

AI-Based Student Performance Prediction and Academic Success Recommendation System Using Random Forest and Explainable Machine Learning

Fardin Umar Mujawar

Department of Artificial Intelligence and Data Science

Anjuman-I-Islam Abdul Razzak Kalsekar Polytechnic
Campus, New Panvel, Navi Mumbai, Maharashtra, India

Guide: **Prof. Ali Karim**

Abstract - Student academic performance is one of the most important indicators of educational success and institutional quality. Traditional methods of evaluating student performance are generally based on periodic examinations and manual analysis, which often fail to identify academically at-risk students at an early stage. Early prediction of student performance enables educational institutions to provide timely interventions, personalized academic guidance, and improved learning support.

This research presents an AI-Based Student Performance Prediction and Academic Success Recommendation System that utilizes Machine Learning techniques to predict student academic performance using multiple educational attributes. The proposed system analyzes attendance, study hours, assignment scores, internal assessment marks, and classroom participation to classify students into three performance categories: High Performer, Average Performer, and At-Risk Student.

The system employs the Random Forest Classifier because of its robustness, high prediction capability, and ability to handle multiple decision variables effectively. To improve transparency and interpretability, Explainable Artificial Intelligence (XAI) techniques are incorporated through feature importance analysis, enabling educators to understand the contribution of each academic parameter toward prediction outcomes.

A novel Academic Success Probability Score (ASPS) is introduced as an additional evaluation metric to estimate the overall probability of student success based on weighted academic indicators. Furthermore, the proposed system generates personalized academic

recommendations that assist students in improving attendance, study habits, assignment completion, and classroom engagement.

The system was implemented using Python, Jupyter Notebook, Pandas, NumPy, Scikit-learn, Matplotlib, and Seaborn. Experimental evaluation conducted on a dataset containing 1,000 student records achieved an overall prediction accuracy of 89%, demonstrating the effectiveness of the proposed approach in supporting intelligent educational decision-making.

The proposed framework provides educational institutions with an efficient, interpretable, and scalable solution for student performance prediction while facilitating early intervention strategies that enhance academic outcomes and reduce failure risk.

Keywords: Artificial Intelligence, Machine Learning, Student Performance Prediction, Random Forest Classifier, Explainable AI, Academic Success Probability Score, Educational Data Mining, Predictive Analytics.

1. INTRODUCTION

The rapid advancement of Artificial Intelligence (AI) and Machine Learning (ML) has significantly transformed various sectors, including education. Educational institutions generate large volumes of academic data such as attendance records, internal assessment scores, assignment performance, classroom participation, and examination results. Analyzing this data manually is often time-

consuming and may fail to identify students who require academic support at an early stage. Consequently, intelligent predictive systems have become increasingly important for improving educational decision-making.

Student performance prediction is a key application of Educational Data Mining (EDM) and Learning Analytics. By analyzing historical academic records, machine learning models can estimate future student performance, identify academically at-risk students, and support educators in implementing timely interventions. Early prediction allows institutions to reduce dropout rates, improve academic outcomes, and provide personalized guidance to students.

Traditional student evaluation methods mainly rely on semester examinations and manual assessment conducted after academic activities are completed. Although these methods provide an overall evaluation of student achievement, they generally lack predictive capabilities and do not assist faculty members in identifying learning difficulties before final examinations. As a result, students experiencing academic challenges may not receive appropriate guidance until it is too late.

Artificial Intelligence offers an effective solution by enabling predictive analysis using multiple educational factors simultaneously. Machine learning algorithms can recognize hidden relationships between academic variables and generate reliable predictions regarding student success. Such systems not only automate the prediction process but also provide valuable insights that help educators understand the major factors affecting academic performance.

The proposed AI-Based Student Performance Prediction and Academic Success Recommendation System utilizes a Random Forest Classifier to analyze five important academic attributes: attendance, study hours, assignment scores, internal assessment marks, and classroom participation. Based on these factors, the system classifies students into three categories: High Performer, Average Performer, and At-Risk Student.

To improve the interpretability of prediction results, the proposed system incorporates Explainable Artificial Intelligence (XAI) through feature importance analysis. This enables educators to identify which academic factors contribute most significantly to student performance. In addition, the system introduces an Academic Success Probability Score (ASPS) that estimates the likelihood of academic success using weighted educational indicators. A personalized recommendation module further assists

students by suggesting practical improvements based on their individual academic profiles.

The complete system has been implemented using Python and Jupyter Notebook with libraries including Pandas, NumPy, Scikit-learn, Matplotlib, and Seaborn. Experimental evaluation on a dataset consisting of 1,000 student records achieved an overall prediction accuracy of 89%, demonstrating that the proposed framework can effectively support educational institutions in improving student learning outcomes through intelligent predictive analytics.

2. Literature Review

Student performance prediction has become an important research area within Educational Data Mining and Artificial Intelligence. Researchers have proposed numerous predictive models using machine learning techniques to analyze academic data and identify students who may require additional educational support. These studies demonstrate that predictive analytics can improve teaching strategies, optimize learning environments, and enhance institutional decision-making.

Several existing studies employ traditional classification algorithms such as Decision Trees, Naïve Bayes, Logistic Regression, Support Vector Machines (SVM), and K-Nearest Neighbors (KNN) for academic performance prediction. These approaches utilize features including attendance, examination marks, assignment scores, demographic information, and behavioral characteristics. While many of these models achieve satisfactory prediction accuracy, their performance often depends on dataset quality and feature selection.

Recent research has shown that ensemble learning algorithms, particularly Random Forest, provide superior prediction performance due to their ability to combine multiple decision trees and reduce overfitting. Random Forest has become one of the most widely adopted algorithms for educational prediction because of its robustness, scalability, and capability to handle complex relationships among multiple academic variables.

Explainable Artificial Intelligence (XAI) has also received considerable attention in educational applications. Instead of providing only prediction results, XAI techniques help educators understand the reasoning behind machine learning decisions. Feature importance analysis enables institutions to determine which academic factors most strongly influence student success, thereby improving trust, transparency, and practical decision-

making.

Several researchers have also explored recommendation systems that provide personalized academic guidance based on prediction results. These systems encourage students to improve attendance, increase study time, complete assignments regularly, and participate more actively in classroom activities. Such recommendations contribute to continuous academic improvement rather than simply reporting prediction outcomes.

Despite these advancements, many existing systems focus primarily on prediction accuracy while providing limited interpretability and post-prediction guidance. Few studies integrate performance prediction, explainable machine learning, academic success scoring, and personalized recommendations within a single intelligent educational framework.

The proposed AI-Based Student Performance Prediction and Academic Success Recommendation System addresses these limitations by combining Random Forest-based prediction, Explainable AI through feature importance analysis, the novel Academic Success Probability Score (ASPS), and a personalized recommendation engine into one unified platform. This integrated approach not only predicts academic performance with high accuracy but also provides meaningful insights and actionable recommendations that support both students and educators in achieving improved learning outcomes.

3. PROBLEM STATEMENT

Educational institutions continuously collect academic data related to student attendance, internal assessments, assignment performance, classroom participation, and examination results. However, this valuable information is often used only for record-keeping and final evaluation rather than for proactive academic decision-making. Conventional assessment methods generally identify poor-performing students only after examinations have been completed, leaving limited opportunities for timely intervention.

Manual analysis of academic records becomes increasingly difficult as the number of students grows. Furthermore, traditional evaluation approaches do not effectively reveal the relationships among multiple academic factors that influence student performance. As a result, educational institutions require an intelligent system capable of predicting student performance at an early stage while providing meaningful explanations and

personalized recommendations for academic improvement.

The primary challenge addressed in this research is the development of an AI-based predictive framework that accurately classifies student performance using multiple educational attributes while maintaining transparency, interpretability, and practical usefulness for educators and students.

4. OBJECTIVES

The objectives of the proposed AI-Based Student Performance Prediction and Academic Success Recommendation System are:

1. To develop an intelligent machine learning system capable of predicting student academic performance using multiple educational parameters.
2. To classify students into three performance categories: High Performer, Average Performer, and At-Risk Student based on academic behavior and performance indicators.
3. To improve prediction transparency through Explainable Artificial Intelligence by identifying the relative importance of each academic feature.
4. To introduce the Academic Success Probability Score (ASPS) for estimating the likelihood of student academic success using weighted educational factors.
5. To generate personalized academic recommendations that assist students in improving their overall academic performance through targeted interventions.

5. PROPOSED SYSTEM

The proposed system integrates Educational Data Mining, Machine Learning, Explainable Artificial Intelligence, and Recommendation Techniques into a unified academic decision-support framework.

Initially, the student dataset is collected and preprocessed to remove inconsistencies and prepare the data for model training. The dataset consists of five major academic features: attendance, study hours, assignment scores,

internal assessment marks, and classroom participation. These features collectively represent the academic engagement and learning behavior of individual students.

The processed dataset is divided into training and testing subsets. A Random Forest Classifier is then trained using the training dataset to learn the relationships between academic attributes and student performance categories. The trained model predicts whether a student belongs to the High Performer, Average Performer, or At-Risk category.

To improve interpretability, feature importance analysis is performed after model training. This Explainable AI component enables educators to understand which academic variables contribute most significantly to prediction outcomes.

The proposed framework further computes an Academic Success Probability Score (ASPS), which provides an additional quantitative indicator representing the student's overall academic success potential based on weighted academic attributes.

The system generates personalized recommendations by analyzing individual student weaknesses. Students with low attendance receive attendance improvement suggestions, while those with insufficient study hours, assignment performance, or classroom participation receive targeted academic guidance. This transforms the system from a simple prediction model into an intelligent educational support platform.

6. SYSTEM ARCHITECTURE



7. METHODOLOGY

The proposed AI-Based Student Performance Prediction and Academic Success Recommendation System follows a systematic machine learning workflow consisting of data collection, preprocessing, feature selection, model training, prediction, evaluation, and recommendation generation. The methodology is designed to ensure accurate prediction, model interpretability, and practical usability for educational institutions.

Step 1: Data Collection

A structured student academic dataset containing 1,000 student records was prepared for experimental analysis. Each record represents the academic profile of an individual student using multiple educational parameters. The dataset was organized in CSV format and imported into the Python environment using the Pandas library.

Step 2: Data Preprocessing

The collected dataset was preprocessed before training the machine learning model. Data preprocessing included:

- Verification of missing values
- Validation of feature consistency
- Removal of duplicate records
- Data formatting
- Feature selection

The dataset was examined using descriptive statistical analysis to ensure its quality and suitability for machine learning.

Step 3: Feature Selection

Five major academic attributes were selected as independent variables for prediction:

Feature	Description
Attendance	Student attendance percentage
Study Hours	Average daily study duration
Assignment Score	Performance in assignments
Internal Marks	Internal assessment marks
Participation	Classroom participation score

The target variable was Student Performance, which was categorized into:

- High Performer
- Average Performer
- At-Risk Student

Step 4: Data Splitting

The dataset was divided into training and testing subsets using the Train-Test Split technique.

- Training Dataset: 80%
- Testing Dataset: 20%

This division enables the machine learning model to learn from historical academic records while evaluating its performance on previously unseen data.

Step 5: Model Training

The Random Forest Classifier was selected as the primary prediction algorithm due to its high accuracy, robustness, and resistance to overfitting.

The model constructs multiple decision trees using different subsets of training data and combines their predictions through majority voting. This ensemble learning approach significantly improves prediction stability and classification performance.

Step 6: Performance Prediction

After model training, the system predicts student performance into one of three predefined academic categories:

- High Performer
- Average Performer
- At-Risk Student

The prediction process is entirely automated and requires only the academic feature values as input.

Step 7: Model Evaluation

The trained model was evaluated using standard machine learning performance metrics including:

- Accuracy Score
- Confusion Matrix
- Classification Report
- Feature Importance Analysis

These evaluation techniques provide quantitative evidence regarding the reliability and effectiveness of the prediction model.

Step 8: Explainable Artificial Intelligence

To improve transparency, Explainable AI (XAI) techniques were incorporated using Feature Importance Analysis. This enables educators to understand which academic attributes contribute most significantly toward prediction outcomes.

The feature importance analysis revealed that Attendance and Internal Marks were the two most influential factors affecting student academic performance.

Step 9: Academic Success Probability Score (ASPS)

A custom Academic Success Probability Score (ASPS) was introduced to estimate the overall academic success potential of each student.

The ASPS was calculated using weighted contributions from multiple academic indicators:

- Attendance
- Assignment Score
- Internal Marks
- Study Hours
- Classroom Participation

This score provides educators with an additional quantitative measure beyond simple categorical prediction.

Step 10: Personalized Recommendation System

Based on prediction results and ASPS values, the system automatically generates personalized academic recommendations.

Examples include:

- Improve attendance percentage.
- Increase daily study hours.
- Complete assignments regularly.
- Participate more actively during classroom sessions.
- Maintain consistent academic performance.

The recommendation module transforms the prediction model into an intelligent educational decision-support system.

8. DATASET DESCRIPTION

The experimental dataset consists of **1,000 student academic records** representing multiple educational performance indicators.

The dataset includes the following attributes:

Attribute	Data Type	Description
Student ID	Integer	Unique identifier
Attendance	Numeric	Attendance percentage

Attribute	Data Type	Description
Study Hours	Numeric	Daily study duration
Assignment Score	Numeric	Assignment evaluation marks
Internal Marks	Numeric	Internal assessment marks
Participation	Numeric	Classroom participation score
Performance	Categorical	High, Average, At-Risk

The dataset was stored in CSV format and processed using the Pandas library within Jupyter Notebook. Prior to model training, the dataset was validated to ensure the absence of missing values and duplicate records.

The dataset was designed to simulate realistic academic conditions while maintaining consistency across all educational parameters. Such preprocessing improves model reliability and enhances prediction accuracy.

9. RANDOM FOREST CLASSIFIER

Random Forest is an ensemble supervised machine learning algorithm widely used for classification and prediction problems. It constructs multiple decision trees during the training phase and combines their outputs through majority voting to generate the final prediction.

Unlike a single decision tree, Random Forest significantly reduces overfitting by introducing randomness during both data sampling and feature selection.

The algorithm was selected for this research because of the following advantages:

- High prediction accuracy
- Robust performance on structured educational datasets
- Ability to process multiple academic variables simultaneously
- Reduced overfitting compared to individual decision trees
- Built-in feature importance estimation for Explainable AI

In the proposed system, Random Forest successfully classified students into High Performer, Average Performer, and At-Risk categories while achieving an overall prediction accuracy of 89%.

The feature importance generated by the algorithm also enabled interpretation of academic factors influencing prediction outcomes, making the system more transparent and reliable for educational applications.

Software / Tool	Purpose
Python 3.12	Programming Language
Jupyter Notebook	Development Environment
Pandas	Data Processing
NumPy	Numerical Computing
Scikit-learn	Machine Learning
Matplotlib	Data Visualization
Seaborn	Statistical Visualization
Git	Version Control
GitHub	Source Code Repository
Windows 11	Operating System

REFERENCES

- [1] C. Cortes and V. Vapnik, "Support-Vector Networks," Machine Learning, vol. 20, no. 3, pp. 273–297, 1995.
- [2] L. Breiman, "Random Forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.
- [3] T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning, 2nd ed. New York, NY, USA: Springer, 2009.
- [4] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, Data Mining: Practical Machine Learning Tools and Techniques, 4th ed. Burlington, MA, USA: Morgan Kaufmann, 2017.
- [5] F. 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," in Proceedings of the 9th Python

in Science Conference (SciPy 2010), Austin, TX, USA, 2010, pp. 56–61.

- [7] J. D. Hunter, "Matplotlib: A 2D Graphics Environment," Computing in Science & Engineering, vol. 9, no. 3, pp. 90–95, 2007.
- [8] F. Chollet, Deep Learning with Python, 2nd ed. Shelter Island, NY, USA: Manning Publications, 2021.
- [9] 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.
- [10] 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: Applications and Reviews, vol. 40, no. 6, pp. 601–618, Nov. 2010.
- [11] C. Molnar, Interpretable Machine Learning, 2nd ed., 2022.
- [12] Python Software Foundation, Python Programming Language. Available: <https://www.python.org>
- [13] NumPy Developers, NumPy Documentation. Available: <https://numpy.org>
- [14] Pandas Development Team, Pandas Documentation. Available: <https://pandas.pydata.org>
- [15] Scikit-learn Developers, Scikit-learn Documentation. Available: <https://scikit-learn.org>