that can be used by the scientists to model and drive real-world insights [34][35][36]. Linear and Logistic regression analysis are commonly the major algorithms scientists study in predictive modeling. Hence many analysts rationalize that the above algorithms are the only form of regressions because of the frequent use of these algorithms in the analysis. But actually there are numerous forms of regression models with slight variations and specific conditions that can be used for the effective analysis. In this research paper, we apply the following regression models to predict the student's grade: 1) Boosted trees regression, 2) Decision trees regression, 3) Linear regression 4) Random forest regression. In the field of information technology, predictive model is most widely used to predict the future insights.

A. Boosted trees regression (BRT)

Boosted trees regression (BRT) is one of the numerous methods that aim to increase the execution of a single model by using many models and merging them for prediction. BRT applies two algorithms such as regression trees are from the classification and regression tree cluster of models, and boosting builds and combines a group of models [37].

B. Decision trees regression

A decision tree actually represents rules for isolating data into groups. The first rule isolates the whole dataset into some number of fragments, and then another rule may be applied to single fragment, different rules to different fragments, establishing a second generation of fragments. Typically, a fragment may be either split or left isolated to form a final group. The tree depicts the first split into fragments as branches originating from a root and the succeeding splits as branches originating from nodes on older branches. The unsplit nodes are the final groups which are present in the leaves of the tree. For some willful reason, trees are always drawn upside down. For a tree to be valuable, the data in a leaf must be when compared to the target measure, so that the tree shows isolation of a mixture of data into refined groups [38].

C. Random forest regression

Random forest regression is a mixture of tree predictors such that each tree depends on the values of a random vector experimented independently and with the same circulation for all trees in the forest. The generalization error for forests touches so as to a limit as the number of trees in the forest becomes enormous. The generalization error of a forest of

tree classifiers depends on the potency of the individual trees in the forest and the correlation between them [39].

D. Linear regression

In predictive modeling, one of the most commonly used modeling technique is the linear regression and also researchers use this regression analysis when learning predictive analysis for the first time. Here in this regression analysis, the dependent variable is continuous, independent variables can be isolated, and the nature of the regression line is only straight. Linear regression analysis uses the Least Square method for fitting a regression line and calculates the best-fit line for the experimental data by reducing the sum of the squares of the vertical deviations from each data point to the line. Hence, the deviations are first shaped and when added there is no cancelling out between positive and negative values [40].

V. OVERALL PROCESS

The student marks from 4 semesters are collected as an anonymous dataset, which is used to create a regression to predict the student final exam marks. This dataset includes 6 different internal exam marks as features, which will be used to predict the final exam marks.

A. Analyze the Dataset

The sample dataset is analyzed to get fine details about the various features and their relationships. The individual grades and their relationship with the target is given in the below box plot Fig,1. The box plot gives the finite details for the target based on the mean, second and third quartiles. Fig 2., shows the linear relationship between the mid exam and the final exam marks based on their final grades.

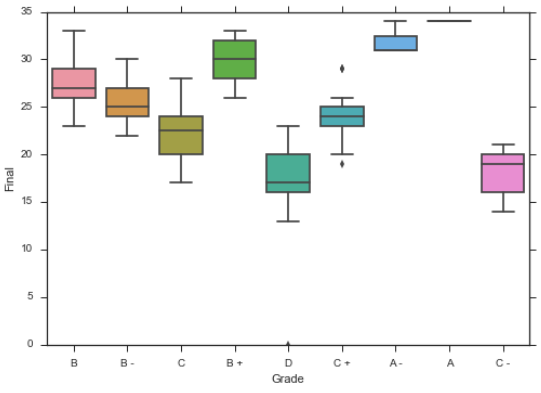

Fig. 1 Represent the target based on grades

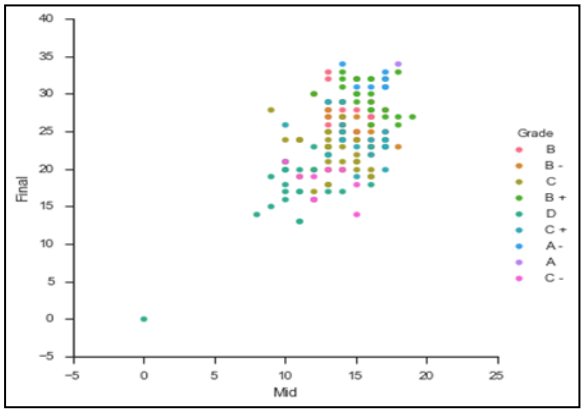

Fig. 2 Relationship between Mid and Final exam marks grouped by grades

B. Feature Selection Process

The first task is to perform the feature extraction process, the following fig 3. shows the relationship between the various features and the target. All possible feature combinations are used to perform the regression analysis to find the most significant features. These identified features will be used to train the model.

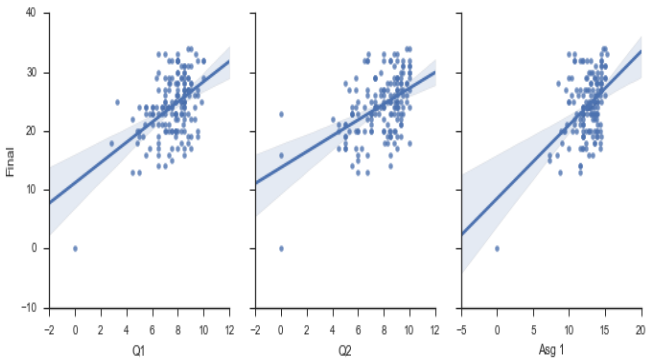

Fig. 3a

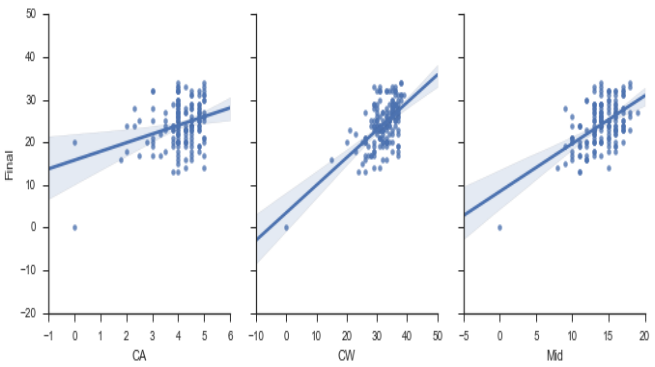

Fig. 3b

The figures 3a and 3b represents the relationship between the features and the target.The best method to select the viable features is the Pearson correlation coefficient. This is used as a measure of the strength of the relationship between two features. If there is no proper correlation between the features, then Pearson's coefficient produces a less strength of relationship between them. The Pearson's correlation coefficient 'r' is used to measure the correlation in a sample. Pearson's 'r' value can range from -1 to 1. If r = -1 indicates a perfect negative linear relationship between the features, if

r = 0 indicates no linear relationship between features, and if r = 1 gives a perfect positive linear relationship between the features [41,42]. The Pearson's correlation coefficient 'r' is calculated for all the features with the target feature. If the 'r' value is optimal, then those features are selected to predict the target. The correlation between one of the feature mid exam mark with the target final exam mark is depicted in Fig 4.

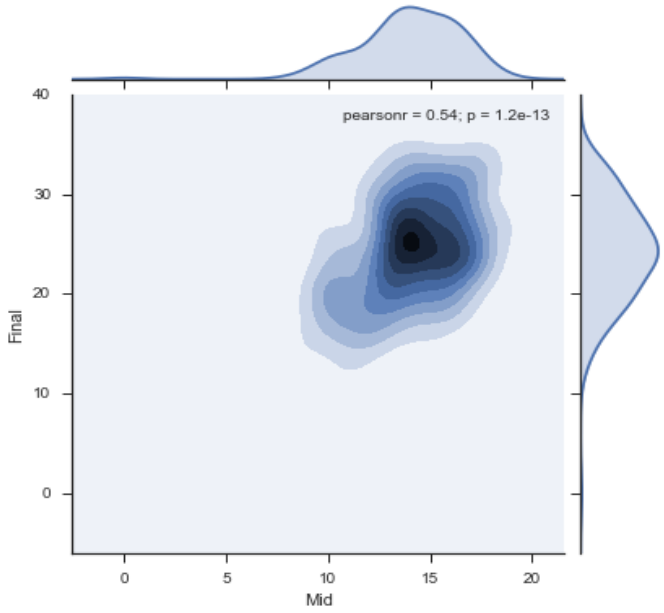

Fig.4 Pearson's correlation coefficient between mid-exam mark and final exam mark

The following table shows the Pearson's correlation coefficient 'r' for all the possible features. From the observation of table 1, the features Mid and Class work(CW) are having a higher positive value for 'r', therefore only these two features are selected for training the models.

Features selected	Pearson's correlation coefficient 'r'
Mid - Final	0.54
Quiz1 - Final	0.48
Quiz2 – Final	0.48
Assignment – Final	0.46
Class Activity – Final	0.31
Course Work - Final	0.58

Table 1: Pearson's correlation coefficient for all features.

C. Training Phase

The whole dataset is split into two, one is training dataset with 80% of the student records, and the remaining 20% is used as the testing dataset. The training data is used to create four different models using Boosted trees regression, Decision tree regression, Linear regression and Random forest regression. The same set of training data is given as input to all models to train the entire system. During the training process various results like maximum error, Root Mean Square Value (RMSE), time for training the model were recorded. The boosted trees regression uses 10 trees, with a maximum depth of 6 to train the entire training data. The decision tree regression also uses the maximum depth of 6 trees during the training process. For the Linear regression training phase used the Newton method as the solver and the number of iterations carried out is 10. For the Random forest regression the training process used maximum of 10 trees with a depth 6 trees and with 10 iterations.

D. Testing Phase

Once the four models are created the efficiency of the models should be tested, for this testing process the 20 % of the data separated from the original dataset will be used to test with all the models. The various parameters used to compare the performance during the testing phase are, errors during the testing phase and the time taken for training the model.

VI. PERFORMANCE EVALUATION

The four models were implemented, trained and tested with the open source tool python. The following table 2, shows the performance evaluation of all the four regression models based on their training, testing errors and the time taken for the training process. The Max error is the maximum error in the predication of the target value. RMSE is the square root of the mean of the square of all of the error. This is the most commonly used measure to find the error metrics in numerical predictions. The values of all these parameters should be very less for the efficient prediction of student marks.

Type of Regression	Testing		Training		Training time (in sec)
	Max error	RMSE	Max error	RMSE	
Boosted trees regression model	8.30	4.11	9.07	3.27	0.125
Decision tree regression model	24.09	16.12	24.8	17.48	0.007
Linear regression model	7.45	3.58	10.58	4.07	0.001
Random forest regression model	9.07	4.23	8.78	4.26	0.029

Table 2. Performance evaluation of various models

The training RMSE for the boosted trees regression model is 3.27 which is very less when compared with other models. The Random forest regression model produced the maximum error rate of 8.78, which is very less when compared to other models. In the testing phase Linear regression model produce very less error rate with maximum error is 7.45 and RMSE 3.58, which is very less when compared with other models. The Linear regression model took only 0.001 seconds for the

training process which is very less when compared with other models when executed with I3 processor.

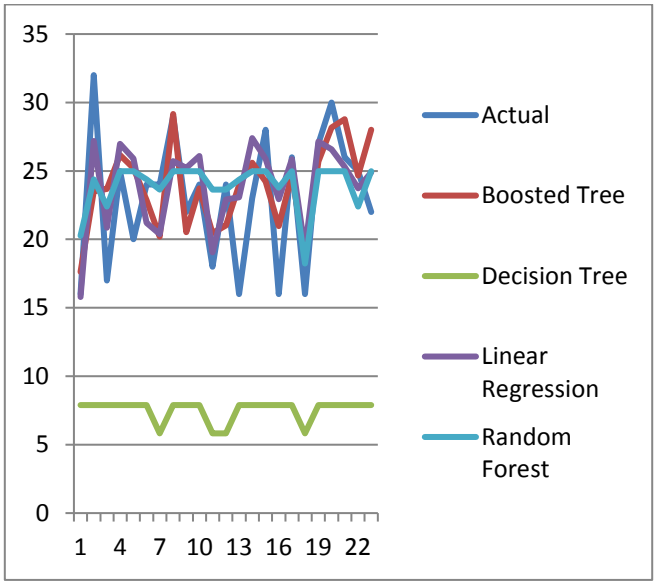

Fig 5. Final exam marks predicted by all models for the testing data

The above Fig. 5 shows the marks predicted by all the four models from the testing data along with the actual marks. From all these observations of testing error, time and accuracy of prediction the Linear regression model is best suited model to predict the student marks based on their previous performance. As it is observed that, there is a linear relationship between the features and the target which made the Linear regression model to outperform other models.

VII. Conclusion

Free and open source analytics software plays a key role in this current era to predict useful information from the raw datasets. The free and open source software tools like python, R are driving force in analytical modernizations and help researchers to predict better insights from the huge datasets. Data Analytics and predictive modelling is gaining its acceptance in almost all applications of real world. One of the data analytics techniques i.e., regression model is an interesting topic to the researchers as it is precisely and competently extract the data for future insights and interpretations. This research work extracted the dataset related to the students' marks for a particular course. The Pearson's correlation coefficient method was used to find the most significant features. The same training dataset was used to train the four regression models and a separate testing set was used to evaluate the performance of mentioned models. The experimental results in this research, identified the linear regression model as the best regression model for predicting the student final exam marks. This will really help students and teachers to improve the performance of the students.

References

[1] R. Baraniuk, "*Open education: New opportunities for signal processing*," Plenary Speech, 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2015.

[2] *"Openstax college,"* http://openstaxcollege.org/, accessed: 2015-05-07.

[3] L.Earl. *Assessment of Learning, for Learning, and as Learning.* Thouand Oaks,CA,Corwin Press, 2003.

[4] Fabio Kon, Paulo Meirelles, Nelson Lago , Antonio Terceiro, Christina Chavez, Manoel Mendonça *"Free and Open Source Software Development and Research: Opportunities for Software Engineering*.

[5] N. R. Kuncel and S. A. Hezlett, *"Standardized tests predict graduate students success,"* Science, vol. 315, pp. 1080–1081, 2007.

[6] E. Cohn, S. Cohn, D. C. Balch, and J. Bradley, "*Determinants of undergraduate gpas: Sat scores, high-school gpa and high-school rank," Economics of Education Review*, vol. 23, no. 6, pp. 577–586, 2004.

[7] E. R. Julian, "*Validity of the medical college admission test for predicting medical school performance*," Academic Medicine, vol. 80, no. 10, pp. 910–917, 2005.

[8] P. A. Gallagher, C. Bomba, and L. R. Crane, *"Using an admissions exam to predict student success in an adn program*," Nurse Educator, vol. 26, no. 3, pp. 132–135, 2001.

[9] W. L. Gorr, D. Nagin, and J. Szczypula, *"Comparative study of artificial neural network and statistical models for predicting student grade point averages*," International Journal of Forecasting, vol. 10, no. 1, pp. 17–34, 1994.

[10] N. T. Nghe, P. Janecek, and P. Haddawy, *"A comparative analysis of techniques for predicting academic performance,"* in Frontiers In Education Conference-Global Engineering: Knowledge Without Borders, Opportunities Without Passports, 2007. FIE'07. 37th Annual. IEEE, 2007, pp. T2G–7.

[11] R. D. Goldman and R. E. Slaughter, "*Why college grade point average is difficult to predict*." Journal of Educational Psychology, vol. 68, no. 1, p. 9, 1976.

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

[13] E. Osmanbegovi´c and M. Sulji´c, "*Data mining approach for predicting student performance*," Economic Review, vol. 10, no. 1, 2012.

[14] B. K. Baradwaj and S. Pal, "*Mining educational data to analyze students' performance*," arXiv preprint arXiv:1201.3417, 2012.

[15] S. K. Yadav, B. Bharadwaj, and S. Pal, "*Data mining applications: A comparative study for predicting student's performance*," arXiv preprint arXiv:1202.4815, 2012.

[16] A. B. E. D. Ahmed and I. S. Elaraby, "*Data mining: A prediction for student's performance using classification*

method," World Journal of Computer Application and Technology, vol. 2, no. 2, pp. 43–47, 2014.

[17] P. Cortez and A. M. G. Silva, "*Using data mining to predict secondary school student performance*," 2008.

[18] L. H. Werth, *Predicting student performance in a beginning computer science class*. ACM, 1986, vol. 18, no. 1.

[19] J. L. Turner, S. A. Holmes, and C. E. Wiggins, "*Factors associated with grades in intermediate accounting,"* Journal of Accounting Education, vol. 15, no. 2, pp. 269–288, 1997.

[20] A. Y. Wang and M. H. Newlin, "*Predictors of web-student performance: The role of self-efficacy and reasons for taking an on-line class*," Computers in Human Behavior, vol. 18, no. 2, pp. 151–163, 2002.

[21] S. Kotsiantis, C. Pierrakeas, and P. Pintelas, *"Predicting students'performance in distance learning using machine learning techniques,*" Applied Artificial Intelligence, vol. 18, no. 5, pp. 411–426, 2004.

[22] C. G. Brinton and M. Chiang, "*Mooc performance prediction via clickstream data and social learning networks,*" in 34th INFOCOM IEEE. 2015, To appear.

[23] C. Romero, M.-I. L´opez, J.-M. Luna, and S. Ventura, "*Predicting students' final performance from participation in on-line discussion forums,*" Computers & Education, vol. 68, pp. 458–472, 2013.

[24] M. I. Lopez, J. Luna, C. Romero, and S. Ventura, "*Classification via clustering for predicting final marks based on student participation in forums."* International Educational Data Mining Society, 2012.

[25] K.Goda and T.Mine. *Analysis of student' learning activities through qualifying time-series comments.* Proc. of the KES 2011, Part 2, LNAI 6882, Springer-Verlag Berlin Heidelberg, pp.154-164, 2011.

[26] Jingyi Luo, Shaymaa E, and Fukuoka "*Predicting Student Grade based on Free-style Comments using Word2Vec and ANN by Considering Prediction Results Obtained in Consecutive Lessons*".

[27] M. D. Calvo-Flores, E. G. Galindo, M. P. Jim´enez, and O. P. Pineiro, *"Predicting students marks from moodle logs using neural network models*," Current Developments in Technology-Assisted Education, vol. 1, pp. 586–590, 2006.

[28] D. Garcıa-Saiz and M. Zorrilla, *"A promising classification method for predicting distance students performance."* EDM, pp. 206–207, 2012.

[29] C. Romero, S. Ventura, P. G. Espejo, and C. Herv´as, "*Data mining algorithms to classify students*." in EDM, 2008, pp. 8–17.

[30] B. Minaei-Bidgoli, D. A. Kashy, G. Kortemeyer, and W. F. Punch, "*Predicting student performance: an application of data mining methods with an educational web-based system*," in Frontiers in education, 2003. FIE 2003 33rd annual, vol. 1. IEEE, 2003, pp. T2A–13.

[31]http://searchdatamanagement.techtarget.com/definition/data-analytics accessed on 21-10-2016.

[32] https://en.wikipedia.org/wiki/Data_analysis accessed on 22-10-2016.

[33]http://searchdatamanagement.techtarget.com/definition/predictive-modeling accessed on 25-10-2016.

[34] http://www.predictiveanalyticstoday.com/predictive-modeling/ accessed on 21-10-2016.

[35] Sunghae Jun, Seung-Joo LeeandJea-Bok Ryu "A Divided Regression Analysis for Big Data" International Journal of Software Engineering and Its Applications Vol. 9, No. 5 (2015), pp. 21-32 http://dx.doi.org/10.14257/ijseia.2015.9.5.03

[36] Glenn De'ath, Katharina E. Fabricius, "Classification and Regression Trees: A Powerful yet Simple Technique for Ecological Data Analysis" DOI: 10.1890/0012-9658(2000)081[3178:CARTAP]2.0.CO;Volume 81, Issue 11November 2000 Pages 3178–3192.

[37] http://www.enotes.com/research-starters/regression-analysis accessed on 24-10-2016.

[38] *A working guide to boosted regression trees* J. Elith , J. R. Leathwick and T. Hastie3

[39] Decision Trees for Predictive Modeling Padraic G. Neville SAS Institute Inc. 4 August 1999

[40] Machine Learning, 45, 5–32, 2001 2001 Kluwer Academic Publishers. Manufactured in The Netherlands. Random Forests LEO BREIMAN Statistics Department, University of California, Berkeley, CA 94720

[41] https://www.analyticsvidhya.com/blog/2015/08/comprehensive-guide-regression/ accessed on 21-10-2016.

[42] http://onlinestatbook.com/2/describing_bivariate_data/pearson.html accessed on 27-10-2016