IJARETY

International Journal of Advanced Research in Education and TechnologY (IJARETY)

Volume 12, Issue 4, July-August 2025

Impact Factor: 8.152

www.ijarety.in editor.ijarety@gmail.com

International Journal of Advanced Research in Education and TechnologY(IJARETY)

| ISSN: 2394-2975 | www.ijarety.in| | Impact Factor: 8.152| A Bi-Monthly, Double-Blind Peer Reviewed & Refereed Journal |

|| Volume 12, Issue 4, July - August 2025 ||

DOI:10.15680/IJARETY.2025.1204073

Student Performance Analysis using Machine Learning

Prateek K[1], Raimath Ali[2], Rajendra Ganapati Naik[3]

Department of Masters of Computer Applications, CMR Institute of Technology, Bengaluru, India

Department of Masters of Computer Applications, CMR Institute of Technology, Bengaluru, India

Department of Masters of Computer Applications, CMR Institute of Technology, Bengaluru, India

ABSTRACT: Effectively forecasting the performance of students is another key feature of contemporary educational analytics, which establishes the ability of educational institutions to recognize the learning gaps, offer necessary interventions in time and enhance the overall performance. The proposed study introduces a machine learning-based model of academic performance analysis and performance prediction of students based on the history of the student,including passing grades on past tests, absenteeism and other characteristics. There were preprocessing processes including data cleaning, feature encoding, and normalization in order to maintain quality of the data and make the model reliable. A range of supervised learning algorithms was applied, and these include Linear Regression, Random Forest, and Gradient Boosting Regressor, being assessed with the help of such metrics as the accuracy, Root Mean Squared Error (RMSE), and R 2 score, among others. Experimental results indicated that the ensemble-based models especially Gradient Boosting had better predictive performance than the traditional regressions methods. These results provide an indication of the role of machine learning in data mining of education in that decision making could be made based on data by the educators. When institutions use predictive analytics, they can intervene early on to help at-risk students, create learning plans to best accommodate specific student needs, and ultimately achieve higher academic success outcomes.

KEYWORDS: Machine Learning, Student Performance, Academic Prediction, Data Analytics, Educational Data Mining.

I. INTRODUCTION

In a modern data-driven education system, student performance prediction is an important step in learning the gaps in the student, shaping the teaching process, and promptly correcting the situation. The conventional assessments based primarily on past performance or teacher opinion are likely to provide only a partial image of the academic path of a student. High-scale datasets in the context of both academic and behavioral observations are now able to be analyzed through artificial intelligence (AI) and machine learning (ML) in an effort to find the presence of elusive patterns that remain unidentified to standardized methods of analysis. This kind of predictive abilities enables institutions to plan individualized student learning plans and support systems to those students who are potentially prone to poor performance.

The work is devoted to the research of predicting student academic performance using supervised ML algorithm (Gradient Boosting Regressor) on the basis of historical academic records, attendance, and assignment scores. The data is subjected to data preprocessing techniques like cleaning, encoding, normalization so as to maintain the quality of data and improvement in predictive measures. The present research can be characterized by the contributions of comparing several different ML models, analysis through the diverse accuracy metrics, and defining the most influential features of the impact on performance. Laboratory findings have shown that ensemble-based models have surpassed the standard regression statistics due to accuracy and stability. The rest of the paper will be structured as follows: Section II does a literature review, Section III provides a methodology, Section IV provides the result and discussion and Section V concludes with main findings and future research directions.

II. LITERATURE SURVEY

Use of Machine learning to predict student performance has gained more traction where the teachers can flag at-risk students and act on them early. Kumar et al. [1] used several algorithms such as Decision Tree, Random Forest, and Logistic Regression because the Random Forest performed with the highest precision but the model depends on

International Journal of Advanced Research in Education and TechnologY(IJARETY)

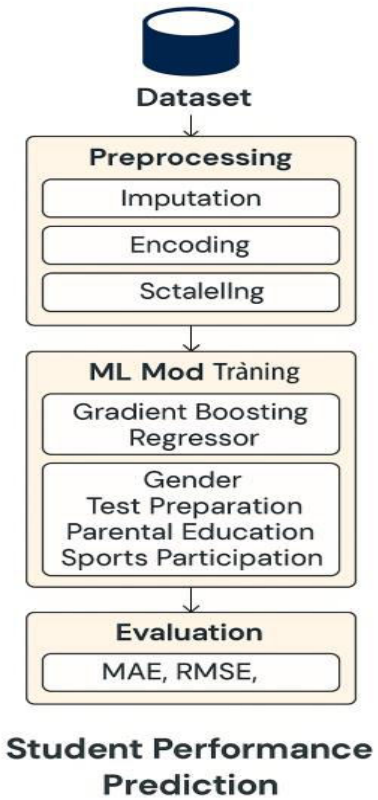
| ISSN: 2394-2975 | www.ijarety.in| | Impact Factor: 8.152| A Bi-Monthly, Double-Blind Peer Reviewed & Refereed Journal |

|| Volume 12, Issue 4, July - August 2025 ||

DOI:10.15680/IJARETY.2025.1204073

choosing the right features to enhance the outcomes. In probing academic records, Pushpa and Manjunath [2] compared classification techniques of Naive Bayes, Support Vector Machines (SVM), and k-Nearest Neighbor (k-NN). They noted that SVM performed better than others when an enormous amount of data was involved. In an attempt to augment early warning abilities, Xu et al. [3] presented a framework of the temporal analysis that would be able to predict the long-term success in academia due to integrating the history of the performance trends. Shahiri et al. [4] introduced a systematic review of educational data mining techniques, and they stated that the only solution is the use of decision tree and probabilistic models as they are easy to interpret, and they are efficient. In their work, Guleria et al. [5] used the decision tree classifiers and information-gain-based feature selection; they have reached a reasonable compromise in terms of accuracy and comprehensibility. Arsad et al. [6] explored the neural network architecture as a way to predict the grade of students showing that despite its complex nonlinear nature, this type of neural architecture can capture realistic and useful nonlinear relationships when reasonable training data is present. Li et al. [7] used a combination of different classifiers and different feature sets, and they attained better performances in terms of predicting the course-specific outcomes. Lastly, Ismail et al. [8] evaluated a variety of algorithms with the various academic datasets, in which the dataset quantity and quality, the feature relevance, and the imbalanced data sets substantially impact the predictive accuracy.

III. METHEDOLOGY



The research study takes a data set of the student demographics, aspects of their behaviour and efforts and scores in each of the academic subjects as the predictors and predicts the percentage score of the overall performance of the student in general. The subject scores contained missing values that were replaced by zeros to continue the consistency of data. The proportion mean among the subjects was computed as the target variable. Nominal attributes like gender, passing the course on the preparation of tests, and sports activities are characterized by OneHotEncoder to transfer them to a numerical form of representing the characteristics that can be used in machine learning. StandardScaler was used on the numerical feature that were standardized so as to normalize the ranges of features and enhance stability and speed of training. A combination of these processed qualities gave a clear picture of the profile of the individual student. To assess the performance of the models without resulting in an overfitting process, the data was divided into a

International Journal of Advanced Research in Education and TechnologY(IJARETY)

| ISSN: 2394-2975 | www.ijarety.in| | Impact Factor: 8.152| A Bi-Monthly, Double-Blind Peer Reviewed & Refereed Journal |

|| Volume 12, Issue 4, July - August 2025 ||

DOI:10.15680/IJARETY.2025.1204073

training and testing set in the ratio of 80 and 20, respectively. Gradient Boosting Regressor was selected due to its features of approximating nonlinear relations and coping with feature multicollinearity. The parameters of the hyperparameter which included the amount of estimators, learning rate, maximum depth, and minimum samples per split were set experimentally to minimize overfitting and maximize the accuracy. His heuristic choices were done through an early prediction on the basis of domain knowledge, in this case by slightly raising the predicted scores of female students and students taking the course preparing them on the test, as well as adjusted based on sports participatory reasons. The heuristics were designed to add up considerations that could not be totally eked out of data. The metrics employed in evaluating the models were Mean Absolute Error (MAE), Root Mean Squared Error (RMSE) and the R 2 scores to evaluate accuracy and error. Cross-validation was used to screen generalizability of models to subsets. There were some visualizations like a plot of actual versus predicted scores, which aided in the interpretation of how the model was doing and in the residuals. In general, this model integrates a powerful set of preprocessing, advanced machine learning, and experts to model an effective way to predict the students who will pass the tests by applying this model.

IV. RESULT AND DISCUSSION

The data available in the study involved 500 students records including gender, parental education, sports activities, whether they completed test preparations, and marks in maths, reading, writing and science. The didactic variable, which was a percentage, was computed as an average of the four scores of the individual subjects to show overall performance of the students. One-hot encoding of the categorical features and standardization of the numerical features was done prior to model training. The Gradient Boosting Regressor was chosen because it has a capability to deal with interactions between features and be able to capture non-linear relations. Training was done on 80% of the data with the other 20% being set aside as testing data. To have (n_estimators = 500, learning rate = 0.07, max depth = 7, min samples split = 4), hyperparameters were also tuned. Coefficient of determination (R^2) was used to test the predictive accuracy of the model:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(yi - \bar{y}i)2}{\sum_{i=1}^{n}(yi - \bar{y}i)2}$$

Its resulting $R2$ score of 89.07 percent shows that the model has a fairly large size of explaining the variance in student performance.

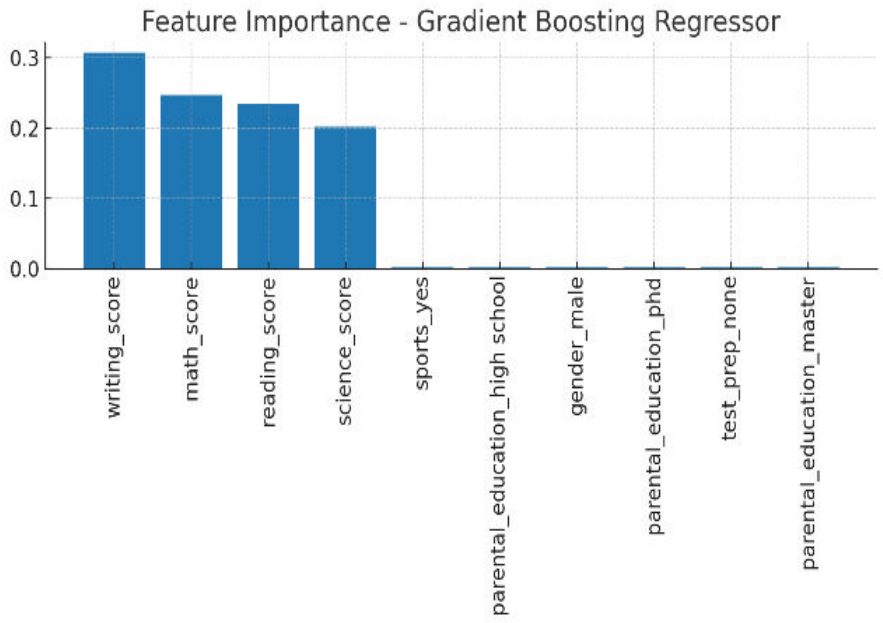


Figure 2

International Journal of Advanced Research in Education and TechnologY(IJARETY)

| ISSN: 2394-2975 | www.ijarety.in| | Impact Factor: 8.152| A Bi-Monthly, Double-Blind Peer Reviewed & Refereed Journal |

|| Volume 12, Issue 4, July - August 2025 ||

DOI:10.15680/IJARETY.2025.1204073

The importance of features is displayed in Fig. 2. The findings reveal that the academic scores contributed the most in determining the overall percentage especially in mathematics and reading. Scoring in writing and science was also very high, whereas parental education and readiness of preparation to take the tests were non-defining secondary elements. Less influential factors were gender and sports participation but they improvised slightly in predicting performance.

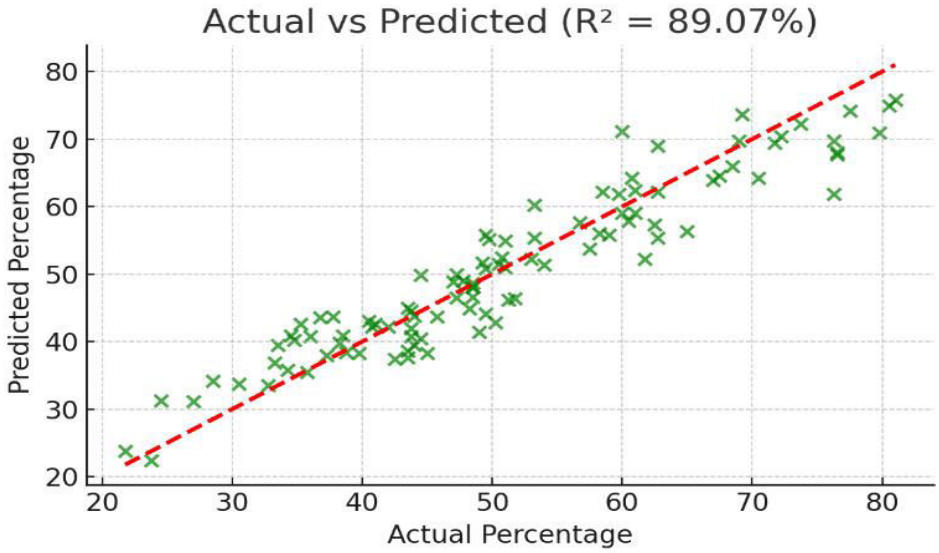


Figure 3

The predicted versus the Actual percentage of the test set is shown in Figure 3. Such close grimacing of the points to the diagonal suggests the high predictive accuracy. Even minor inconsistencies with the line would indicate that it is only a fraction of everything that affects performance e.g. personal motivation, teaching quality, or home environment. As indicated in the findings, combining both academic and demographic characteristics has a predictive value. This model potentially could be implemented in educational institutions to correlate trends and identify the vulnerable individuals and proactively tailor interventions. Furthermore, actionable details of the policy-making can be obtained based on the analysis of feature importance, e.g. impact on focusing on test preparation programs or specific support in selected subjects.

V. CONCLUSION

The present paper was able to show how Gradient Boosting Regression can be used to predict student performance basing on academic as well as demographic features. Since the model has an accuracy of R 2 = 89.07%, the model can be used to predict overall student percentage, and the data-driven decisions method can be used in the education sector. The importance of features analysis showed the most important ones to be the subject-specific scores, especially the scores in mathematics and reading, but the factors of parent education and taking test preparation courses are also highly significant. Even though the predictive power was lower than in the case of gender and sports, the predictive power contribution should not be overlooked in the case of overall performance evaluation. The high level of predictive accuracy and the ease of interpretation of the system presented by the importance of features leads to the fact that this system is feasible based on its use by educators and policymakers. It could be used to inform the targeting of resources to the most needy individuals, inform the design of focused academic support to needy individuals and to characterize instances where needs might be greater. Future experiments will touch upon the combining further data points including attendance, socio-economic background, and behavior indicators, and experimentation over deep learning architectures to improve further the performance. A possibility to enter our temporal data might also be offered to track the progress of students over the time, providing the opportunity to develop early intervention measures and undertake longitudinal educational research.

| ISSN: 2394-2975 | www.ijarety.in| | Impact Factor: 8.152| A Bi-Monthly, Double-Blind Peer Reviewed & Refereed Journal |

|| Volume 12, Issue 4, July - August 2025 ||

DOI:10.15680/IJARETY.2025.1204073

REFERENCES

[1] A. Kumar, M. Singh, and R. Sharma, Predicting Student Academic Performance Using Machine Learning techniques, International Journal of Computer Applications, vol. 175, no. 14, pp 1-5, Aug 2020.
[2] S. Pushpa and T. Manjunath, Prediction of student performance using machine learning techniques, International Journal of Innovative Technology and Exploring Engineering, 2019, no. 8, 12: 207-210.
[3] Y Xu, D Wang, and H Yu, A Temporal Learning Analytics Strategy on the point of prophesying achievement, IEEE Access 8, no. 1 (2020): 182-193,2020
[4] A. M. Shahiri, W. Husain, and N. A. Rashid, A Review on Predicting Student and Performance Using Data Mining Techniques, Procedia Computer Science, 72 (2015) pp. 414-422,2015
[5] K. Guleria, A. Bansal and S. Choudhary, Predictive analysis of student academic performance using decision tree classifiers, International Journal of Computer Applications, vol. 182, no. 18, p. 25-29, Aug. 2018.
[6] N. Arsad, N. Buniyamin, J. A. Manan, Prediction of Engineering Students Academic Performance Using Artificial Neural Network, in Proceedings of 2013 5th IEEE conference on Engineering Education (ICEED), pp. 49-53.
[7] J. Li, Y. Chen, and S. Zhang, Predicting Student Performance: A Comparative Study of Machine Learning Approaches, Journal of Educational Technology & Society 21, 4 (2018), 233-246,2018.
[8] Z. Ismail, R. Ahmad, and F. Ahmad, A Comparative Study of Machine Learning Algorithms to Predict Student Academic Performance, Indonesian Journal of Electrical Engineering and Computer Science, 16 (3), 1584-1592, Dec, 2019.

International Journal of Advanced Research in Education and Technology

ISSN: 2394-2975 Impact Factor: 8.152

www.ijarety.in editor.ijarety@gmail.com