*Grenze International Journal of Engineering and Technology, June Issue*

# Machine Learning based Student Academic Performance Prediction

Mr. J. Nagaraju[1], P. Amarnath Reddy[2], P. Maheshwar Reddy[3], A. Vijaychand[4] and N. Yaswanth[5]
[1-5]Department of AI&DS, Lakireddy Bali Reddy College of Engineering, Mylavaram, Andhra Pradesh, India
Email: jnraju@lbrce.ac.in, amarnathpalakollu123@gmail.com, mahisince2002@gmail.com, vijaychanduathanti4@gmail.com, yaswanthreddynaredla9912@gmail.com

***Abstract*—The primary objective of educational institutions is to ensure students receive a high-quality education. In modern times, a significant portion of their resources and efforts are directed towards assessing student performance. Through thorough analysis, they can pinpoint groups of students who may require additional support and interventions to improve their academic outcomes. Recently, researchers have introduced various machine learning techniques to forecast academic success. This study employs Linear Regression and Random Forest algorithms to anticipate students' academic achievements. Evaluating the models using metrics such as confusion matrix, accuracy, precision, recall, and F1 score reveals that the Random Forest algorithm outperforms others in predicting academic performance, as indicated by the findings.**

***Index Terms*— Academic performance, Machine learning, Confusion Matrix, Random Forest.**

## I. INTRODUCTION

Education plays a vital role in individual development, fostering knowledgeable citizenship. It enables us to grasp and enhance societal functioning. It involves a teacher-student context, like Zoom or Teams meetings. For students, it spans primary, middle, and high school, followed by college or university. Free higher education is offered based on performance post-high school. Machine Learning algorithms aid in predicting student performance.

Machine Learning (ML) is branch of Artificial Intelligence which makes the computer tasks to improve their performance automatically by themselves through the knowledge they gain through experience. Machine Learning algorithms can be classified into three categories most popular among the is Supervised Learning which works on an outcome-based learning, the next is Unsupervised Learning which doesn't have a specified outcome mostly used for clustering and finally. Reinforcement Learning which makes machines to learn themselves by using trial and error method

## II. LITERATURE REVIEW

In the realm of academic data analysis, a forward-looking model has been devised to forecast scholars' achievements, aiming to identify those students facing potential challenges. This drawback holds considerable weight due to the multifaceted nature of academic performance, influenced by various factors such as grades, statistical metrics, psychological traits, cultural background, educational history, and academic environment.

*Grenze ID: 01.GIJET.10.2.642*

Machine learning techniques like Random Forest, AdaBoost, and Support Vector Machine (SVM) are employed to anticipate student success. A condensed overview of the relevant research findings is provided in Table I.

TABLE I. OUTLINE OF RESEARCH WORK DONE

| S.No | Paper | Features | Size of Dataset | ML Algorithms Used | Best fitted Algorithm |
|---|---|---|---|---|---|
| 1 | Camilo et a1, 2015[3] | Student Grades | 635 | Naïve Bayes, Decision Tree | Naïve Bayes |
| 2 | Santan a, 2017[4] | Student Grades | 400 | Decision Tree, SVM, ANN, Naïve Bayes | Support Vector machine |
| 3 | Xu et al 2014[7] | Students Grades, Environment | 1170 | Linear Regression, Logistic, Proposed Progressive Prediction Algorithm, Regression, Random Forest, KNN | Proposed Progressive Prediction algorithm |
| 4 | 2014 [6] | Ability, Learning Strategies, Persona | 914 | Regression SVM, DT, ANN, KNN, Naïve Bayes | Bayes, Support Vector Machine, KNN |
| 5 | Alhabri et al, 2016[10] | Performance, Student Statistical Characteristics, Student modules | 2687 | Logistic Regression, ANN, BN, Decision Tree, DA, DL, and Ensemble approach | All gave almost same results |
| 6 | Meier et al, 2015[9] | Student Grades | 700 | KNN, Linear regression, SVM Logistic regression | Proposed Algorithm |
| 7 | Livieri s et al, 2012[1] | Student Grades | 279 | Decision Tree, Rule learning, SVM, ANN | SVM, ANN |
| 8 | Sarker, 2014[5] | Student Satisfaction and integration, Student Statistical characteristics and personal Information | 149 | Logistic Regression Artificial Neural Network | Logistic regression |

*A. Data Set*

In our data set we have 38 attributes likely encompass a range of factors related to a student's academic performance. These could include demographic information (such as age, gender), socioeconomic status indicators (like parental income or education level), educational background (previous grades or test scores), and behavioral or psychological factors (attendance, sports, project works).

TABLE II. ATTRIBUTE METADATA

| School | Sex | Age | Address |
|---|---|---|---|
| Famsize | Pstatus | Medu | Fedu |
| Mjob | Fjob | Reason | guardian |
| traveltime | studytime | failures | schoolsup |
| famsup | paid | activities | nursery |
| higher | internet | romantic | famrel |
| freetime | gout | Dalc | Walc |
| Health | Absences | G1 | G2 |
| G3 | Sports | projects | volunteer |
| clubs | cumulative | | |

Source: https://archive.ics.uci.edu/ml/datasets/student+performance

## III. PROPOSED WORK

*A. System Components*

*Data collection*

A Student database Data Set is obtained from online repository. This information approach understudy accomplishment in auxiliary training of two Portuguese schools. The dataset 396 values and 38 attributes.

*Data pre-processing:*

Data preprocessing is a crucial step in preparing raw data for analysis. It involves several tasks such as handling missing values, dealing with outliers, standardizing or normalizing numerical features, encoding categorical variables, and scaling the data to ensure uniformity and comparability. Additionally, data preprocessing may

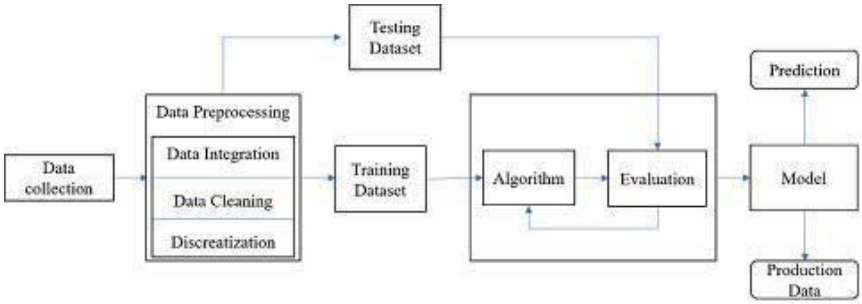


Figure 1. Components of the system

include feature selection or dimensionality reduction techniques to remove irrelevant or redundant attributes. The goal of data preprocessing is to clean and transform the data into a format that is suitable for analysis by machine learning algorithms, enhancing the accuracy and effectiveness of predictive models. Proper preprocessing ensures that the resulting models are robust, generalizable, and free from biases introduced by the original data.

*Training Data*

- Training data is the foundational set of examples used to teach a machine learning model how to perform a specific task. It comprises input-output pairs or features with corresponding labels. These data points serve as the model's learningmaterial,
- Enabling it to discern patterns and relationships within the data. High-quality and diverse training data is crucial for the model to generalize well to unseen instances. Data pre-processing, including cleaning, normalization, and feature engineering,is often performed to refine the training data beforefeeding it into the model. The effectiveness and accuracy of the trained model heavily depend on the quality and representativeness of the training data.

*Testing Data*

Testing data refers to a subset of data used to evaluatethe performance and generalization ability of a machine learning model after it has been trained on the training data. It's crucial for assessing how well the model can make predictions on new, unseen examples. Testing data should be separate from the training data to ensure unbiased evaluation. Typically, testing data is withheld from the model during the training process and only used for evaluation afterward. It helps in detecting overfitting and provides insights into the model's performance metrics such as accuracy, precision, recall, and F1-score,among others. The quality and representativeness of the testing data are vital for obtaining reliable assessments of the model's performance.

*B. Methodology*

*Linear Regression Method*

Linear regression serves as a machine learning technique employed for statistical analysis, particularly in understanding the relationship between a dependent variable and one or more independent variables. Its primary objectives include assessing the significance of predictors, making predictions, and forecasting trends.

In this model, a straight line is utilized to represent the data. The dependent variable (y) is influenced by changes in the independent variable (x). When the independent variable increases, a positive relationship is observed, resulting in an increase in the dependent variable (y), represented as $y = \beta0 + \beta1x$. Conversely, a negative relationship occurs when an increase in the independent variable leads to a decrease in the dependent variable, expressed as $y = \beta0 - \beta1x$.

To establish the line of regression, predicted observations are plotted, and errors are calculated for each actual observation. The regression line is then determined as the one with the least error, aiming to minimize discrepancies. The slope of the regression line is denoted as $\beta1$, while the y-intercept is represented by $\beta0$. The optimal fit line, or the best match line, is obtained by determining the values of $\beta0$ and $\beta1$

*Linear regression accuracy is assessed using metrics such as R-squared (R2) and Root Mean Squared Error (RMSE), which gauge how well the model fits the data.*

*Random Forest Method*

Random forest is a machine learning algorithm that uses decision trees to construct its model. It can be used to solve both classification and regression issues. It's more adaptable, and the production is a mix of multiple decision trees.

Step1: we select the random samples from the original dataset and create a bootstrap dataset by the random samples.
Step2: Create decision trees using the bootstrapped dataset. Considering the random subset of variables at each step. Then go back to step1 and again performing step2. Now we have several decision trees for our dataset. we take a few decision trees that will satisfy all the conditions of the problem
Step3: we observe each decision tree and its output. Because the random forest is made up of multiple decision trees. So we consider every decision tree to predict the best value.
Step4: we observe every decision and decides the result. then which is having the highest votes we declare it as our final output.

This algorithm is used in our project to predict student performance with high accuracy.

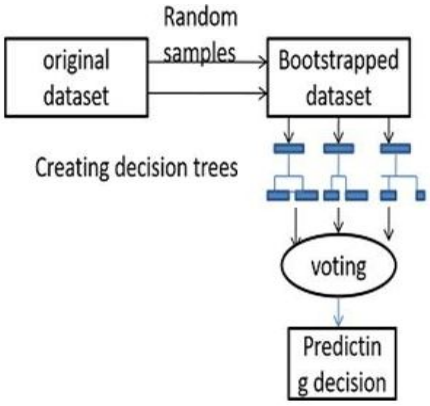


Figure 2. Random Forest Methodology

TABLE III. COMPARISON CHART FOR THE PERFORMANCE METRICS

| Algorithm | Accuracy | Precision | Recall | F1 Score |
|---|---|---|---|---|
| Linear Regression | 0.84 | 0.77 | 0.79 | 0.78 |
| Ada boast | 0.94 | 0.73 | 0.93 | 0.82 |
| SVM | 0.94 | 0.78 | 0.88 | 0.82 |
| Random Forest | 0.95 | 0.77 | 0.87 | 0.82 |

*Support Vector Machine(SVM)*
Support Vector Machine (SVM) is a versatile machine learning technique utilized for both classification and regression tasks. Employing various kernels such as linear, polynomial, or radialbasis function, SVM efficiently separates data pointsin a high-dimensional space. Typically functioning as a binary classifier, SVM excels in scenarios with two potential outcomes. However, its training process can be computationally intensive, particularly with large datasets, and its performancemay degrade in such instances. Despite these challenges, SVM remains a robust choice for many classifications regression problems, offering flexibilityand effectiveness in diverse domains

*Adaboost*
AdaBoost, short for Adaptive Boosting, is a powerful ensemble learning technique in machine learning. It iteratively adjusts the weights of misclassified instances, assigning higher weights to those instances that are harder to classify correctly. This adaptability enables AdaBoost to focus on the difficult instances, gradually improving the overall performance of the ensemble model. Unlike traditional boosting methods, AdaBoost constructs a series of weak learners, each refining the errors made by its predecessors. Through this iterative process, weak learners are progressively transformed into strong ones, collectively enhancing the predictive accuracy of the model. In essence, AdaBoost efficiently harnesses the collective wisdom of weak learners to achieve impressive classification results.

## IV. RESULT ANALYSIS

The machine learning algorithm models are created and tested, SVM (support vector machine), Ada Boost, linear regression, and Random Forest.

Performance Measures:
The developed Prediction model is compared with other existing model based on four factors. Out of which Two True values True Positive (TP), True Negative (TN) and two are False Positive (FP) and false negative (FN).
Accuracy
Accuracy is calculated by 1) ACC = TP+TN/TP+TN+FN+FP
Recall is calculated by 2) REC = TP/TP+FN = TP/P
Precision

Precision is calculated by 3) PREC = TP/TP+FP

4) F1 Score = 2 * Precision * Recall /Precision + recall

Here the linear regression gives 92% accuracy, Ada boost algorithm gives 97% accuracy, SVM gives 97% accuracy and Random Forest gives 98% accuracy Based upon the performance metric values like precision, recall, f1score,

F1 Score

F1 Score is calculated by based on the precision and recall of the test.

and accuracy the Random Forest gives the better results of the student's academic performance. Table 2 provides Comparison chart for the performance metrics. Table 3 provides accuracy and performance measures comparison for different models

*Experimental Analysis*

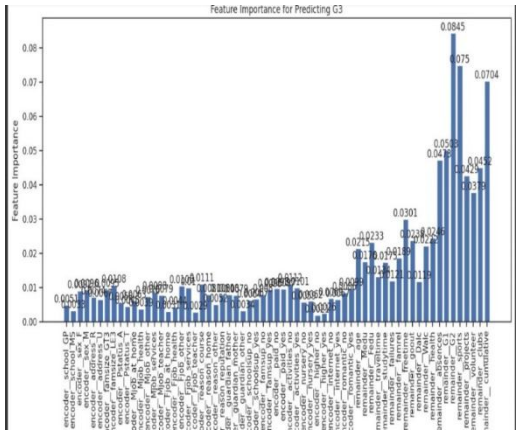


Figure 3. Feature Importance

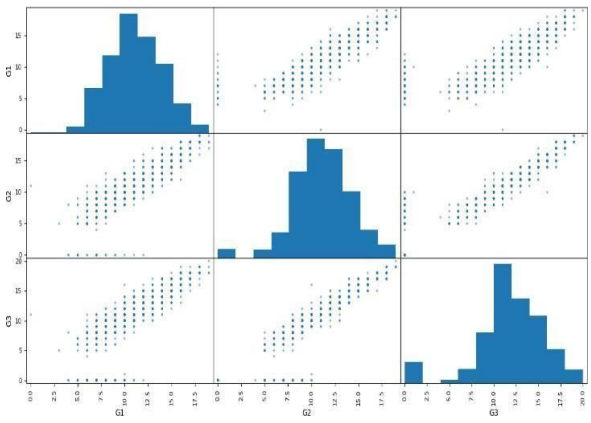


Figure 4. Relationship between G1,G2 and G3

G3(Grade 3) is highly reliant on G1(Grade 1) and G2(Grade 2). G1 and G2 have a significant impact onG3's pattern.

*Relationship between G1,G2 and G3*

The following are the confusion matrixes of Linear Regression, AdaBoast, SVM and Random Forest

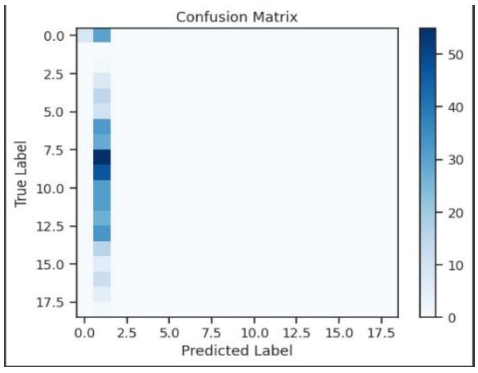


Figure 5. CM value for Linear Regression

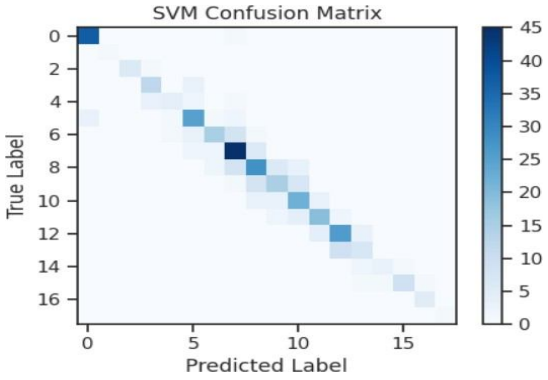


Figure 6. CM value for SVM

## V. Conclusion & Future Work

Predicting student academic achievement poses a formidable challenge within educational systems, prompting the exploration of novel methodologies. In this study, researchers proposed and evaluated two distinct machine learning techniques tailored for forecasting the success rates of students enrolled in academic courses. Their primary objective encompassed not only the identification of pivotal factors impacting student performance but also the facilitation of student self-assessment, leveraging past academic records for improvement. The study's underscore efficacy of the proposed model in accurately forecasting students' future academic outcomes based

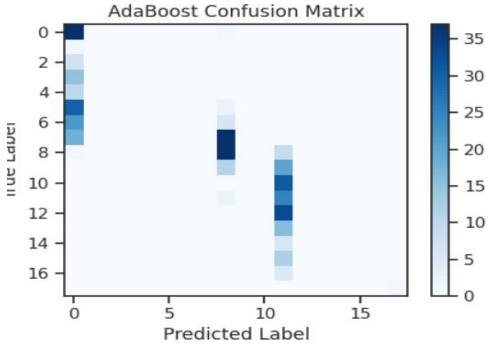


Figure 7. CM value for AdaBoost

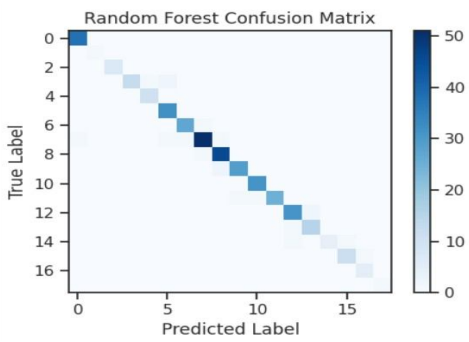


Figure 8. CM value for Random Forest

on their prior test performances. Noteworthy results revealed that linear regression achieved an impressive 92% accuracy in categorizing occurrences, while Random Forest exhibited an even higher accuracy rate of 98%, affirming the robustness of the predictive framework. To enhance future studies on this dataset, the integration of natural language processing techniques is recommended, as they have shown promise in providing deeper insights and more accurate.

## REFERENCES

[1] Aydoğdu, Ş. Predicting student final performance using artificial neural networks in online learning environments. Educ Inf Technol 25, 1913–1927 (2020). https://doi.org/10.1007/s10639-019- 10053-x.

[2] Basha, S., & Subrahmanyam, M., “Research of various data mining Techniques for IoT applications”, International Journal of Techniques Recent Technology and Engineering, Vol.8, Special Issue 11, 2019, pp. 1083-1086.

[3] C. E. López Guarín, E. L. Guzmán and F. A. González, "A Model to Predict Low Academic Performance at a Specific Enrollment Using Data Mining," in IEEE Revista Iberoamericana de Tecnologias del Aprendizaje, vol. 10, no. 3, pp. 119-125, Aug. 2015, doi:10.1109/RITA.2015.2452632.

[4] Evandro B. Costa, Baldoino Fonseca, et.,al “Evaluating the effectiveness of educational data mining techniques for early prediction of students' academic failure in introductory programming courses”, Computers in Human Behavior, Volume73, 2017, Pages 247-256, ISSN 0747-5632, https://doi.org/10.1016/j.chb.2017.01.047.

[5] F. Sarker, T. Tiropanis and H. C. Davis, "Linked data, data mining and external open data for better prediction of at-risk students," 2014 International Conference on Control, Decision and Information Technologies (CoDIT), 2014, pp. 652-657, doi: 10.1109/CoDIT.2014.6996973.

[6] G. Gray, C. McGuinness and P. Owende, "An application of classification models to predict learner progression in tertiary education," 2014 IEEE International Advance Computing Conference (IACC), 2014, pp. 549-554, doi: 10.1109/IAdCC.2014.6779384.

[7] J. Xu, K. H. Moon and M. van der Schaar, "A Machine Learning Approach for Tracking and Predicting Student Performance in Degree Programs," in IEEE Journal of Selected Topics in Signal Processing, vol. 11, no. 5, pp. 742-753, Aug. 2017, doi: 10.1109/JSTSP.2017.2692560

[8] N. Buniyamin, U. b. Mat and P. M. Arshad, "Educational data mining for prediction and classification of engineering students achievement," 2015 IEEE 7th International Conference on Engineering Education (ICEED), 2015, pp. 49-53, doi:10.1109/ICEED.2015.7451491.

[9] Y. Meier, J. Xu, O. Atan and M. van der Schaar, "Predicting Grades," in IEEE Transactions on Signal Processing, vol. 64, no. 4, pp. 959-972, Feb.15, 2016, doi: 10.1109/TSP.2015.2496278.