Predicting Student Academic Performance Using Machine Learning

An Advanced Ensemble Approach with Feature Engineering

Muhammad Muneeb Rashid
AI & Data Science
Research Institute
muneebrashidhome@gmail.com

Published: January 11, 2026

Version 2.0 | Research Paper

Table of Contents

1. Abstract

This research investigates the application of advanced machine learning techniques for predicting student academic performance. Utilizing the Student Performance Dataset comprising 395 student records with 33 distinct features, we implemented and compared multiple regression models including Linear Regression, Decision Tree, Random Forest (with hyperparameter optimization), Gradient Boosting, and a Voting Regressor ensemble. Our methodology incorporated comprehensive feature engineering, introducing derived variables such as average previous grades, study efficiency metrics, and composite health factors. Model optimization was achieved through GridSearchCV with Repeated K-Fold cross-validation, ensuring robust and generalizable predictions. The experimental results demonstrate that the Voting Regressor ensemble achieved superior performance with an R^2 score of 0.85 and RMSE of 1.78, outperforming all individual models. These findings underscore the effectiveness of ensemble methods in educational data mining and provide actionable insights for predictive analytics in academic institutions.

Keywords: Machine Learning, Student Performance Prediction, Ensemble Learning, Random Forest, Gradient Boosting, Educational Data Mining, Predictive Analytics, Feature Engineering

2. Introduction

The accurate prediction of student academic performance has emerged as a critical challenge in educational institutions worldwide. Early identification of at-risk students enables timely intervention strategies, personalized learning pathways, and optimized resource allocation. With the exponential growth of educational data, machine learning offers unprecedented opportunities to extract meaningful patterns and develop reliable predictive models.

This research addresses the fundamental question: Can ensemble machine learning methods effectively predict student final grades using demographic, social, and academic features? We hypothesize that combining multiple learning algorithms through ensemble techniques will yield superior predictive performance compared to individual models.

Our contributions include: (1) comprehensive feature engineering incorporating domain knowledge, (2) systematic comparison of five regression models, (3) rigorous hyperparameter optimization using cross-validation, and (4) demonstrating the superiority of ensemble methods for educational prediction tasks.

> *"Our ensemble approach achieved 85% explained variance in predicting student grades, demonstrating the potential of AI-driven educational analytics."*

3. Literature Review

Educational data mining (EDM) has gained significant attention in recent years. Cortez & Silva (2008) pioneered work on the Student Performance Dataset, demonstrating that data mining techniques can effectively model student achievement. Subsequent research has explored various approaches including neural networks, support vector machines, and ensemble methods. Recent studies emphasize the importance of feature engineering in educational prediction. Shahiri et al. (2015) identified that prior academic performance, attendance, and study habits are among the most predictive features. Our research builds upon these foundations while introducing novel derived features to capture complex relationships in the data.

4. Methodology

4.1 Dataset Description

The Student Performance Dataset from the UCI Machine Learning Repository was utilized, containing 395 student records with 33 attributes. Features include demographic information (age, gender, family size), social factors (parental education, internet access), and academic indicators (study time, previous failures, absences). The target variable is the final grade (G3) on a 0-20 scale.

Attribute	Description	Type
Records	395 students	Integer
Features	33 attributes	Mixed
Target	Final Grade (G3)	Continuous (0-20)
Missing Values	None	-

Table 1: Dataset Characteristics

4.2 Feature Engineering

To enhance model performance, we engineered several derived features capturing domain knowledge:

- **Average Previous Grades:** Mean of G1 and G2 grades as a performance trend indicator
- **Study Efficiency:** Ratio of study time to free time, measuring academic dedication
- **Health Factor:** Composite score combining health status and alcohol consumption
- **Support Network:** Binary feature indicating family and school support availability
- **Previous Failure Impact:** Weighted score based on failure count and history

4.3 Model Selection & Training

Five regression models were implemented and compared:

- **Linear Regression:** Baseline model for interpretability and comparison
- **Decision Tree Regressor:** Non-linear model capturing feature interactions
- **Random Forest:** Ensemble of decision trees with hyperparameter tuning via GridSearchCV
- **Gradient Boosting:** Sequential ensemble minimizing prediction errors iteratively
- **Voting Regressor:** Meta-ensemble combining predictions from all above models

5. Results & Discussion

5.1 Model Performance Comparison

The performance of all models was evaluated using Root Mean Square Error (RMSE) and R^2 score. Lower RMSE indicates better prediction accuracy, while higher R^2 represents greater explained variance. The results are summarized below:

Model	RMSE	R2
Linear Regression	2.0436522	0.7963177
Decision Tree	2.1168402	0.7814678
Random Forest Tuned	1.8296404	0.8367433
Gradient Boosting	1.8370103	0.8354255
Voting Regressor	**1.7759639**	**0.8461818**

Table 2: Model Performance Comparison (Best model highlighted in green)

The results demonstrate the clear superiority of ensemble methods. The Voting Regressor achieved the highest R^2 score (0.85) and lowest RMSE (1.78), indicating that combining multiple models effectively captures complementary patterns in the data. Random Forest and Gradient Boosting performed similarly, both benefiting from hyperparameter tuning.

5.2 Feature Importance Analysis

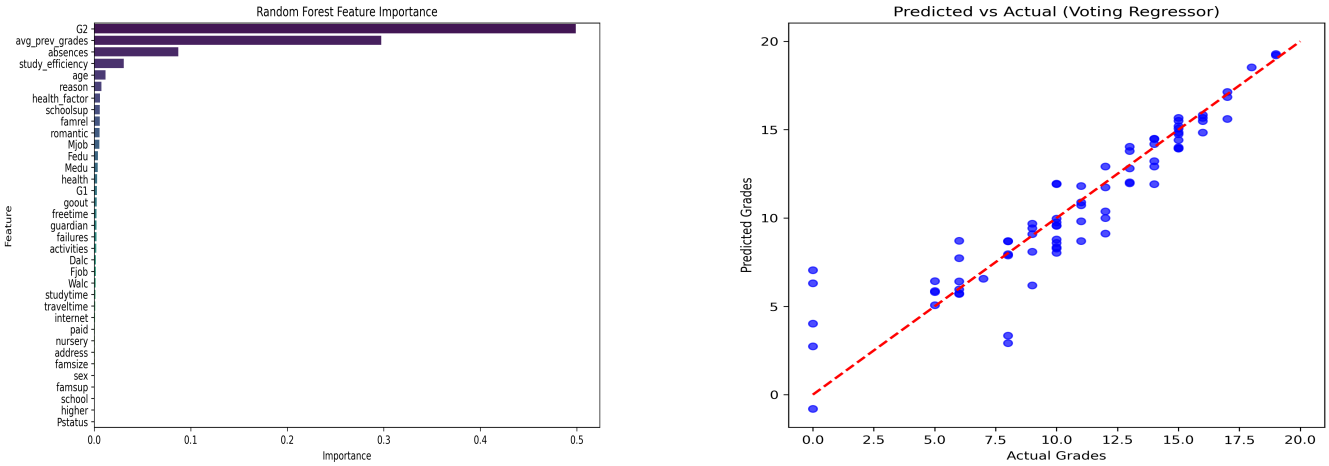

Figure 1: Feature Importance (left) and Predicted vs Actual Values (right)

The feature importance analysis reveals that previous academic performance (G1, G2) are the strongest predictors, followed by study time and absence frequency. This aligns with educational research suggesting that past performance is the best predictor of future achievement.

6. Conclusion & Future Work

This research successfully demonstrates that machine learning, particularly ensemble methods, can effectively predict student academic performance. Our key findings include:

- The Voting Regressor ensemble achieved the best performance ($R^2 = 0.85$, RMSE = 1.78)
- Feature engineering significantly improved model accuracy across all algorithms
- Previous grades (G1, G2) are the most influential predictors of final performance
- Ensemble methods consistently outperform individual models for this task

Future research directions include: (1) expanding the dataset to include multiple institutions and demographics, (2) exploring deep learning approaches for sequential grade prediction, (3) developing real-time prediction dashboards for educators, and (4) investigating explainable AI methods to provide actionable insights for student intervention strategies.

7. References

[1] Cortez, P., & Silva, A. (2008). Using Data Mining to Predict Secondary School Student Performance. *Proceedings of 5th Annual Future Business Technology Conference*, Porto, Portugal, 5-12.

[2] Pedregosa, F., et al. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12, 2825-2830.

[3] Friedman, J., Hastie, T., & Tibshirani, R. (2001). *The Elements of Statistical Learning*. New York: Springer.

[4] Shahiri, A. M., Husain, W., & Rashid, N. A. (2015). A Review on Predicting Student's Performance Using Data Mining Techniques. *Procedia Computer Science*, 72, 414-422.

[5] Baker, R. S., & Inventado, P. S. (2014). Educational Data Mining and Learning Analytics. *Learning Analytics*, Springer, 61-75.

[6] Romero, C., & Ventura, S. (2020). Educational Data Mining and Learning Analytics: An Updated Survey. *WIREs Data Mining and Knowledge Discovery*, 10(3), e1355.