International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 10 Issue: 04 | April - 2026 SJIF Rating: 8.659 ISSN: 2582-3930

Student Performance Prediction Using Machine Learning

Rakhi Meshram[1], Bhakti kale[2], Mayuri Kapadnis[3] , Nayan Patil[4]

[1]*Department of Artificial Intelligence & Data Science, AISSMS Institute of Information Technology, Pune, India*
[1]*Department of Artificial Intelligence & Data Science, AISSMS Institute of Information Technology, Pune, India*
[1]*Department of Artificial Intelligence & Data Science, AISSMS Institute of Information Technology, Pune, India*
[1]*Department of Artificial Intelligence & Data Science, AISSMS Institute of Information Technology, Pune, India*

Abstract - Student academic performance prediction has become an important application of machine learning in the field of educational data mining. Educational institutions collect large amounts of student data such as attendance records, assignment scores, previous academic results, and study behavior. Analyzing this data can help identify patterns that influence student success. This study proposes a machine learning based approach to predict student academic performance using historical academic and behavioral data. The dataset is first preprocessed by handling missing values, cleaning data, and transforming features. Exploratory Data Analysis (EDA) is performed to understand relationships between different variables. Machine learning algorithms such as Random Forest, Support Vector Machine (SVM), and XGBoost are applied to build predictive models. The trained models are evaluated using performance metrics such as accuracy and precision. The results show that machine learning techniques can effectively predict student performance and help identify weak students at an early stage. This system can assist educators in providing timely support and improving overall student learning outcomes.

Key Words: Machine Learning, Student Performance Prediction, Educational Data Mining, Random Forest, SVM, XGBoost.

1.INTRODUCTION

In recent years, the use of data analytics and machine learning in the education sector has increased significantly. Educational institutions generate large amounts of data related to student attendance, assignments, examination results, and classroom participation. Analyzing this data can provide useful insights into student learning behavior and academic performance.

Predicting student performance is an important task for educators because it helps identify students who may face academic difficulties. Early prediction allows teachers and institutions to provide additional support such as mentoring, extra classes, and personalized learning strategies. This can help improve overall student performance and reduce failure or dropout rates.

Machine Learning (ML) techniques enable computers to learn patterns from historical data and make predictions about future outcomes. By applying machine learning algorithms to student data, it is possible to develop models that can accurately predict student academic performance. In this study, a machine learning based approach is proposed to predict student performance using features such as attendance, assignment marks, previous academic scores, and study hours. Algorithms such as Random Forest, Support Vector Machine (SVM), and XGBoost are used to build predictive models. The objective of this system is to help educators identify weak students early and take necessary actions to improve their academic outcomes.

2. BODY OF PAPER

2.1 Dataset Description

The dataset used in this study contains various features related to student academic performance and learning behavior. These features help in identifying patterns and trends that influence student success. The dataset includes attributes such as attendance percentage, assignment marks, previous semester grades, study hours, and participation in class activities.

In addition to these primary features, the dataset may also include other relevant factors such as internal assessment scores, class participation, and consistency in academic performance over time. These attributes provide a comprehensive view of a student's academic profile and help improve the accuracy of the prediction model.

The dataset is collected from academic records and is structured in tabular format, where each row represents an individual student and each column represents a specific feature. The target variable in the dataset indicates the final outcome of the student, such as "Pass" or "Fail", or performance categories like "Good", "Average", or "Poor".

The collected data is used to train machine learning models that can predict whether a student is likely to perform well or poorly in examinations. Before applying machine learning algorithms, proper data preprocessing is performed, including handling missing values, removing inconsistencies, and transforming categorical data into numerical form. These steps ensure that the dataset is

clean, consistent, and suitable for building accurate and reliable prediction models.

2.2 Data Preprocessing

Data preprocessing is a crucial step in building an effective machine learning model. Raw data collected from real-world sources often contains missing values, noise, and inconsistencies that can negatively impact model performance. Therefore, preprocessing is required to clean and transform the data into a suitable format for analysis and model training.

In this project, the following preprocessing techniques are applied:

- **Removal of missing or null values:** Missing data is handled by either removing incomplete records or replacing missing values using appropriate methods such as mean or median imputation.
- **Conversion of categorical variables into numerical values:** Since machine learning algorithms require numerical input, categorical features such as labels or categories are converted into numerical form using encoding techniques.
- **Data normalization and scaling:** The dataset is normalized to ensure that all features are on a similar scale. This helps improve the performance of algorithms that are sensitive to feature magnitude, such as Support Vector Machine.
- **Splitting the dataset into training and testing sets:** The dataset is divided into training and testing subsets. The training data is used to build the model, while the testing data is used to evaluate its performance on unseen data.

These preprocessing steps ensure that the dataset is clean, consistent, and suitable for training machine learning models. Proper preprocessing improves model accuracy and helps in achieving reliable and meaningful predictions.

2.3 Exploratory Data Analysis

Exploratory Data Analysis (EDA) is performed to understand the structure and distribution of the dataset. Visualization techniques such as bar charts, histograms, and correlation heatmaps are used to analyze the relationships between different variables.

EDA helps in identifying important features that strongly influence student performance. For example, attendance and previous academic scores often show a strong correlation with student success.

2.4 Machine Learning Model

In this project, several machine learning algorithms are used to predict student academic performance. These algorithms are selected based on their ability to handle classification problems and provide accurate predictions using structured data.

Random Forest:
Random Forest is an ensemble learning algorithm that builds multiple decision trees during training and combines their outputs to improve prediction accuracy. Each tree is trained on a random subset of the data, which helps reduce overfitting and increases model robustness. It is particularly effective for handling large datasets with multiple features and can also provide insights into feature importance, helping identify which factors most influence student performance.

Support Vector Machine (SVM):
Support Vector Machine is a supervised machine learning algorithm used for classification tasks. It works by finding the optimal hyperplane that separates different classes in the dataset with the maximum margin. SVM is effective in high-dimensional spaces and performs well when there is a clear margin of separation between classes. It is also useful in cases where the dataset is not linearly separable by using kernel functions.

XGBoost:
XGBoost (Extreme Gradient Boosting) is an advanced boosting algorithm that builds models sequentially, where each new model focuses on correcting the errors of the previous ones. It uses gradient boosting techniques and includes regularization to prevent overfitting. XGBoost is known for its high performance, speed, and efficiency, making it one of the most popular algorithms for structured data problems. It often provides better accuracy compared to traditional models.

Table -1: Dataset Structure

Student ID	Attendance (%)	Assignment Score	Previous Marks	Study Hours	Final Result
101	85	78	82	4	Pass
102	65	60	58	2	Fail
103	90	88	91	5	Pass
104	70	66	64	3	Pass

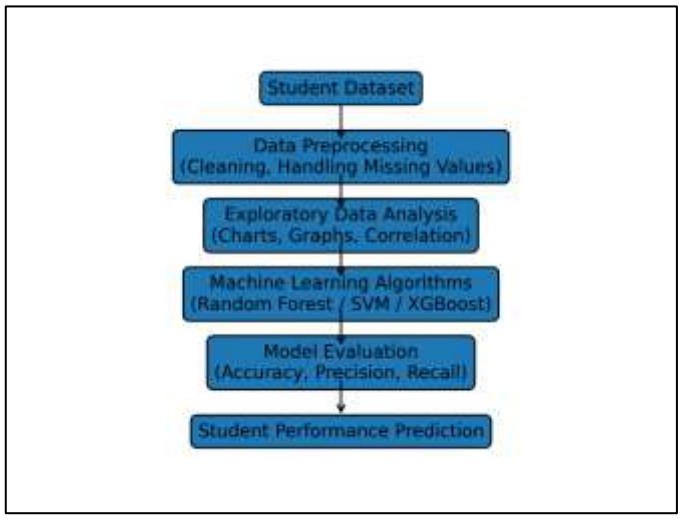

Fig -1: Workflow of the Student Performance Prediction System

Fig-2: Dashboard visualization of student performance

Algorithm	Accuracy (%)	Precision	Recall
Random Forest	89%	0.88	0.87
Support Vector Machine (SVM)	84%	0.83	0.82
XGBoost	91%	0.90	0.89

Fig-3: Accuracy Comparison of Machine Learning Models

3. CONCLUSIONS

In this study, a machine learning based approach was used to predict student academic performance using historical academic and behavioral data. Various features such as attendance, assignment scores, previous marks, and study hours were analyzed to understand their impact on student performance.

The dataset was preprocessed and analyzed using Exploratory Data Analysis techniques to identify patterns and relationships between different variables. Machine learning algorithms such as Random Forest, Support Vector Machine (SVM), and XGBoost were applied to build predictive models.

The results show that machine learning techniques can effectively predict student performance and identify students who may require additional academic support. Early identification of weak students can help educators take necessary actions such as providing extra guidance, mentoring, or personalized learning support.

This system can assist educational institutions in improving student success rates and making data-driven decisions to enhance the learning process.

ACKNOWLEDGEMENT

The authors would like to express their sincere gratitude to the Principal and faculty members of *AISSMS Institute of Information Technology, Pune,* for providing the necessary support and academic environment to carry out this work. We would also like to thank the *Department of Artificial Intelligence and Data Science* for their guidance, encouragement, and valuable insights throughout the development of this project.

REFERENCES

1. Romero, C., & Ventura, S. (2010). Educational Data Mining: A Review of the State of the Art. *IEEE Transactions on Systems, Man, and Cybernetics*, 40(6), 601-618.

2 Cortez, P., & Silva, A. (2008). Using Data Mining to Predict Secondary School Student Performance. *Proceedings of the 5th Future Business Technology Conference*, Portugal.

3 Kotsiantis, S. (2012). Use of Machine Learning Techniques for Educational Purposes: A Decision Support System for Forecasting Students' Grades. *Artificial Intelligence Review*, 37(4), 331-344.

4 Baker, R. S., & Yacef, K. (2009). The State of Educational Data Mining in 2009: A Review and Future Visions. *Journal of Educational Data Mining*, 1(1), 3-17.

5 Han, J., Kamber, M., & Pei, J. (2011). *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publishers.

6 Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1), 5-32.

7 Cortes, C., & Vapnik, V. (1995). Support Vector Networks. *Machine Learning*, 20(3), 273-297.

8 Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785-794.

9 Witten, I. H., Frank, E., & Hall, M. A. (2016). *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann.

10 Al-Barrak, M. A., & Al-Razgan, M. (2016). Predicting Students Final GPA Using Decision Trees: A Case Study. *International Journal of Information and Education Technology*, 6(7), 528-533.