ISSN 2278-3091
Volume 8, No.4, July – August 2019
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse75842019.pdf
https://doi.org/10.30534/ijatcse/2019/75842019

Predicting Student's Performance by using Classification Methods

Amita Dhankhar [1], Kamna Solanki[2], Arvind Rathee[3], Ashish[4]
[1,2,3,4]University Institute of Engineering and Technology, Maharshi Dayanand University, Rohtak-124001, India
[1]Amita.infotech@gmail.com, [2]Kamna.mdurohtak@gmail.com

ABSTRACT

Most of the developing countries are facing the problem of ever-rising low-quality population. To convert this low-quality population into a high-quality one, efforts are required to be laid down. These efforts include investment in Research and development and in the education sector. If the people living in an area will be educated then they will be productive for the nation and eventually contribute towards its GDP. The advancement of technology helps educational institutions to turn raw data into actionable insights to achieve desirable results. This study has worked towards prediction models so that student's performance be evaluated timely so that necessary steps be taken in due time to improve their performance. In this study, the record of 265 Computer Science and Engineering students at UIET, MDU Rohtak are used for prediction task by using five algorithms namely Simple Linear Regression, Random Forest, Decision Table, SMOreg, LWL. Algorithms are compared by using WEKA tool. The results showed that out of these five algorithms, Simple Linear Regression gave the best prediction accuracy with the highest correlation coefficient of 0.78, lowest Mean Absolute Error of 4.97 and lowest Root Mean Squared Error of 6.75. This study has laid the foundation in selecting efficient algorithms for predicting the results of students.

Key words: Decision table, Random forest, Simple regression, Student's performance, WEKA.

1. INTRODUCTION

Education is an important element to strengthen economic growth and stability. Formal education is an example of education that can be attained through schools and colleges. In most of the educational institutes, a conventional classroom is still the main learning method. The drawback of such a method is that it makes hard to understand each student because the number of students is large and there is a limited number of meetings. Especially the first-year students of the college experiences difficulties that are sufficient to cause them to drop out. So, the early prediction of student's academic performance helps to get an idea of student's level of learning, to identify student's success or failure in the course registered and to provide timely intervention for student's at risk.

2. RELATED WORK

Many research studies have been conducted by the researchers that provide a comprehensive overview of the prediction methods. This section enlists some of them that have been considerably discussed by the numerous researchers. Yu *et al.* identified at-risk students by using sentimental analysis on self-evaluation comments. To predict student's performance, they have used Convolutional Neural Networks (CNN) and Support Vector Machine (SVM) [1]. Sandoval *et al.* used LR, RLR, and Random forest to predict students' performance and concluded that RF performed better than LR and RLR [2]. Okubo *et al.* used Recurrent Neural Networks (RNN) for predicting final grades of students. They also performed a comparative analysis between RNN and Multiple Regression analysis and concluded that for early prediction of final grades RNN is effective [3]. Costa *et al.* evaluate the effectiveness of four Educational Data Mining Namely Naive Bayes, Neural Networks, Support Network Machine and Decision tree [4]. Babic applied three machine learning methods namely Neural Networks, Classification Tree and Support Vector Machine and concluded that Neural Networks performed better than the other two methods [5]. Li *et al.* implemented Fuzzy Clustering and Multi-Variable Regression for predicting student's academic performance [6]. Kumar performed the comparative analysis of five classifiers namely Rotation Forest (ROF), Naive Bayes, Sequential Minimum Optimization, Radial Basis Function, Multilayer Perceptron using Weka for prediction of student's performance and concluded that ROF has produced superior classification performance [7]. Xing *et. al.* integrated Learning Analytics, Educational Data Mining and theory to solve the problem of predicting student's performance through Genetic Programming [8]. Mayilvaganan *et. al.* compared the performance of classification techniques namely C4.5, Naive

Bayes, AODE, Multi- labeled, K- Nearest Neighbour for predicting the student's performance using Weka tool [9]. Hidayah *et al.* proposed the student's classification model to predict a student's academic performance. It applied the neuro-fuzzy concept that is the combination of fuzzy if-then rules and neural network's ability to learn [10]. Marquez-Vera *et al*. used Genetic programming and different data mining approaches for predicting student's failure at school [11]. Huang *et al.* Predicted student's academic performance in engineering dynamics by comparing four mathematical models namely SVM, Radial Basis Function Network model, multilayer perception network model and multiple linear regression [12]. Kotsiantis *et al.* proposed a technique for predicting student's performance in distance education. The proposed technique combines three classifiers namely Naive Bayes, 1-NN and WINNOW [13]. Zafra and Ventura predicted whether a student will pass or fail a certain course by applying grammar guided genetic programming algorithm G3P-MI [14]. Oladokun *et al.* developed and tested an Artificial Neural Network Model based on multilayer perceptron topology for predicting student's academic performance [15]. Ayan *et al.* applied linear and logistic regression for prediction of university student's academic achievements [16]. Ibrahim and Rusli measured the academic performance of the students by their cumulative grade point average. They concluded that ANN, Decision tree and Linear regression produce more than 80% accuracy and ANN outperforms the other two models [17]. Hsu *et al.* Proposed hybrid model for predicting student's course performance. The hybrid model combines a genetic algorithm and the Apriori algorithm. Further, they compared the proposed model with genetic algorithm and concluded that the proposed model has higher computation efficiency and prediction accuracy [18]. Tsai *et. al.* proposed a two-phase fuzzy mining and learning algorithm [19].

3. METHODOLOGY

In this paper, the grades are used for the analysis and measurement of the performance. The first and foremost step is to collect the dataset required for the prediction of student's performance. In this study, the dataset of 290 students of B.Tech 1st year (Computer Science and Engineering) at University Institute of Engineering & Technology (UIET), Maharishi Dayanand University (MDU) Rohtak, Haryana, India were collected. The data contained: Name, Gender, Registration Number , Individual subject marks of class 10 (Grade Point 1-10), Individual subject marks of class 12 (Each subject score out of 100), Joint Entrance Examination Score (Each subject score out of 360), Individual Subject marks of semester 1 (Each subject score out of 150 or 50), Individual subject marks of semester 2 (Each subject score out of 150, 50 or 75). After data cleaning, this dataset is reduced to 265 in number. Then comes the data filtering phase in which a large amount of data available to us is reduced by removing the unwanted data attributes. In this study, student name, gender and student registration number carry no significance, so we removed it from the dataset. Finally, comes the data transformation phase in which new attributes are derived from available attributes to assist a better interpretation of information. In this study, we converted the score of 10th, 12th, JEE Main, 1st Semester and 2nd Semester from individual subject-wise marks to an overall percentage. After these the following five attributes are left:

I. Class 10th percentage
II. Class 12th percentage
III. JEE percentage
IV. 1st Semester percentage
V. 2nd Semester percentage

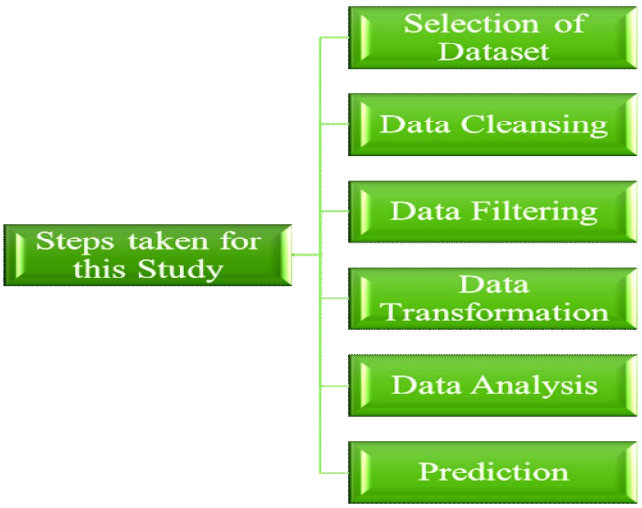
These five attributes were then used for processing and data analysis. The image below displays the attributes in the .arff file.

```
@relation ScorePredictor

@attribute 10th Numeric
@attribute 12th Numeric
@attribute jee Numeric
@attribute 1stSem Numeric
@attribute 2ndSem Numeric
```

In this study, data analysis is performed on various trained models for the prediction of students 2nd Semester overall percentage. Algorithms used for training purpose are Simple Linear Regression, SMOreg, LWL, Decision Table, Random Forest. Further ensemble learning by combining the predictions of two or more than two algorithms by using the meta Vote classifier in WEKA is also applied. We got even better predictions. All possible combinations of the above five mentioned algorithms are applied to train our models in the combination of two, three, four and five algorithms at once in Vote classifier and evaluated their results on four combination rules i.e, Average of Probability, Minimum Probability, Maximum Probability, and Median.

The following image depicts the steps followed in this study for prediction of student's performance.



4. RESULTS AND DISCUSSION

The classification performances of above mentioned five algorithms were analyzed for standard performance parameters namely Correlation coefficient, Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE). Both MAE and RMSE are common metrics used to measure the accuracy of continuous variables. Figure 1 shows the values of the correlation coefficient, Mean Absolute Error and Root Mean Squared Error for all the five algorithms separately for the selected dataset. Out of all five algorithms, Decision table has RMSE of 7.37 and MAE of 5.39 and is the worst-performing algorithm whereas Simple Linear Regression has RMSE of 6.75 and MAE of 4.97 and is best performing algorithm as shown in figures 2 and 4. Another important metric in measuring the performance of an algorithm is the correlation coefficient. A correlation coefficient is a binary class classifier performance evaluation parameter. Its value may vary from -1 to +1. +1 represents the perfect fit and -1 represents the worst fit.

Name	Correlation coefficient	MAE	RMSE
Decision Table	0.73	5.39	7.37
Random Forest	0.73	5.35	7.27
LWL	0.71	5.21	7.53
SMOreg	0.77	5.16	6.91
Simple Linear Regression	0.78	4.97	6.75

Figure 1: Values of the correlation coefficient, MAE and RMSE for Algorithms

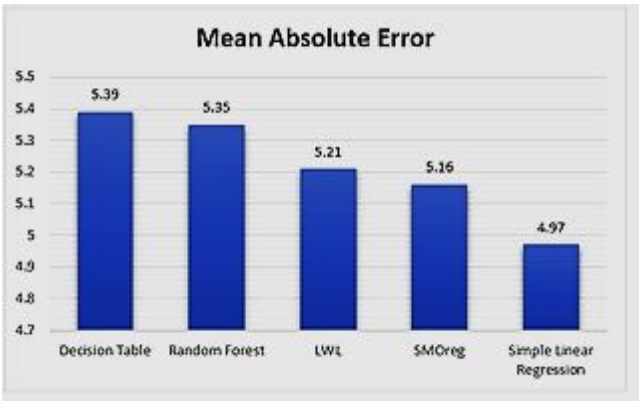


Figure 2: MAE for Algorithm

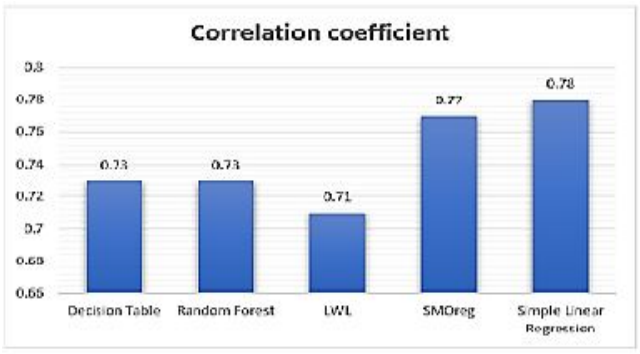


Figure 3: Correlation Coefficient for Algorithm

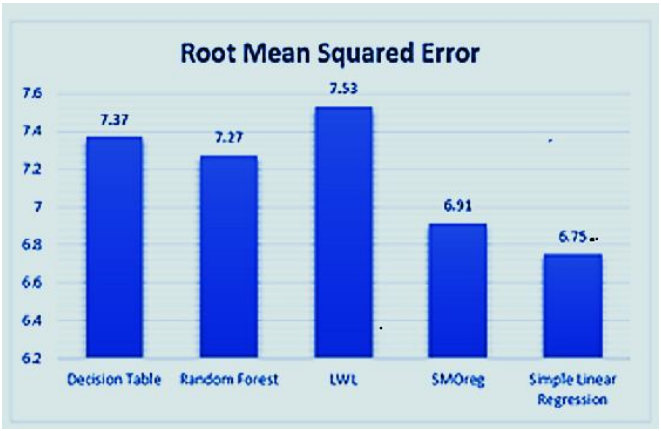


Figure 4: RMSE for Algorithm

It can be observed from Figure 1, that the correlation coefficient is highest for Simple Linear Regression, which also has the least MAE as shown in figure 2. Ensemble learning is used to improve machine learning results by combining several models. This approach helps in producing a better predictive performance as compared to a single model. For this study, voting is used as the ensemble learning method. Voting is tested for four different combinations of classifiers and on different combination rules. Figure 5 contains the results of different combination rules for four different combinations of classifiers. In Figure 5 it is observed that out of Median, Maximum Probability, Minimum Probability and Average of Probability, the MAE was best for Average of probability as combination rule.

In this, an average of probability has best MAE and RMSE results for every combination. Root Mean Squared Error (RMSE) is a good metric to measure performance, but in our case, we observed that two combinations of classifiers had same RMSE values. Both 3- classifier combination and 4-classifier combination had 6.69 as RMSE value. Mean Absolute Error (MAE) also had values 4.78 and 4.79 for 3 classifier combination and 4-classifier combination respectively (as shown in figure 4 and figure 5). 3-combination classifiers are better by 0.01. From observation above stated it can be concluded that the 3-classifier combination i.e. SLR, DT, and LWL have the best performance with 4.78 as MAE and 6.69 as RMSE. Also, it has the highest correlation coefficient of 0.78. From table 3 it can be observed that the correlation coefficient is the same but both RMSE and MAE value vary. Ensemble learning gave better results, it reduced MAE value by 0.19 (as shown in Figure 8).

Name	Average of Probability			Minimum Probability			Maximum Probability			Median		
	Correlation coefficient	MAE	RMSE	Correlation coefficient	MAE	RMSE	Correlation coefficient	MAE	RMSE	Correlation coefficient	MAE	RMSE
SLR + LWL	0.78	4.8	6.7	0.77	4.94	6.86	0.77	5.24	7.44	0.77	4.94	6.86
SLR + LWL + DT	0.78	4.78	6.69	0.75	5.41	7.36	0.74	5.55	7.87	0.74	5.41	7.36
SLR + LWL + DT + SMOreg	0.78	4.79	6.69	0.76	5.24	7.15	0.75	5.58	7.78	0.77	4.96	6.78
SLR + LWL + DT + SMOreg + RF	0.77	4.87	6.76	0.76	5.51	7.3	0.75	5.66	7.94	0.77	5.02	6.83

Figure 5: All possible combinations of Algorithm

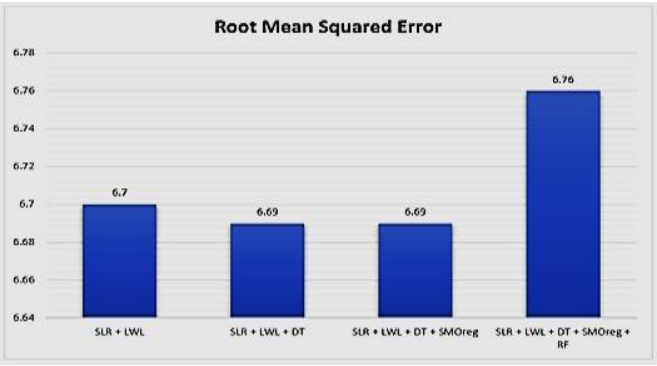


Figure 6: RMSE for the combination of Algorithm

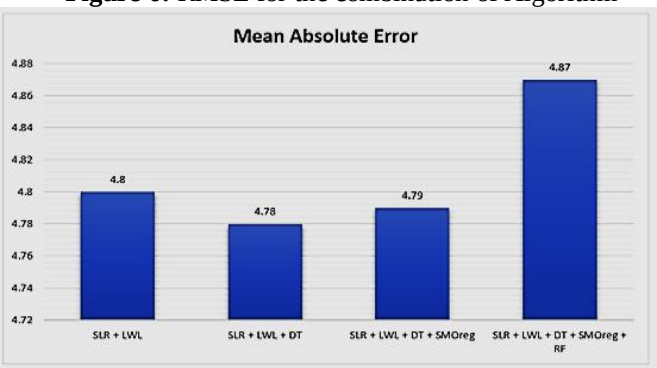


Figure 7: MAE for the combination of Algorithm

Name	SLR + LWL + DT	Simple Linear Regression	Difference
Correlation coefficient	0.78	0.78	0
MAE	4.97	4.78	0.19
RMSE	6.75	6.69	0.06

Figure 8: Comparison of SLR + LWL + DT and SLR

5. CONCLUSION AND FUTURE SCOPE

Predicting student's performance is beneficial to both educator and learner as it helps them to improve their teaching and learning process. In this study, the performance of five algorithms is compared in the prediction of student's performance. Results demonstrated that Simple Linear Regression gave the best prediction accuracy with the highest correlation coefficient of 0.78, the lowest MAE of 4.97 and lowest RMSE of 6.75. Further, it is also concluded that Vote with a combination of three algorithms i.e, Simple Linear Regression, Locally Weighted Learning and Decision Table with the combination rule of Average of Probability having lowest MAE of 4.78 and is the best among all. This study has attempted to find an effective algorithm for predicting student's performance. This will help both the learner and educator and will eventually contribute in enhancing the performance of the students by taking appropriate measures timely. Further study in this direction will try to perform a vaster analysis using a large dataset to generalize the results and will try to design more means and measures to predict students' performance.

REFERENCES

1. L. C. Yu, C. W. Lee, H. I. Pan, C. Y. Chou, P. Y. Chao, Z H. Chen, S. F. Tseng, C. L. Chan, K. R. Lai, **"Improving early prediction of academic failure using sentiment analysis on self□evaluated comments"**, *Journal of Computer Assisted Learning*, Vol. 34, No. 4, pp. 358-65, Aug. 2018. https://doi.org/10.1111/jcal.12247
2. A. Sandoval, C. Gonzalez, R. Alarcon, K. Pichara and M. Montenegro, **"Centralized student performance prediction in large courses based on low-cost variables in an institutional context"**, *The Internet and Higher Education*, Vol. 37, pp.76-89, 2018. https://doi.org/10.1016/j.iheduc.2018.02.002
3. F. Okubo, T. Yamashita, A. Shimada, and H. Ogata, "**A neural network approach for students' performance prediction",** *Learning Analytics & Knowledge Conference,* ACM, pp. 598-599, 2017.

https://doi.org/10.1145/3027385.3029479

4. E. B. Costa, B. Fonseca, M. A. Santana, F. F. de Araújo, and J. Rego, **"Evaluating the effectiveness of educational data mining techniques for early prediction of student's academic failure in introductory programming courses",** *Computers in Human Behavior*, vol. 73, pp. 247-256, Aug. 2017. https://doi.org/10.1016/j.chb.2017.01.047
5. I. B. Durdevic, **"Machine learning methods in predicting the student academic motivation",** *Croatian Operational Research Review*, vol. 8, no. 2, pp. 443-461, Dec. 2017. https://doi.org/10.17535/crorr.2017.0028
6. Z. Li, S. Changjing and S. Qiang, **"Fuzzy-clustering embedded regression for predicting student academic performance",** In *2016 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pp. 344-351, 2016.
7. M. Kumar, **"Superiority of Rotation Forest Machine Learning Algorithm in Prediction of Student's Performance",** *International Journal of Computer Applications*, vol. 137, no. 2, 2016. https://doi.org/10.5120/ijca2016908712
8. W. Xing, R. Guo, E. Petakovic, and S. Goggins, **"participation-based student final performance prediction model through interpretable Genetic Programming: Integrating learning analytics, educational data mining and theory",** *Computers in Human Behavior*, vol. 47, pp. 168-181, 2015. https://doi.org/10.1016/j.chb.2014.09.034
9. M. Mayilvaganan, D. and Kalpanadevi, **"Comparison of classification techniques for predicting the performance of students' academic environment",** In *2014 International Conference on Communication and Network Technologies*, IEEE, pp. 113-118, Dec. 2014. https://doi.org/10.1109/CNT.2014.7062736
10. I. Hidayah, A. E. Permanasari and N. Ratwastuti, **"Student classification for academic performance prediction using neuro fuzzy in a conventional classroom",** In *2013 International Conference on Information Technology and Electrical Engineering (ICITEE)*, pp. 221-225, Oct. 2013. https://doi.org/10.1109/ICITEED.2013.6676242
11. C. Marquez-Vera, A. Cano, C. Romero and S. Ventura, **"Predicting student failure at school using genetic programming and different data mining approaches with high dimensional and imbalanced data",** *Applied Intelligence*, vol. 38, no. 3, pp. 315-330, April. 2013. https://doi.org/10.1007/s10489-012-0374-8
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, Feb. 2013. https://doi.org/10.1016/j.compedu.2012.08.015
13. S. Kotsiantis, K. Patriarcheas and M. Xenos, **"A combinational incremental ensemble of classifiers as a technique for predicting students' performance in distance education",** *Knowledge-Based Systems*, vol. 23, no. 6, pp. 529-535, Aug. 2010. https://doi.org/10.1016/j.knosys.2010.03.010
14. A. Zafra and S. Ventura, **"Predicting Student Grades in Learning Management Systems with Multiple Instance Genetic Programming",** *International Working Group on Educational Data Mining*, July 2009. https://doi.org/10.1109/ISDA.2009.108
15. V. O. Oladokun, A. T. Adebanjo, and O. E. Charles-Owaba, **"Predicting students' academic performance using artificial neural network: A case study of an engineering course",** *The Pacific Journal of Science and Technology*, vol. 9, no. 1, pp. 72-79, 2008.
16. M. N. R. Ayan, and M. T. C. García, **"Prediction of university students' academic achievement by linear and logistic models",** *The Spanish journal of psychology*, vol. 11, no. 1, pp. 275-288, May 2008. https://doi.org/10.1017/S1138741600004315
17. Z. Ibrahim and D. Rusli, **"Predicting student's Academic Performance: Comparing artificial neural network, decision tree and linear regression"**, *The 21 Annual SAS Malaysia Forum, Kuala Lumpur, Malaysia*, pp. 1–6, Sept. 2007.
18. P. L. Hsu, R. Lai and C. C. Chiu, **"The hybrid of association rule algorithms and genetic algorithms for tree induction: an example of predicting the student course performance",** *Expert Systems with Applications*, vol. 25, no. 1, pp. 51-62, July 2003. https://doi.org/10.1016/S0957-4174(03)00005-8
19. C. J. Tsai, S. S. Tseng and C. Y. Lin, **"A two-phase fuzzy mining and learning algorithm for adaptive learning environment",** *Computational Science, Berlin, Springer* pp. 429-438, May 2001. https://doi.org/10.1007/3-540-45718-6_47