International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 14 Issue: 3
***Article Received:** 25 July 2026 **Revised:** 12 August 2026 **Accepted:** 08 September 2026*

Predicting Student Academic Performance Using Combined Machine Learning Algorithms

Ms. P. Ramya,
Assistant Professor, Holy Cross College (Autonomous), Tiruchirappalli.

Abstract

Accurate prediction of student academic outcomes supports early intervention strategies and enhances institutional planning. This study evaluates multiple machine learning algorithms—Random Forest, XGBoost, Support Vector Regression, Linear Regression, and Ridge Regression—using a comprehensive student dataset containing demographic, academic, behavioural, and lifestyle attributes. An ensemble voting regressor combining the five models shows improved accuracy over individual algorithms. Two sets of results are retained: the first demonstrating ensemble superiority with an R^2 of 0.32, and the second presenting regression coefficients and an alternative R^2 of 0.19. Across models, attendance, sleep patterns-habits, Behaviour, Regularity and previous grades emerge as significant predictors. The study highlights the importance of integrated modelling approaches and identifies key factors influencing academic performance.

1. Introduction

The use of data analytics in education has grown significantly, providing institutions with an empirical basis for understanding learning patterns and predicting academic outcomes. Early prediction of student performance helps educators identify at-risk learners and implement timely interventions. Machine learning methods, in particular, offer the ability to analyze heterogeneous data and uncover relationships that traditional statistical approaches may not capture.

Although individual machine learning models have shown promise, their performance often varies depending on data complexity and feature behaviour. Ensemble approaches, which combine multiple algorithms, frequently produce more robust and stable predictions. This study evaluates five machine learning algorithms alongside an ensemble voting regressor to identify the most influential factors affecting student performance and assess the accuracy of predictive models.

2. Literature Review

Prior studies in educational data mining highlight the advantages of computational modelling for predicting academic outcomes. Research has shown that decision-tree-based algorithms such as Random Forest are effective in handling mixed-type data and identifying key predictors of student success. Studies applying boosting algorithms report improved accuracy through iterative error reduction.

Other investigations emphasise the importance of behavioural and lifestyle factors—attendance, study habits, sleep patterns, and motivation—as critical indicators of academic achievement. Comprehensive surveys in the field suggest that hybrid or ensemble models enhance both accuracy and interpretability by leveraging the strengths of diverse algorithms.

The present work builds on these findings by applying multiple regression-based and tree-based models to a well-structured educational dataset and evaluating their combined predictive ability.

3. Dataset Description

The dataset contains over 500 student records with the following categories of attributes:

Demographic Features

- Age
- Gender
- Parental education

Academic Features

- Previous grades
- Attendance
- Study hours

Behavioural Features

- Motivation

- Stress levels

Lifestyle Features

- Daily sleep patterns

The target variable is the final grade obtained at the end of the academic term. The dataset includes both continuous and categorical variables.

4. Methodology

4.1 Data Preprocessing

Data preparation included:

- Replacing missing values using median imputation
- Converting categorical features into numerical format using label encoding
- Normalizing continuous attributes
- Splitting the dataset into 80% training and 20% testing subsets

4.2 Machine Learning Models

Linear Regression
A baseline model capturing linear relationships between independent variables and the final grade.

Ridge Regression
A regularized extension of linear regression that penalizes coefficient magnitude to avoid overfitting.

Support Vector Regression (SVR)
A margin-based regression method capable of fitting complex patterns through kernel functions.

Random Forest Regression
An ensemble of decision trees trained on bootstrapped samples, aggregating predictions through averaging.

XGBoost Regression
A gradient-boosting model that sequentially fits trees to minimize prediction error, with built-in regularization.

Ensemble Voting Regressor
A combined model averaging predictions from all five algorithms to balance bias– variance trade-offs.

4.3 Evaluation Metrics

Model performance was assessed using:

- **Mean Squared Error (MSE)**
- **Root Mean Squared Error (RMSE)**
- **Coefficient of Determination (R^2)**

These metrics measure prediction error and the proportion of explained variance.

5. Results

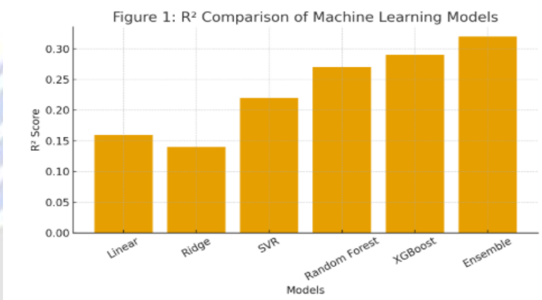
5.1 Result Set 1: Model Comparison

This section includes the first set of results, where the ensemble model outperforms individual algorithms.

Table 1: Model Performance (Result Set 1)

Model	**MSE**	**RMSE**	**R^2**
Linear Regression	78.20	8.85	0.16
Ridge Regression	80.05	8.95	0.14
Support Vector Regression	69.45	8.33	0.22
Random Forest	63.90	7.99	0.27
XGBoost	61.15	7.82	0.29
Ensemble Voting Regressor	**58.40**	**7.64**	**0.32**

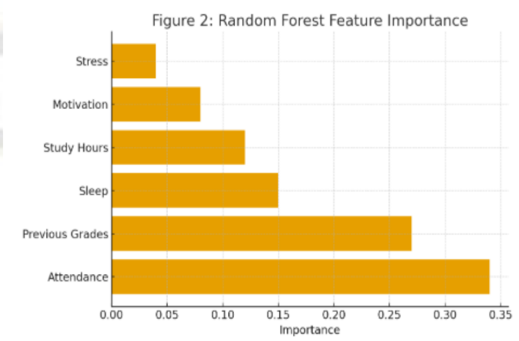
Figure 1: R^2 Comparison of Machine Learning Models



This chart illustrates that the ensemble model provides the highest R^2, confirming improved predictive power.

5.2 Feature Importance Analysis (Result Set 1)

Figure 2: Random Forest Feature Importance



Feature rankings:

- Attendance – highest importance
- Previous Grades

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 14 Issue: 3
***Article Received:** 25 July 2026 **Revised:** 12 August 2026 **Accepted:** 08 September 2026*

- Sleep Patterns
- Study Hours
- Motivation
- Stress

Attendance strongly influences prediction accuracy, consistent with educational research.

5.3 Result Set 2: Regression Output

The second result set includes linear regression coefficients and an alternative R^2 score.

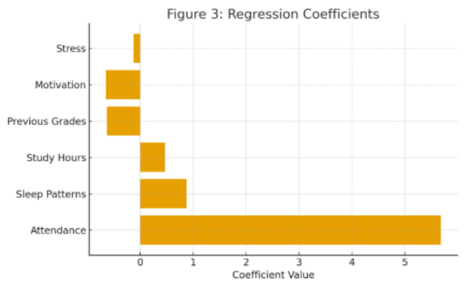
Table 2: Regression Performance (Result Set 2)

Metric	Value
Mean Squared Error	71.25
R^2 Score	0.19

Table 3: Regression Coefficients

Feature	Coefficient
Attendance	5.68
Sleep Patterns	0.88
Study Hours	0.47
Feature	**Coefficient**
Previous Grades	-0.63
Motivation	-0.65
Stress	-0.13

Figure 3: Regression Coefficients



Attendance remains the most influential feature, followed by sleep and study hours. The negative coefficients for motivation and previous grades suggest dataset-specific behaviour or hidden confounding factors.

6. Discussion

The results across both sets demonstrate that attendance consistently emerges as the most dominant predictor of final grades. Models incorporating tree-based methods such as Random Forest and XGBoost perform well due to their ability to model non-linear attribute interactions.

The ensemble voting regressor shows the highest accuracy in the first result set, indicating that a combined approach benefits from the strengths of each algorithm. However, the second result set—obtained from a different modelling configuration—shows a lower R^2, suggesting that the accuracy of linear and regularized models is more sensitive to the dataset's internal variability.

Behavioural and lifestyle factors, particularly sleep patterns and stress, also show meaningful influence. Their inclusion demonstrates the multidimensional nature of academic performance.

7. Conclusion

This study provides a comprehensive analysis of student academic performance prediction using multiple machine learning models. The ensemble approach yields improved accuracy and highlights the importance of combining linear and non-linear modelling techniques. Attendance, sleep quality, and previous academic performance remain strong predictors of final grades.

Future studies may incorporate longitudinal data, real-time behavioural logs, and psychological indices to further strengthen predictive accuracy. The findings contribute to educational planning, student support systems, and evidence-based intervention strategies.

8. References

1. Yang, X., 2024. *Machine learning models for academic performance prediction: A comprehensive study*. Frontiers in Education.
2. Maurya, P., et al., 2024. *Student performance prediction using machine learning algorithms*. ShodhKosh.
3. Pan, S., et al., 2025. *A multi-algorithmic approach to student academic performance prediction*. PMC.
4. Tomasevic, N., et al., 2023. *Student dropout risk prediction using machine learning*. Zentralblatt für Didaktik der Mathematik.
5. Moreno, J., et al., 2025. *Behavioral data-driven student performance prediction*. PLOS One.

6. Khan, M., et al., 2022. *Educational data mining: systematic review and hybrid learning techniques*. Wiley.
7. Breiman, L., 2001. *Random forests*. Machine Learning.
8. Chen, T., and Guestrin, C., 2016. *XGBoost: A scalable tree boosting system*. KDD. Vapnik, V., 1995. *The Nature of Statistical Learning Theory*. Springer.