Musik in bayern
ISSN: 0937-583x Volume 90, Issue 8 (Aug -2025)
https://musikinbayern.com DOI https://doi.org/10.15463/gfbm-mib-2025-437

Prediction of Student Performance using Machine Learning Algorithm

Dr. Ananthi Sheshasaayee
Department of Computer Science, Quaid-e-Millath Government College for Women, India
ananthi.research@gmail.com

Mrs. Suganya. A
Department of Computer Science, Quaid-e-Millath Government College for Women, India
suganyaa.research@gmail.com

To Cite this Article
Dr. Ananthi Sheshasaayee, Mrs. Suganya. A **"Prediction of Student Performance using Machine Learning Algorithm"** *Musik In Bayern, Vol. 90, Issue 8, Aug 2025, pp196-207*

Article Info
Received: 22-06-2025 Revised: 26-07-2025 Accepted: 06-08-2025 Published: 28-08-2025

Abstract

Educational data has become a significant area of research in recent times. The analysis of educational data is essential for improving the educational system for students. According to research, the main focus of current studies has been on the study of student performance data. To help students improve for their final exam, academic performance analysis is crucial in determining the areas in which they fall short. Exam outcomes are improved when students utilize the analysis of their past performance to pinpoint areas in which they are weak. Machine learning algorithms are used to analyze such educational data to make more precise and better predictions about the students' final performance. Several machines learning methods, including Random Forest, Gradient Boosting, Support Vector Regression, Logistic Regression, and XGBRF, are used in our study. To effectively construct the model, the hyperparameter is tuned and the feature selection is carried out using the training data. The developed model is applied to test data where the XGBRF algorithm outperforms other algorithms in terms of accuracy when it comes to forecasting student performance.

Keywords: Educational data, Academic performance, Machine learning, Feature selection, Student marks.

1. Introduction

One important factor that has a big impact on our society is education. Particularly in the sphere of education, information and communication technology has had a significant influence on numerous study fields. As an example, in reaction to the recent COVID-19 pandemic, some nations have made use of a variety of e-Learning settings [1]. The acknowledgment of the significance of e-learning technology is commendable, particularly in light of the obstacles presented by the COVID-19 pandemic. E-learning holds great promise for expanding educational opportunities and raising service standards. The influence on over 990 million students has been acknowledged by UNESCO, which emphasizes how critical it is to use technology to support remote learning [2]. After the COVID-19 pandemic, education has evolved into a new way to instruct students both online and offline. However, it's critical to uphold the standards in this new educational model and ensure that students attain the same levels of comprehension and academic performance.

The framework in which modern educational institutions operate is complicated and fiercely competitive. Therefore, some of the issues that most colleges now confront are analyzing performance, offering high-quality education, developing ways for evaluating the students' performance, and recognizing future needs. [3]. For many stakeholders, including students, instructors, and academic institutions, predicting students' performance in a particular course or program is crucial. Applications of student performance prediction have shown promise in forecasting dropout rates and at-risk students. It is also utilized to create personalized recommendation systems and early warning systems to enhance the educational experience for students [4]. Various researchers started to use artificial intelligence and machine learning algorithms to predict student performance.

Artificial intelligence has lately emerged as a successful method for assessing and analyzing student performance due to the quick improvements in technology. Machine learning was used by numerous researchers to forecast student achievement [5].

In the study of educational data, particularly in the analysis of student performance data, machine learning has reached unprecedented levels. Because educational institutions rely on information-driven choices, machine learning algorithms are employed to analyze the massive amounts of educational data in order to find patterns and derive valuable insights. A variety of techniques and data, including historical data, the students' prior academic achievement,

attendance, assignments, and socioeconomic characteristics, are used in the machine learning algorithm's examination of student performance data. In many different areas of education, machine learning algorithms are utilized to analyze educational data and produce useful findings that can enhance and develop methodologies for instruction.

2. Related Work

This section appears into how prior researchers used a variety of basic categorisation methods to forecast student academic achievement. In today's rapidly evolving educational landscape, the need for effective assessment methods has never been more critical. Student performance prediction has emerged as a vital tool for enhancing academic outcomes and promoting student satisfaction. By leveraging advanced machine learning algorithms, educators can predict students' grades based on various performance indicators, enabling addressing the students lacking in earlier and improving their future grades.

Numerous researchers have focused on analysing student performance data using effective machine learning methods. The study also addresses the various machine learning approaches used to analyse educational statistics, with the goal of identifying trends that influence academic success and supporting early intervention measures. Notable research cited includes those that use intelligent algorithms for mood classification and tools such as WEKA to analyse academic achievement. The study emphasises machine learning's ability not only to forecast student outcomes, but also to inform personalised educational interventions, thereby improving the whole learning experience [6].

Several studies have examined using of data mining techniques to predict students' academic performance where researchers have implemented algorithms such as decision trees, neural networks, and support vector machines to achieve accurate forecasts of academic outcomes. These investigations highlight the value of early prediction for enabling timely interventions and personalized instructional approaches. Challenges frequently noted include issues with data quality, appropriate feature selection [7].

Another research article uses artificial neural networks (ANNs) to forecast students' final outcomes during online learning. It analyses data from 3,518 university students, taking into account gender, content scores, time spent on content, number of accesses, assignment scores, attendance, and session duration. The ANN model's prediction accuracy of 80.47% [8].

Recent literature studies underscore the increasing application of data mining and learning analytics methodologies to forecast student performance. Decision trees, neural networks, and support vector machines are frequently applied, with decision trees often standing out for their predictive accuracy. Researchers draw on diverse data sources, such as student demographics and behavioral records from educational platforms, to build effective models. Despite promising results, studies point to ongoing issues, including incomplete reporting of sample sizes and subject areas. Nonetheless, the application of these advanced analytical approaches holds significant potential for enabling data-driven interventions and improving educational outcomes [9].

Numerous researchers have analysed the student performance dataset pertaining to the Bachelor's program in Computer Science and Information Technology, focussing on data from 2017 to 2019. The logistic algorithm demonstrated superior predictive accuracy for classifying pass percentages compared to other algorithms [10].

Decision tree analysis has shown to be a strong and easy-to-understand way to guess how well students will do in college. By utilizing various input variables such as previous academic records, demographic information, and behavioral data, decision trees offer clear and actionable insights into factors affecting student success. This approach not only achieves accurate classification of students' academic outcomes but also facilitates early identification of those at risk, enabling timely support and intervention. The transparency and simplicity of decision tree models make them especially valuable for educators and administrators aiming to enhance student achievement and retention [11].

A recent study used a probabilistic neural network to analyse student mark prediction, revealing that the multilayer perceptron algorithm achieved the highest accuracy at 90.1%, while the sensitivity of the probabilistic neural network was the greatest among all algorithms at 76.9% [12]. Another researcher utilised LSTM RNN for analysing student performance prediction and achieved 91% accuracy [13].

3. Methodology

It provides a detailed summary of the methodology used to forecast student achievement, which employs an advanced analytical tool and algorithm to examine student data. The entire

Musik in bayern
ISSN: 0937-583x Volume 90, Issue 8 (Aug -2025)
https://musikinbayern.com DOI https://doi.org/10.15463/gfbm-mib-2025-*437*

procedure is divided into several phases, as shown in Figure. 1. Each part of this revolutionary process is described in detail, providing insight into the intricate steps required.

3.1 Data Collection:

Data collection is a key component in predicting student performance since the quality and relevance of the data influence the accuracy of the prediction models. During this procedure, numerous student-related attributes are acquired, such as college name, age, gender, assignments, attendance, and various test scores.

Figure 1: Proposed Framework for Student Performance Prediction

Table 1 shows the converted attributes. These characteristics were discovered to have a crucial influence on understanding student academic performance. The student performance dataset, comprising various attributes, was collected from students who were taught using the FBT methodology.

Attribute	Description & Conversion
College Name	Women's Christian College -1, SRM-2, Shri Krishnaswamy College -3, Anna Adarsh College -4 , Vel Tech-5 , Patrician College -6 , SDNB Vaishnav College-7, B.S. Abdur Rahman Crescent -8, St. Anne's Arts and Science College-9, Mahalakshmi Women's College -10
Age	Below 20 -1, Above 20 -2
Sex	Male – 1, Female -0
MED	Secondary School -1, Undergraduate -2, Postgraduate -3, Others -4
FED	Secondary School -1, Undergraduate -2, Postgraduate -3, Others -4
Family size	Greater Than 3 -1, Lesser Than 3 -2
Family Support	Yes -1, No-0
Internet	Yes -1, No-0
Travel Time (to college)	Less than 30 min -1,30 min to 1 hour- 2, 1 hour to 2 hour -3
Study Time (per day)	Upto 1 hours -1, Upto 2 hours -2, Upto 3 hours -3

Musik in bayern
ISSN: 0937-583x Volume 90, Issue 8 (Aug -2025)
https://musikinbayern.com DOI https://doi.org/10.15463/gfbm-mib-2025-437

Activities	Sports-1, Clturals-2, Others-3
Arrears	Yes -1, No-0
Higher Education	Yes -1, No-0
Online Source (For Learning)	Yes -1, No-0
Mobile Usage (per day)	Less than 30 min -1,30 min to 1 hour- 2, 1 hour to 2 hour -3
Health	Good-1, Fair-2, Poor-3
Attendance	Least Attendance -1, Below Average Attendance-2, Average Attendance -3, Above Average Attendance-4, Full Attendance -5
Assignments	Least Submission -1, Below Average Submission -2, Average Submission -3, Above Average Submission -4, Full Submission -5.
T1 - T5 (Regular Assessments)	Excellent (80 to 100) -1, Good (60 to 79) -2, Average (40 to 59) -3, Below Average (20 to 39) -4, Poor (0 to 19) -5.

Table 1: Student Performance Dataset Description

3.2 Removal of Missing Values using KNN Imputation

The student performance data is processed using an advanced K-Nearest Neighbors (KNN) imputation technique to accommodate missing values. This KNN-based imputation anticipates missing data points by evaluating the nearest 'k' nearby samples and utilizing a similarity measure like the Euclidean distance to estimate proximity. The number of neighbors (k) and distance measure are important considerations for adjusting the imputation procedure. The end result is a completely populated dataset ready for additional analysis or modelling. The dataset with no missing value is shown in Figure 2.

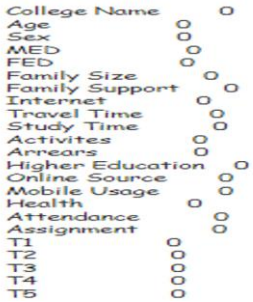

Figure 2: No Missing Values After KNN Imputation

3.3 ZScore

Musik in bayern
ISSN: 0937-583x Volume 90, Issue 8 (Aug -2025)
https://musikinbayern.com DOI https://doi.org/10.15463/gfbm-mib-2025-437

Z-scores, also known as standard scores, indicate how much a data point deviates from a distribution's mean. Positive Z-Scores show values that are higher than the mean, while negative scores suggest values that are lower. In data preprocessing, observations with Z-Scores greater than +3 or less than -3 are frequently identified as outliers and eliminated to reduce their impact on the study. This standardisation not only shifts the data's central location but also normalises its variability, allowing for more precise detection and identification of outliers. In the student performance dataset features like MED, FED , Family size were found to be outliers which is shown in Figure 3.

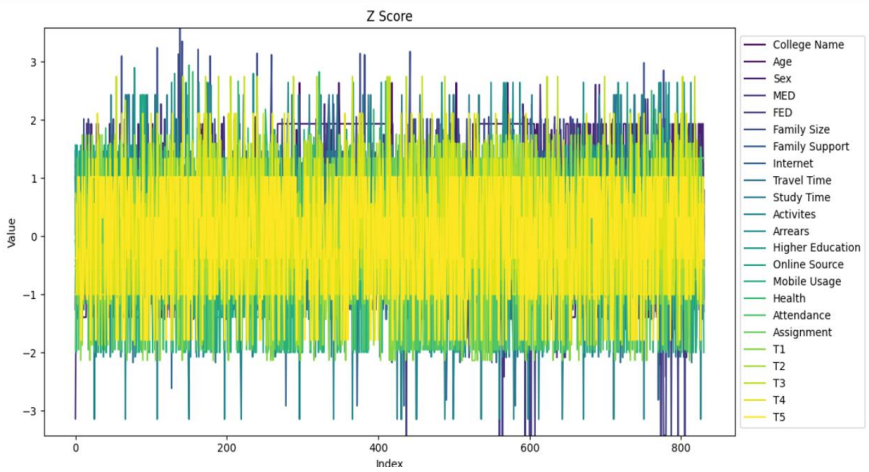

Figure 3: Z-Score Results

3.4 Feature Selection

SelectKBest, paired with F-regression, is used to forecast student performance and identify the most relevant elements affecting academic results. This method evaluates the linear correlations between numerous student qualities and performance, identifying the features with the highest impact on prediction accuracy. Targeted feature selection increases predictive models' precision and efficiency by focussing on the most significant variables. It also lowers interference from less significant features, which improves the model's overall interpretability. In summary, using SelectKBest with F-regression improves the prediction process and allows for better informed decision-making in educational institutions.

3.5 Hyperparameter Tunning

When predicting student performance, Grid Search combined with Cross-Validation (Grid Search CV) is an effective strategy for fine-tuning machine learning model settings, making them more accurate and dependable. It works by experimenting with different parameter

combinations and testing the model's performance on different data splits. This rigorous testing helps to avoid overfitting, ensuring that the model works well not only on previously known data but also on fresh, unseen data. Grid Search CV aids in the development of better-predicting models by altering essential aspects such as learning rate, tree depth, and regularisation.

4. Result and Discussion

This section focusses at several machine learning techniques used to analyse and forecast student performance data. It also compares the accuracy levels attained by various algorithms when used on the student performance dataset.

4.1 XGBRF Regressor

The XGBRF regressor is a powerful combination of the XGBoost algorithm and Random Forest approaches that was specifically built for regression jobs. Unlike the classic XGBoost technique, which creates trees one after the other to correct earlier errors, XGBRF constructs many trees at the same time by randomly selecting features—similar to a Random Forest. This unusual combination combines the strong, varied predictions of Random Forests with the fine-tuned accuracy of gradient boosting. This implies it can detect complicated, nonlinear patterns in student data, making it particularly useful for forecasting academic performance. By delivering accurate and dependable projections, XGBRF assists educators in identifying students who may be struggling early on and tailoring support to fit their needs, thereby enhancing educational results.

Mathematically, XGBRF can be formulated as:

$$\hat{y} = \sum_{m=1}^{M} w_m \cdot T_m(x)$$

where: M is the number of boosting rounds.

w_m is the weight associated with the m^{th} tree.

$T_m(x)$ is the m^{th} decision tree built using Random Forest's method.

4.2 Random Forest

Random Forest (RF) is a supervised machine learning method for forecasting student performance. It is very good in classification tasks like categorising students' academic results. It generates numerous decision trees by randomly selecting subsets of student data and attributes. Each tree makes an individual prediction, and the ultimate conclusion is selected by majority vote, increasing overall prediction accuracy and stability. This ensemble technique enables the model to capture complicated, nonlinear interactions between many student variables, including attendance, study habits, and grades. As a result, Random Forest is a powerful and dependable tool for discovering patterns in student performance and assisting with educational decision-making.

4.3 Gradient Boosting

Gradient boosting is a sophisticated ensemble learning technique that has gained popularity in predicting student academic achievement. It works by iteratively building a sequence of weak predictive models, usually decision trees, with each succeeding model attempting to reduce the residual errors of its predecessors. This sequential learning approach improves overall prediction accuracy by successfully minimising bias and volatility. Gradient boosting models in educational data mining use a variety of student-related characteristics, such as prior academic records, attendance, and engagement indicators, to provide reliable projections of student outcomes like grades or dropout probability. Gradient boosting outperforms typical prediction algorithms in capturing complicated, nonlinear interactions across heterogeneous datasets.

4.4 Support Vector Regressor

Support Vector Regression (SVR) is a sophisticated supervised learning technique that is often used to predict student performance. SVR seeks to identify a function that approximates the link between input variables (such as academic records and behavioural data) and continuous desired outcomes while keeping model complexity within a defined tolerance. SVR's capacity to handle nonlinear relationships through kernel functions and preserve robustness against overfitting makes it particularly ideal for educational data, which frequently exhibits complicated patterns.

4.5 Logistic Regression

Musik in bayern
ISSN: 0937-583x Volume 90, Issue 8 (Aug -2025)
https://musikinbayern.com DOI https://doi.org/10.15463/gfbm-mib-2025-*437*

The logistic function converts a linear combination of input features to a probability value. This probability function calculates the likelihood that a student belongs to a specific performance category. The model's interpretability and efficiency make it an appropriate choice for educational contexts, facilitating informed decision-making. By linking student data to outcome probabilities, logistic regression enables proactive interventions to improve academic success.

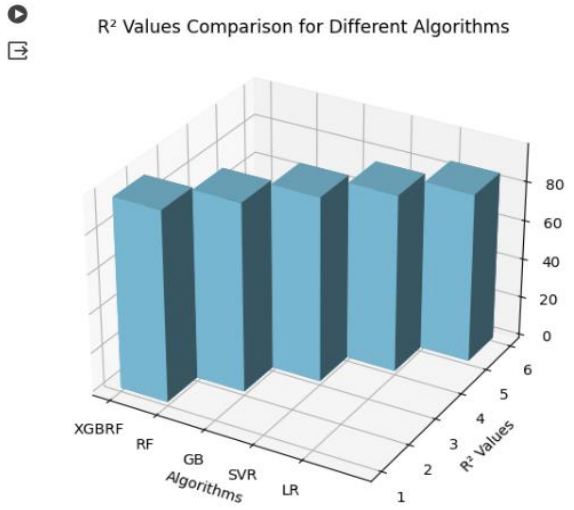

Figure – 4 : R2 Values for Different Algorithms

Regressor	MAE	RMSE	R-Squared
XGBRF	0.107	0.134	98.2
RF	0.154	0.193	96.3
GB	0.196	0.245	94
SVR	0.250	0.313	90.2
LR	0.299	0.374	86

Table – 2: Results of MAE, RMSE and R^2 of Various Algorithms

A comparison investigation of the various regression algorithms shown in Figure 4 used to predict student performance demonstrates a clear hierarchy in model effectiveness, with the The Table 2 clearly shows the MAE , RMSE , R-Squared value of all the algorithms used in prediction of student performance .XGBRF regressor appearing as the best option. With the lowest Mean Absolute Error (MAE) of 0.107 and Root Mean Square Error (RMSE) of 0.134, as well as the highest R-squared value of 98.2%, the XGBRF model demonstrates extraordinary precision and robustness in capturing the underlying trends of student academic data. These outstanding performance measures emphasise its capacity to reduce prediction errors and

explain the great majority of the diversity in student outcomes, making it an extremely accurate predictor of academic success. The Random Forest (RF) model has significant predictive potential, with an MAE of 0.154, RMSE of 0.193, and R-squared of 96.3%. This demonstrates that RF, while slightly less accurate than XGBRF, is still quite successful in modelling complicated, nonlinear interactions among student-related data. Its ensemble structure, which includes many decision trees, allows it to generalise effectively and provide consistent predictions across a wide range of student characteristics.

Gradient Boosting (GB), Support Vector Regression (SVR), and Logistic Regression (LR) models, on the other hand, exhibit a progressive drop in both accuracy and explanatory power with time. GB performs reasonably well, with an MAE of 0.196, RMSE of 0.245, and R-squared of 94%, although the ensemble-based approaches exceed it. SVR, a more traditional technique, with an MAE of 0.250, RMSE of 0.313, and R-squared of 90.2%, showing lesser precision and less ability to capture complexity than tree-based models. Logistic Regression (LR) has the highest error measurements (MAE 0.299, RMSE 0.374) and the lowest R-squared value of 86%, indicating its limits in dealing with the nonlinear, multidimensional patterns common in educational datasets.

5. Conclusion

This study demonstrates the tremendous potential of XGBRF to accurately predict student achievement. These models stand out for their ability to manage a wide range of complicated and diverse student data, including academic records, attendance, and behavioural characteristics. This allows them to detect subtle patterns and associations that simpler models may overlook, making them extremely useful tools for educational analytics where understanding the complexities of student performance is critical. By making precise and trustworthy predictions, these algorithms help educators and organisations identify children who may want more assistance. This enables the deployment of specific interventions suited to individual student needs, which ultimately improves academic performance. Adoption of these advanced models paves the opening for more personalised, data-driven education, hence creating learning environments that promote student achievement and informed decision-making.

Reference

[1] Giannakas, F., Troussas, C., Voyiatzis, I., & Sgouropoulou, C. (2021). A deep learning classification framework for early prediction of team-based academic performance. *Applied Soft Computing*, *106*, 107355.

[2] Dhawan, S. (2020). Online learning: A panacea in the time of COVID-19 crisis. *Journal of educational technology systems*, *49*(1), 5-22.

[3] Albreiki, B., Zaki, N., & Alashwal, H. (2021). A systematic literature review of student'performance prediction using machine learning techniques. *Education Sciences*, *11*(9), 552.

[4] Alamri, R., & Alharbi, B. (2021). Explainable student performance prediction models: a systematic review. *IEEE Access*, *9*, 33132-33143.

[5] Alshabandar, R., Hussain, A., Keight, R., & Khan, W. (2020, July). Students performance prediction in online courses using machine learning algorithms. In *2020 International Joint Conference on Neural Networks (IJCNN)* (pp. 1-7). IEEE.

[6] Sekeroglu, B., Dimililer, K., & Tuncal, K. (2019, March). Student performance prediction and classification using machine learning algorithms. In *Proceedings of the 2019 8th international conference on educational and information technology* (pp. 7-11).

[7] López-Zambrano, J., Torralbo, J. A. L., & Romero, C. (2021). Early prediction of student learning performance through data mining: A systematic review. Psicothema, 33(3), 456.

[8] Aydoğdu, Ş. (2020). Predicting student final performance using artificial neural networks in online learning environments. *Education and Information Technologies*, *25*(3), 1913-1927.

[9] Namoun, A., & Alshanqiti, A. (2020). Predicting student performance using data mining and learning analytics techniques: A systematic literature review. Applied Sciences, 11(1), 237.

[10] Hashim, A. S., Awadh, W. A., & Hamoud, A. K. (2020, November). Student performance prediction model based on supervised machine learning algorithms. In IOP conference series: materials science and engineering (Vol. 928, No. 3, p. 032019). IOP Publishing.

[11] Hamoud, A., Hashim, A. S., & Awadh, W. A. (2018). Predicting student performance in higher education institutions using decision tree analysis. International Journal of Interactive Multimedia and Artificial Intelligence, 5, 26-31.

[12] Mason, C., Twomey, J., Wright, D., & Whitman, L. (2018). Predicting engineering student attrition risk using a probabilistic neural network and comparing results with a backpropagation neural network and logistic regression. Research in Higher Education, 59, 382-400.

ISSN: 0937-583x Volume 90, Issue 8 (Aug -2025)
https://musikinbayern.com DOI https://doi.org/10.15463/gfbm-mib-2025-*437*

[13] Kukkar, A., Sharma, A., Singh, P. K., & Kumar, Y. (2023). Predicting Students Final Academic Performance Using Deep Learning Techniques. In IoT, Big Data and AI for Improving Quality of Everyday Life: Present and Future Challenges: IOT, Data Science and Artificial Intelligence Technologies (pp. 219-241). Cham: Springer International Publishing.