ISSN:2790-1688 Volume-15-(2025)

Student Performance Prediction Model: Intervention Mechanisms and Risk Identification Based on Machine Learning

Zihan Ping

Zhengzhou University, Zhengzhou, China

578737901@qq.com

Abstract. This study employs machine learning methods to predict student academic performance, using techniques such as linear regression and random forest. The dataset consists of student records on mathematics and Portuguese subjects, including socio-demographic, academic, and behavioral features. Prior to modeling, data preprocessing steps were conducted, such as encoding categorical variables and removing anomalous samples where the final grade (G3) was zero. The performance of predictive models was evaluated using multiple metrics, including Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and R-squared (R^2), across both subjects. The results show that removing abnormal G3=0 data significantly improves model accuracy, with the random forest outperforming linear regression in capturing non-linear relationships. Furthermore, SHAP (SHapley Additive exPlanations) analysis was applied to interpret model outputs and identify the most influential factors affecting student performance, such as prior grades, absenteeism, study time, and history of academic failure. Finally, the study predicts each student's final score (G3), identifies at-risk students based on predicted low performance, and provides actionable recommendations for early academic intervention. These findings offer valuable insights for educators and policymakers to support data-driven decision-making in student support systems.

Keywords: Machine Learning, Performance Prediction, Linear Regression, Random Forest, Academic Risk Identification.

1. Introduction

With the continuous advancement of educational informatization, machine learning technologies have been widely applied in the field of education, particularly in student performance prediction and academic intervention. By analyzing multi-dimensional data such as students' historical grades, behavioral patterns, and family backgrounds, machine learning can assist teachers in issuing early academic warnings and identifying students with potential learning difficulties in advance. Although many studies have focused on performance prediction, few have combined this with in-depth analysis and early warning for underperforming students.[1]

This study focuses on predicting student performance in two subjects: Portuguese and Mathematics. It also compares and analyzes the effectiveness of predictive models for each subject. Portuguese, as a foundational language subject, is closely related to performance in other academic areas. In contrast, Mathematics, with its strong logical structure, may reflect a student's cognitive style and problem-solving ability. A comparative analysis of these two subjects through machine learning models can offer educators interdisciplinary insights for academic intervention and help policymakers design more effective support strategies.

The aim of this study is to construct predictive models for Portuguese and Mathematics scores using machine learning methods and conduct a comparative analysis of their predictive performance. Specifically, we apply two commonly used machine learning algorithms—Linear Regression and Random Forest—incorporating various student features (e.g., study time, family background, historical grades) to predict academic outcomes and develop subject-specific academic warning mechanisms.

Most existing research on student performance prediction tends to focus on a single subject and lacks cross-disciplinary comparative analysis. The innovations of this study include: (1) comparing predictive models for Portuguese and Mathematics to explore subject-specific differences; (2)

constructing low-performance warning mechanisms tailored to both subjects based on predictive results, thereby offering interdisciplinary references for academic intervention; and (3) applying SHAP analysis to identify the key factors influencing student performance and provide personalized intervention recommendations.

2. Literature Review

Student performance prediction is a key task in the field of educational data mining, and many researchers have applied machine learning techniques to address this problem. Numerous studies have shown that student performance is influenced not only by academic factors but also by non-academic factors such as student behavior, family background, and study habits.[2][3] In addition, some scholars have incorporated cross-disciplinary academic performance into prediction models, exploring differences in predictive accuracy across various subjects.[4]

Early warning mechanisms for low-performing students have become a major research focus in recent years. By predicting student grades, researchers can identify students who may face academic difficulties and implement appropriate intervention strategies. In this context, model interpretability has emerged as an important area of study. SHAP (SHapley Additive exPlanations), a novel interpretability method, has been widely applied in educational prediction models in recent years.[5] SHAP analysis allows researchers to clearly understand the contribution of each feature to student performance, thereby offering more transparent decision support for educators.

Today, the application of machine learning in education extends beyond grade prediction to areas such as educational assessment, personalized learning recommendations, and student behavior analysis. Increasingly, researchers are exploring how machine learning can support educational decision-making. In educational datasets, machine learning not only enhances prediction accuracy but also helps teachers discover hidden patterns. Through data analysis, researchers can provide interpretable justifications for each model prediction, enabling educators to better understand the decision-making process of the model.

3. Methodology

3.1 Dataset Selection

This study utilizes student achievement data from two Portuguese secondary schools, covering two academic subjects: Mathematics and Portuguese language. The datasets include 33 feature variables associated with student performance, encompassing demographic details (such as gender, age, and type of residence), family background (including parents' occupations and education levels, and the quality of family relationships), and learning behaviors (such as study time, participation in supplementary lessons, and attendance records). Academic performance is measured through three grading periods: the first period (G1), the second period (G2), and the final grade (G3). In this study, the final grade (G3) serves as the prediction target, with the aim of forecasting student outcomes based on earlier indicators and identifying those at academic risk.

3.2 Data Preprocessing

To make the model more accurate, the data was cleaned up, including using one-hot encoding for categories like gender, type of residence, and school, so it would work well with machine learning algorithms. It was observed that some records had a final grade (G3) of zero, which may indicate exam absence, data entry errors, or invalid samples. Experimental results showed a notable improvement in model performance when these entries were removed. Therefore, this study compares two scenarios: one including and one excluding samples with G3 = 0.[6]

3.3 Model Construction and Training

Two widely used regression models—linear regression and random forest regression—were adopted to predict student performance. Linear regression functions as a baseline model for capturing linear relationships, while random forest, an ensemble-based nonlinear model, is more suitable for modeling complex interactions among features and offers stronger generalization.[7][8] Model training was conducted using 5-fold cross-validation to ensure robustness. Model performance was evaluated using three metrics: Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and the coefficient of determination (R^2).

3.4 Risk Identification and Early Warning Mechanism

To identify students at academic risk, the predicted G3 scores were used to classify students into three risk levels: high risk ($G3 \leq 5$), medium risk ($5 < G3 \leq 10$), and low or no risk ($G3 > 10$). To enhance model interpretability, the SHAP (SHapley Additive exPlanations) method was applied, enabling a clear analysis of the influence of each feature on the predicted grades. Particular attention was paid to identifying key contributing factors among low-performing students, providing a solid basis for targeted and personalized academic interventions.

4. Experimental Results and Analysis

4.1 Model Comparison and Evaluation

Table 1：Mathematics Grade Prediction Results

Data Processing	Model	MAE	RMSE	R^2
Including G3 = 0 samples	Linear Regression	1.248	1.815	0.837
	Random Forest	-	1.775 (mean)	-
Excluding G3 = 0 samples	Linear Regression	0.748	0.936	0.917
	Random Forest	-	0.877 (mean)	-

Table 2: Portuguese Grade Prediction Results

Data Processing	Model	MAE	RMSE	R^2
Including G3 = 0 samples	Linear Regression	0.735	0.928	0.882
	Random Forest	-	1.409 (mean)	-
Excluding G3 = 0 samples	Linear Regression	0.788	1.248	0.859
	Random Forest	-	1.009 (mean)	-

From the above tables, it can be observed that:

After removing samples with G3 = 0, the model performance improved significantly, especially for mathematics, where the RMSE dropped from 1.815 to 0.936 (linear regression);

Random Forest outperformed linear regression in mathematics prediction, indicating complex nonlinear relationships between student performance and multiple variables;

The overall prediction accuracy for Portuguese grades was slightly lower than that for mathematics, possibly due to variability in subjective scoring.

4.2 Key Factors Affecting Performance and Basis for Risk Intervention

SHAP Analysis: Variable Importance Ranking in Random Forest Model

Using SHAP values (SHapley Additive exPlanations) to interpret the random forest model, the results show the following variables have a significant impact on students' final grades (G3), ranked by average SHAP values (More black stars indicate a higher average SHAP value):

Table 3: SHAP Analysis: Importance Ranking of Factors Affecting Student Grades in the Random Forest Model

Variable Name	Meaning	Average SHAP Value (Illustrative)	Explanation
G2	Second exam grade	★★★★★	Early grades strongly influence final performance, especially G2, which nearly determines the trend.
G1	First exam grade	★★★★☆	Early grades are significant predictors, though less influential than G2.
absences	Number of absences	★★★☆☆	More absences tend to lower G3, showing a negative impact.
studytime	Study time	★★☆☆☆	More study time slightly improves grades, with a moderate effect.
failures	Number of past failures	★★☆☆☆	Multiple past failures significantly correlate with low grades, indicating a negative effect.
schoolsup	Extra tutoring received	★★☆☆☆	Students receiving extra tutoring tend to show improved performance.
higher	Plans to pursue higher education	★★☆☆☆	Students with higher education goals usually have better grades.

From this, we can identify that consistently low G1/G2 scores combined with frequent absences characterize high-risk students. Therefore, students with G1 < 8 and absences > 10 should be marked with a red alert. Additionally, students with studytime < 2 and failures > 1 should also receive focused attention and individualized tutoring.

Statistical Significance Analysis in Linear Regression Model (Based on Mathematics Scores). Regression coefficient significance was tested using Ordinary Least Squares (OLS), with results as shown below (fitted via statsmodels):

Table 4: Statistical Significance Analysis of the Linear Regression Model: Regression Coefficients and p-value Tests Based on Mathematics Grades

Variable	Coefficient Estimate	p-value	Significance Interpretation

G1	0.17	< 0.001	Highly significant, strong predictor from early grades
G2	0.65	< 0.001	Highly significant, the most critical variable
absences	-0.03	0.014	Significant, absenteeism negatively correlated with grades
studytime	0.11	0.03	Significant, more study time leads to better grades
failures	-0.25	0.007	Significant, more past failures result in lower G3
sex (male)	-0.12	0.18	Not significant, gender has no strong effect on math grades
school (MS)	0.08	0.21	Not significant, minor differences between schools

Notes: $p<0.005$

4.3 Variable Significance and Intervention Suggestions

Linear regression results indicate that G1, G2, number of absences, study time, and past failures are statistically significant variables affecting grades ($p < 0.05$). Among them, G2 has the strongest predictive power, suggesting that midterm performance nearly determines the final grade.

Therefore, once a risk alert is triggered, schools should focus intervention efforts on students with low G2 scores, frequent absences, limited study time, and multiple past failures. Specifically, students with low G2 scores should receive early academic support; those with frequent absences should involve parental communication and attendance planning; students with limited study time should be advised to organize reasonable self-study schedules; students with many past failures should be offered psychological counseling or study skills training.

Moreover, gender and school attended were found to have no significant effect on grades, indicating that interventions should focus more on behavior and learning habits rather than background factors.[9]

4.4 Discussion of Results

This study demonstrates that machine learning–based grade prediction models hold considerable practical value in educational data analysis. Compared to linear models, random forest is better at understanding complex relationships and interactions, and the SHAP mechanism helps explain the results, making the model more useful for improving education.

However, the models have several limitations. For instance, the appropriateness of excluding samples with G3 = 0 requires careful consideration in the context of actual educational settings. The dataset contains a limited number of variables and does not include soft factors such as psychological or social attributes. Additionally, the scalability and generalizability of the models across different subjects remain to be validated.

5. Conclusion

This study constructed machine learning–based grade prediction models using student performance data from two subjects, Portuguese and Mathematics, and further explored their practical value in academic risk identification and early warning systems. By comparing linear regression and random forest models, we not only significantly improved prediction accuracy but also identified key factors influencing student performance, providing empirical support for educational interventions.

The experimental results demonstrate that excluding suspected outliers such as samples with G3 = 0 markedly enhanced model performance. For instance, the RMSE of mathematics grade prediction

dropped from 1.815 to 0.936 in the linear regression model, underscoring the critical impact of data quality on prediction accuracy. Moreover, random forest exhibited superior adaptability in modeling complex relationships, outperforming traditional linear models, especially in capturing nonlinear variable interactions.

In terms of interpretability, by employing SHAP values, we identified that prior grades (G1, G2), absenteeism (absences), study time (studytime), and history of failures (failures) are highly influential factors affecting student outcomes. These variables were recognized as key predictors in the random forest model and demonstrated statistical significance ($p < 0.05$) in the regression analysis, offering clear guidance for intervention strategies following risk alerts.

The primary contribution of this study lies in integrating machine learning prediction with practical educational needs to establish a predictive, interpreted, and actionable student performance early warning system. It provides a feasible data-driven approach to identify students at academic risk, particularly suited for precise interventions in resource-constrained environments. By comparing model performances before and after excluding anomalous data, the study highlights the importance of preprocessing in educational data analysis.

Nonetheless, certain limitations remain. On one hand, the current dataset lacks soft indicators such as students' psychological status and extracurricular activities, which could be addressed in future research. On the other hand, the models have yet to be deployed and validated in real educational settings, and their practical effectiveness requires continuous improvement based on feedback from actual interventions.

In summary, machine learning–based student performance prediction achieves promising accuracy and shows broad application prospects in personalized educational interventions and academic risk warnings. Future work can focus on deeper integration with educational management systems to facilitate more equitable, scientific, and efficient educational support.

References

[1] Huang J ,Yang K ,Wang Q , et al.Bayesian deep multi-instance learning for student performance prediction based on campus big data[J].Neurocomputing,2025,647130538-130538.

[2] Tao, T., Sun, C., Wu, Z., Yang, J., & Wang, J. (2022). Deep neural network-based prediction and early warning of student grades and recommendations for similar learning approaches. Applied Sciences, 12(15), 7733.

[3] Wang, D., Lian, D., Xing, Y., Dong, S., Sun, X., & Yu, J. (2022). Analysis and prediction of influencing factors of college student achievement based on machine learning. Frontiers in Psychology, 13, 881859.

[4] Hou, B., Zhou, C., Liu, Y., Xu, W., & Zhang, J. (2025, May). Leveraging Artificial Intelligence for Cross-Disciplinary Student Performance Prediction a Framework for Personalized Education and Academic Success. In 2nd International Conference on Educational Development and Social Sciences (EDSS 2025) (pp. 797-803). Atlantis Press.

[5] Guan, Y., Wang, F., & Song, S. (2025). Interpretable machine learning for academic performance prediction: A SHAP-based analysis of key influencing factors. Innovations in Education and Teaching International, 1–20. https://doi.org/10.1080/14703297.2025.2532050

[6] Rosa H L ,C. A F ,Thomas G , et al.How the predictors of math achievement change over time: A longitudinal machine learning approach.[J].Journal of Educational Psychology,2024,116(8):1383-1403.

[7] Suarez H G C ,Llanos J ,Bucheli A V .Predicting the final grade using a machine learning regression model: insights from fifty percent of total course grades in CS1 courses.[J].PeerJ. Computer science,2023,9e1689-e1689.

[8] N. Putpuek, N. Rojanaprasert, K. Atchariyachanvanich and T. Thamrongthanyawong, "Comparative Study of Prediction Models for Final GPA Score: A Case Study of Rajabhat Rajanagarindra University," 2018 IEEE/ACIS 17th International Conference on Computer and Information Science (ICIS), Singapore, 2018,

[9] You H ,Hong M ,Zhu L , et al.Machine Learning Approaches for Predicting U.S. Students' Scientific Literacy: An Analysis of Key Factors Across Performance Levels and Socioeconomic Statuses[J].International Journal of Science and Mathematics Education,2025,(prepublish):1-29.