Predictive Analysis of Student Performance using Machine Learning Models

Aarti Rathod[1], Neha Vora[2]

[1]Student, [2]Assistant Professor, SVKM'S Usha Pravin Gandhi College of Arts, Science and Commerce, Mumbai, India.

E-mail: [1]rathodaarti331@gmail.com, [2]nehavora2501@gmail.com

Abstract

In order to recognize the students who are not performing well, it is of great significance to predict student performance with the highest accuracy possible. In this study, the functions of machine learning techniques, such as Random Forest, Gradient Boosting, XGBoost, and Support Vector Classifier are employed in the prediction of student outcomes depending on studied hours, attendance, activities, and parental education level. Accordingly, after the dataset was pre-processed and the models were assessed using the machine learning models, the Gradient Boosting gave the best accuracy output measuring 97% while the Random Forest was remarkably close, producing almost identical results in terms of accuracy. The effectiveness of drawing on data to identify students who are at high risk of dropping out of an institution is brought out in this study.

Keywords: Student Performance, Machine Learning, Random Forest, Gradient Boosting, XG Boost, Support Vector Classifier, Educational Data Mining.

1. Introduction

Educational institutions are facing challenges to improve student success rates by identifying key factors that impact the academic performance of students. However, generally used analysis approaches may not provide accurate results since they do not consider the multifactorial nature of the problem and/or multiplicity and interactions among defined factors such as study habits, attendance behavior's, family backgrounds etc.

DOI: https://doi.org/10.36548/jitdw.2025.1.002
Received: 08-02-2025. received in revised form: 04-03-2025, accepted: 21-03-2025, published: 03-04-2025

The implementation of machine learning models allows an effective approach when a large number of features need to be considered for prediction problems. Moreover, intervention needs can be detected with high accuracy in order to prevent failures within a given student population in a timely manner. In this research machine learning models to predict whether a student will pass or fail using several features obtained from a student performance dataset is compared with each other. The main aim of the study is to discover knowledge from the student data that helps to predict student academic performance using a classification algorithm in this perspective, the key objective of the research is to find a model that yields the maximum performance with respect to accuracy, precision, recall, and F1- score. Moreover, the importance of several features has been analyzed in order to understand which variables play a critical role in predicting student success[9-12]. The insights from this study can help education institutions make evidence-based interventions, manage student performance more effectively, and optimize resource utilization. This research not only adds to the growing field of educational data mining but also carries practical overtures for educators and administrators. Predictive models deployed in schools and universities offer options for personalized support, thereby improving overall academic success rates. Focusing on the early identification of students at risk of failure, an institution can certainly make its, educational climate more supportive for students to overcome obstacles and succeed in their pursuit of academics. Thus, by identifying if a student will pass or fail based on certain criteria, predictive analytics can support timely interventions by institutions. For example, a student with low attendance, minimum study hours, or who remains disconnected from extracurricular activities may raise a red flag for extra support, counseling, or academic resources. By referencing the algorithms developed from historical data, machine learning can highlight, from a variety of input variables, unseen patterns and relationships to present educators with informed decisions based on real-time forecasts[13-15].

2. Related Work

Machine learning approaches used for the prediction of student performance have gained a lot of attention over the last few years because they are expected to help educational institutions identify students who can fail. Many studies have attempted to explore various models and approaches toward proper prediction and accurate determination of student outcomes based on academic and non-academic features.

Yadav et al. [1] discussed the application of algorithms like Artificial Neural Networks (ANNs), Naïve Bayes, Decision Trees, and SVM for predicting student performance. Highlighted these algorithms' utility in detecting underperforming students and proposed a framework using rule-based recommender systems, considering demographic, academic, and psychological factors. Utilized models such as Decision Trees, Random Forest, and Naïve Bayes to enhance prediction accuracy by integrating diverse features.

Priya et al. [2] tested models including Random Forest and Gradient Boosting for predicting student performance. Recommended adding more features like socioeconomic and psychological variables to enhance model performance. Esmael et al. [3] utilized decision tree algorithms to extract patterns and predict student outcomes from academic records. Demonstrated that the approach is feasible even in smaller datasets for educational environments.

Fiseha et al. [4] emphasized on the strength of Random Forest and Gradient Boosting in achieving high accuracy using small datasets. Lubna et al. [5] applied educational data mining techniques using Random Forest and XGBoost to predict student performance based on historical data for accurate results. Chen et al. [6] Classified algorithms such as SVM, Naïve Bayes, and Decision Trees for predicting student performance. Emphasized the significance of variables like prior academic performance, attendance, and extracurricular participation in improving the accuracy of the model. Hayder et al. [7] presented a literature review on the application of data mining techniques for student performance prediction and recommended the addition of psychological and family background features for better precision predictions. Shahiri et al. [8] in his experiments demonstrated the effectiveness of Random Forest, Gradient Boosting, and Decision Tree algorithms in student performance prediction. They further stress in incorporating additional attributes other than academic data, for enhanced precision prediction.

3. Methodology

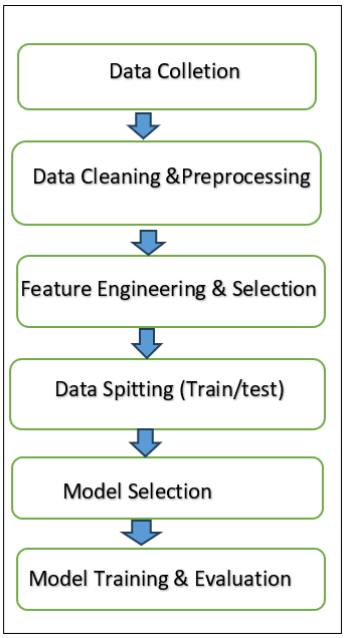

Figure 1. Flowchart of the Proposed Student Performance Analysis

3.1 Dataset Preparation

The methodology of the research, which was performed by applying and comparing different machine learning algorithms to predict student performance based on various academic and non- academic features, has been presented in Figure 1 . The dataset used in this research is a primary dataset; it was collected by using a Google form. The dataset includes 174 students with a target variable representing pass or failure outcomes of each student from the previous academic year. These features can be listed as Student ID,Study hours per week, Attendance Rate, Previous Grades, Extracurricular Activities, Participation, and Parental Education. The overview of the dataset is depicted in Table 1.

Table 1. The Overview of the Dataset

Variable	Description	Data Type	Example Values
Student_ID	Unique identifier for a student	String/Integer	S12345, 1001
Study_Hours_Per_Week	Average number of hours a student studies weekly	Integer/Float	10, 15.5
Attendance_Rate	Percentage of classes attended	Float (0-100)	85.5, 92.0
Previous_Grades	Student's past academic performance (GPA or marks)	Float/Integer	3.8, 85
Extracurricular_Activities	Participation in activities (sports, clubs, etc.)	Boolean/String	Yes/No
Participation	Level of engagement in class discussions, projects	Integer (Scale 1-5)	3, 4
Parental_Education	Highest education level of parents	String	High School, Bachelor's Ph.D, Other

3.2 Data Preprocessing

Preprocessing is a very important process to make the model understand the data in a better way. First, row which didn't contribute anything in terms of value prediction were dropped, such as the "Student ID" field, which has nothing much to do with prediction regarding the target variable. Numerical features like study hours, attendance rate, and previous grades were cleaned by filling in the missing values for each column with its median value. Then, categorical features such as extracurricular activities or parental education were converted to numerical data. An example could be "Yes" and "No" responses in extra-curricular participation encoded as 1 and 0, respectively. The parental education was one-hot encoded to make the categories interpretable by each of the machine learning models. One-hot encoding for parental education level since it is a categorical feature with multiple classes (e.g., High School, Bachelor's, PhD). Machine learning models need numerical inputs, and one-hot encoding converts each category into an individual binary column (0 or 1), which becomes simpler for the model to understand.

3.2.1 Feature Engineering

It converts raw data into useful inputs for machine learning models by

- Removal of non-informative columns: 'Student ID' row was removed as it does not help in prediction.
- Handling missing values: Median imputation was applied for numerical features, which include study hours, attendance, and previous grades, respectively.
- Conversion of categorical features
- One-hot encoding was used to convert the feature about parental education levels

3.2.2 Feature Selection

Feature selection plays an important role in the improvement of model performance, as it reduces noise and computational complexity. The features selected were mainly derived from their predictive power and contribution to model accuracy. Some of the features selected include study hours, attendance rate, past academic grades, participation in extracurricular activities, and parents' education level.

3.2.3 Feature Importance

Random Forest and Gradient Boosting showed study hours and past academic grades to be the most influencing features, followed by an attendance rate and level of parental education. Extracurricular participation had only a moderate effect but was positively directed at increasing performance. The focus on such essential features improved their accuracy during the prediction of student performance.

3.3 Experimental Setup

The experiment was done in Python 3. The environment that was used to implement the whole work was Google Colab with Pandas, NumPy, and Scikit-learn (sklearn), which are very important libraries in the processing of data and other machine learning functions.

3.4 Model Used

Random Forest Classifier - Random Forest is an ensemble learning model that creates multiple decision trees and combines their outputs to improve accuracy. It works by averaging

the predictions from different trees, which reduces overfitting and increases robustness. In the context of student performance prediction, it helps analyze multiple factors such as study hours, attendance, and parental education to classify students into pass or fail categories. This model performed very well, with a 97% accuracy, making it one of the most reliable classifiers for predicting student outcomes.

Gradient Boosting Classifier - Gradient Boosting is another ensemble technique that builds decision trees sequentially, where each tree corrects the errors of the previous one. It minimizes bias and variance, making it a powerful tool for structured data. During the experiment, the Gradient Boosting model proved well-suited for identifying complex patterns in student performance by correlating extracurricular activities with subject performance. With an accuracy of 97 percent, it offered better prediction properties than the competing models, as its learning process was enhanced at each round.

XGBoost classifier - XGBoost, an improved gradient boosting model designed for efficient computation and optimal regularization, is particularly useful for large datasets and prevents overfitting through its built-in mechanisms. In this study, XGBoost provided fast and accurate predictions of student performance, considering factors such as attendance and past grades. While achieving a slightly lower accuracy (94%) compared to Random Forest and Gradient Boosting, its speed and efficiency make it a strong candidate for real-time applications in educational analytics

Support Vector Classifier (SVC)- Support Vector Classifier (SVC) works by constructing a hyperplane that best separates data points into different classes. It is particularly effective in high-dimensional spaces and works well with smaller datasets. However, in this study, SVC underperformed compared to tree-based models, achieving only 80% accuracy. This could be due to the model's sensitivity to kernel choice and parameter tuning. Although SVC can be used for classification tasks, the current dataset is not the best suited for student performance prediction. Table 2 presents the details of the hyperparameters used.

Table 2. Hyperparameter Used

Model	Hyperparameter	Tuned Values Tested	Optimal Value
Random Forest Classifier	n_estimators	50, 100, 200	200

	max_depth	3, 5, 7, None	None (Full depth)
	min_samples_split	2, 5, 10	2
	min_samples_leaf	1, 2, 4	1
AdaBoost Classifier	n_estimators	50, 100, 200	200
	learning_rate	0.01, 0.1, 1.0	0.1
Gradient Boosting Classifier	learning_rate	0.01, 0.1, 0.2	0.1
	n_estimators	50, 100, 200	200
	max_depth	3, 5, 7	3
XGBoost Classifier	learning_rate	0.01, 0.1, 0.2	0.2
	n_estimators	50, 100, 200	200
	max_depth	3, 5, 7	3
	subsample	0.6, 0.8, 1.0	0.6
Support Vector Classifier	kernel	Linear, RBF, Polynomial	RBF
	C	0.1, 1, 10	10
	gamma	scale, auto	scale

3.5 Model Evaluation Methods and Metrics

The following are the methods and metrics that are used in evaluating the machine learning model**:**

3.5.1 Methods

- The dataset was divided into 80% training and 20% testing for evaluating performance on unseen data by any model.
- GridSearchCV was applied for hyperparameter tuning with 5-fold cross-validation to determine the optimal model parameters.
- Models were fine-tuned through the adjustment of key hyperparameters to enhance the accuracy and prevent overfitting.

3.5.2 Performance Metrics

- **Accuracy**: Measures of the percentage of students correctly classified as pass/fail.

- **Precision**- It is the ratio of correctly predicted positive observations to the total predicted positives. This shows the proportion of predicted positive instances that were actually positive.

- **Recall**- Recall, or sensitivity or True positive rate, is the ratio of correctly predicted positive observations to all the observations in the actual class. It gives the model's capability of observing all the actual positives.

- **F1 score**- The F1 score is the measure of a harmonic mean of precision and recall.

This gives a balance between precision and recall, and it happens to be particularly useful in a case where the class distribution is imbalanced

4. Results and Discussion

Random Forest and Gradient Boosting models achieved high accuracy, both around 97%. XGBoost followed closely with 94%, while the Support Vector Classifier (SVC) performed the poorest at 80%. These results indicate that tree-based ensemble models are best suited for predicting student performance with this dataset. Specifically, Gradient Boosting and Random Forest demonstrated high precision and recall, accurately classifying students as passed or not passed. Despite its general popularity, SVC struggled to generalize on this particular data, likely due to its sensitivity to kernel and hyperparameter selection. Figures 2 to 6 and Table 3 illustrates the performance of the machine learning models.

Model: RandomForestClassifier

	precision	recall	f1-score	support
0	1.00	0.91	0.95	11
1	0.96	1.00	0.98	24
accuracy			0.97	35
macro avg	0.98	0.95	0.97	35
weighted avg	0.97	0.97	0.97	35

Accuracy: 0.9714285714285714

Figure 2. Performance of Random Forest Classifier

Model: GradientBoostingClassifier

	precision	recall	f1-score	support
0	1.00	0.91	0.95	11

1	0.96	1.00	0.98	24
accuracy			0.97	35
macro avg	0.98	0.95	0.97	35
weighted avg	0.97	0.97	0.97	35

Accuracy: 0.9714285714285714

Figure 3. Performance of GradientBoost Classifier

Model: XGBClassifier

	precision	recall	f1-score	support
0	1.00	0.82	0.90	11
1	0.92	1.00	0.96	24
accuracy			0.94	35
macro avg	0.96	0.91	0.93	35
weighted avg	0.95	0.94	0.94	35

Accuracy: 0.9428571428571428

Figure 4. Performance of XGBoost Classifier

Model: SVC

	precision	recall	f1-score	support
0	0.75	0.55	0.63	11
1	0.81	0.92	0.86	24
accuracy			0.80	35
macro avg	0.78	0.73	0.75	35
weighted avg	0.79	0.80	0.79	35

Accuracy: 0.8

Figure 5. Performance of SVC

AdaBoostClassifier:
Accuracy: 0.8285714285714286

	precision	recall	f1-score	support
0	0.86	0.55	0.67	11
1	0.82	0.96	0.88	24
accuracy			0.83	35
macro avg	0.84	0.75	0.78	35
weighted avg	0.83	0.83	0.82	35

Figure 6. Performance of AdaBoost Classifier

Table 3. Comparison of Model Performance

Model	Accuracy	Precision	Recall	F1-Score
Random Forest Classifier	97%	0.98	0.97	0.97
Gradient Boosting Classifier	97%	0.98	0.97	0.97
XGBoost Classifier	94%	0.96	0.94	0.94
Support Vector Classifier	80%	0.79	0.8	0.79

5. Conclusion

This study demonstrates the efficiency of various machine learning models in predicting student performance based on factors such as study hours, attendance, previous grades, extracurricular activities, and parental education. The Random Forest and Gradient Boosting models exhibited high accuracy, precision, recall, and F1 score, indicating their reliability in accurately identifying students who require additional support. These models can effectively assist educational institutions in making proactive decisions and implementing timely interventions to improve student outcomes. However, this study has limitations. The dataset, comprising only 174 entries, may not adequately represent the diverse student population across different educational settings. Furthermore, the variables considered were limited to academic and participation factors, neglecting other influential factors such as psychological, social, and economic conditions. The models were evaluated on a static dataset, potentially failing to capture the dynamic nature of student behavior. Future research should address these limitations by utilizing larger, more diverse datasets to enhance model generalizability. Incorporating features like socioeconomic status, mental health indicators, and learning styles can provide a more comprehensive understanding of student performance. Longitudinal studies could also be conducted to track changes in student behavior and performance over time, enabling the development of more dynamic, real-time predictive models. Additionally, future investigations will explore the implementation of these models within educational institutions as early warning systems, providing educators with improved tools for identifying and supporting at-risk students.

References

[1] Yadav, Nitin Ramrao, and Sonal Sachin Deshmukh. "Prediction of Student Performance Using Machine Learning Techniques: A Review." In International Conference on Applications of Machine Intelligence and Data Analytics (ICAMIDA 2022), pp. 735-741. Atlantis Press, 2023.

[2] Priya, S., T. Ankit, and D. Divyansh. "Student performance prediction using machine learning." In Advances in parallel computing technologies and applications, pp. 167-174. IOS Press, 2021.

[3] Esmael Ahmed- Student Performance Prediction Using Machine Learning Algorithms-Information System, College of Informatics, Wollo University, Dessie 7200, Ethiopia

[4] Fiseha Berhanu & Addisalem Abera (MSC) College of Engineering & Technology Lecturer at Computer Science Department Dilla University, Dilla, Ethiopia, International Journal of Computer Applications, December 2015

[5] Lubna Mahmoud Abu Zohair-Prediction of Student's performance by modelling small dataset size, Abu Zohair International Journal of Educational Technology in Higher Education (2019)

[6] Chen, Ziling, Gang Cen, Ying Wei, and Zifei Li. "Student performance prediction approach based on educational data mining." IEEE Access 11 (2023): 131260-131272.

[7] Hayder, Alabbas. "Predicting student performance using machine learning: A comparative study between classification algorithms." (2022)

[8] .Shahiri, Amirah Mohamed, Wahidah Husain, and Nur'aini Abdul Rashid. "A review on predicting student's performance using data mining techniques." procedia computer science 72 (2015): 414-422.

[9] Nguyen Thai-Nghe, et al. "Matrix and Tensor Factorization for Predicting Student Performance." International Conference on Computer Supported Education 2 (2011): 69-78

[10] Khan, Anupam, and Soumya K. Ghosh. "Student performance analysis and prediction in classroom learning: A review of educational data mining studies." Education and information technologies 26, no. 1 (2021): 205-240.

[11] Pandey, Mrinal, and Vivek Kumar Sharma. "A decision tree algorithm pertaining to the student performance analysis and prediction." International Journal of Computer Applications 61, no. 13 (2013): 1-5.

[12] Leelaluk, Sukrit, et al. "Predicting Student Performance Based on Lecture Materials Data Using Neural Network Models." Proceedings of the 4th Workshop on Predicting Performance Based on the Log Data (2022): 11-15

[13] Hu, Qian, and Huzefa Rangwala. "Academic Performance Estimation with Attention-Based Graph Convolutional Networks." arXiv preprint arXiv:2001.00632 (2019)

[14] Li, Haotian, Huan Wei, Yong Wang, Yangqiu Song, and Huamin Qu. "Peer-inspired student performance prediction in interactive online question pools with graph neural network." In Proceedings of the 29th ACM international conference on information & knowledge management, pp. 2589-2596. 2020.

[15] Wang, Yinkai, et al. "Graph-Based Ensemble Machine Learning for Student Performance Prediction." arXiv preprint arXiv:2112.07893 (2021).