iJRASET
International Journal For Research in Applied Science and Engineering Technology

# INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

**Volume:** 14 **Issue:** IX **Month of publication:** September 2026

**DOI:** https://doi.org/10.22214/ijraset.2026.84762

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
*ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538*
*Volume 14 Issue IX Sep 2026- Available at www.ijraset.com*

# Student Performance Prediction Using Machine Learning

Shalini S.[1], Sahana R.[2], Poorvika R.[3], Aaliya Ashme A.W.[4], Rubali R.[5]
[1, 2, 3, 4]*U.G. Student, Department of Information Technology, Annai Womens College, Karur, TamilNadu, India*
[5]*Assistant Professor, Department of Computer Applications, Annai Womens College, Karur, TamilNadu, India*

***Abstract:*** ***Student academic performance is an important factor in evaluating the effectiveness of the learning process and identifying students who may require additional academic support. Traditional methods of evaluating student performance mainly depend on examination marks and teacher observations, which may not provide sufficient information for early identification of students at academic risk. This paper presents a machine learning-based approach for predicting student performance using relevant academic and personal attributes. The proposed system involves data preprocessing, feature selection, model training, and performance evaluation. Machine learning algorithms such as Linear Regression, Decision Tree, Random Forest, and Support Vector Machine can be applied to identify patterns in student data and predict their expected academic performance. The performance of the models can be compared using suitable evaluation metrics such as accuracy, Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE), depending on the prediction task. The proposed approach can assist educational institutions in identifying students who may need additional guidance and support. The system demonstrates the potential of machine learning in supporting data-driven academic decision-making.***
***Keywords: Machine Learning, Student Performance, Prediction, Data Mining, Random Forest, Educational Data Mining, Classification.***

## I. INTRODUCTION

Education plays an important role in the development of individuals and society. Student performance is commonly evaluated through examinations, assignments, attendance, and classroom activities. However, identifying students who are likely to experience academic difficulties at an early stage can be challenging. Machine learning provides techniques for discovering patterns from historical data and using those patterns to make predictions. In educational environments, machine learning can be used to analyze student-related information such as previous academic marks, attendance, study time, participation, and other relevant factors.

Student performance prediction can help teachers and educational institutions understand the factors that influence academic achievement. Early prediction can also support personalized guidance and intervention for students who require additional assistance. This research proposes a machine learning-based framework for predicting student academic performance. The system processes student-related data, trains different machine learning models, evaluates their performance, and identifies a suitable model for prediction.

## II. PROBLEM STATEMENT

Educational institutions often identify poor academic performance only after the completion of examinations. This makes early intervention difficult. A predictive system is therefore required to analyze available student data and estimate academic performance before final examination results.

The main problem addressed in this research is:

How can machine learning techniques be used to predict student academic performance based on available academic and behavioral factors?

## III. OBJECTIVES

The major objectives of the proposed system are:

*1)* To collect and preprocess student academic data.
*2)* To identify important factors associated with student performance.
*3)* To apply suitable machine learning algorithms for prediction.
*4)* To compare the performance of different machine learning models.
*5)* To identify a suitable model for student performance prediction.
*6)* To support early identification of students who may require academic assistance.

*ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538*
*Volume 14 Issue IX Sep 2026- Available at www.ijraset.com*

## IV. LITERATURE REVIEW

Educational Data Mining and Machine Learning have increasingly been used to analyze student-related data. Previous studies have explored factors such as attendance, previous examination marks, study time, parental background, and student participation to predict academic outcomes. Classification algorithms such as Decision Tree, Random Forest, Support Vector Machine, and K-Nearest Neighbors have been applied to categorize students into different performance levels. Regression-based techniques have also been used when the objective is to predict numerical marks. Among these approaches, ensemble methods such as Random Forest are useful because they combine multiple decision trees and can capture complex relationships between different input features. However, model performance depends on the quality of the dataset, feature selection, preprocessing methods, and evaluation strategy. This study focuses on comparing machine learning approaches for student performance prediction and determining an effective model based on experimental evaluation.

## V. PROPOSED SYSTEM

The proposed system consists of five major stages:
Data Collection → Data Preprocessing → Feature Selection → Model Training → Performance Evaluation

### *A. Data Collection*

The dataset contains student-related attributes that may influence academic performance. Possible attributes include:

- Previous examination marks
- Attendance
- Study time
- Assignment performance
- Internal assessment marks
- Participation
- Previous academic performance

The dataset should be obtained from a suitable public educational dataset or collected with appropriate permission.

### *B. Data Preprocessing*

The collected data may contain missing values, inconsistent values, and categorical attributes. Data preprocessing is performed to improve the quality of the dataset.
The preprocessing stage may include:

- Handling missing values
- Removing duplicate records
- Encoding categorical variables
- Detecting and handling outliers
- Feature scaling where required
- Splitting the dataset into training and testing sets

### *C. Feature Selection*

Feature selection identifies the attributes that contribute significantly to the prediction task. Removing irrelevant features can reduce model complexity and improve prediction performance.

### *D. Model Training*

The preprocessed dataset is used to train multiple machine learning algorithms. The following algorithms can be considered:

- Linear Regression
- Decision Tree
- Random Forest
- Support Vector Machine
- K-Nearest Neighbors

The trained models learn relationships between input features and student performance.

*E. Performance Evaluation*

The trained models are evaluated using appropriate metrics.

For classification:

- Accuracy
- Precision
- Recall
- F1-Score
- Confusion Matrix

For numerical score prediction:

- Mean Absolute Error (MAE)
- Mean Squared Error (MSE)
- Root Mean Squared Error (RMSE)
- $R^2$ Score

## VI. SYSTEM ARCHITECTURE

The architecture of the proposed system can be represented as:

Student Dataset
↓
Data Preprocessing
↓
Feature Selection
↓
Train-Test Split
↓
Machine Learning Models
↓
Model Evaluation
↓
Best Performing Model
↓
Predicted Student Performance

## VII. ALGORITHMS

*A. Decision Tree*

A Decision Tree uses a tree-like structure to make predictions. It divides the dataset according to selected features and creates decision rules. It is relatively easy to understand and interpret.

*B. Random Forest*

Random Forest is an ensemble learning algorithm that combines multiple decision trees to produce a prediction. It generally provides better robustness than a single decision tree and can handle multiple input features.

*C. Support Vector Machine*

Support Vector Machine identifies an optimal decision boundary between classes. It can be useful for classification problems involving different student performance categories.

*D. K-Nearest Neighbors*

K-Nearest Neighbors predicts the class of a new observation based on the classes of its nearest observations in the feature space.

## VIII. IMPLEMENTATION

The proposed system can be implemented using **Python**. Libraries such as Pandas and NumPy can be used for data processing, while Scikit-learn can be used for machine learning algorithms and evaluation.

A typical implementation workflow is:

*1)* Import the dataset.
*2)* Analyze and clean the data.
*3)* Select relevant features.
*4)* Encode categorical variables.
*5)* Split the dataset into training and testing data.
*6)* Train multiple machine learning models.
*7)* Generate predictions.
*8)* Calculate evaluation metrics.
*9)* Compare the models.
*10)* Select the best-performing model.

## IX. RESULTS AND DISCUSSION

The performance of the proposed models should be evaluated using the selected dataset. The results should be reported using actual experimental values rather than assumed values.

Table I
Performance Comparison of Machine Learning Models

| S. No. | Machine Learning Model | Accuracy (%) | MAE | RMSE | $R^2$ Score |
|---|---|---|---|---|---|
| 1 | Linear Regression | — | — | — | — |
| 2 | Decision Tree | — | — | — | — |
| 3 | Random Forest | — | — | — | — |
| 4 | Support Vector Machine (SVM) | — | — | — | — |
| 5 | K-Nearest Neighbors (KNN) | — | — | — | — |

The experimental results will determine which algorithm provides the most suitable performance for the selected dataset. A confusion matrix and graphical comparison can also be included to improve the analysis.

## X. ADVANTAGES

The proposed system provides the following advantages:

*1)* Enables early prediction of student performance.
*2)* Helps identify students who may require additional support.
*3)* Reduces dependence on manual analysis.
*4)* Enables comparison of multiple machine learning algorithms.
*5)* Supports data-driven educational decision-making.
*6)* Can be extended with additional student-related features.

## XI. LIMITATIONS

The accuracy of the proposed system depends strongly on the quality and quantity of the dataset. Student performance is influenced by many factors that may not be available in the dataset. Therefore, predictions should be considered as decision-support information rather than a definitive assessment of a student's ability.

Privacy and responsible handling of student information are also important considerations when developing such systems.

## XII. FUTURE ENHANCEMENT

The proposed system can be enhanced by incorporating larger and more diverse datasets. Future work may include deep learning techniques, explainable machine learning, real-time academic data, and personalized recommendation systems.

The system could also be integrated into an educational platform to provide early alerts and personalized learning recommendations to students and teachers.

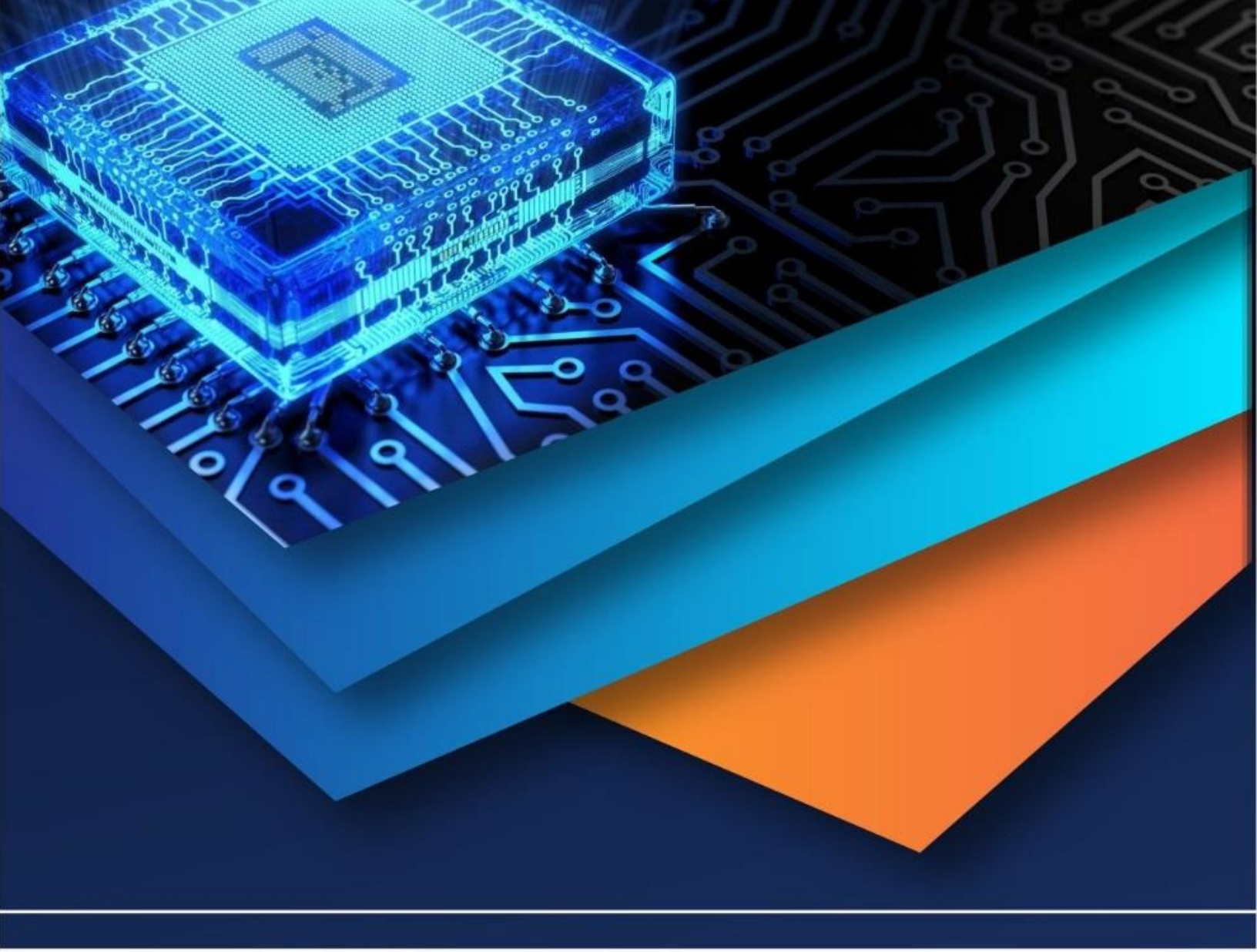
## XIII. CONCLUSION

This paper presents a machine learning-based approach for predicting student academic performance. The proposed framework involves data collection, preprocessing, feature selection, model training, and evaluation. Different machine learning algorithms can be compared to identify an effective model for the prediction task.

Such a system can assist educational institutions in identifying students who may require additional academic support and can contribute to data-driven decision-making. With improved datasets and responsible use of student information, machine learning can provide useful support for educational performance analysis.

## REFERENCES

[1] C. Romero and S. Ventura, “Educational Data Mining: A Review of the State of the Art,” IEEE Transactions on Systems, Man, and Cybernetics, Part C, vol. 40, no. 6, pp. 601–618, 2010.
[2] R. S. Baker and K. Yacef, “The State of Educational Data Mining in 2009: A Review and Future Visions,” Journal of Educational Data Mining, vol. 1, no. 1, pp. 3–17, 2009.
[3] T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning, Springer, 2009.
[4] C. M. Bishop, Pattern Recognition and Machine Learning, Springer, 2006.
[5] F. Pedregosa et al., “Scikit-learn: Machine Learning in Python,” Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

# INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY