International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 10
***Article Received:** 25 July 2023 **Revised:** 12 September 2023 **Accepted:** 30 October 2023*

Student Performance Prediction and Classification Using Learning Analytics

Devi S [1]
Research Scholar
Department of Computer Science
Dr. APJ Abdul Kalam University Indore, India
Email : devi.aakar@gmail.com

Dr. Arpana Bharani [2]
Research Supervisor
Department of Computer Science
Dr. APJ Abdul Kalam University Indore, India
Email : arpanabharani@gmail.com

Abstract—For a productive and a good life, education is a necessity and it improves individuals' life with value and excellence. Also, education is considered a vital need for motivating self-assurance as well as providing the things are needed to partake in today's World. Throughout the years, education faced a number of challenges. Different methods of teaching and learning are suggested to increase the learning quality. In today's world, computers and portable devices are employed in every phase of daily life and many materials are available online anytime, anywhere. Technologies like Artificial Intelligence had a surprising evolution in many fields especially in educational teaching and learning processes. Higher education institutions have started to adopt the use of technology into their traditional teaching mechanisms for enhancing learning and teaching. In this paper datasets have been considered for the prediction the error ratio of student performance respectively using five machine learning algorithms. Eighteen experiments have been performed and preliminary results suggest that performances of students might be predictable and classification of these performances can be increased by applying pre-processing to the raw data before implementing machine learning algorithms.

Key words : Student performance prediction, Deep learning for education, Educational data mining, Learning analytics.

I. INTRODUCTION

The student performance prediction problem has been partly studied within the learning analytics and educational data mining communities in the form of the student dropout (or completion) prediction problem (which is an important subclass problem of the student performance prediction problem). To evaluate the higher education institution with respect to the parameters for student performance implementation. The amount of data in educational environment maintained in electronic format has seen a dramatic increase in recent time. The data can be collected from historical and operational data reside in the databases of educational institutes. The task to manage the large amount of data and determine the relationships among variables in the data is not easy to be done. The prediction of student's performance in an institutions is one of the most vital issues in higher education[3]. The task to develop effective predictors of academic success is a critical issue for educators. Performance predicate is dependent upon motivation, attitudes, peer influence, curriculum and by the continued real-time monitoring of student's performance using a simple rapid response system and as noted predicts correctly which student may need some attention or reinforcements in the course of their education[4]. The various such as age, gender, school related factors, environment of the home, or the support given by the parents and other family members are responsible for failure of student's performance.

2. RELATED WORK

Most previous works can be divided into two approaches:The first traditional approach principally relies on generalized linear models, including logistic regression, linear SVMs and survival analysis [1] Each model considers different types of behavioral and predictive features extracted from various raw activity records (e.g., clickstream, grades, forum, grades).The second emerging approach involves an exploration of neural networks (NN). Few prior works explore deep neural network (DNN) model , recurrent neural network (RNN) model [6] and convolutional neural networks (CNN) followed by RNN [2]. However, all of these new models, so far, have shown primitive performance. This is mainly because the models still rely on feature engineering to reduce input dimensions which appears to limit one to develop larger (i.e., better) NN models. Lui study used NN-back propagation (BP) NNfeed forward recurrent models. The study aimed to explore the impacting features for the English course score using the college students' data . The data set contains 101 freshmen's NCEE (National College Entrance Examination) score. The results indicate that the most important for the student academic performance in the first semester final grade is the NCEE marks, the learning attitude, and the gender. In contrast, age has a small effect on the scores. After training the model 20 times, the relative error is smaller, and the accuracy is higher. Furthermore, a study used classifiers such as J48, SVM, RF, NB, and MLP to predict students' performance levels using the data set of Kalboard 360 [3]. The result showed that the MLP outperformed other classifiers with 0.760 accuracy. Similarly, another study [6] was performed using the same data set, applying three classifiers such as DT, NB, and ANN. The study achieved 10 to 15% increase in the performance prediction when compared with results after removing some features. They focused on students' preparation and the impact of parents' participation in the learning process.

3. PREDICTION PERFORMANCE

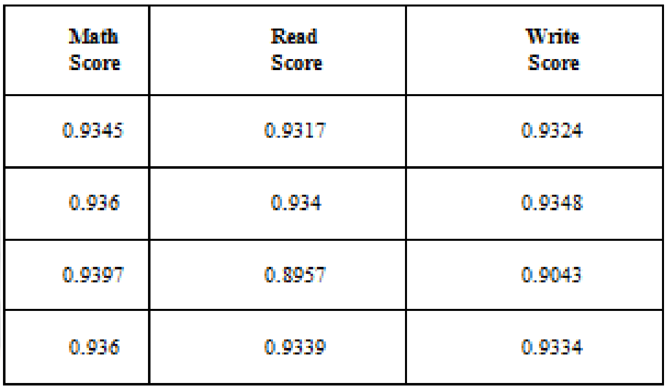
3.1 AN EFFICIENT VALUES BASED ON COGNITIVE FACTORS DISCUSS WITH CLASSIFIERS

Effectiveness cognitive factors are Math score, Reading Score, and Writing Score identified from our dataset collected from kaggle database. This dataset has eight attributes. Such as (i)Gender (ii) Race/Ethnicity (iii) Parent level of Education (iv) Lunch (v)Test Preparation Course (vi) Math Score (vii) Reading Score (viii) Writing Score. Data pre-processing is must and feature selection is very important that InfoGainAttributeEval with Ranker method was used. Seven attributes are selected such as 8,7,6,2,3,4,5. Supervised learning selection using 10 cross-validation[8]. 1000 records are having this dataset. 518 are Female records and Male 482 records.

3.1.1 Correlation Coefficient Analysis

According tothe reference research various discussion in an analysis for correlation coefficient of different variables with the student's dataset. With the dataprocessing methodology using ,Python, Spearman correlation coefficients can be simulated as shown in Figure using the data setin kaggle. Spearm an correlation coefficients between the classifiers with variables in the data set are shown. These factors vary from every student with our environmental factors, all contributing to the formation of the overall.This coefficients analysis reveals the general trend and significant factors on the student score.

Table 1.Correlation Coefficient of different student's performance



Math Score	Read Score	Write Score
0.9345	0.9317	0.9324
0.936	0.934	0.9348
0.9397	0.8957	0.9043
0.936	0.9339	0.9334

Select the strong attribute from the spearman correlation coefficient. Students' performance valuable attributes are math scores, reading the score, and writing the score. These are very strong attributes of the dataset. *A variety of mathematical techniques, such as multivariate linear regression, neural networks, Bayesian networks, decision trees, and genetic algorithm* have been engaged to develop surface various models to predict student academic performance. Multivariate linear regression is among the most widely in work mathematical techniques.

Select the strong attribute from the spearman correlation coefficient. Students' performance valuable attributes are math scores, reading the score, and writing the score. These are very strong attributes of the dataset.

3.1.2 MultivariaTe Regression Analysis

As the technique mentioned above suggests, an empirical evaluation based on the statistics set is performed to check multiple factors and their effect at the median as a response variable. Inside the first region, facts evaluation is performed on scores, in the experience that the impudence of the number of college students on the general performance is analyzed, which may be seen inside the above parent. In determine , the horizontal axis represents the common wide variety of college students' overall performance, at the same time as the vertical axis represents the median.

3.1.3 Mean Absolute Error

Mean Absolute Errors (MAE) is a loss characteristic used for regression. Use MAE whilst doing regression and don't want outliers to play a big function. The loss is an implying over the absolute differences between actual and expected values deviations in either direction from the actual value are handled the identical manner. After testing the performance of the built models using the K-fold cross-validation method.

Table 2 Accuracy from mean absolute error

CLASSIFIER	MAE	ACCURACY
GUASSIAN	4.3161	95.69%
LINEAR REGRESSION	4.2543	95.75%
MLR	5.2163	94.78%
SMO Regression	4.234	95.77%

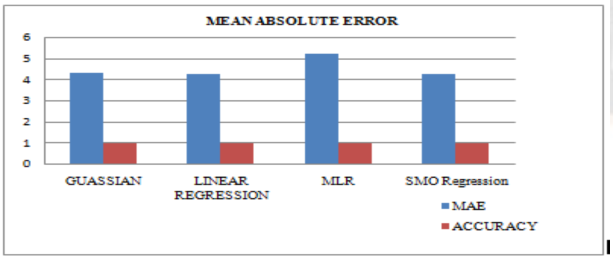


Fig.1 Visualization of Mean Absolute Error

3.1.4. Root Mean Square Error

Ensure the accuracy of classifiers with the aid of a statistical approach called root suggest mistakes. Technique needs a difference among discovered frequency and predicted frequency values determined. Man or woman variations are considered as residuals and RMSE serves to combination the vbnm into an unmarried measure of predictive authority.

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}.$$

Table 3. Accuracy from root mean square error

CLASSIFIER	RMSE	ACCURACY
GUASSIAN	5.4537	94.54%
LINEAR REGRESSION	5.3353	94.66%
MLR	6.4424	93.55%
SMO Regression	5.3357	94.67%

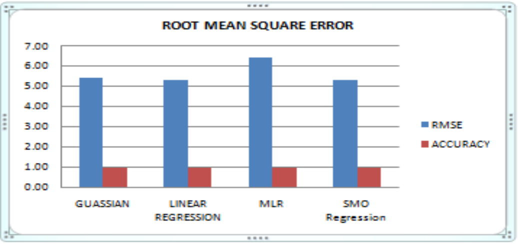


Fig.2 Visualization of root *mean square error*

3.1.5 Relative Absolute Error

Relative Absolute errors (RAE) are a way to measure the overall performance of a predictive version. It's in the main utilized in device gaining knowledge of, records mining, and operations control. RAE is not to be stressed with relative errors[3], which is a general measure of precision or accuracy for units like clocks, rulers, or scales.

The error is made relative for simple predictor the common of the real values from the training records. After testing the performance of the built models using the K-fold cross-validation method. It have obtained for each model the values of the metrics RMSE, R-squareand MAE.

$$\frac{|p_1 - a_1| + \ldots + |p_n - a_n|}{|\bar{a} - a_1| + \ldots + |\bar{a} - a_n|}$$

With
Actual target values a1,a2,...an.
Predicted target values p1,p2....pn.

Table *4* Accuracy from root absolute error

CLASSIFIER	RAE	ACCURACY
GUASSIAN	35.91%	64.09%
LINEAR REGRESSION	35.39%	64.61%
MLR	43.40%	56.60%
SMO Regression	35.22%	64.78%

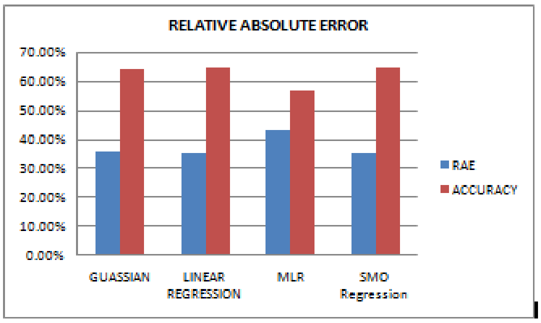


Fig.3 VisualizationRelative Absolute Error

3.1.6 Root Relative Squared Error

The basis Relative Squared error (RRSE) is described as the square root of the sum of squared mistakes of a predictive model[4] normalized by way of the entire squared errors of a simple model. In different words, the square root of the Relative Squared error (RSE).

$$\frac{(p_1-a_1)^2+\ldots+(p_n-a_n)^2}{(\bar{a}-a_1)^2+\ldots+(\bar{a}-a_n)^2}$$

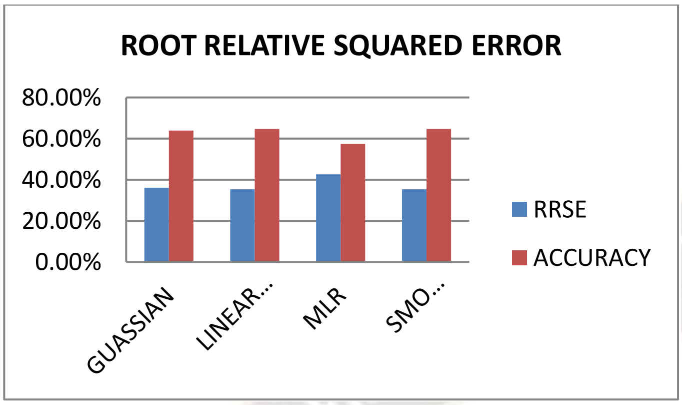


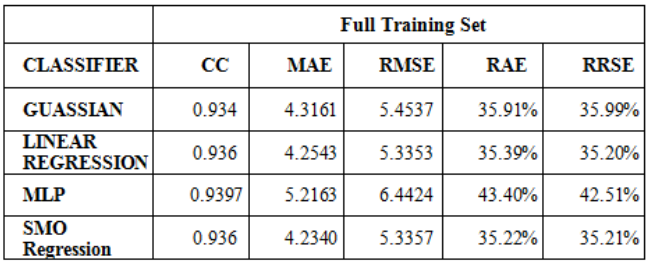
Fig. 4 Visualization Root Relative Squared Error

3.2 EVALUATION MEASURE

The best result of accuracy in calculations of the mean absolute error, root mean squared error, root relative squared error, and relative absolute error in machine learning performance of experimental way of approach. The MAE is better when the lower value of error. It is not very sensitive to outliers in comparison to MSE[3]. It does not punish huge errors that performance is measured on continuous variable data. It gives a linear value that averages the weighted individual differences equally. MSE commonly used metrics but are least useful when a single bad prediction. The entire model predicts abilities when the dataset contains a lot of noise. It is most useful when the dataset contains outliers or unexpected values. RMSE is squared before they are averaged which assigned a higher weight to larger errors that are much more useful when large errors are present and they drastically affect the model's performance. Avoid taking the absolute value of error and this trait is useful in many mathematical calculations. This value is better when a lower error.

The experimental result of pupil dataset parameter metrics Correlation Coefficient (CC) suggests Absolute errors (MAE), Root mean square blunders (RMSE), Relative Absolute error (RAE), and Root Relative Squared mistakes *(RRSM).*

Table 5. Description of full training set



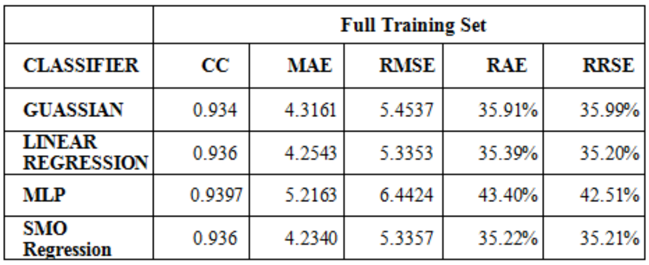
	Full Training Set				
CLASSIFIER	CC	MAE	RMSE	RAE	RRSE
GUASSIAN	0.934	4.3161	5.4537	35.91%	35.99%
LINEAR REGRESSION	0.936	4.2543	5.3353	35.39%	35.20%
MLP	0.9397	5.2163	6.4424	43.40%	42.51%
SMO Regression	0.936	4.2340	5.3357	35.22%	35.21%

The results produced by way of the classifier technique on the training dataset hired, comprising a test dataset of one thousand data. The experimental final results stated that a fixed of 518 college students, i.e. 90% of being graduated. Likewise, the 155 college students contain the 80% possibility of being graduated and so on.

3.2.1Guassian Process Regression (GPR)

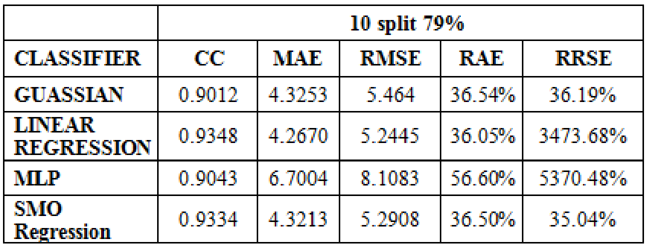
The Gaussian method possibility distribution over prediction made by means of the corresponding Bayesian neural network computation in artificial neural community (ANN). The Gaussian procedures Classifier is a classification machine gaining knowledge of algorithm. Gaussian approaches are a generalization of the Gaussian possibility distribution and can be used as the basis for stylish non-parametric gadget gaining knowledge of algorithms for type and regression.

Table 6 Description of test set



	Full Training Set				
CLASSIFIER	CC	MAE	RMSE	RAE	RRSE
GUASSIAN	0.934	4.3161	5.4537	35.91%	35.99%
LINEAR REGRESSION	0.936	4.2543	5.3353	35.39%	35.20%
MLP	0.9397	5.2163	6.4424	43.40%	42.51%
SMO Regression	0.936	4.2340	5.3357	35.22%	35.21%

Table 7 Description of test set



	10 split 79%				
CLASSIFIER	CC	MAE	RMSE	RAE	RRSE
GUASSIAN	0.9012	4.3253	5.464	36.54%	36.19%
LINEAR REGRESSION	0.9348	4.2670	5.2445	36.05%	3473.68%
MLP	0.9043	6.7004	8.1083	56.60%	5370.48%
SMO Regression	0.9334	4.3213	5.2908	36.50%	35.04%

3.2.2 Linear Regression (LR)

The regression line is an instant line. Whereas logistic regression is for type issues, which predicts a possible range between zero to one. The regression line is a sigmoid curve. It defines the relationship among the two variables by way of becoming a regression line to the data. One of the two variables depends variable that's depending on some other variable called an unbiased variable. With the exceptional fit regression line to the information, the error price among the predicted and actual values may be minimized.

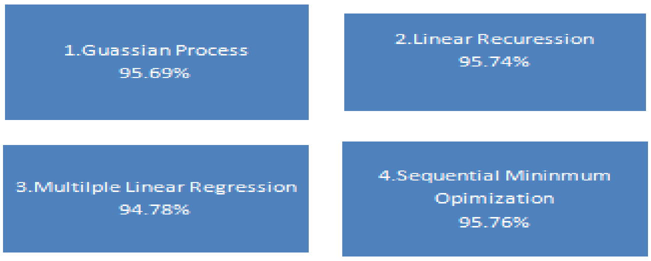
It is an arithmetic regression method that is used for prediction analysis. It is a standard method that works on regression and demonstrates the relationship among the constant variables. It finds the function which predicts for given X predicts Y where Y is continuous. F(X) → Y. Many types of functions can be used. The simplest type of function is a linear function. X can comprise a single feature or multiple features. The basic concept of linear regression is to find a line that best fits data as shown in Figure 4.6. The best fit line means the total prediction error for all data points is as small as possible. The error is the distance between the points to the regression line.

3.2.3 SMO Regression

Sequential minimum Optimization computing and retaining the threshold cost. Its miles solving the regression hassle the use of SVM [4].

4. PERFORMANCE METRICS OF STUDENT ACADMECIC PERFORMANCE THROUGH FEATURE EXTRACTION METHOD

Prediction of student performance is measures like errors calculated the setting of the learning rate runs through the learning process of the entire network. When training is allowed, choosing the right one can improve the protection of the system and ultimately reduce errors. There is no perfect method for the selection of the learning rate and only by frequently trying different learning rates to achieve a balance between training efficiency and network error. In the end a new classification was carried out based on different degrees of stability to evaluate and compare the result of the models used. Feature extraction is performed on a dataset and classification experiments are performed using a simple classifier. At the same time, the performance of the deep learning algorithm is compared with other similar algorithms. Test data are from a kaggle dataset. The model is trained during training and gets better results from validation.



1.Guassian Process
95.69%

2.Linear Recuression
95.74%

3.Multilple Linear Regression
94.78%

4.Sequential Mininmum Opimization
95.76%

Fig.5 Display the accuracy of machine learning regression function

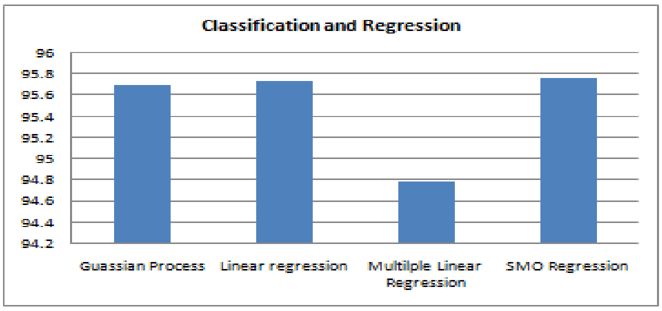


Fig 6. Classification and Regression

5. CONCLUSION

Instructional institutes require revolutionary techniques to enhance the first-class of training to acquire the first-class consequences and decrease the failure rate. Academic achievement of a student is of the highest priority for any institute or university across the globe. Using various methods to predict the performance of the student accurately would be highly required. Predicting the performance would also enable the institutions to focus more on students having more probability of performing lower in order to improve their performance. Deep learning in education performs a great feature because it empowers research and is awaiting the overall performance of inexperienced persons so that critical measures can be taken to decorate the mastering way. Essentially describes the supervised studying algorithms are categorized or regressed. Recursion strategies are displayed and in the long run, multiple linear Perceptron contained a decrease charge of errors.

References

[1] J. Whitehill, K. Mohan, D. Seaton, Y. Rosen, and D. Tingley. Delving deeper into mooc student dropout prediction. arXiv preprint arXiv:1702.06404, 2017.

[2] W. Wang, H. Yu, and C. Miao. Deep model for dropout prediction in moocs. In Proceedings of the 2nd International Conference on Crowd Science and Engineering (ICCSE 2017), pages 26–32, Beijing, China, 2017.

[3]G. Sujatha, S. Sindhu and P. Savaridassan "Predicting student's performance using personalized analytics", IJARTE Volume.119 Issue. 12, Page 221-231 2018.

[4] S. K. Shevade, "Improvements to the SMO Algorithm for SVM Regression", IEEE TRANSACTIONS ON NEURAL NETWORKS, VOL. 11, Issue. 5, Page 1188- 1193,2000.

[5] S.Ayesha, et al., "Data Mining Model for Higher Education System", European Journal of Scientific Research, Vol. 43, No.1, pp.24 – 29, 2010.

[6]Emaan Abdul Majeed, and Khurun Nazir Junejo, "Grade Prediction Using Supervised Machine Learning Techniques," Internation journal of advanced research in computer science, Vol. 9, Issue 23, Page 112-120, 2016.

[7]Ermiyas Birihanu Belachew, and Feidu Akmel Godena, "Student Performance Prediction Model using Machine Learning Approach: case of Wolkite University", in International Journal if Advanced Research in Computer Science and Software Engineering: Vol. 7, Issue 2, Page.550-563, 2017.

[8] S. Kotsiantis, et al." Preventing student dropout in distance learning systems using machine learning techniques", Applied Artificial Intelligence, Vol.18, Issue 15, Page 411-426,2017.

[9] C.E.Moucary,M. Khair and W. Zakhem, "Improving student's performance using data clustering and neural networks in foreign-language based higher education" Research Bulletin of Jordan ACM, Vol. 2, Issue 3,Page 27-34,2011.

[10] E.Moucary,M. Khair and W. Zakhem, "Improving student's performance using data clustering and neural

***Article Received:** 25 July 2023 **Revised:** 12 September 2023 **Accepted:** 30 October 2023*

networks in foreign-language based higher education" Research Bulletin of Jordan ACM, Vol. 2, Issue 3,Page 27-34,2011

[11] O, O., & P, C. " Predicting Students' Academic Performances – A Learning Analytics Approach using Multiple Linear Regression, international journal of Computer Application, Vol.157, Issue 4, Page 37-44, 2020.