Predicting a New Student's Mathematics Achievement Based on Prior Student Performance

Iyad Suleiman, Tel Hai Academic College, Israel
Rozan Abbas, The Max Stern Yezreel Valley College, Israel
Amani Attallah, The Max Stern Yezreel Valley College, Israel
Rula Jiryis, Kinneret Academic College, Israel

The IAFOR International Conference on Education in Hawaii 2026
Official Conference Proceedings

Abstract

Mathematics achievement is widely recognized as a critical predictor of academic success and future participation in science, technology, engineering, and mathematics (STEM) fields. However, many educational systems struggle to identify students at risk of underperformance early enough to provide effective intervention. This study proposes a data-driven approach for predicting a new student's mathematics achievement based on historical student performance data. Using a publicly available dataset, two predictive modeling techniques—multiple linear regression and random forest regression (see Breiman, 2001; Kuhn & Johnson, 2013)—are applied to estimate mathematics scores and identify key contributing factors. The modeling pipeline includes data preprocessing, categorical encoding, feature scaling, feature selection, and cross-validation. Model performance is evaluated using mean squared error (MSE) and the coefficient of determination (R^2). Results indicate that multiple linear regression outperforms random forest regression in both predictive accuracy and interpretability, with reading and writing scores emerging as the most influential predictors. The findings highlight the potential of interpretable machine learning models to support educational decision-making and targeted pedagogical interventions (Breiman, 2001; OECD, 2019; Siegler et al., 2012).

Keywords: mathematics achievement, educational data mining, predictive modeling, linear regression, random forest

iafor
The International Academic Forum
www.iafor.org

Introduction

Student achievement in mathematics has long been a central concern in educational research and policy due to its strong association with academic progression, employability, and long-term socioeconomic outcomes. Despite increased investment in education, substantial performance gaps in mathematics persist across diverse student populations. Traditional assessment methods often identify learning difficulties only after they have become entrenched, limiting the effectiveness of remedial interventions.

Recent advances in educational data mining and machine learning provide new opportunities to analyze large-scale educational datasets and uncover patterns not readily identifiable through conventional statistical approaches. Predictive modeling enables the early identification of students who may require academic support by leveraging historical performance indicators (see Baker & Inventado, 2014; Breiman, 2001; Kuhn & Johnson, 2013; Romero & Ventura, 2020).

The objectives of this study are twofold: (a) to predict the mathematics achievement of a new student based on prior student data and (b) to identify the most influential features contributing to mathematics performance. By comparing a parametric linear model with a nonparametric ensemble model, this research seeks to balance predictive accuracy with interpretability—an essential requirement in educational contexts.

Theoretical Background

Predictive Modeling in Education

Predictive modeling has become an increasingly influential paradigm in educational research, particularly within the broader field of educational data mining (EDM) and learning analytics. These approaches aim to leverage historical educational data to identify patterns, predict academic outcomes, and support data-informed decision-making at both instructional and policy levels. Prior research demonstrates that predictive models can effectively identify students at risk of academic failure, enabling early interventions that are timelier and more targeted than traditional assessment practices (Baker & Inventado, 2014; Romero & Ventura, 2020).

In the context of mathematics education, predictive analytics has been widely applied to examine how prior academic performance, demographic variables, and learning-related factors contribute to future achievement. Large-scale assessments such as PISA consistently reveal strong correlations between mathematics achievement and literacy-related competencies, including reading comprehension and written expression, suggesting shared underlying cognitive processes such as abstraction, symbolic reasoning, and problem representation (OECD, 2019; Siegler et al., 2012). These findings provide a theoretical justification for incorporating reading and writing scores as key predictors in models of mathematics achievement. (OECD, 2019).

Multiple Linear Regression in Educational Research

Multiple linear regression (MLR) remains one of the most widely used statistical techniques in educational research due to its interpretability, statistical rigor, and strong theoretical foundations. MLR models the linear relationship between a dependent variable and multiple

independent variables, allowing researchers to estimate both the direction and magnitude of each predictor's contribution. In educational contexts, this transparency is particularly valuable, as stakeholders such as educators, policymakers, and school administrators often require clear explanations of model behavior rather than purely predictive accuracy (Field, 2018; Gelman et al., 2020).

Despite its advantages, MLR relies on several assumptions, including linearity, independence of errors, homoscedasticity, and absence of multicollinearity. Violations of these assumptions may limit its ability to capture complex or nonlinear relationships inherent in educational data. Nevertheless, numerous studies have shown that linear models often perform competitively with more complex machine learning methods when relationships are approximately linear and when predictors are strongly correlated with the outcome variable (Gelman et al., 2020). This balance between performance and interpretability makes MLR a strong baseline model for educational prediction tasks.

Random Forest Regression and Ensemble Learning

Random forest regression is a nonparametric ensemble learning method that constructs a large number of decision trees using bootstrap sampling and random feature selection, and aggregates their predictions to produce a final estimate (Breiman, 2001). This approach is particularly effective in capturing nonlinear relationships and complex interactions between predictors without requiring explicit model specification. As a result, random forests have been widely adopted in educational data mining for tasks such as performance prediction, dropout detection, and behavioral modeling (Cutler et al., 2012; Kotsiantis et al., 2013).

However, the increased flexibility of random forest models comes at the cost of reduced interpretability. While feature importance measures provide some insight into predictor relevance, these measures may be biased toward variables with higher variability or larger numbers of distinct values (Strobl et al., 2007). In educational settings, where explainability and fairness are critical considerations, such limitations must be carefully weighed against potential gains in predictive accuracy. Consequently, comparing random forest regression with interpretable linear models offers valuable insight into the trade-offs between complexity and transparency.

Feature Selection and Interpretability

Feature selection plays a central role in predictive modeling by identifying the most informative predictors while reducing dimensionality, overfitting, and computational complexity. In educational research, effective feature selection also enhances interpretability, allowing researchers to align model outputs with established theoretical frameworks of learning and achievement. Techniques such as recursive feature elimination (RFE), permutation importance, and model-based importance scores are commonly employed to assess predictor relevance (Guyon & Elisseeff, 2003; Kuhn & Johnson, 2013).

Importantly, feature selection should not be treated as a purely algorithmic process. Domain knowledge and theoretical considerations are essential to ensure that selected features are meaningful, available at prediction time, and ethically appropriate. In this study, feature selection is used not only to improve predictive performance but also to support theoretically grounded interpretations of the factors influencing mathematics achievement.

Method

Dataset

The dataset was obtained from the Kaggle platform and includes 1,000 student records comprising demographic attributes, academic background variables, and standardized test scores. The outcome variable is mathematics achievement. Predictor variables include gender, race/ethnicity, parental level of education, lunch type, test preparation course, reading score, and writing score. A student identification variable was excluded from modeling.

Data Preprocessing

Numerical variables were standardized using z-score normalization. Categorical variables were encoded using label encoding for feature selection and one-hot encoding for regression modeling. Although no missing values were present, mean imputation was applied to numerical features to enhance robustness.

Experimental Design

The dataset was split into training (80%) and testing (20%) sets using a fixed random seed. Two models were trained: multiple linear regression and random forest regression with 200 trees. Model performance was evaluated using mean squared error (MSE) and R^2. Five-fold cross-validation was conducted to assess generalization stability (Breiman, 2001).

Results

Predictive Performance

Multiple linear regression achieved superior performance (MSE = 27.33, R^2 = .88) compared with random forest regression (MSE = 38.94, R^2 = .84), indicating predominantly linear relationships between predictors and mathematics achievement (Breiman, 2001).

Table 1

Predictive Performance of Models on the Test Set

Model	MSE ↓	R^2 ↑
Multiple Linear Regression	27.33	0.88
Random Forest Regression	38.94	0.84

Note. MSE = mean squared error; R^2 = coefficient of determination. Lower MSE values and higher R^2 values indicate better predictive performance.

Cross-Validation

Cross-validation results confirmed the robustness of both models, with linear regression demonstrating lower variance across folds.

Table 2
Five-Fold Cross-Validation Results

Model	CV MSE (M ± SD)	CV R^2 (M ± SD)
Multiple Linear Regression	31.32 ± 3.22	.866 ± .013
Random Forest Regression	40.02 ± 3.32	.829 ± .016

Feature Importance

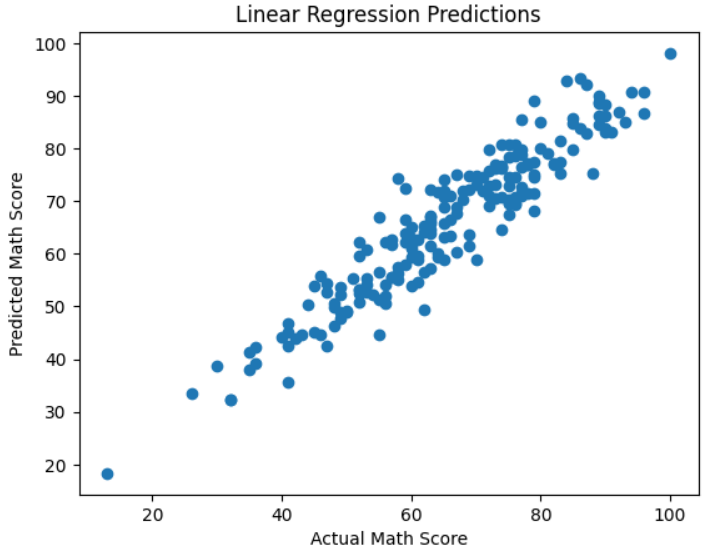
Recursive feature elimination for linear regression consistently selected reading score, writing score, gender, and test preparation course. Random forest feature importance analysis revealed that reading and writing scores accounted for the majority of predictive contribution.

Table 3
Top Feature Importances Identified by Random Forest Regression

Feature	Importance
Reading score	0.606
Writing score	0.165
Gender (male)	0.119
Student ID	0.039
Test preparation course (none)	0.014
Lunch type (standard)	0.013
Race/ethnicity (Group E)	0.01
Parental education (some college)	0.006
Race/ethnicity (Group C)	0.005
Parental education (some high school)	0.005

Graphical Interpretation (APA Style)

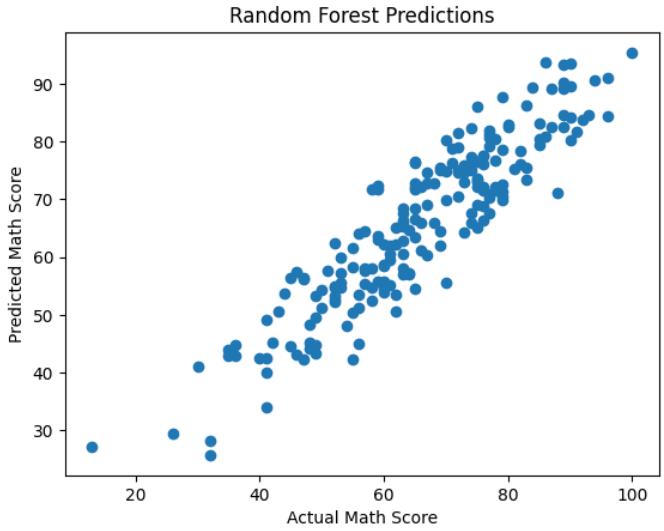
Figure 1
Observed vs. Predicted Mathematics Scores Using Multiple Linear Regression



Note. Each point represents a student observation. The diagonal line indicates perfect prediction. The strong alignment along the diagonal reflects high predictive accuracy.

Figure 2

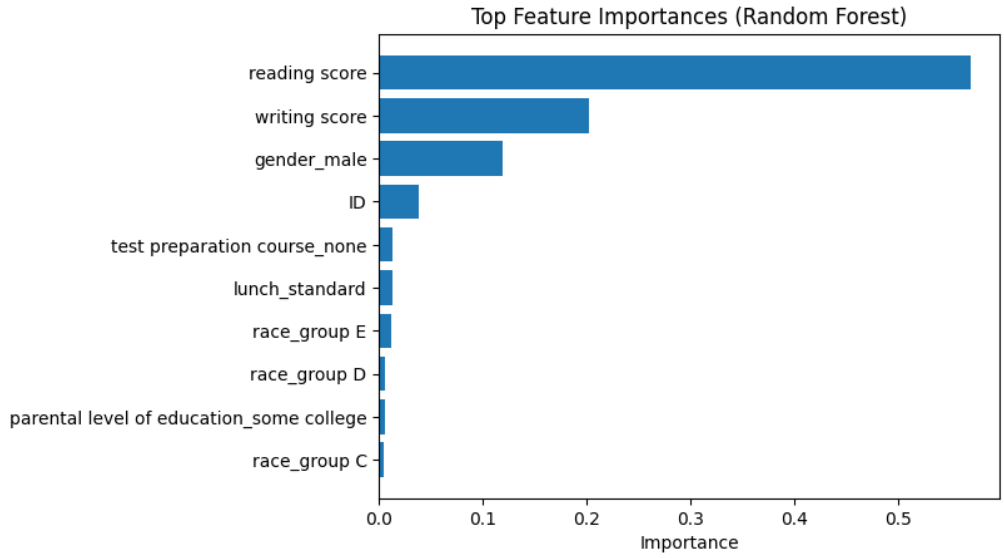
Observed vs. Predicted Mathematics Scores Using Random Forest Regression



Note. Predictions show greater dispersion compared with linear regression, indicating reduced accuracy and increased variance.

Figure 3

Feature Importance Rankings From the Random Forest Model



Note. Reading and writing scores contribute the majority of predictive power, highlighting the central role of literacy-related skills in mathematics achievement.

Discussion

The results reinforce existing educational research demonstrating strong associations between literacy-related skills and mathematics achievement. Although random forest regression captured complex patterns, its added complexity did not yield superior predictive performance. The interpretability and stability of linear regression make it more suitable for educational applications where transparency is essential (Breiman, 2001).

Conclusion

This study demonstrates the effectiveness of interpretable machine learning models for predicting mathematics achievement. Multiple linear regression, combined with feature selection and cross-validation, provides a reliable and transparent approach for early identification of students at risk. Future research may incorporate longitudinal data, hybrid nonlinear models, or causal inference techniques.

Declaration of Generative AI and AI-Assisted Technologies in the Writing Process

The authors declare that ChatGPT (OpenAI), a generative AI–assisted language tool, was used during the preparation of this manuscript. The use of the tool was limited to language editing, stylistic refinement, structural reorganization, and assistance with formatting according to conference submission guidelines.

ChatGPT was not used to generate original research ideas, data, analyses, results, or interpretations. All research design, methodological decisions, data analysis, findings, and conclusions are the sole work and responsibility of the authors. The authors reviewed and edited all AI-assisted content to ensure accuracy, academic integrity, and compliance with scholarly standards.

References

Baker, R. S., & Inventado, P. S. (2014). Educational data mining and learning analytics. In J. A. Larusson & B. White (Eds.), *Learning analytics: From research to practice* (pp. 61–75). Springer. https://doi.org/10.1007/978-1-4614-3305-0_4

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Cutler, A., Cutler, D. R., & Stevens, J. R. (2012). Random forests. In C. Zhang & Y. Ma (Eds.), *Ensemble machine learning* (pp. 157–175). Springer. https://doi.org/10.1007/978-1-4419-9326-7_5

Field, A. (2018). *Discovering statistics using IBM SPSS statistics* (5th ed.). Sage.

Gelman, A., Hill, J., & Vehtari, A. (2020). *Regression and other stories*. Cambridge University Press. https://doi.org/10.1017/9781139161879

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research, 3*, 1157–1182.

Kotsiantis, S. B., Zaharakis, I., & Pintelas, P. (2013). Machine learning: A review of classification and combining techniques. *Artificial Intelligence Review, 26*(3), 159–190. https://doi.org/10.1007/s10462-007-9052-3

Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling*. Springer. https://doi.org/10.1007/978-1-4614-6849-3

Organisation for Economic Co-operation and Development. (2019). *PISA 2018 results*. OECD Publishing. https://doi.org/10.1787/5f07c754-en

Romero, C., & Ventura, S. (2020). Educational data mining and learning analytics: An updated survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 10*(3), e1355. https://doi.org/10.1002/widm.1355

Siegler, R. S., Duncan, G. J., Davis-Kean, P. E., Duckworth, K., Claessens, A., Engel, M., Susperreguy, M. I., & Chen, M. (2012). Early predictors of high school mathematics achievement. *Psychological Science, 23*(7), 691–697. https://doi.org/10.1177/0956797612440101

Strobl, C., Boulesteix, A.-L., Zeileis, A., & Hothorn, T. (2007). Bias in random forest variable importance measures. *BMC Bioinformatics, 8*, 25. https://doi.org/10.1186/1471-2105-8-25