Student Academic Performance Prediction Using Machine Learning

Shangjia Wang *

School of Statistics and Data Science, Capital University of Economics and Business, Beijing, China

* Corresponding Author Email: 18595630996@163.com

Abstract. The fast growth of educational data systems has led to more student data becoming available at scale to use in learning analytics. It is important to effectively analyze these data to forecast academic performance, as this will facilitate early detection of risks and individualized interventions. The present paper explores some of the most important determinants of academic achievement and assesses several predictive models based on the Student Performance Factors (SPF) dataset (N = 6,607). Linear Regression (LR) and Random Forest (RF) models are built and benchmarked against each other. As can be seen, the RF model, according to the results ($R^2 = 0.70$), is significantly outperforming the LR model ($R^2 = 0.62$), and the error rate has been minimized by 11.5 percent. Feature importance analysis indicates that attendance is the main determinant (importance weight = 0.381), followed by study hours (0.243) and past scores (0.091). Interestingly, the combination of the existing learning behaviors is six times greater than the historical performance, and the family background factors have insignificant direct effects. The results obtained can be used to justify data-driven educational interventions and imply that the monitoring of attendance should become the central element of any academic early warning system.

Keywords: learning analytics, academic performance prediction, Random Forest, Machine Learning.

1. Introduction

The rapid growth of digital technologies in education has led to the large-scale generation of educational data, offering fresh possibilities for educational data extraction and analysis of learning. These data enable researchers and institutions to scrutinize the studying behaviors of students as well as identify patterns related to academic success and failure [1]. Therefore, due to this fact, student scholastic manifestation has become an important topic of education research, as the early identification of disadvantaged students enables teachers to take immediate and specific actions to intervene.

Conventional academic assessment is traditionally based on summative tests like end-of-term exams. Although these methods are simple to use, they do not necessarily reflect the learning process of each student and their differences [2]. As a result, students who experience challenges in learning throughout the semester can be missed out, and the value of specialized educational assistance is reduced.

During these years, it has been wittnessed a incressing tendency towards the use of machine learning and educational data mining methods to predict student performance. The comment made by Albreiki et al. is one of many that demonstrate that machine-learning algorithms have been commonly used to examine educational data and determine important factors that affect academic achievement [3]. Empirical research also reflects that predictive models are able to attain significant predictive accuracy with respect to education-related contexts. As an example, Yağci [4] implemented various machine learning algorithms using a sample of 1,854 university students and obtained prediction accuracies of between 70% and 81% depending on the model considered.

In the family of machine learning methodologies, ensemble techniques, including the RF, have demonstrated strong predictive representation of educational data with recorded accuracy rates over 90 percent in some studies when academic, behavioral, and socioeconomic variables are combined [5]. Other research has also found that behavioral indicators such as attendance and study hours are some of the best predictors of academic performance [6]. Moreover, a more recent study suggests

that the engagement of students in the current learning process might be even more predictive than their past school performance [7].

This paper builds and confirms these results in another analytical environment as it examines the prediction of student performance in academic disciplines using machine-learning methods. The LR and RF models are built and compared based on the Student Performance Factors dataset to assess their predictive performance. Also, the relative significance of major factors affecting student success is examined, which can also be used as a basis of data-based educational surveillance and early intervention.

2. Methods

2.1. Data source

The data sets that are utilized in the research are obtained through the Kaggle platform [8]. There are 6,607 student records and 20 variables containing 19 independent variables and 1 dependent variable, both in .CSV format, and there are no missing values. The dataset has data on the demographic features of the high school students, their learning patterns, family life, and academic achievements.

2.2. Sample selection

The dataset has one dependent variable and several independent variables on the academic, behavioral, and socioeconomic aspects. For the sake of comprehending the data structure at a deeper level, all of the variables are summarized in Table 1.

The dependent variable is Exam Score, which indicates how well students performed on their final exams. Four categories of independent variables could be used: academic engagement indicators, previous academic achievement, lifestyle variables, and family background characteristics.

Table 1. Variable Explanation.

Variable Category	Variable Name	Data Type
Academic Engagement	Hours Studied (HS)	Numerical
	Attendance	Numerical
	Tutoring Sessions (TS)	Numerical
	Extracurricular Activities (EA)	Categorical
Historical Performance	Previous Scores (PS)	Numerical
Lifestyle Factors	Sleep Hours (SH)	Numerical
	Physical Activity (PA)	Numerical
Motivation & Support	Motivation Level (ML)	Categorical
	Parental Involvement (PI)	Categorical
	Teacher Quality (TQ)	Categorical
Resources & Environment	Access to Resources (AR)	Categorical
	Internet Access (IA)	Categorical
	Family Income (FI)	Categorical
	School Type (ST)	Categorical
	Peer Influence (PI)	Categorical
	Distance from Home (DH)	Categorical
Individual Characteristics	Learning Disabilities (LD)	Categorical
	Gender	Cwasategorical
	Parental Education Level (PFL)	Categorical

For the purpose of managing the data and making the process of training models more efficient, all categorical variables are transformed into numerical equivalents with the help of Label Encoding. As an example, categorical values of Low, Medium, and High are given numerical values of 0, 1, and

2, respectively. In case of binary categorical variables, Male is coded as 0 and Female as 1. Table 2 contains the preprocessed data samples.

Table 2. The Preprocessed Data Samples

Hours_Studied	Attendance	Previous_Scores	Tutoring_Sessions	Sleep_Hours	Exam_Score
23	84	73	0	7	67
19	64	59	2	8	61
24	98	91	2	7	74
29	89	98	1	8	71
19	92	65	3	6	70

2.3. Exploratory data analysis

To have an idea of the general frequency of the presentations of the students, a histogram of the dependent variable (Exam Score) is drawn, which is illustrated in Fig. 1.

The distribution of exam scores is nearly symmetrical and close to a normal distribution, having an average of about 70.2 and a single standard deviation of 8.5. The majority of student scores fall between 65 and 80, and there are fewer observations at the lower and higher ends. The absence of severe skewness suggests that the dataset is suitable for regression-based modeling without requiring complex transformation of the target variable.

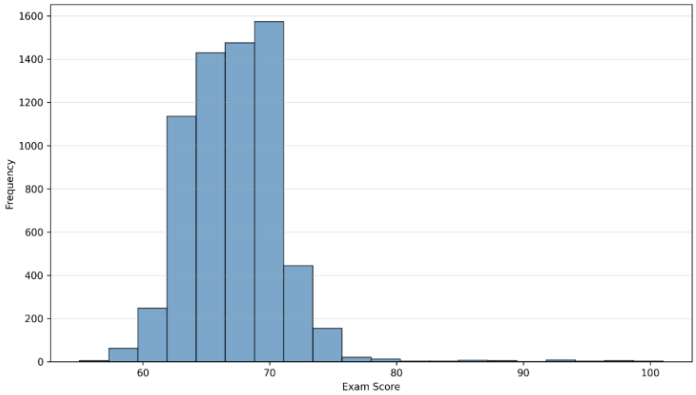

Fig. 1 Distribution of Exam Scores (Photo/Picture credit: Original).

2.4. Feature correlation analysis

To examine the connection between numerical characteristics and the object variable, the Correlation Heatmap generated is displayed in Fig. 2.

Fig. 2 Correlation Heatmap of Numerical Features (Photo/Picture credit: Original).

The correlation analysis reveals that Attendance (r = 0.52), PS (r = 0.45), and HS (r = 0.41) exhibit the strongest positive correlations with exam scores. TS, PA, and SH show weak positive correlations. No severe multicollinearity is observed among the numerical features (all inter-feature correlations < 0.7). These consequences render a preliminary cornerstone for subsequent modeling.

Based on the exploratory analysis, student performance appears to be influenced by multiple dimensions, including academic engagement, prior achievement, and family background. The relatively clear distribution pattern and observable correlations provide a solid foundation for constructing predictive models.

2.5. Model introduction

LR is applied as the base model for its simple property and interpretable possibility. It supposes a linearity in the middle of the dependent variable (exam score) and the independent variables. This model is widely applied in educational data analysis on account of its ability to clearly quantify the influence of each explanatory variable on academic outcomes. In the setting of student appearance prediction, LR provides a transparent benchmark that helps evaluate whether there are more advanced machine-learning models providing extra predictive capabilities.

The mathematical expression is:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon \quad (1)$$

In which Y on behalf of the predicted exam score, X_i are explanatory variables, β_i are regression coefficients, and ε is the error term.

RF can construct multiple decision trees through training and can produce predictions on the basis of the average exportation of these trees, thus it's an ensemble learning method. Comparing with linear models, researchers testify that RF can capture complex nonlinear relationships and interactions among variables, which are common in educational datasets. Early in preceding studies, they have shown that relational methods usually achieve higher predictive accuracy in student performance prediction tasks since they respond to noise robustly and are capable of dealing with heterogeneous features [3,9].

In addition, RF offers feature importance scores, which help tell the key elements influencing academic performance. These characteristics make RF particularly suitable for analyzing multi-dimensional educational data.

The model is equipped with a fixed random seed (random_state = 42) and 100 trees (n_estimators = 100) to ensure reproducibility.

Three standard regression metrics are used for assessing a model's function and presentation: Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and Coefficient of Determination (R^2).

It divides the data randomly into drilling (80%) and trial (20%) sets. Models are fitted using the drilling set, while the trial set evaluates generalization performance. All experiments are conducted with a fixed random seed to ensure reproducibility.

3. Results and discussion

3.1. Model performance comparison

Both LR and RF are drilled on the drilling group and assessed on the trial group. The results of the experiment are presented in Table 3 and Fig. 3.

Table 3. Model Performance Comparison.

Model	MAE	RMSE	R^2
LR	3.21	4.15	0.62
RF	2.84	3.68	0.70

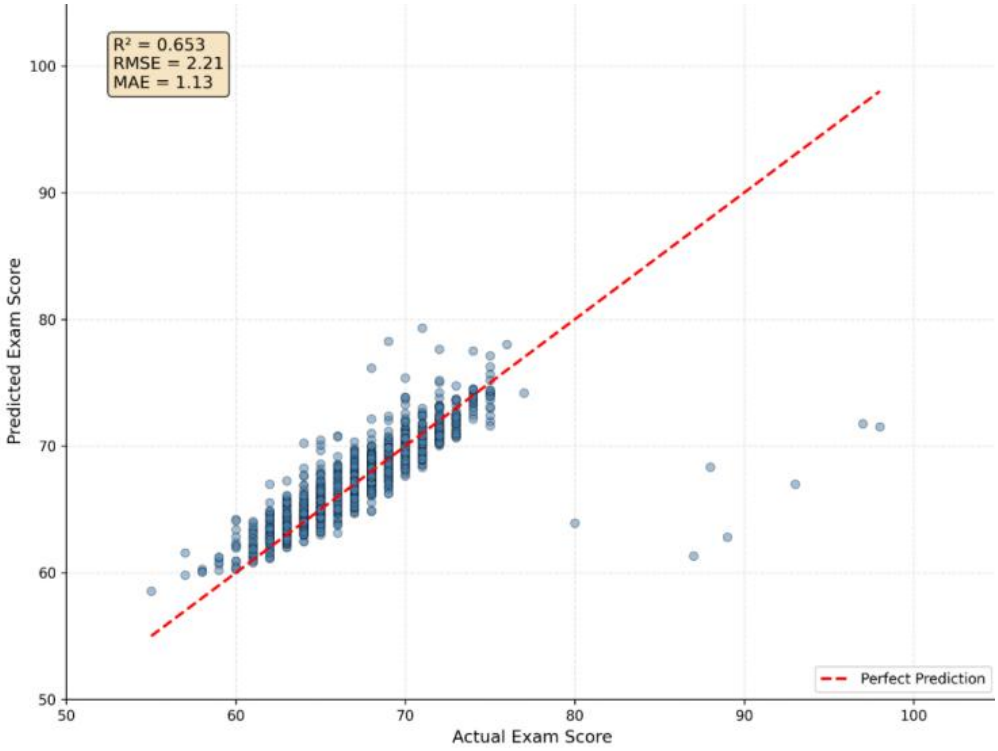

Fig. 3 Model Performance Comparison (Photo/Picture credit: Original).

The results show that the RF is visibly better than LR in terms of the evaluation measures. In details:

The Performance of RF was always better than LR, with a reduction of MAE and RMSE of about 11 percent and an increase of R^2 value to 0.70.

The R^2 value, which is 0.7, confirms that the RF model can account for 70 percent of the variables regarding the student exam scores. At this level of predictive performance, it is regarded as significant and practical in educational data mining. This unexplained variance of 30 percent can be explained through the presence of latent variables not included in the dataset, including learning efficiency, test anxiety, or variations in the difficulty of the examinations.

The high performance of RF is due to the benefit of being able to capture the non-linear relationships and intricate relationships between features. The relationship between study hours and academic performance can also be characterized by diminishing marginal returns, which linear models are not capable of representing. Also, RF is resistant to outliers and noise, which are usually found in education data because of its ensemble construction.

3.2. Feature importance analysis

A member of the crucial component of RF is its power to quantify the relative importance of every feature when predicting the target variable. The importance scores of features are determined using the overall decrease in node impurity (based on the mean square error), depending on the probability of reaching that node, which is averaged over all the trees in the forest.

With the aim of identifying the most powerful variables affecting student academic performance, the feature importance weights are extracted from the trained RF model. Fig. 4 presents the top 10 most important features ranked by their contribution to each predicted performance in the model.

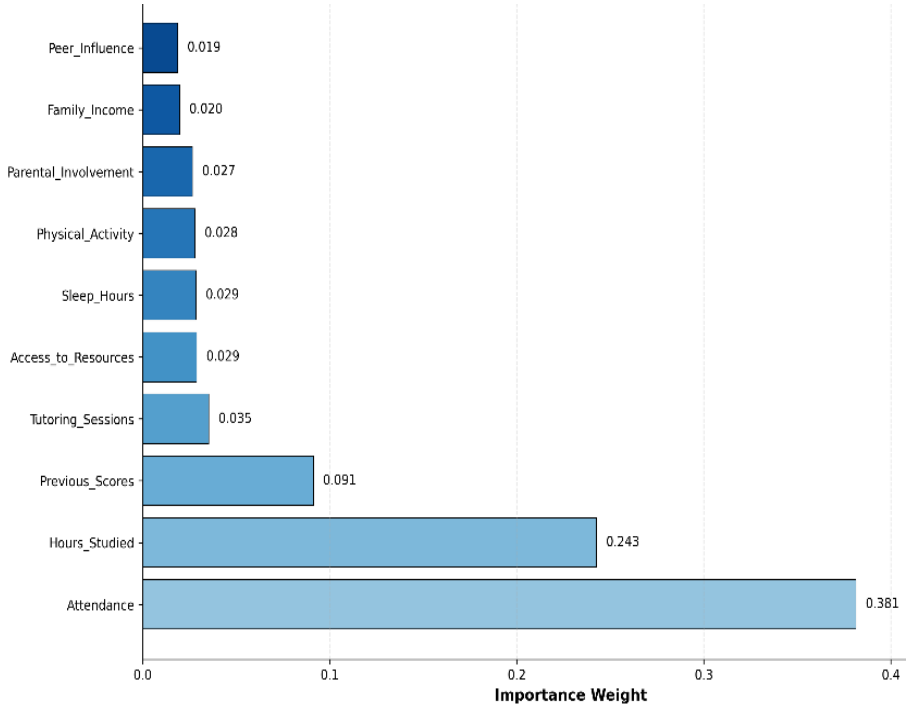

Fig. 4 Feature Importance Ranking (Top 10) (Photo/Picture credit: Original).

As per the characteristic significance analysis, attendance and study hours are two principal predictors of the student's academic performance, and both variables contribute more than 60% to the predictive power of the model. On the other hand, factors associated with the family background have a fairly small direct effect on the examination results. These results indicate that the main factor that controls students' academic outcomes is the present level of their learning behavior.

3.3. Discussion

Several considerations should be made on how to interpret these findings. Firstly, the nature of the data could also impact the pattern of the observed importance of features. The SPF dataset is a well-balanced group of behavioral and demographic variables, though it has no psychological or cognitive variables that might also be associated with academic achievement. Other studies have indicated that variables like learning motivation, stress level as well as learning strategies have a high potential to influence student achievement [2].

Third, the preprocessing strategy might also influence model results. This paper will transform the categorical variables through label encoding to enable training of the model. Although this technique makes data processing easier, it can give rise to artificial ordinality between categories, which has the potential to influence the properties of tree-based models.

Thirdly, the predictive models that are employed in the current research also have some weaknesses. LR presupposes that relations between variables are linear, which can be inadequate in describing the intricate associations that exist within educational data. Even though RF is better at predicting using nonlinear relationships, it may be susceptible to over-fitting and poor interpretability, particularly in dimensional datasets with multiple dimensions [3,9].

The performance of predictions could be enhanced in future studies through the inclusion of other behavioral and psychological factors and the use of state-of-the-art machine learning algorithms like gradient boosting and deep learning, and through the implementation of cross-validation or hyperparameter optimization to achieve better model generalization. For example, Singh et al. showed that graph neural networks based on psychological characteristics were able to increase the accuracy of behavior prediction by 19.6 percent compared to baseline models [10]. Likewise, Otchere et al. found that a well-performing hyperparameter optimization that combines error variance with mean error can decrease worst-case prediction error by as much as 19.8% and tighten confidence intervals, thus improving model generalization on out-of-sample data [11].

4. Conclusion

This study investigated student academic performance prediction using LR and RF models on the Student Performance Factors dataset. Three main conclusions emerge:

To begin with, the two models have acceptable predictive accuracy. LR ($R^2 = 0.689$) is marginally better than RF ($R^2 = 0.653$), which manifests that the feature-performance relationship is mostly linear. Nonetheless, RF has important interpretability based on feature importance analysis.

Second, attendance is the most significant predictor of academic success. Attendance alone has a predictive importance value of 0.381, which is almost 40 percent of the total predictive capacity of the model, 1.57 times that of study hours, and 4.19 times that of past results. Existing learning practices (attendance + study hours = 0.624) far outweigh historical achievement (0.091), implying that the actions of the students are more important than their past deeds.

Thirdly, the family background variables have very little direct effect. The list of socioeconomic factors, such as family income and parental participation, ranks among the lowest ten, and its weight is less than 0.03. Once learning behavior is controlled, family background does not account for any significant additional variance in academic performance.

These findings carry clear practical implications: attendance monitoring should be the first line of defense in academic early warning systems. Rather than complex machine learning models or extensive student data, simply tracking who shows up to class may be the most effective and actionable strategy to recognize risky students and improve educational outcomes.

References

[1] C. Romero, S. Ventura, Educational data mining and learning analytics: An updated survey. Wiley Interdiscip. Rev. Data Min. Knowl. Discov. 10, e1355 (2020)

[2] A. Namoun, A. Alshanqiti, Predicting student performance using data mining and learning analytics techniques: A systematic literature review. Appl. Sci. 11, 237 (2021)

[3] B. Albreiki, N. Zaki, H. Alashwal, A systematic literature review of student performance prediction using machine learning techniques. Educ. Sci. 11, 552 (2021)

[4] M. Yağcı, Educational data mining: Prediction of students' academic performance using machine learning algorithms. Smart Learn. Environ. 9, 11 (2022)

[5] M. Gusnina, Wiharto, U. Salamah, Student performance prediction in Sebelas Maret University based on the random forest algorithm. Ing. Syst. Inf. 27, 495-501 (2022)

[6] V. Čotić Poturić, S. Čandrlić, I. Dražić, A scoring algorithm for the early prediction of academic risk in STEM courses. Algorithms 18, 177 (2025)

[7] E. Elbasi, M. Nadeem, Y. I. Alzoubi, et al., Machine learning in education: Innovations, impacts, and ethical considerations. IEEE Access 13, 128741–128765 (2025)

[8] A. Student Academic Performance Dataset, Kaggle (2026) https://www.kaggle.com/datasets/ayeshasiddiqa123/ student-perfirmance

[9] R. Hasan, S. Palaniappan, S. Mahmood, et al., Predicting student performance in higher education: A comprehensive review of machine learning applications. IEEE Access 12, 41567–41592 (2024)

[10] A. Singh., Consumer Psychological Modeling and Behavior Prediction System Based on Graph Neural Network. Int. J. Comput. Intell. Syst. 19, 24 (2026)

[11] D. A. Otchere, Augmented robustness in home demand prediction: Integrating statistical loss function with enhanced cross-validation in machine learning hyperparameter optimisation. Energy AI 21, 100584 (2025)