

AI-Based Student Performance Prediction Using Machine Learning

Mrs. Mercy Smilin¹, N. Vigneshwari², A. Naseeha³, M. Madhushree⁴, V. Pon Azhagi⁵,
P. Raghunathan⁶, I. Manikandan⁷

¹Assistant Professor, J.P. College of Engineering, Ayikudy, Tenkasi-627852, Affiliated to Anna University, Chennai-6000252

²⁻⁷Student, Department Of Information Technology, J.P.College Engineering, Ayikudy, Tenkasi-627852, Affiliated To Anna University, Chennai-6000252

Abstract: Student performance prediction is an important application of Artificial Intelligence in the education sector. Academic performance can be influenced by attendance, internal assessment marks, assignment performance, study habits, previous examination results, and other learning-related factors. Traditional evaluation mainly depends on examinations and manual analysis, which may not identify students who require academic support at an early stage. This research proposes an AI-based student performance prediction system using machine learning techniques. The system collects relevant student information, preprocesses the data, selects useful features, trains machine learning models, and predicts expected academic performance. Algorithms such as Decision Tree, Random Forest, Logistic Regression, and Support Vector Machine can be applied and compared. The models can be evaluated using suitable performance measures such as accuracy, precision, recall, F1-score, and confusion matrix. The proposed system can help teachers identify students who may need additional guidance and can support data-driven academic planning. The system is intended as a decision-support tool and not as a replacement for teachers. With reliable data and proper evaluation, machine learning can provide useful information for early academic intervention and improved learning support.

Keywords: Artificial Intelligence, Machine Learning, Student Performance, Prediction, Educational Data Mining, Academic Performance, Classification

I. INTRODUCTION

Education produces a large amount of student-related data. Academic records may contain attendance, internal marks, assignment scores, previous examination marks, study time, participation, and other learning-related information. When this information is analyzed systematically, it can provide useful insights into student progress.

In many institutions, student performance is evaluated mainly through examinations and periodic assessments. Although these methods are important, they may not provide an early indication that a student is experiencing academic difficulties. A student may show changes in attendance, assignment completion, or internal marks before the final examination. Identifying such patterns can help teachers provide support at the right time.

Machine Learning is a branch of Artificial Intelligence that enables computers to learn patterns from data

and make predictions. In educational applications, machine learning can be used to analyze historical student information and predict academic outcomes. The prediction may be represented as a performance category such as low, medium, or high, or as a numerical score.

The proposed research focuses on developing an AI-based approach for predicting student performance using machine learning. The system follows a sequence of steps: data collection, data preprocessing, feature selection, model training, model evaluation, and prediction.

1.1 Need for the Study

A systematic prediction system can help institutions analyze large student datasets more efficiently. It can provide early information about possible academic difficulties and support teachers in planning additional guidance, remedial activities, or personalized learning support.

1.2 Scope of the Study

The study focuses on software-based prediction using student-related data. It does not require special hardware. The system can be implemented using Python and commonly used machine learning libraries. The final prediction depends on the quality and relevance of the available dataset.

1.3 Research Contribution

- Develops a structured machine learning workflow for student performance prediction.
- Considers multiple academic and learning-related features.
- Provides a framework for comparing different machine learning algorithms.
- Supports early identification of students who may require additional academic attention.

II. PROBLEM STATEMENT

Educational institutions collect many types of student information, but manually analyzing all records can be difficult when the number of students is large. Teachers may not always be able to identify patterns that indicate a possible decline in academic performance at an early stage.

Therefore, there is a need for a systematic approach that can analyze available student data and predict expected academic performance. The proposed system aims to use machine learning algorithms to learn patterns from historical data and generate performance predictions for new student records.

III. OBJECTIVES

- To collect relevant student academic and learning-related data.
- To clean and preprocess the collected dataset.
- To identify useful features associated with academic performance.
- To train suitable machine learning models.
- To evaluate and compare the trained models.
- To predict student performance using the selected model.
- To support teachers in identifying students who may require additional guidance.

IV. RESEARCH QUESTIONS

- Which student-related features are useful for predicting academic performance?
 - Which preprocessing steps are required before machine learning?
 - How do different machine learning algorithms perform on the selected dataset?
 - Which evaluation measures are suitable for the prediction task?
 - How can prediction results support early academic intervention?

V. EXPECTED OUTCOME

The expected outcome is a software-based machine learning system that accepts relevant student information and produces a performance prediction. The system should also provide evaluation results that allow different algorithms to be compared. The output can be used as supporting information by teachers and educational institutions.

VI. LITERATURE REVIEW

Educational Data Mining is an important research area that applies data analysis and machine learning methods to educational information. Researchers have explored prediction of academic performance, student retention, learning behavior, and other educational outcomes.

Romero and Ventura described Educational Data Mining as a field that applies data mining techniques to educational settings. Their work highlighted the importance of discovering useful patterns from educational data. Baker and Yacef also discussed educational data mining and its potential for understanding learning processes.

Machine learning algorithms have been applied to student datasets containing academic, demographic, attendance, and behavioral information. Decision Trees are frequently considered because their rules can be relatively easy to understand. Random Forest combines multiple trees and can provide robust predictions. Logistic Regression is suitable for classification problems, while Support Vector

Machine can be used for classification and regression.

Previous research also indicates that data preparation has a major effect on model performance. Missing values, inconsistent records, irrelevant features, and class imbalance can affect prediction quality. Therefore, preprocessing, feature selection, and appropriate evaluation are important parts of a reliable prediction system.

6.1 Research Gap

Student performance is affected by several interacting factors, and no single feature can completely explain academic outcomes. Some studies focus on marks alone, while others include attendance, study habits, or demographic information. A useful research approach should consider multiple relevant features and compare suitable machine learning techniques.

6.2 Summary of Literature

The literature supports the use of machine learning for educational prediction tasks. It also shows that the choice of dataset, features, preprocessing methods, algorithm, and evaluation metric can influence the final result. These observations form the basis of the methodology proposed in this research.

VII. SYSTEM WORKFLOW

The proposed system is a software-based student performance prediction framework. It receives student information as input, prepares the data, trains machine learning models, evaluates the models, and produces a performance prediction.

7.1 System Workflow

Student Data → Data Preprocessing → Feature Selection → Train/Test Split → Model Training → Model Evaluation → Selected Model → Performance Prediction

7.2 Input

The input may contain attendance percentage, internal assessment marks, assignment marks, previous semester marks, study time, participation,

and other relevant attributes. The exact features depend on the dataset used for implementation.

7.3 Processing

During preprocessing, missing values and inconsistent records are handled. Categorical attributes can be encoded into numerical form. Numerical features can be scaled when required. After preprocessing, important features are selected for model training.

7.4 Output

The output may be a performance category such as Low, Medium, or High. If the research is designed as a regression problem, the output can instead be a predicted numerical score. The output can also include model evaluation results.

7.5 System Architecture

- Data Collection: Obtain a suitable student performance dataset.
- Data Preprocessing: Clean and transform the raw data.
- Feature Selection: Identify relevant input attributes.
- Model Training: Train selected machine learning algorithms.
- Model Evaluation: Measure prediction performance.
- Prediction: Generate performance results for new records.

7.6 Benefits of the Proposed System

The proposed system can reduce the effort required for large-scale manual analysis and provide early analytical information. It can support teachers in identifying students who may benefit from additional guidance. It can also be extended with new features and improved algorithms in future work.

VII. METHODOLOGY

1.1 Data Collection

A suitable student performance dataset is collected from an educational dataset or institutional source. The dataset should contain enough records and relevant attributes for machine learning. Data

collection must follow appropriate privacy and ethical practices.

8.2 Data Preprocessing

Raw data may contain missing values, duplicate records, inconsistent formats, and categorical values. Missing values can be handled using suitable methods. Duplicate and invalid records should be removed or corrected. Categorical features can be encoded, and numerical values can be normalized or standardized when required

8.3 Feature Selection

Feature selection is used to identify attributes that contribute useful information for prediction. Possible features include attendance, internal marks, assignment scores, previous examination marks, study time, and participation. Removing irrelevant features can reduce unnecessary complexity.

8.4 Data Splitting

The prepared dataset is divided into training and testing portions. The training set is used to learn patterns, while the testing set is used to evaluate the model on unseen data. A validation set or cross-validation can also be used when appropriate.

8.5 Model Training

The selected algorithms are trained using the training data. Multiple algorithms can be trained using the same features so that their performance can be compared fairly

8.6 Model Evaluation

After training, each model is evaluated using suitable metrics. For classification, accuracy, precision, recall, F1-score, and confusion matrix are useful. For regression, Mean Absolute Error and Root Mean Squared Error can be used.

8.7 Prediction

After evaluation, an appropriate model can be selected based on the research objective and evaluation results. New student records can then

be supplied to the model to generate predicted performance

IX. MACHINE LEARNING ALGORITHMS

9.1 Decision Tree

A Decision Tree is a supervised machine learning algorithm that represents decisions using a tree structure. Each internal node represents a condition, branches represent outcomes, and leaf nodes represent the final prediction. The method is useful because its decision rules can be presented in an understandable form.

9.2 Random Forest

Random Forest is an ensemble learning algorithm that combines multiple decision trees. Each tree learns from a different sample of the training data and selected features. The combined prediction can be more stable than the prediction of a single tree.

9.3 Logistic Regression

Logistic Regression is commonly used for classification. It estimates the probability of an observation belonging to a class. For student performance prediction, it can be used when the target is represented by categories such as low, medium, and high.

9.4 Support Vector Machine

Support Vector Machine is a supervised learning algorithm that can be used for classification and regression. In classification, it finds a suitable decision boundary between classes. It can perform well when the data has useful separating patterns.

9.5 Algorithm Comparison Criteria

The algorithms can be compared using the same testing dataset and evaluation metrics. The comparison should consider prediction performance, interpretability, computational requirements, and suitability for the dataset.

X.

Feature	Type	Example
Attendance	Numerical	85%
Internal Marks	Numerical	72/100
Assignment Score	Numerical	80/100
Previous Result	Numerical	75/100
Study Time	Numerical/Categorical	2–4 hours
Performance	Target	Low/Medium/High

DATASET AND IMPLEMENTATION

10.1 Dataset Features

The dataset may include attendance percentage, internal examination marks, assignment marks,

previous semester marks, study time, participation, and other learning-related attributes. The target variable can be a performance category or final academic score.

Algorithm	Main Use	Key Feature
Decision Tree	Classification/Regression	Easy-to-understand rules
Random Forest	Classification/Regression	Combines multiple trees
Logistic Regression	Classification	Probability-based classification
SVM	Classification/Regression	Decision boundary

10.2 Software Requirements

- Python programming language.
- Jupyter Notebook, Google Colab, or another Python development environment.
- Pandas for data processing.
- NumPy for numerical operations.
- Scikit-learn for machine learning algorithms and evaluation.
- Matplotlib for visualization.

10.3 Implementation Steps

- Load the student dataset.
- Inspect and clean the data.
- Handle missing values and encode categorical features.
- Select relevant input features and target variable.
- Split the dataset into training and testing data.
- Train the selected machine learning algorithms.
- Evaluate each model using suitable metrics.
- Use the selected model for performance prediction.

10.4 Ethical Considerations

Student information should be handled responsibly. Personal identifiers should be protected, and only

relevant information should be used. Predictions should not be treated as permanent labels. Teachers should consider the wider academic context before taking decisions based on model output.

XI. PERFORMANCE EVALUATION, RESULTS AND DISCUSSION

11.1 Evaluation Metrics

Accuracy represents the proportion of correctly predicted observations. Precision indicates how many observations predicted as a particular positive class are actually positive. Recall indicates how many actual positive observations are correctly identified. F1-score combines precision and recall into a single measure.

A confusion matrix shows the number of correct and incorrect predictions for each class. These measures help researchers understand not only the overall performance but also how the model behaves for individual performance categories.

11.2 Result Presentation

After implementation, the results of each algorithm should be recorded in a table. The table may contain

accuracy, precision, recall, and F1-score. Graphs can be used to visually compare model performance.

Model	Accuracy	Precision	Recall	F1-Score
Decision Tree	To be calculated	To be calculated	To be calculated	To be calculated
Random Forest	To be calculated	To be calculated	To be calculated	To be calculated
Logistic Regression	To be calculated	To be calculated	To be calculated	To be calculated
SVM	To be calculated	To be calculated	To be calculated	To be calculated

11.3 Discussion

The final results should be discussed based on the actual dataset and experimental measurements. If two models show different results, possible reasons may include feature relationships, dataset size, class distribution, and algorithm characteristics. The discussion should avoid claiming that one algorithm is universally better because model performance depends on the dataset and experimental setup.

Incorrect predictions should also be examined. Student performance can be influenced by factors that are not included in the dataset. Therefore, a prediction error does not necessarily indicate a failure of the overall approach. It may indicate the need for additional features, better preprocessing, more data, or a different model.

11.4 Expected Practical Use

The prediction output can be used as supporting information for academic mentoring. Teachers may review the prediction together with attendance, marks, classroom participation, and other relevant information. This can help create timely academic support plans.

XII. ADVANTAGES AND LIMITATIONS

12.1 Advantages

- Provides early prediction of possible academic performance.
- Helps teachers identify students who may need additional guidance.
- Supports analysis of large amounts of educational data.
- Reduces dependence on only manual analysis.

- Supports data-driven academic planning.
- Allows comparison of multiple machine learning algorithms.
- Can be extended with additional features and improved models.

12.2 Limitations

- Prediction quality depends on the quality and size of the dataset.
- Missing or incorrect data can reduce model performance.
- Not all factors affecting academic performance can be measured.
- A model trained on one dataset may not perform equally on another dataset.
- Predictions should not be treated as the only basis for academic decisions.

XIII. FUTURE SCOPE

- Use larger datasets collected from different institutions.
- Include additional academic and learning-related features.
- Compare advanced machine learning and deep learning methods.
- Develop personalized academic recommendations.
- Create teacher dashboards for continuous monitoring.
- Use explainable AI to make predictions easier to understand.
- Update models with new data to improve long-term usefulness.

XIV. CONCLUSION

AI-based student performance prediction using machine learning provides a useful approach for analyzing educational data. By using relevant information such as attendance, academic marks, assignment performance, study-related factors, and previous results, machine learning models can identify patterns associated with student performance.

The proposed system includes data collection, preprocessing, feature selection, model training, evaluation, and prediction. Decision Tree, Random Forest, Logistic Regression, and Support Vector Machine can be evaluated using suitable metrics. The resulting system can provide useful analytical information for teachers and institutions.

The system should be considered a decision-support tool rather than a replacement for human judgment. With reliable data, appropriate evaluation, and responsible use, the proposed approach can support early academic guidance and data-driven educational planning.

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. F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.
4. L. Breiman, "Random Forests," Machine Learning, vol. 45, pp. 5–32, 2001.
5. T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning, 2nd ed. Springer, 2009.
6. C. M. Bishop, Pattern Recognition and Machine Learning. Springer, 2006.
7. T. M. Mitchell, Machine Learning. McGraw-Hill, 1997.
8. I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, Data Mining: Practical Machine Learning Tools and Techniques, 4th ed. Morgan Kaufmann, 2016.