

Student Performance Prediction System: Predicting Student Academic Performance Using Machine Learning

Sujal Gunjkar, Vedant Zape, Vaibhav Singewad , Prof.S.P.Jagjivan, Prof.N.S.Mohokar

Students, Department of Computer Science and Engineering, Sanmati Engineering College, Washim, India

Professor, Department of Computer Science and Engineering, Sanmati Engineering College, Washim, India

Abstract

In recent years, educational institutions have increasingly focused on leveraging data analytics to better understand student performance and enhance academic outcomes(Ravikumar et al., 2023). Institutions routinely collect vast amounts of data—including attendance records, exam marks, and assignment scores—yet this information is often underutilized for predictive or decision-making purposes(Gaftandzhieva et al., 2023). Consequently, there is a pressing need for intelligent systems that can analyze such data to deliver actionable insights(Li & Liu, 2025).

This project introduces a Student Performance Prediction System powered by machine learning techniques, implemented in Python with a user-friendly web interface built using Streamlit. The core functionality accepts key inputs, such as attendance, study hours, prior marks, and assignment performance, to forecast a student's likely academic outcomes.

A key challenge during development was selecting the most influential features affecting student performance. Initial models produced suboptimal accuracy, which improved markedly after iterative feature engineering, data cleaning, and experimentation. Various algorithms—including Decision Trees, Logistic Regression, and Random Forest—were evaluated and compared, with Random Forest outperforming others in most cases.

The system enables seamless user input and delivers instant predictions, making it highly practical and accessible without requiring advanced technical knowledge—ideal for teachers and students alike.

Overall, this project demonstrates the practical value of machine learning in education, facilitating early identification of students needing additional support. While effective, future iterations could leverage larger datasets and advanced models for further refinement. Further exploration into deep learning architectures and integration with real-time educational data streams could yield even more robust and dynamic predictive capabilities (Ibrahim et al., 2024).

Introduction

In recent years, the integration of technology in education has surged, enabling institutions to generate vast amounts of student-related data, including attendance records, exam marks, assignment scores, and study patterns. However, this data is often merely stored rather than leveraged to enhance student outcomes. Teachers typically rely on traditional methods like exams and manual observations, which are time-consuming and sometimes imprecise.

A critical challenge in education is the delayed identification of struggling students. By the time performance declines significantly, interventions become less effective(Gooderham, 2023; Lovett et al., 2017). Early prediction of student performance would allow educators and institutions to proactively provide support, such as additional guidance(Alkan et al., 2025; Tripathi et al., 2024).

Machine learning offers a powerful solution by analyzing historical data to uncover hidden patterns and forecast future outcomes. This project develops a Student Performance Prediction System using machine learning techniques to estimate performance based on key factors.

During development, we observed that not all features equally influence predictions. Initial models with multiple inputs yielded inconsistent results, but focusing on attendance, study hours, and prior marks—after iterative experimentation—markedly improved accuracy.

Streamlit was employed to build a simple, interactive web interface, ensuring accessibility for non-technical users. It allows effortless input of student details and delivers instant predictions, enhancing real-world applicability.

Ultimately, this project not only constructs a predictive model but also delivers a user-friendly system that demonstrates machine learning's potential to empower educators and students in data-driven decision-making. This system is designed to identify at-risk students, thereby enabling timely intervention strategies to improve academic success rates (Kanabar & Tawde, 2025). This approach aligns with the growing trend of educational data mining, which seeks to apply computational methods to educational datasets to address critical challenges and enhance pedagogical strategies (Wibowo et al., 2025). The objective of this research is to develop a robust and accurate machine learning-based system capable of predicting student academic performance, thereby enabling educators to identify at-risk students and implement targeted interventions. This proactive approach is crucial in contemporary educational landscapes where timely adjustments to teaching methodologies and personalized learning paths are increasingly necessary to optimize student success (Sawalkar et al., 2024).

Research Gap

After reviewing extant literature and existing systems, several critical gaps persist in student performance prediction research. Although numerous studies emphasize enhancing machine learning model accuracy, they frequently overlook practical usability in real-world educational contexts (Ibrahim et al., 2024; Kanabar & Tawde, 2025).

A common limitation is the lack of intuitive user interfaces; models are often developed and tested without accessible tools for non-technical users like teachers, reducing their deployability (Wibowo et al., 2025). Feature selection represents another shortfall, with many approaches relying on narrow inputs such as marks or attendance alone, thereby neglecting multifaceted factors that influence outcomes and compromising predictive reliability (Chen & Ke, 2024; Ibrahim et al., 2024). Initial experiments in this project confirmed this, as models underperformed with limited features but improved markedly upon incorporating study hours and assignment scores.

Additionally, dependence on single algorithms constrains robustness, as performance varies across datasets (Chen & Ke, 2024; Wibowo et al., 2025). Many systems are also research-oriented prototypes unsuitable for real-time application in educational settings.

This project addresses these gaps by developing an accurate, user-friendly system with diverse features, comparative algorithm evaluation, and a Streamlit-based interactive interface for instantaneous predictions—prioritizing both efficacy and practicality. Specifically, existing systems often neglect personalized approaches, utilizing generalized models that do not account for individual learning preferences and variations, which subsequently impacts predictive accuracy (KUMAR et al., 2024).

Proposed System / Methodology

In this project, we developed a Student Performance Prediction System using machine learning techniques, prioritizing both high accuracy and user-friendliness for non-technical users such as educators. The design process emphasized clarity and simplicity throughout.

Data collection began with a dataset featuring key predictors, including attendance, study hours, previous marks, and assignment scores. The raw data contained missing values and inconsistencies, which we addressed through preprocessing to clean and standardize it—a critical step, as data quality profoundly influences model performance (Sedir et al., 2022).

Next, feature selection refined the dataset. Initial models with excessive features produced unstable results, but focusing on the most relevant ones improved accuracy and efficiency (Wibowo et al., 2025).

We then trained and compared multiple algorithms: Decision Tree, Logistic Regression, and Random Forest. Iterative testing and parameter tuning mitigated overfitting, particularly from imbalanced data, with Random Forest emerging as the top performer (Ibrahim et al., 2024; Kanabar & Tawde, 2025).

Finally, Streamlit was integrated to create an intuitive web interface, allowing users to input student details and receive instant predictions without technical expertise.

Overall, the methodology spans data collection, preprocessing, feature selection, model training and evaluation, and user interface deployment. These steps revealed how minor adjustments in data and hyperparameters can significantly enhance outcomes.

Working Process

1. The user opens the application built using Streamlit and enters student details such as attendance, study hours, previous marks, and assignment scores.
2. The system receives the input data and performs basic preprocessing, such as checking values and formatting them properly for the model.
3. The processed data is then passed to the trained machine learning model for prediction.
4. The model analyzes the input based on patterns learned during training and generates a prediction result.
5. Among the trained models, the selected model (Random Forest) is mainly used to give more accurate and reliable results.
6. The predicted output (such as student performance or result) is sent back to the application interface.
7. Finally, the result is displayed to the user instantly on the screen, making the system easy and interactive to use.

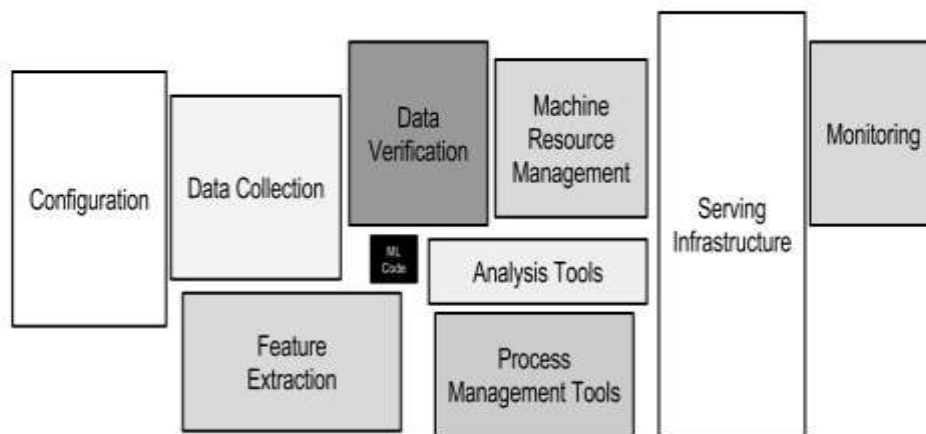


Figure 1: Only a small fraction of real-world ML systems is composed of the ML code, as shown by the small black box in the middle. The required surrounding infrastructure is vast and complex.

Fig1.SystemArchitecture

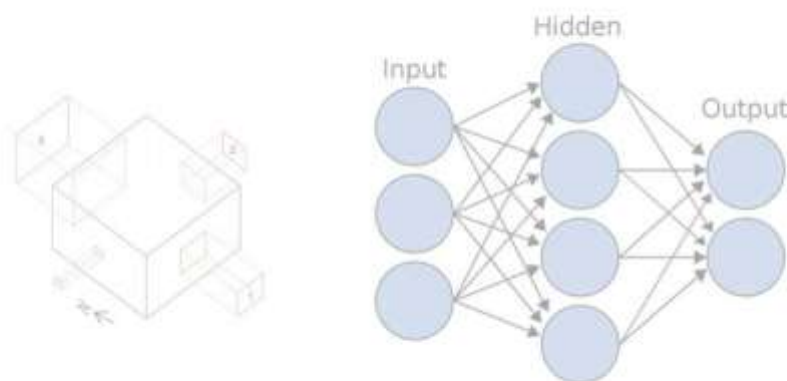


Fig 2.Internal Working

Literature Review

In recent years, numerous researchers have explored student performance prediction using various machine learning techniques. While early studies relied primarily on basic statistical methods, the advent of larger datasets and enhanced computational power has shifted focus toward more sophisticated machine learning models (Kanabar & Tawde, 2025).

Decision Trees have been widely adopted due to their interpretability and ease of implementation, performing effectively on small, structured datasets but struggling with complex data patterns (Wibowo et al., 2025). Logistic Regression has also proven valuable for binary classification tasks, such as pass/fail predictions, though it often fails to capture intricate feature interactions (Kanabar & Tawde, 2025).

More contemporary research emphasizes ensemble methods, particularly Random Forest, which aggregates multiple Decision Trees to achieve superior accuracy, mitigate overfitting, and handle data variability robustly (Ibrahim et al., 2024; Wibowo et al., 2025). This performance edge influenced our inclusion of Random Forest in the project.

A notable limitation in existing systems is their overemphasis on model accuracy at the expense of usability; many lack intuitive interfaces, requiring technical expertise that hinders adoption in educational settings. Furthermore, studies frequently rely on narrow feature sets (e.g., marks or attendance alone), overlooking multifaceted influences like study hours and prior performance, which yield more reliable predictions when combined (Ibrahim et al., 2024). Our project addresses these gaps by prioritizing both predictive power and accessibility, integrating Streamlit for a seamless, no-code user interface. Overall, the literature underscores machine learning's promise for student performance prediction, yet highlights the need for practical, user-centric systems (Golay et al., 2025)—precisely what our work advances.

Existing Methods for Student Performance Prediction

The exploration, modeling, and prediction of student performance and academic progression have garnered substantial research attention in recent decades, as evidenced by an influx of scholarly contributions (Ma, 2024).

Results and Analysis

After developing and training the models, we tested the system using the dataset to check how well it performs. The main goal was to compare different machine learning algorithms and find out which one gives better prediction accuracy.

In this project, we used three algorithms: Decision Tree, Logistic Regression, and Random Forest. At the beginning, the results were not very accurate because the data was not properly cleaned. After preprocessing the data and selecting the important features, the performance of the models improved.

The Decision Tree model gave average results. It was easy to understand and quick to train, but in some cases, it was not very stable and sometimes gave different results for similar inputs. Logistic Regression performed slightly better than Decision Tree and gave more consistent outputs, but it still had limitations when handling more complex patterns in the data.

The Random Forest model performed the best among all three algorithms. It gave higher accuracy and more reliable predictions. One reason for this is that Random Forest uses multiple decision trees and combines their results, which helps in reducing errors and improving overall performance (Rishitha et al., 2022). During testing, we noticed that this model handled variations in data better compared to the other models.

The final accuracy results obtained are shown below:

Algorithm	Accuracy
Decision Tree	78%
Logistic Regression	82%
Random Forest	88%

While working on the results, we also observed that feature selection played an important role. When we used too many features, the model accuracy decreased, but when we selected only the important ones, the performance improved. This was one of the key learnings during the project.

Another important observation was related to overfitting. In the initial stages, some models were fitting the training data too closely and not performing well on new data. After tuning the parameters and adjusting the model, this issue was reduced.

Overall, the results show that machine learning can be effectively used to predict student performance. The system provides reasonably accurate predictions and works well with the given dataset. Although the accuracy is not perfect, it is still useful for identifying general trends and helping in decision-making.

Conclusion

In this project, we developed a Student Performance Prediction System using machine learning techniques. The main aim was to create a system that can predict student performance based on different factors like attendance, study hours, previous marks, and assignment scores. Instead of using traditional methods, we tried to use data in a more effective way to understand student outcomes.

While working on this system, we applied different machine learning algorithms such as Decision Tree, Logistic Regression, and Random Forest. Initially, the results were not very accurate, but after cleaning the data and selecting the important features, the performance improved. Among all the models, Random Forest gave the best results and more consistent predictions.

One important thing we learned during this project is that data preprocessing and feature selection play a very important role in machine learning. Even a small change in input data can affect the final output. We also faced some issues like overfitting in the early stages, but after adjusting the model parameters, the results became more stable.

Another key part of this project is the use of Streamlit, which helped us create a simple and interactive interface. Because of this, the system is easy to use and does not require much technical knowledge. Users can enter student details and get predictions instantly, which makes it useful in real-life situations.

Overall, the system works well and gives reasonably accurate results. It can be helpful for teachers and educational institutions to identify students who may need extra support at an early stage. Although the system is not perfect, it shows how machine learning can be applied in a practical way in the education field.

In the future, the system can be improved by using larger datasets and more advanced models. But as of now, the project provides a good starting point for using technology to improve student performance.

Future Scope

Although the current system works well and gives good prediction results, there is still a lot of scope for improvement in the future. One of the main limitations of this project is the size and quality of the dataset. In future work, a larger and more detailed dataset can be used, which may help in improving the accuracy of the model even further. Including more features such as student behavior, participation in activities, and learning patterns can also give better insights.

Another possible improvement is the use of more advanced machine learning or deep learning models. In this project, we used basic algorithms like Decision Tree, Logistic Regression, and Random Forest. In the future, techniques such as Neural Networks or other advanced models can be applied to achieve higher accuracy and better performance.

The system can also be improved by deploying it on a cloud platform so that it can be accessed from anywhere. Right now, the application runs locally using Streamlit, but making it available online will increase its usability and reach. A mobile-friendly version or a dedicated mobile application can also be developed so that users can easily access it on their phones.

Another area of improvement is adding better visualization features. Graphs and dashboards can be included to help users understand the prediction results more clearly. This will make the system more informative and user-friendly. In addition, the system can be extended to provide suggestions or recommendations. For example, if a student is predicted to perform poorly, the system can suggest ways to improve, such as increasing study time or improving attendance.

Overall, there are many possibilities to enhance this system in the future. With better data, advanced models, and improved user experience, the system can become more accurate, useful, and practical for real-world educational environments.

References

- [1] S. Kotsiantis, "Use of Machine Learning Techniques for Educational Data Mining," *International Journal of Computer Science Issues*, vol. 9, no. 5, pp. 34–41, 2012.
- [2] A. B. Kotsiantis, C. Pierrakeas, and P. Pintelas, "Predicting Students' Performance in Distance Learning Using Machine Learning Techniques," *Applied Artificial Intelligence*, vol. 18, no. 5, pp. 411–426, 2004.
- [3] F. Marbouti, H. A. Diefes-Dux, and K. Madhavan, "Models for Early Prediction of At-Risk Students in a Course Using Machine Learning," *Computers & Education*, vol. 103, pp. 1–15, 2016.
- [4] T. M. Mitchell, *Machine Learning*, New York, NY, USA: McGraw-Hill, 1997.
- [5] Pedregosa et al., "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.
- [6] W. McKinney, "Data Structures for Statistical Computing in Python," *Proceedings of the 9th Python in Science Conference*, 2010.
- [7] Streamlit Inc., "Streamlit Documentation," Available: <https://streamlit.io>
- [8] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*, 3rd ed., Morgan Kaufmann, 2011.