IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 9, September 2026

DOI: 10.17148/IJARCCE.2026.15913

AI-BASED STUDENT PREFORMANCE PREDICTION

Dr.B. Rajesh Kumar[1], S. Shailesh[2], Mohammed Aadhil.I[3]

Head & Associate Professor, Dept of IT & Cognitive, Sri Krishna Arts and Science College[1]

Computing Science, Sri Krishna Arts and Science College[2,3]

Abstract: Education is increasingly generating large amounts of digital data through examinations, assignments, attendance systems, online learning platforms, and internal assessments. This data can provide valuable information about student learning and academic progress.

Student Performance Prediction using Machine Learning is a system that analyzes historical student information and predicts the likely academic performance of students. The system can consider several factors, including attendance, previous examination marks, assignment performance, internal assessment marks, study hours, classroom participation, and other relevant attributes.

The proposed system consists of multiple stages, including data collection, data preprocessing, feature selection, model training, model testing, and performance prediction. Machine learning algorithms such as Decision Tree, Random Forest, Logistic Regression, and Support Vector Machine can be evaluated for the prediction task.

The system can help teachers identify students who may require additional academic support. It can also help educational institutions understand patterns associated with student achievement and improve academic planning.

However, predictions should be treated as supporting information rather than as a final judgment about students. Privacy, fairness, transparency, and human supervision are important when applying machine learning to educational data.

Keywords: Student Performance, Machine Learning, Educational Data Mining, Prediction, Classification, Academic Analytics, Data Science.

INTRODUCTION

Education is an important part of social and economic development. Educational institutions continuously evaluate students to understand their learning progress and academic achievement.

Student performance is influenced by many factors. Examination marks are important, but they are not the only indicators. Attendance, assignments, previous academic results, study habits, participation, internal assessments, and learning activities can also provide useful information.

In a traditional education system, teachers generally evaluate students using examinations, assignments, classroom activities, and personal observations. While these methods are valuable, manually analyzing all available information becomes increasingly difficult when there are many students.

Machine Learning provides a way to analyze large datasets automatically. A machine learning algorithm learns patterns from historical data and uses those patterns to make predictions about new data.

For example, if historical data shows a relationship between attendance, previous marks, assignment performance, and final results, a machine learning model can learn these relationships.

Importance of the System

The proposed system can provide an **early indication of academic performance**. Teachers can use this information to identify students who may need additional support, such as:

- Remedial classes
- Additional assignments
- Academic counseling

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 9, September 2026
DOI: 10.17148/IJARCCE.2026.15913

- Individual guidance
- Learning resources
- Attendance support

The goal is not to replace teachers but to provide an additional analytical tool for academic decision-making.

BACKGROUND OF THE STUDY

Educational Data Mining

Educational Data Mining is the process of applying data mining, statistics, and machine learning techniques to educational information.

Educational institutions may have data related to:

- Student attendance
- Examination marks
- Assignment scores
- Internal assessments
- Course participation
- Learning activities
- Previous academic results

Analyzing this information can help discover useful patterns.

Learning Analytics

Learning Analytics focuses on collecting and analyzing learner-related information to understand and improve learning and educational outcomes.

Difference Between Traditional Analysis and Machine Learning

Traditional Analysis	Machine Learning
Often manually performed	Automated or semi-automated
Can be time-consuming	Can process large datasets efficiently
Usually focuses on known indicators	Can identify patterns in multiple features
Limited scalability	More scalable for large datasets
Mainly descriptive	Can also support predictive analysis

Machine learning therefore provides an opportunity to move from simply recording student results toward **predictive academic analytics**.

PROBLEM STATEMENT

Educational institutions collect a large amount of student data, but identifying students who may struggle academically at an early stage remains challenging.

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 9, September 2026
DOI: 10.17148/IJARCCE.2026.15913

Major Problems

1. Late Identification of Weak Performance

Students may be identified as academically weak only after examination results are published.

2. Large Amount of Data

Teachers may have to analyze attendance, tests, assignments, and examination records for many students.

3. Multiple Factors

Student performance depends on several factors. Exam marks alone may not provide a complete picture.

4. Manual Monitoring

Continuous manual monitoring can be difficult and time-consuming.

5. Lack of Early Warning

Without predictive information, it may be difficult to determine which students are likely to need support before final examinations.

Proposed Problem

Develop a system that uses historical student information to predict academic performance and provide an early indication of students who may need additional support.

Example

If a student has:

- Low attendance
- Low previous marks
- Low assignment scores
- Low internal assessment results

the model may identify the student as having a higher likelihood of low academic performance.

This prediction can trigger further review by teachers.

OBJECTIVES

Main Objective

To design and develop a machine-learning-based system for predicting student academic performance using relevant educational data.

Specific Objectives

Data-Related Objectives

1. Collect relevant student performance data.
2. Identify suitable academic and behavioral features.
3. Clean and preprocess the dataset.
4. Handle missing and inconsistent values.

Machine Learning Objectives

5. Select suitable machine learning algorithms.
6. Train models using historical student data.

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 9, September 2026
DOI: 10.17148/IJARCCE.2026.15913

7. Test models using unseen data.
8. Compare model performance.

Educational Objectives

9. Identify students who may require additional academic support.
10. Provide useful information for teachers and academic counselors.
11. Support data-driven academic planning.
12. Encourage early intervention.

Expected Outcome

The expected outcome is a prediction model capable of classifying students into performance categories or estimating their academic score, depending on the selected approach.

LITERATURE SURVEY

Several researchers have explored the use of machine learning and data mining for analyzing student performance.

Studies in Educational Data Mining have investigated the relationship between academic performance and factors such as:

- Previous grades
- Attendance
- Assignment performance
- Study habits
- Demographic information
- Assessment results
- Learning activities

Commonly Used Algorithms

Decision Tree

Decision Trees are popular because their decision-making process can be relatively easy to interpret.

Random Forest

Random Forest combines multiple decision trees and is often effective for structured datasets.

Logistic Regression

Logistic Regression is a commonly used classification technique that can provide probabilities for class membership.

Support Vector Machine

SVM can classify data by identifying a suitable separating boundary.

Neural Networks

Neural networks can model complex patterns but may require more data, tuning, and computational resources.

Research Gap

Different institutions have different student populations, courses, assessment systems, and learning environments. Therefore, a model that performs well in one dataset may not automatically perform equally well elsewhere.

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 9, September 2026
DOI: 10.17148/IJARCCE.2026.15913

A reliable system should therefore:

- Use appropriate data.
- Evaluate multiple models.
- Validate predictions carefully.
- Consider fairness and privacy.
- Keep educators involved in decision-making.

EXISTING SYSTEM

In a conventional educational environment, student performance is evaluated through several methods.

Existing Methods

Examination-Based Evaluation

Students are evaluated through periodic and final examinations.

Assignment Evaluation

Teachers assess student understanding through assignments and projects.

Attendance Monitoring

Attendance is recorded and sometimes used as an indicator of student engagement.

Teacher Observation

Teachers use classroom participation and behavior to understand student progress.

Disadvantages

1. Time Consuming

Analyzing information for hundreds or thousands of students can require substantial effort.

2. Delayed Intervention

Students may receive additional support only after poor results become visible.

3. Limited Prediction

Traditional methods generally describe current or previous performance rather than predicting future outcomes.

4. Human Error

Manual data entry and analysis can introduce errors.

5. Difficult Pattern Detection

Relationships between multiple variables may not be obvious through manual analysis.

PROPOSED SYSTEM

The proposed system uses machine learning to analyze student data and predict academic performance.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 9, September 2026

DOI: 10.17148/IJARCCE.2026.15913

System Workflow

Student Data Collection

↓

Data Cleaning

↓

Data Preprocessing

↓

Feature Selection

↓

Training and Testing

↓

Machine Learning Model

↓

Model Evaluation

↓

Performance Prediction

↓

Result Display

Input Data

The system can use:

- Attendance percentage
- Previous examination marks
- Assignment scores
- Internal assessment
- Study hours
- Participation
- Assignment completion
- Previous academic performance

Output

The model can produce a classification such as:

High Performance

Medium Performance

Low Performance

or a numerical predicted score.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 9, September 2026

DOI: 10.17148/IJARCCE.2026.15913

Main Benefit

The system provides an early analytical indicator that can help teachers decide whether a student should receive additional attention.

DATASET AND FEATURE ENGINEERING

Dataset

The dataset contains historical information about students and their academic outcomes.

A suitable dataset can be obtained from:

- Institutional records, with appropriate authorization and privacy protection
- Public educational datasets
- Research datasets

Example Features

Feature	Example Value
Attendance	90%
Previous Examination Mark	78
Assignment Score	85
Internal Assessment	82
Study Hours	3 hours/day
Participation	High
Assignment Completion	95%
Performance Category	High

Feature Engineering

Feature engineering involves transforming available data into useful model inputs.

For example:

Average Assessment Score

=
(Test 1 + Test 2 + Assignment + Internal Assessment) / Number of Assessments

Another feature could represent **attendance category**:

- Below 60% → Low
- 60–75% → Medium
- Above 75% → High

The actual thresholds should be determined according to the dataset and institutional context rather than being assumed universally.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 9, September 2026

DOI: 10.17148/IJARCCE.2026.15913

METHODOLOGY

Step 1 — Data Collection

Collect relevant historical student information.

Step 2 — Data Exploration

Examine:

- Number of records
- Number of features
- Data types
- Missing values
- Distribution of target classes

Step 3 — Data Cleaning

Remove duplicates and correct invalid or inconsistent records.

Step 4 — Missing Value Handling

Appropriate techniques can be used to handle missing information.

Examples include:

- Removing unsuitable records
- Filling numerical values using statistical methods
- Filling categorical values using appropriate categories

Step 5 — Encoding

Convert categorical values into numerical representations when required.

Step 6 — Feature Selection

Select relevant variables and remove unnecessary or potentially harmful attributes.

Step 7 — Dataset Splitting

Separate data into training and testing sets.

Step 8 — Model Training

Train selected algorithms on the training dataset.

Step 9 — Model Evaluation

Evaluate the trained models on unseen test data.

Step 10 — Prediction

Use the selected model to predict the performance of new students.

MACHINE LEARNING ALGORITHMS

1. Decision Tree

A Decision Tree uses a tree-like structure to make decisions.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 9, September 2026

DOI: 10.17148/IJARCCE.2026.15913

Example

Attendance > 75%?

|

YES

|

Previous Mark > 60?

|

YES

|

High Performance

Advantages

- Easy to understand
- Easy to visualize
- Useful for classification
- Can work with different types of features

2. Random Forest

Random Forest combines multiple decision trees.

Advantages

- Robust for many tabular problems
- Can handle several input variables
- Less dependent on one decision tree
- Can provide feature importance estimates

3. Logistic Regression

Logistic Regression estimates the probability of a class.

For example:

Input Student Data → Probability → Performance Category

It is useful as a simple baseline model.

4. Support Vector Machine

SVM attempts to find a boundary separating different classes.

It can be effective when the data has suitable characteristics and preprocessing.

5. Model Selection

The best model should not be selected only because it has the highest accuracy.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 9, September 2026

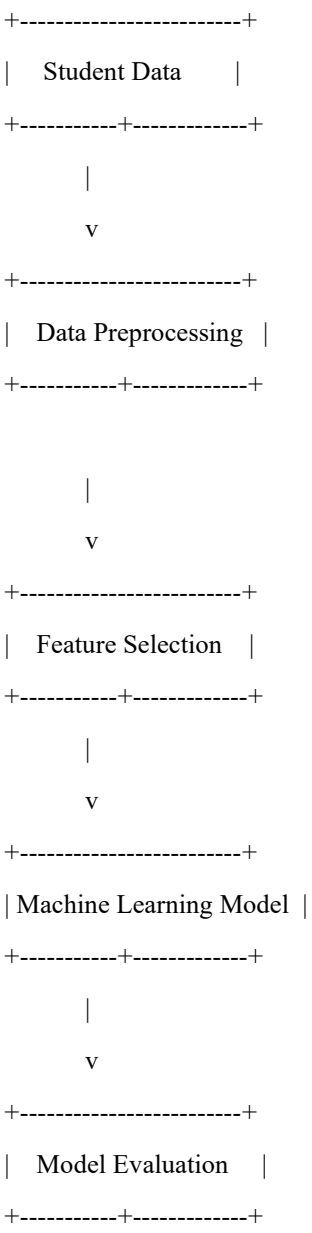
DOI: 10.17148/IJARCCE.2026.15913

Other considerations include:

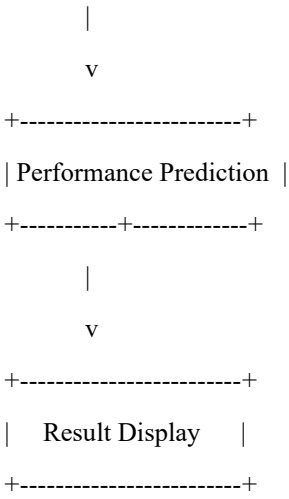
- Precision
- Recall
- F1-score
- Interpretability
- Dataset size
- Computational requirements
- Consequences of incorrect predictions

SYSTEM ARCHITECTURE AND IMPLEMENTATION

System Architecture



International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 9, September 2026
DOI: 10.17148/IJARCCE.2026.15913



Technologies

Programming Language

Python

Libraries

- Pandas
- NumPy
- Scikit-learn
- Matplotlib
- Seaborn

Development Environment

- Jupyter Notebook
- Google Colab
- Visual Studio Code or another Python environment

Storage

- CSV files
- Relational database
- Other authorized educational data sources

Implementation

The implementation begins by loading the dataset. The data is explored and cleaned before relevant features are selected.

Machine learning models are then trained and evaluated. The best-performing suitable model can be saved and used to generate predictions for new records.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 9, September 2026

DOI: 10.17148/IJARCCE.2026.15913

RESULTS, ADVANTAGES, LIMITATIONS AND FUTURE SCOPE

Model Evaluation

The model can be evaluated using:

Accuracy

Percentage of correctly classified records.

Precision

Measures how many predicted members of a class are actually members of that class.

Recall

Measures how many actual members of a class are successfully identified.

F1-Score

Provides a balance between precision and recall.

Confusion Matrix

Shows correct and incorrect predictions for each class.

Results Presentation

The final project can include:

- Model accuracy comparison
- Precision comparison
- Recall comparison
- F1-score comparison
- Confusion matrix
- Feature importance
- Actual vs predicted performance

Actual values should be obtained from your own dataset and experiments.

Advantages

1. Early identification of potentially struggling students.
2. Efficient analysis of large student datasets.
3. Supports academic decision-making.
4. Can analyze multiple factors.
5. Helps teachers provide targeted support.
6. Can be integrated into educational applications.

Limitations

1. Results depend on data quality.
2. Historical data can contain bias.
3. Not every factor influencing performance is measurable.

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 9, September 2026
DOI: 10.17148/IJARCCE.2026.15913

4. Predictions are probabilistic and uncertain.
5. Privacy must be protected.
6. Predictions should not be used as the sole basis for high-impact decisions.

Future Scope

- Real-time prediction
- Web-based dashboard
- Mobile application
- Personalized learning recommendations
- Explainable AI
- Integration with Learning Management Systems
- Continuous model monitoring
- Larger and more diverse datasets
- Early-warning academic support systems

CONCLUSION

Student Performance Prediction using Machine Learning provides a practical approach for analyzing educational data and identifying patterns associated with academic outcomes.

The proposed system can combine information such as attendance, previous marks, assignments, internal assessments, and other relevant features to generate performance predictions.

Machine learning can help teachers identify students who may need additional academic support before final examination results are available.

The system can also help institutions analyze the factors associated with academic performance and improve planning.

However, machine learning should be used responsibly. A prediction does not represent a student's fixed ability or guarantee a future result. Human judgment remains essential, particularly when decisions may significantly affect a student's education.

Therefore, the most appropriate approach is to use machine learning as a **supporting tool for teachers and educational institutions**, combined with responsible data management, privacy protection, fairness, and transparent decision-making.

REFERENCES

[1]. Romero, C. and Ventura, S., research on Educational Data Mining and Learning Analytics.
[2]. UCI Machine Learning Repository, Student Performance Dataset.
[3]. Pedregosa, F. et al., Scikit-learn: Machine Learning in Python, Journal of Machine Learning Research.
[4]. Research publications on Student Performance Prediction using Machine Learning.
[5]. Research literature on Educational Data Mining.
[6]. Research literature on Machine Learning Classification Algorithms.
[7]. Documentation related to Python, Pandas, NumPy, Scikit-learn, Matplotlib, and Seaborn.