Article

Forecasting Student Academic Performance Using Machine Learning

Danial Polat [1] and Azamat Serek* [2]

[1]Department of Computer Science, SDU University, Kaskelen, Kazakhstan
[2]School of Information Technologies and Engineering, Kazakh-British Technical University (KBTU), Almaty, Kazakhstan

DOI: 10.47344/w85rct27

Abstract

Educational data mining depends on accurate student academic outcome forecasting to detect students who need help early and receive specific support. Traditional linear models have been used extensively yet they fail to detect the intricate non-linear patterns which exist in student achievement data. The evaluation of machine learning algorithms and their features for student outcome prediction in Portuguese secondary education remains insufficient because of missing systematic assessments. The research investigates how Linear Regression and Random Forest and K-Nearest Neighbors perform when predicting Portuguese language grades from 649 student records containing 30 demographic and social and academic attributes. The evaluation of model performance used three established metrics which included Mean Squared Error (MSE) and R-Squared (R^2) and Mean Absolute Error (MAE). The results showed Linear Regression produced the most accurate predictions through its lowest MSE (9.00) and MAE (2.30) values but its weak R^2 value (0.01) indicated poor explanatory power. The error rates of Random Forest matched those of Linear Regression (MSE = 9.48 and MAE = 2.34) yet its negative R^2 (-0.04) indicated poor generalization because of irrelevant features and suboptimal hyperparameters. The KNN model showed the worst results (MSE = 11.10 and MAE = 2.57 and R^2 = -0.21) because it failed to detect important patterns without additional optimization. The results show that educational prediction tasks require both optimal feature selection and parameter adjustment for successful results. The research shows that linear models perform better than complex methods in specific situations yet optimized non-linear models demonstrate superior ability to understand student achievement complexity. The research provides essential guidelines for developing better feature engineering and machine learning approaches to predict educational results.

Keywords: machine learning education, education artificial intelligence, edtech, AI in edtech, predictive power education.

Email (1): 210107170@stu.sdu.edu.kz ORCID: 0009-0006-1576-5120
*Corresponding author:
Email: a.serek@kbtu.kz ORCID: 0000-0001-7096-6765

Received: September 23, 2025. Reviewed: September 24, 2025. Accepted: September 30, 2025.

I. Introduction

The achievements that students achieve in secondary education influence their individual development and have beneficial repercussions on their communities. Students who are successful in their academic endeavors throughout this time frame will have access to both future work prospects and high-quality university education. This enhances economic security, long-term health results, and the quality of society. Due to its impact on students, teachers, and administrative personnel, accurate student performance prediction is therefore a very important study topic.

Through the use of student grade projections, teachers can identify difficult students early on and provide them with targeted assistance that will enhance their academic performance. Predictive data assists educators in developing personalized lesson plans, which reduces student dropout rates and maximizes the use of educational resources. Data-driven policymaking allows for evidence-based decisions to be made for improving student achievement rather than relying on conjecture.

The prediction problems show significant efficacy for machine learning (ML). Because machine learning algorithms can handle high-dimensional and non-linear interactions in data, they are far more effective than traditional statistical approaches for identifying complex patterns in educational datasets. Because ML algorithms can handle a variety of variables, such as demographic data, sociological and behavioral aspects, and academic measures, they are highly useful for modeling purposes given the many circumstances that affect students' successes. Researchers and practitioners can better anticipate student outcomes and gain a better understanding of school performance metrics by using machine learning techniques.

In this work, we study into the prediction of Portuguese language grades in the Portuguese secondary school system using machine learning techniques. The used dataset comprises 649 student records with 30 attributes including academic achievement measures, sociocultural factors, and demographic data [1]. We assess three predictive machine learning models—Linear Regression, Random Forest, and K-Nearest Neighbors (KNN)—based on our analysis. The three models demonstrate different analytical techniques, with KNN operating on local data inner structures, Random Forest assisting in handling non-linear patterns through ensemble learning, and Linear Regression acting as a statistical baseline. Together, the three models enable evaluation of the models' methodological stability and predictive power.

Although machine learning has been widely applied to educational prediction tasks , most prior studies either focus on broad achievement outcomes or use datasets from non-Portuguese contexts. As a result, there is limited empirical evidence on how different models perform when applied specifically to Portuguese language grades in the secondary school system. This research addresses that gap by systematically benchmarking traditional and advanced models on this underexplored dataset.

The aim of the study. is therefore to evaluate the effectiveness and limitations of commonly used machine learning algorithms in predicting Portuguese language grades, while providing insights into the methodological challenges of educational data mining.

There are 3 main objectives of our study:

1) To thoroughly compare Random Forest, KNN, and Linear Regression for the job of predicting grades in Portuguese.
2) To evaluate model performance using common evaluation measures, such as Mean Absolute Error (MAE), R-Squared (R^2), and Mean Squared Error (MSE), in order to ascertain generalizability and predictive accuracy.
3) To explore methodological and practical implications of applying machine learning to educational prediction tasks, with particular attention to feature selection, model interpretability, and risks of overfitting.

The novelty of our work lies not in the algorithms themselves, which are well established, but in their systematic application to Portuguese secondary school data. By explicitly benchmarking simple and advanced models against a baseline, we reveal the limitations of widely used demographic and sociocultural predictors and highlight the conditions under which complex models such as Random Forest fail to outperform simpler ones. This study therefore contributes unique empirical evidence about the boundaries of current machine learning approaches in educational prediction.

Through these contributions, the study aims to advance the growing field of educational data mining and learning analytics. Beyond theoretical significance, the findings are intended to provide actionable insights for educators and policymakers, supporting evidence-based strategies to enhance student success and equity in secondary education.

II. Literature Review

The research examines how student achievement prediction methods have progressed from basic statistical methods to contemporary machine learning approaches. The initial research used linear regression models for prediction yet these models failed to detect non-linear educational data relationships because they lack interpretability. The field now uses Random Forests and K-Nearest Neighbors (KNN) algorithms for superior performance because these models detect hidden patterns and model non-linear relationships in educational data. The evaluation strategy needs to include both traditional metrics MSE and R^2 and ethical factors

Received: September 23, 2025. Reviewed: September 24, 2025. Accepted: September 30, 2025.

such as fairness and interpretability to achieve accurate model performance. The literature review concludes by detailing advanced predictive machine learning models and future research paths that encompass multimodal and contextual data integration and emphasize the continuing difficulties and opportunities to develop effective predictive systems.

A. Traditional Approaches to Student Performance Prediction

Initally, researchers of academic outcome prediction systems utilized linear regression mathematical models to investigate how student grades relate to various factors such as demographic factors and academic background and socioeconomic status [2]. The utility of linear models in educational datasets analysis is still a widespread utility because they are easy to understand but these models fail to detect the non-linear educational patterns that exist in the datasets [3]. Linear models predict that every one-unit change in prior academic achievement will produce the same effect on the outcome regardless of the student's starting performance level. The model fails to recognize essential non-linear patterns because it does not detect when students reach a point where their efforts stop producing meaningful progress or when they need to reach a specific grade to start improving their performance. The complex student success patterns require alternative modeling approaches because linear models fail to capture these intricate relationships.

B. Emergence of Machine Learning in Educational Prediction

Researchers now use machine learning methods including decision trees and Random Forests and Support Vector Machines (SVM) and K-Nearest Neighbors (KNN) to address these research challenges. The models achieve better prediction results because they detect complex student data relationships and hidden patterns in student datasets according to ref4 and ref5. The **Random Forest** model uses multiple decision trees to generate predictions through training each tree on different subsets of data and features. The ensemble method provides stability through tree bias reduction because it averages out individual tree errors [6]. The KNN algorithm makes predictions for new instances through the analysis of their distance relationships in feature space. The KNN algorithm uses the 'k' most similar students who share characteristics like grades and course load and attendance patterns to detect specific achievement trends in student data [7]. The success of these methods depends on the quality and relevance of the selected features because they need to identify non-obvious patterns in the data according to [8].

C. Feature Selection in Predictive Modeling

The selection of appropriate features stands as a fundamental problem in predictive modeling because ML algorithms achieve their best results when they use relevant and high-quality predictor variables. The selection of inadequate features leads to decreased accuracy and reduced interpretability because it brings in unneeded noise and useless information (ref9). Researchers have developed multiple solutions to handle this problem. The RFE method removes features step by step starting with the least important ones while training models on the remaining features until it reaches the best combination of features. The Boruta algorithm uses statistical comparison between real features and their 'shadow' counterparts to identify all relevant features. Research indicates that the addition of behavioral and emotional features to academic and demographic data leads to better predictive results according to recent studies [10], [11]. The predictive machine learning model benefits from three types of student datasets that encompass their platform engagement factors and textual assignment duration and their self-esteem and motivation assessment results of surveys. The inclusion of various non-standard factors aids researchers to investigate behaviour of students via sophisticated perspectives instead of relying solely on traditional academic records.

D. Evaluation Metrics in Educational Prediction

The performance of predictive models in education whether successful or not is usually evaluated using math metrics such as Mean Squared Error (MSE), R-Squared (R^2), and Mean Absolute Error (MAE) [12]. MSE aids to measure the average of the squared differences between the predicted values and actual values, and penalizing huge errors more strongly. MAE helps to provide a more intuitive measure of the average magnitude of the notion of error, showcasing the average distance between predicted and actual university outcomes. R^2 showcases the proportion of the variance in the dependent variable of the dataset that is predictable from the independent variable(s), suggesting a fast way to comprehend how well the model's predictions aids to explain the data. However, overreliance on these metrics alone may overlook wider aspects of model utility, including fairness, interpretability, and generalizability across various educational and university contexts [13]. A model might get high accuracy on one student dataset

Received: September 23, 2025. Reviewed: September 24, 2025. Accepted: September 30, 2025.

but perform neither good nor bad or exhibit systematic bias against another, potentially leading to serious bad outcomes. Therefore, a comprehensive sophisticated evaluation must consider not just predictive accuracy but also the ethical implications and practical consequences of the model's predictions.

E. Non-Linear Models and Future Directions

Not considering Random Forest and KNN algorithms, advanced non-linear mathematical models such as SVMs and Neural Networks have shown good potential in educational prediction tasks in universities [14]. SVMs can define optimal hyperplanes for classification purposes, utilizing a "kernel trick" to make map of data into higher-dimensional mathematical spaces where non-linear relationships can become linear and thus separable notions. Neural Networks, meanwhile, can define and model sophisticated, hierarchical learning patterns by processing data through multiple layers of interconnected nodes [15]. Each layer of the networks learns increasingly abstract factors, aiding the network to find intricate relationships that are often missed by other machine learning models. Despite their promise, these complex methodologies require very much of computational resources and careful parameter tuning, which may restrict their scalability in large-scale educational and university environments [16].

Future research on educational predictions should prioritize enhancements in feature selection methodologies, the integration of various so called multimodal data sources (for instance, online activity logs, psychological indicators, and behavioral datasets), and the inclusion of contextual mathematical variables to obtain the multifaceted nature of academic performance in universities [17], [18]. These integrative and iterative approaches may provide more complex insights into student learning and aid the development of predictive frameworks with greater practical applicability and utility [19], [20]. For instance, combining a set of student's past grades with their participation in online forums and their response to psychological online and offline surveys could create a much more robust and sophisticated predictive profile.

III. Methods

This section of methods showcases description of the dataset, target variables, applied preprocessing procedures, chosen machine learning models, and evaluation strategy of the models used in this study. The provided methodological design has been developed to ensure both the reliability of the conducted analysis and the interpretability of the obtained findings.

A. Dataset Description

The study utilized the publicly available dataset taken from [1], which contains 649 student rows with 30 columns collected from two Portuguese secondary schools. The described dataset contains demographic information, societal and family background, and school-related features. These attributes were collected through a set of school reports and structured questionnaires, ensuring a comprehensive complex perspective on student characteristics.

Two subject-specific datasets are provided within the described data source: one for math and one for Portuguese language subject. Given the importance of language proficiency in academic achievement in educational institutions and its strong link to other subjects, the Portuguese language dataset was selected for this research work.

B. Target Variables

The utilized dataset of the study entails three subject-specific grade variables used in the study:

- **G1**: First-period grade,
- **G2**: Second-period grade,
- **G3**: Final grade (which is used as the main prediction target).

While G1 and G2 capture intermediate assessments, G3 represents the final outcome of student performance and therefore served as the dependent variable in this study. The input features covered diverse dimensions such as parental education, daily commute time, extracurricular activities, and family background, enabling a multidimensional analysis of factors influencing academic achievement.

C. Preprocessing and Data Partitioning

Prior to model development, the dataset was carefully inspected for missing values, inconsistencies, and anomalies. Appropriate preprocessing steps were applied to ensure data quality and suitability for machine learning analysis. To evaluate model generalizability, the dataset was divided into training (80%) and testing (20%) subsets. The training subset was employed to fit the models, while the testing subset was reserved exclusively for performance evaluation.

Received: September 23, 2025. Reviewed: September 24, 2025. Accepted: September 30, 2025.

D. Model Selection Rationale

Three machine learning algorithms were selected for comparative evaluation: Linear Regression, Random Forest, and K-Nearest Neighbors. The rationale for choosing these models was twofold: (i) methodological diversity, capturing linear, ensemble, and instance-based approaches, and (ii) suitability to the characteristics of educational data.

1) Linear Regression: Linear Regression was employed as the baseline model due to its interpretability and capacity to identify direct, linear associations between predictors and outcomes. The model assumes a linear relationship between the dependent variable y and the independent variables $x_1, x_2, \ldots, x_p$. Formally, the regression function is expressed as

$$\hat{y} = \beta_0 + \sum_{i=1}^{p} \beta_i x_i + \epsilon, \tag{1}$$

where $\hat{y}$ denotes the predicted grade, β_0 is the intercept, β_i are the regression coefficients, and ϵ represents the error term. In the context of educational research, Equation 1 provides insights into how demographic and behavioral variables, such as parental education or study time, contribute to academic outcomes. This actually makes Linear Regression mathematical model as an appropriate good first step in comparative modeling.

2) Random Forest: Random Forest has been chosen in the study as the ensemble learning math method due to its proven robustness in handling heterogeneous data types in data sets and complex sophisticated feature interactions. The model actually constructs an ensemble or so called set of decision trees, each trained on a bootstrap sample of the dataset, while randomly choosing subsets of attributes for splitting purposes. The final prediction of the model is actually obtained by aggregating the outputs of individual trees of decision tree constructs, such that

$$\hat{y} = \frac{1}{T} \sum_{t=1}^{T} h_t(x), \tag{2}$$

where $h_t(x)$ defines the prediction of target of the t-th tree and T is the total number of decision trees in the constructed forest. As Equation 2 showcases, the averaging mechanism employed which reduces variance and aids to mitigate overfitting issue, which is very important in educational datasets and data storages characterized by non-linear and hierarchical relationships among utilizedvariables. Additionally, Random Forest aids feature importance analysis, improving interpretability alongside predictive performance.

3) K-Nearest Neighbors: K-Nearest Neighbors was included as a non-parametric, instance-based learning method. Unlike parametric models, KNN does not assume a predefined functional relationship between predictors and outcomes. Instead, predictions are derived based on the average of the k closest training samples in the feature space, typically measured using Euclidean distance:

$$d(x, x_i) = \sqrt{\sum_{j=1}^{p} (x_j - x_{ij})^2}, \tag{3}$$

where x is the query point, x_i represents a training instance, and p denotes the number of features. The predicted outcome is then calculated as

$$\hat{y} = \frac{1}{k} \sum_{i \in \mathcal{N}_k(x)} y_i, \tag{4}$$

where $\mathcal{N}_k(x)$ denotes the set of k nearest neighbors of x. As shown in Equation 4, KNN leverages local information within the feature space, making it effective in detecting clusters or subgroups of students with similar socio-demographic and academic characteristics. Its flexibility in adapting to irregular decision boundaries provides a useful complement to the other two models.

E. Model Training and Evaluation

Each model was trained on the training dataset and subsequently evaluated on the testing dataset in order to assess predictive accuracy and generalizability. Three widely used evaluation metrics were employed: Mean Squared Error (MSE), R-Squared (R^2), and Mean Absolute Error (MAE). These metrics capture complementary aspects of model performance, ensuring a balanced analysis.

Received: September 23, 2025. Reviewed: September 24, 2025. Accepted: September 30, 2025.

a) Mean Squared Error (MSE).: The MSE quantifies the average squared deviation between the predicted grades $\hat{y}_i$ and the actual observed grades y_i for n students:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2. \tag{5}$$

This metric penalizes larger errors more heavily, making it useful for detecting models that occasionally make extreme mispredictions. Lower values of MSE indicate better predictive accuracy.

b) R-Squared (R^2).: The R^2 statistic measures the proportion of variance in the dependent variable explained by the model, relative to a baseline model that predicts only the mean of the observed values. It is calculated as:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}, \tag{6}$$

where $\bar{y}$ denotes the mean of the observed values. Values of R^2 close to 1 indicate strong explanatory power, while values near 0 or negative suggest weak or no predictive ability compared to the baseline.

c) Mean Absolute Error (MAE).: The MAE provides a measure of the average absolute difference between predicted and actual grades:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \hat{y}_i\right|. \tag{7}$$

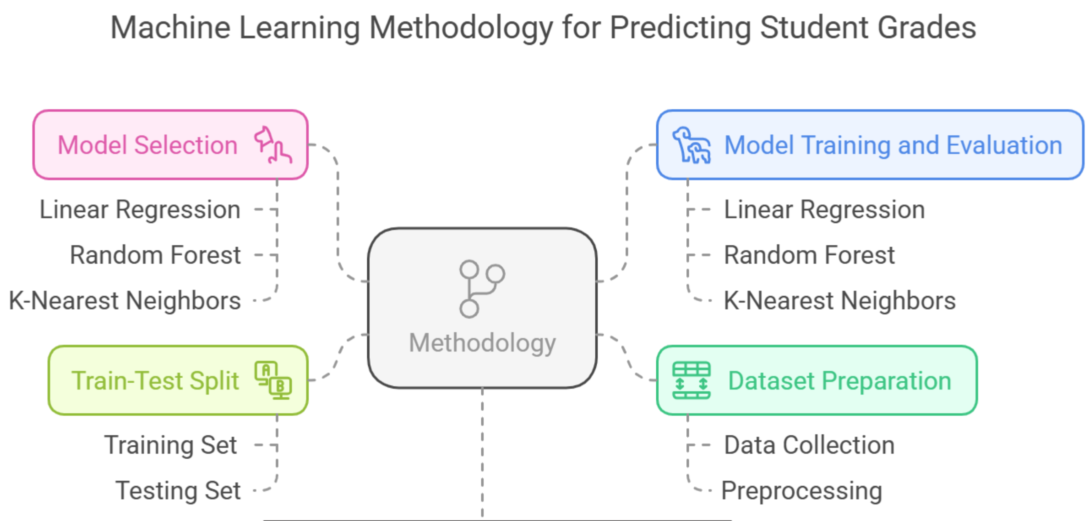
Unlike MSE, this metric treats all errors proportionally, without disproportionately penalizing large deviations. As such, it offers an interpretable measure of average prediction error in grade units.

d) Comparative Evaluation.: By jointly considering the metrics defined in Equations 5–7, the analysis captures different dimensions of predictive performance: overall error magnitude (MSE), explanatory power (R^2), and average deviation (MAE). This triangulation allows for a more comprehensive evaluation of the three models, helping to identify not only the most accurate predictor but also the methodological trade-offs underlying their performance.

F. Methodological Framework

The overall methodology is summarized in Figure 1. The process begins with dataset collection and preprocessing, followed by partitioning into training and testing subsets. The three machine learning models (LR, RF, and KNN) are trained and evaluated using the defined metrics. Finally, the models are compared to determine predictive effectiveness and to derive methodological and practical implications.

Received: September 23, 2025. Reviewed: September 24, 2025. Accepted: September 30, 2025.



Metric	Formula
True positive rate, recall	$\frac{\text{TP}}{\text{TP+FN}}$
False positive rate	$\frac{\text{FP}}{\text{FP+TN}}$
Precision	$\frac{\text{TP}}{\text{TP+FP}}$
Accuracy	$\frac{\text{TP+TN}}{\text{TP+TN+FP+FN}}$
F-measure	$\frac{2 \cdot \text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$

Fig. 1: Proposed methodology to predict grades of students.

IV. Results and Discussion

The evaluation of the three predictive models yielded mixed outcomes, as summarized in Table I. While all models produced relatively low error scores, their explanatory power, measured by R^2, was consistently limited. This indicates that the available features do not capture enough of the underlying variability in student performance to support strong predictions.

TABLE I: Summary of Experimental Results

Model	**MSE**	R^2	**MAE**
Linear Regression	9.00	0.01	2.30
Random Forest	9.48	-0.04	2.34
K-Nearest Neighbors	11.10	-0.21	2.57

The Linear Regression model produced the lowest MSE at 9.00 and MAE at 2.30 which indicated its predictions were slightly more accurate than the other models. The R^2 value of 0.01 from Linear Regression indicates that the model explained less than 1% of the variance in student outcomes. The model generated predictions that matched observed grades but failed to identify any significant connections between predictors and the target variable.

Received: September 23, 2025. Reviewed: September 24, 2025. Accepted: September 30, 2025.

The Random Forest model generated results that matched Linear Regression with an MSE of 9.48 and an MAE of 2.34. The model performed worse than a simple mean-based baseline according to its negative R^2 value of -0.04. The model failed to use its theoretical advantages because the dataset lacked sufficient information to enable the exploitation of ensemble method capabilities. Similar observations have been reported in related studies, where Random Forest underperformed in small or low-dimensional educational datasets that lacked rich behavioral features. In such contexts, the model's strength in capturing complex interactions does not provide an advantage, and overfitting risks increase. The model failed to utilize its theoretical advantages because of poor parameter settings and uninformative features that prevent it from exploiting ensemble method capabilities.

K-Nearest Neighbors delivered the worst results because it produced the highest MSE (11.10) and MAE (2.57) together with the lowest R^2 (-0.21). The results show that student performance cannot be predicted through local feature space similarity. The method shows limited ability to identify general patterns because it depends heavily on neighbor selection and distance metric choices which results in poor generalization performance.

The dataset fails to generate strong predictions because all models show low R^2 values. The selected features which include demographic and social background information fail to represent all factors that influence academic achievement. This finding aligns with previous research emphasizing that demographic features alone are insufficient, and integrating motivational, behavioral, and contextual elements leads to more robust predictors. The analysis indicates that researchers need to acquire additional data which includes behavioral information and motivational elements and contextual elements. The models will improve their ability to explain student outcomes through the addition of attendance records and classroom participation data and socioeconomic status information.

Research should focus on two main methodological improvements for future studies. The combination of feature selection techniques with dimensionality reduction methods will help researchers identify key variables while Gradient Boosting and Neural Networks can handle complex data non-linearities and noise better. These approaches have shown superior performance in comparable educational prediction studies, suggesting their potential applicability here. The process of developing meaningful predictors requires domain knowledge integration because educators should use their expertise to build predictors that reflect actual learning processes.

The best error scores from Linear Regression do not change the fact that using traditional demographic and background information alone remains insufficient for educational prediction. The research requires three essential steps to progress: enhancing data quality and adding new features and developing better models. A key limitation of this study is that it relies solely on Portuguese student records, which may restrict the generalizability of the findings to other educational systems and cultural contexts. Future studies should validate the proposed approaches using diverse datasets to ensure broader applicability. Future studies that address these research areas will create predictive frameworks which deliver practical value beyond basic comparison models.

V. Conclusion and Future Work

The research evaluated the performance of Linear Regression (LR), Random Forest (RF), and K-Nearest Neighbors (KNN) models for predicting Portuguese language course grades of secondary school students. The evaluation of model accuracy and fit used Mean Squared Error (MSE), R-squared, and Mean Absolute Error (MAE) metrics.

The Linear Regression model achieved the best results through its lowest MSE and MAE values which showed minimal prediction errors. The low R-squared value revealed a weak linear connection between student grades and features, which indicated that the linear model did not handle the data complexity effectively. The Random Forest model showed poor performance because its MSE values were slightly higher and its R-squared value was negative, which indicated its inability to handle the data distribution. The K-Nearest Neighbors model demonstrated the worst performance because it produced the highest MSE and most negative R-squared value, which proved its inability to forecast student grades accurately.

The study results show that the selected features do not strongly predict student performance, and the models face problems with overfitting and suboptimal parameter settings. The originality of this study lies in its systematic benchmarking of traditional and advanced machine learning models on Portuguese secondary school data, revealing not only their limited predictive power but also the surprising underperformance of Random Forest in this context. These findings provide practical implications for educators and policymakers: relying solely on demographic and sociocultural features is insufficient, and integrating behavioral and motivational variables is essential for building effective prediction frameworks.

Future research should use non-linear models together with advanced machine learning techniques to provide a better understanding of the factors that influence student achievement. In particular, improving feature selection and expanding datasets across different educational systems will be necessary to enhance both accuracy and generalizability, thereby strengthening the practical impact of predictive models in education.

Received: September 23, 2025. Reviewed: September 24, 2025. Accepted: September 30, 2025.

References

[1] Cortez, P. (2014). Student Performance [Data set]. UCI Machine Learning Repository. Retrieved from https://archive.ics.uci.edu/ml/datasets/Student+Academics+Performance.

[2] Namoun, A., & Alshanqiti, A. (2020). Predicting student performance using data mining and learning analytics techniques: A systematic literature review. *Applied Sciences*, 11(1), 237.

[3] Syed Mustapha, S. M. F. D. (2023). Predictive analysis of students' learning performance using data mining techniques: A comparative study of feature selection methods. *Applied System Innovation*, 6(5), 86.

[4] Kumari, P., Jain, P. K., & Pamula, R. (2018, March). An efficient use of ensemble methods to predict students academic performance. In *2018 4th International Conference on Recent Advances in Information Technology (RAIT)* (pp. 1-6). IEEE.

[5] Alhassan, A., Zafar, B., & Mueen, A. (2020). Predict students' academic performance based on their assessment grades and online activity data. *International Journal of Advanced Computer Science and Applications*, 11(4).

[6] Dutt, A., Ismail, M. A., & Herawan, T. (2017). A systematic review on educational data mining. *IEEE Access*, 5, 15991-16005.

[7] Shao, Z., Bian, H., Chen, Y., Wang, Y., Zhang, J., & Ji, X. (2021). Transmil: Transformer based correlated multiple instance learning for whole slide image classification. *Advances in Neural Information Processing Systems*, 34, 2136-2147.

[8] Altaf, S., Soomro, W., & Rawi, M. I. M. (2019, April). Student performance prediction using multi-layers artificial neural networks: A case study on educational data mining. In *Proceedings of the 2019 3rd International Conference on Information System and Data Mining* (pp. 59-64).

[9] Aljohani, N. R., Fayoumi, A., & Hassan, S. U. (2019). Predicting at-risk students using clickstream data in the virtual learning environment. *Sustainability*, 11(24), 7238.

[10] Aman, F., Rauf, A., Ali, R., Iqbal, F., & Khattak, A. M. (2019, July). A predictive model for predicting students academic performance. In *2019 10th International conference on information, intelligence, systems and applications (IISA)* (pp. 1-4). IEEE.

[11] Amra, I. A. A., & Maghari, A. Y. (2017, May). Students performance prediction using KNN and Naïve Bayesian. In *2017 8th international conference on information technology (ICIT)* (pp. 909-913). IEEE.

[12] Shin, D., & Shim, J. (2021). A systematic review on data mining for mathematics and science education. *International Journal of Science and Mathematics Education*, 19(4), 639-659.

[13] Liu, W., Wu, J., Gao, X., & Feng, K. (2017, December). An early warning model of student achievement based on decision trees algorithm. In *2017 IEEE 6th International Conference on Teaching, Assessment, and Learning for Engineering (TALE)* (pp. 517-222). IEEE.

[14] Raga, R. C., & Raga, J. D. (2019, July). Early prediction of student performance in blended learning courses using deep neural networks. In *2019 International Symposium on Educational Technology (ISET)* (pp. 39-43). IEEE.

[15] Hashim, A. S., Awadh, W. A., & Hamoud, A. K. (2020, November). Student performance prediction model based on supervised machine learning algorithms. In *IOP Conference Series: Materials Science and Engineering* (Vol. 928, No. 3, p. 032019). IOP Publishing.

[16] Altabrawee, H., Ali, O. A. J., & Ajmi, S. Q. (2019). Predicting students' performance using machine learning techniques. *Journal of University of Babylon for Pure and Applied Sciences*, 27(1), 194-205.

[17] Farissi, A., & Dahlan, H. M. (2020, April). Genetic algorithm based feature selection with ensemble methods for student academic performance prediction. In *Journal of Physics: Conference Series* (Vol. 1500, No. 1, p. 012110). IOP Publishing.

[18] Ahammad, K., Chakraborty, P., Akter, E., Fomey, U. H., & Rahman, S. (2021). A comparative study of different machine learning techniques to predict the result of an individual student using previous performances. *International Journal of Computer Science and Information Security (IJCSIS)*, 19(1).

[19] Sánchez-Pozo, N. N., Mejía-Ordóñez, J. S., Chamorro, D. C., Mayorca-Torres, D., & Peluffo-Ordóñez, D. H. (2021, December). Predicting High School Students' Academic Performance: A Comparative Study of Supervised Machine Learning Techniques. In *2021 Machine Learning-Driven Digital Technologies for Educational Innovation Workshop* (pp. 1-6). IEEE.

[20] Romero, C., Ventura, S., Pechenizkiy, M., & Baker, R. S. (Eds.). (2010). *Handbook of Educational Data Mining*. CRC Press.

Received: September 23, 2025. Reviewed: September 24, 2025. Accepted: September 30, 2025.