# Student Performance Prediction Using the UCI Dataset: A Comparison of Interpretable and Ensemble Models

**Article history**
Received
*6 December 2025*
Received in revised form
*6 March 2026*
Accepted
*7 March 2026*
Published online
*26 June 2026*

***Jiajun Guo[a], Chenghang Liu[a], Tianxing Ji[a], Yuyao Li[a], Quan Zhang[a]*,***
*[a]School of Advanced Technology, Xi'an Jiaotong-Liverpool University, Suzhou, China*
**Quan.Zhang@xjtlu.edu.cn*

**Abstract**

AI technologies have not only transformed teaching methods but have also provided novel solutions for education management, assessment, and personalized learning. This study examines whether complex machine learning models consistently outperform simple interpretable approaches in predicting student outcomes. Using the University of California, Irvine (UCI) Student Performance dataset, five key predictors: First period grade (G1), Second period grade (G2), Number of past class failures (Failures), Mother's education level (Medu) and Higher education aspiration (Higher), were extracted from 32 original attributes. Three models, including Linear Regression (LR), Random Forest (RF), and an Ensemble Model (EM), were evaluated across Mathematics and Portuguese subjects using MSE, RMSE, MAE, and $R^2$, with five-fold cross-validation to assess robustness. Experimental results demonstrate that Linear Regression achieved the best overall performance in both subjects, with $R^2$ = 0.779 for Mathematics and $R^2$ = 0.862 for Portuguese, whereas RF and EM did not yield consistent gains. Portuguese is generally more predictable than Mathematics under the same pipeline. Feature influence analysis indicates that early-term grades (G1 and G2) dominate predictive power, suggesting that the approach supports mid-semester/operational prediction rather than start-of-term early-warning. Overall, the findings highlight the practical value of interpretable models for educational analytics when transparency and deployability are important.

**Keywords**: Machine learning, educational data mining, performance prediction, linear regression, feature importance.

## Introduction

The rapid development of artificial intelligence (AI) and machine learning has accelerated their adoption in educational data mining, where predictive analytics are increasingly used to support assessment, academic risk detection, and decision-making in teaching and learning processes (Kuleto et al., 2021; Luan et al., 2020). Between 2017 and 2021, the use of artificial intelligence in educational settings in the United States grew by 47.5% (Kuleto et al., 2021). Recent studies have demonstrated the practical applications of AI technology in several areas, including intelligent tutoring, adaptive learning systems, student behavior analysis, academic performance prediction, personalized learning and tutoring, and online education. For example, intelligent tutoring systems can provide personalized tutoring and instant feedback based on students' learning behaviors and performance outcomes (Chen et al., 2020). By analyzing data on students' learning behaviors, educators can gain insights into their learning habits and requirements, thereby facilitating the optimization of instructional design and management (Luan & Tsai, 2021). The use of virtual reality technology helps create immersive learning environments, enabling a more intuitive understanding of complex knowledge (Zafari et al., 2022). In parallel, the growing availability of educational records has made it feasible to model student outcomes using historical grades and contextual variables, potentially enabling data-informed interventions to support learners who may be at risk (Shafique et al., 2023; Vakhabova et al., 2019).

Despite this progress, a persistent practical question remains insufficiently addressed: do more complex machine learning models consistently outperform simpler and more interpretable models when applied to structured educational datasets? In educational contexts, model choice is not only an accuracy issue but also relates to transparency, ease of maintenance, and stakeholders' trust. However, existing work often focuses either on proposing a specific modelling approach or on applying highly complex models, while systematic comparisons across interpretable and more complex techniques under a controlled experimental setup are still relatively limited (Yağcı, 2022). As a result, it can be difficult for educators and practitioners to determine whether increased model complexity yields meaningful benefits in settings where interpretability and deployability are valued.

To address this gap, this study investigates student performance prediction using the widely adopted

Guo, J., Liu, C., Ji, T., Li, Y., & Zhang, Q. (2026), Student Performance Prediction Using the UCI Dataset: A Comparison of Interpretable and Ensemble Models, *ASEAN Journal of Engineering Education*, 10(1), 24-30. https://doi.org/10.11113/ajee2026.10n1.211

University of California, Irvine (UCI) Student Performance dataset (Cortez, 2008). We analyse two subjects: Mathematics and Portuguese, under the same modelling pipeline, which enables a controlled comparison of model behaviour across subject contexts. Specifically, this study first processed and analyzed the 32 influencing factors in the dataset and selected the five most critical factors to balance predictive performance and interpretability. Then, three regression approaches are compared: Linear Regression (as an interpretable baseline), Random Forest (as a non-linear model capturing interactions), and a simple ensemble model that averages heterogeneous learners. Model performance is assessed using standard error metrics (MSE, RMSE, and MAE) and $R^2$, together with five-fold cross-validation to estimate robustness. Additionally, this study briefly discusses the differences in predicting performance for STEM (Science, Technology, Engineering, and Mathematics) versus non-STEM subjects, which is further discussed in the discussion and conclusion sections.

Overall, this work provides practical interpretations for model selection in educational analysis through controlled comparisons and evaluations of interpretable models and more complex models on standard benchmark datasets, offering a more thorough understanding of which factors most significantly impact student performance. This approach not only improves predictive accuracy but also offers a deeper understanding of the multifaceted nature of academic success.

## Methodology

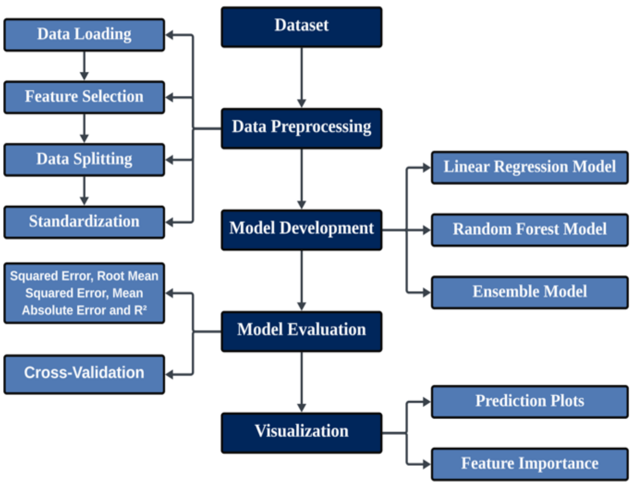


**Figure 1. Methodology flowchart**

Figure 1 summarizes the overall workflow of this study, which consists of five stages: (i) dataset and data loading, (ii) feature selection, (iii) data splitting and preprocessing, (iv) model development, and (v) model evaluation and visualization. The detailed procedures are described below.

*Dataset and Data Loading*

Dataset: This study utilizes the UCI student performance dataset (Cortez, 2008), which contains student performance information for Mathematics (Math) and Portuguese (Por) subjects, as well as 32 features related to student demographics, social and school-related factors, and academic performance. The decision to compare Mathematics and Portuguese subjects was driven by the distinct cognitive demands and teaching methods associated with STEM versus non-STEM subjects. Mathematics, as a core STEM subject, typically requires logical reasoning and problem-solving skills, while Portuguese, a language subject, involves comprehension and linguistic abilities. The two subjects are analyzed separately to examine whether model performance and influential factors remain consistent across different subject areas within the same data source. The Math and Por datasets were loaded into separate data frames for independent preprocessing and modelling.

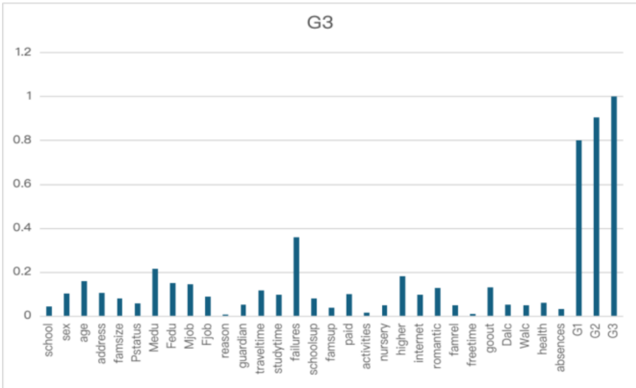


**Figure 2. Feature selection for Math**

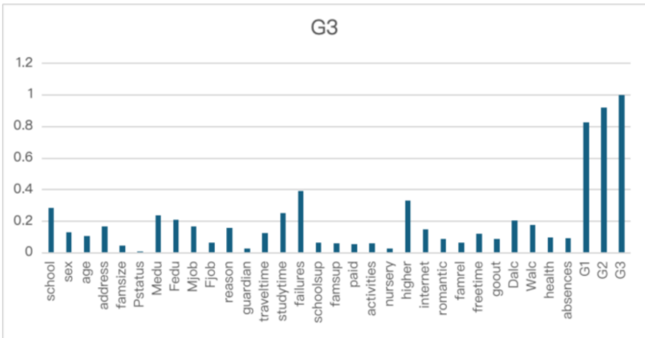


**Figure 3. Feature selection for Por**

Feature Selection: The prediction target is G3, representing the final grade, to reduce model complexity and maintain interpretability, a small set of predictors was selected from the original attributes. Relevant features and the target variable were extracted from the loaded 'DataFrames' to ensure that the models were trained on the most pertinent data. Figure 2 and Figure 3 show the results of the dataset preprocessing. For Math, G1, G2, Failures, Medu, Higher, and Fedu (Father's education level) are most strongly associated with the final grade. For Por, G1, G2, Failures, Higher, School, and Medu are the most important influencing factors. Therefore, the following features were selected as common predictors for both subjects: First period grade (G1), Second period grade (G2), Number of past class failures (Failures), Mother's education level (Medu) and Higher education aspiration (Higher). While some additional variables, such as Fedu and School, show relevance in individual subjects, a unified feature set was adopted to enable a fair and controlled comparison across subjects and models.

Data Splitting: Each dataset was split into training and testing sets using an 80/20 ratio. The training set was used for model fitting and cross-validation, and the held-out test set was used for final evaluation.

Standardization: To ensure all features are on the same scale, standardization was performed using 'StandardScaler'. This process scales the features to have a mean of zero and a standard deviation of one. Standardization is crucial for machine learning algorithms sensitive to the scale of input data, ensuring all features contribute equally to the training process.

*Model Development*

Three regression approaches were developed and compared: Linear Regression (LR), Random Forest (RF), and a simple Ensemble Model (EM).

1) Linear Regression Model (LR): Trained on standardized data, this model assumes a linear relationship between input features and the target variable. Feature coefficients were derived from training data. Performance was evaluated on test data using Mean Squared Error, Root Mean Squared Error, Mean Absolute Error, Coefficient of Determination, and Cross-Validation.

Standard Model Description: Regression assumes a linear relationship between the input features $\boldsymbol{X}$ and the target variable $y$. The model can be expressed as:

$$y = \mathbf{X}\beta + \epsilon \tag{1}$$

where $\boldsymbol{X}$ is the matrix of input 5 features, β is the vector of coefficients learned from the training data, y is the predicted target variable (G3), ϵ is the error term.

Training: Trained on standardized training data using the 'LinearRegression` class from 'scikit-learn'.

Prediction and Evaluation: Predictions were made on the test data, and the model's performance was evaluated using Mean Squared Error, Root Mean Squared Error, Mean Absolute Error, Coefficient of Determination, and Cross-Validation.

Feature Importance: The regression coefficient β represents the influence of each feature on the target variable.

2) Random Forest Model (RF): An ensemble learning method based on decision trees, capturing non-linear relationships and feature interactions. Multiple decision trees were built during training, and the average prediction of individual trees was outputted. Performance was evaluated using the same metrics as LR.

Standard Model Description: Random Forest constructs multiple decision trees during training and outputs the mean prediction of the individual trees. The model's prediction y is given by:

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

where y is the predicted target variable (G3), $T$ is the number of trees, $\hat{y}^{(t)}$ is the prediction from the t-th tree.

The input features $\boldsymbol{X}$ are the same as in the LR.

Training: Trained on the training data using the 'RandomForestRegressor' class from 'scikit-learn'.

Prediction and Evaluation: Predictions were made on the test data, and performance was evaluated using the same metrics as LR.

Feature Importance: Analyzed to understand each feature's contribution to the model's predictions, typically using the mean decrease in impurity.

3) Ensemble Model (EM): To investigate whether combining heterogeneous learners improves predictive performance, a simple ensemble model was constructed by averaging predictions from LR, RF, and Gradient Boosting (GB). Each model was trained individually, and feature importance was averaged from RF and gradient boosting models to gain insights into feature significance. Sequential tree building corrected errors from previous trees. Final predictions were averaged from the three models, and performance was evaluated using the same metrics as the other models. This unweighted averaging strategy was intentionally adopted as a lightweight and transparent baseline ensemble to avoid additional hyperparameters, like stacking weights, that may introduce extra tuning complexity or overfitting risk under limited sample sizes.

Standard Model Description: The ensemble model combined predictions from three models, the final ensemble prediction $\hat{y}_{ensemble}$ is given by:

$$\hat{y}_{ensemble} = \frac{1}{3}(\hat{y}_{LR} + \hat{y}_{RF} + \hat{y}_{GB}) \tag{3}$$

where $\hat{y}_{ensemble}$ is the ensemble prediction of G3. $\hat{y}_{LR}$ is the prediction from the LR. $\hat{y}_{RF}$ is the prediction from the Random Forest mode. $\hat{y}_{GB}$ is the prediction from the GB. The input features $\boldsymbol{X}$ are the same as in the LR.

Training: Each model was individually trained on the training data.

Prediction and Evaluation: The final predictions were obtained by averaging the predictions from the three models.

Feature Importance: Averaged from the Random Forest and Gradient Boosting models to gain insights into the relative importance of each feature.

*Model Evaluation*

The following metrics were used to quantify predictive performance:

Squared Error (MSE): Measures the average squared difference between actual and predicted values.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad (4)$$

It penalizes larger errors more than smaller ones, providing a clear indication of model performance. Lower MSE values indicate better model performance.

Root Mean Squared Error (RMSE): The square root of MSE, converting the error metric back to the same units as the target variable.

$$RMSE = \sqrt{MSE} \quad (5)$$

It helps understand the prediction error in the context of the actual data. Lower RMSE values indicate better model performance.

Mean Absolute Error (MAE): Measures the average absolute difference between actual and predicted values.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad (6)$$

It provides a straightforward interpretation of prediction error. Lower MAE values indicate better model performance.

$R^2$ (Coefficient of Determination): Represents the proportion of variance in the dependent variable that can be predicted from the independent variables.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (7)$$

An $R^2$ value close to 1 indicates that the model explains a large portion of the variance, reflecting better prediction accuracy.

Cross-Validation (CV): To ensure the robustness and reliability of the models, k-fold cross-validation was used. This method involves dividing the training data into five subsets, training the model on four subsets, and validating it on the remaining subset. This process is repeated five times, with each subset serving as the validation set once. Cross-validation metrics (MSE, RMSE, MAE, and $R^2$) are averaged over the five folds to provide a more generalized performance estimate. Notably, during implementation, scikit-learn reports MSE as negative values when using scoring='neg_mean_squared_error'. We report the corresponding positive MSE values in this table for clarity.

*Visualization*

Prediction Plots: Scatter plots of actual versus predicted values were created for both the Math and Portuguese datasets. These plots help visualize the models' performance by showing how closely the predicted values match the actual values.

Feature Importance: Bar charts of feature importance scores were generated to understand the significance of each feature in predicting student performance. The importance scores indicate the contribution of each feature to the predictions, providing insights into the factors most influencing student grades.

**Result**

*Quantitative comparison*

**Table 1. K Cross-Validation (K =5) for Math**

| | MSE | RMSE | MAE | $R^2$ |
|---|---|---|---|---|
| **LR** | 4.532 | 2.129 | 1.281 | 0.779 |
| **CV of LR** | 3.662 | 1.913 | 1.866 | 0.827 |
| **RF** | 4.784 | 2.187 | 1.398 | 0.767 |
| **CV of RF** | 4.535 | 2.087 | 2.087 | 0.784 |
| **EM** | 4.671 | 2.161 | 1.329 | 0.772 |
| **CV of EM** | 4.535 | 2.087 | 2.087 | 0.784 |

**Table 2. K Cross-Validation (K =5) for Por**

| | MSE | RMSE | MAE | $R^2$ |
|---|---|---|---|---|
| **LR** | 1.347 | 1.161 | 0.725 | 0.862 |
| **CV of LR** | 1.705 | 1.166 | 1.277 | 0.804 |
| **RF** | 2.161 | 1.470 | 0.823 | 0.778 |
| **CV of RF** | 2.337 | 1.506 | 1.506 | 0.780 |
| **EM** | 1.633 | 1.278 | 0.767 | 0.833 |
| **CV of EM** | 2.337 | 1.506 | 1.506 | 0.780 |

The results presented in Table 1 and Table 2 demonstrate the performance of three machine learning models—Linear Regression (LR), Random Forest (RF), and an Ensemble Model (EM)—in predicting student grades for Mathematics and Portuguese, using K-fold cross-validation (K=5). Across both subjects, LR achieves the best overall balance between accuracy and robustness, consistently attaining the highest $R^2$ and the lowest error metrics. Notably, the advantage of LR is observed not only in the single train–test split results but also in the CV estimates.

For the Mathematics data set, Linear Regression yielded the lowest MSE of 4.532, and $R^2$ value of 0.779. In comparison, the Random Forest model exhibited higher error metrics, with an MSE of 4.784 and $R^2$ value of 0.767. The Ensemble Model performed moderately, with an MSE of 4.671, and $R^2$ value of 0.772. These results suggest that while the Ensemble Model outperforms the Random Forest, it still falls short of the Linear Regression model's performance. For the Portuguese data set, Linear Regression also outperformed the other models, , with an MSE of 1.347, and a high $R^2$ value of 0.862. The cross-validation results reaffirm this, with an $R^2$ value of 0.804.

Comparing the two subjects, the models achieve higher predictive accuracy on Portuguese than on Mathematics, as reflected by higher $R^2$ and lower absolute error levels. This indicates that, under the same modelling pipeline and predictor set, the Portuguese outcomes are more predictable from the selected features, whereas Mathematics appears to exhibit relatively higher residual variability. This cross-subject difference supports the decision to report results for both datasets separately rather than assuming a single universal performance profile.

Regarding model complexity, RF does not provide a consistent improvement over LR, and the simple averaging ensemble (EM) yields intermediate performance rather than surpassing the best individual model. This suggests that, for this structured dataset with a compact predictor set, adding non-linear model capacity (RF) or combining heterogeneous learners (EM) does not necessarily translate into superior generalisation. Instead, the performance pattern favours a simpler and more interpretable approach.

*Prediction Behaviour*

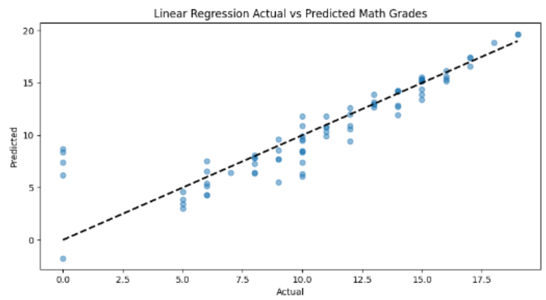


**Figure 4. LR actual vs predicted Math grades**

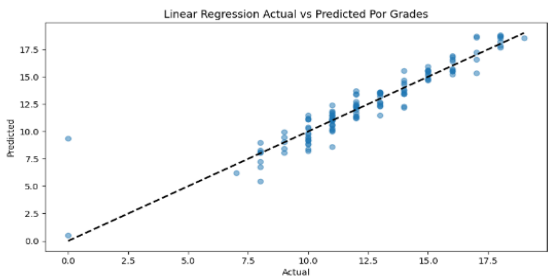


**Figure 5. LR actual vs predicted Por grades**

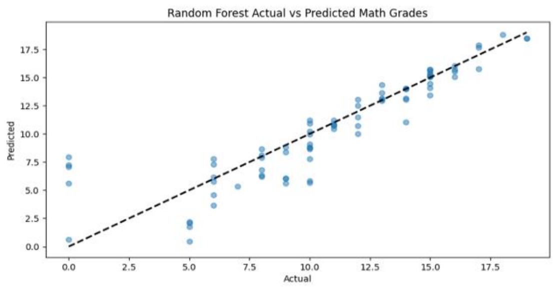


**Figure 6. RF actual vs predicted Math grades**

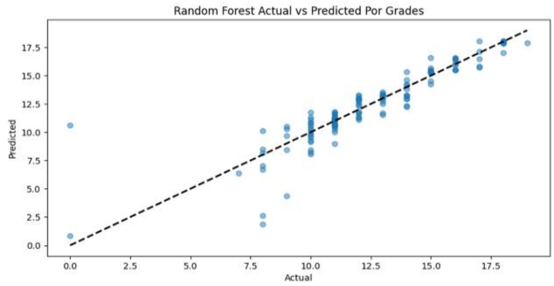


**Figure 7. RF actual vs predicted Por grades**

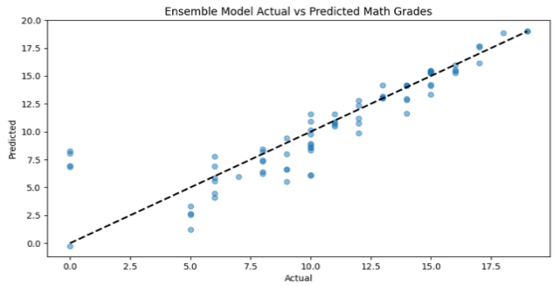


**Figure 8. EM actual vs predicted Math grades**

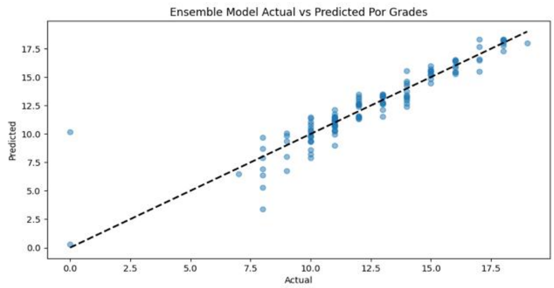


**Figure 9. EM actual vs predicted Por grades**

The results presented from Figure 4 to Figure 9 provide scatter plots of actual versus predicted grades for each model and subject. The x-axis represents the actual grades, and the y-axis denotes the predicted grades. Data points are scattered around a diagonal dashed line, indicating the line of perfect prediction where actual grades match the predicted grades. Overall, all models produce predictions that follow the diagonal trend, indicating meaningful learning of the grade patterns. However, clear differences are visible in the concentration and dispersion of points around the ideal prediction line.

Consistent with the quantitative metrics, LR exhibits tighter clustering around the diagonal, indicating smaller and more uniformly distributed residuals. In contrast, RF shows slightly larger dispersion, particularly for Mathematics, implying that the additional model flexibility may introduce variance without delivering a corresponding gain in predictive accuracy. The ensemble model generally follows the same trend as the base learners and reduces neither dispersion nor systematic deviations sufficiently to outperform LR, which aligns with the tabulated results.

A further qualitative observation from the scatter plots is that prediction errors tend to be more noticeable at the extremes of the grade range, where data are sparser. This behaviour is expected in educational datasets and highlights that predictive performance is strongest in the densest grade regions.

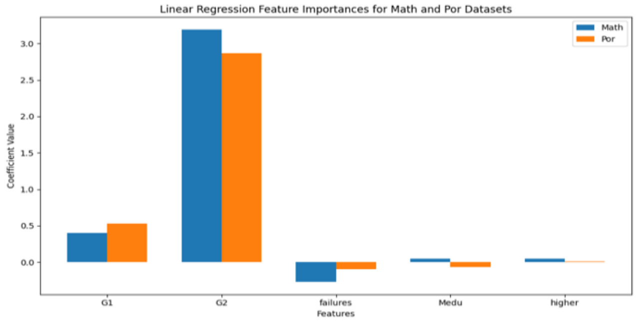


**Figure 10. LR Feature Importances for Math and Por Datasets**

Meanwhile, Figure 10 compares feature importance for predicting grades using Linear Regression. The x-axis lists features: G1, G2, failures, Medu, higher, and the y-axis shows coefficient values, indicating each feature's importance. Blue bars represent math, and orange bars represent Portuguese. A clear and consistent pattern emerges across both subjects: G2 is the dominant predictor, followed by G1, while Failures, Medu, and Higher contribute comparatively smaller effects. This indicates that earlier-term academic performance carries the strongest predictive signal for final outcomes, whereas demographic or aspiration-related variables provide additional but secondary explanatory power. Importantly, the feature influence pattern is broadly consistent across Math and Por, supporting the use of a unified predictor set for fair model comparison. At the same time, the observed subject-level differences in overall predictability (Table 1 and Table 2) suggest that even with similar dominant predictors, the mapping from early indicators to final performance can vary by subject context, motivating subject-specific reporting and interpretation.

## Discussion

This study compared three regression approaches: Linear Regression (LR), Random Forest (RF), and a simple averaging Ensemble Model (EM), for predicting students' final grades (G3) in Mathematics and Portuguese using the UCI Student Performance dataset. Overall, the results show that LR provides the most favourable performance–simplicity trade-off, achieving the strongest predictive accuracy across both subjects while remaining highly interpretable. In contrast, the more complex RF model and the simple ensemble do not consistently improve predictive performance, suggesting that increased model complexity alone does not guarantee better generalisation in this structured educational dataset. The linear regression model has demonstrated certain effectiveness in predicting students' performance, providing educators with a possible tool for designing personalized learning plans and allocating educational resources, ensuring that intervention measures can be implemented in a timely and effective manner to improve educational outcomes in different subjects and student groups.

Comparing the two subjects, the models generally achieved higher predictive performance for Portuguese than for Mathematics, indicating that, under the same modelling pipeline and predictor set, Portuguese outcomes are more predictable from the selected features. This difference should be interpreted cautiously; it does not necessarily imply that one subject is inherently easier to predict in general. Rather, it may reflect subject-specific grading patterns, noise levels, or relationships between early and final assessments within this dataset, which, to some extent, suggests that there may be different features for STEM module and non-STEM module predictions.

A key methodological implication concerns the practical interpretation of "early intervention." Although the use of G1 and G2 improves predictive accuracy, it also means that the model primarily supports a mid-semester or operational prediction scenario, like predicting final grade after early assessments are available, rather than a true "early-warning" system at the very start of the academic term. Future work aimed at earlier-stage prediction would need to rely more on non-grade features, such as engagement proxies, attendance, learning behaviour signals, and formative assessments. It should explicitly evaluate the trade-off between interpretability and accuracy under that constraint.

In addition, there are several limitations that should be acknowledged. First, the study was derived from a specific educational context; therefore, generalisation to other regions or educational systems should be made with caution. Second, the ensemble approach was intentionally kept simple to serve as a transparent baseline. More advanced ensembling strategies, for instance, stacking with learned weights, could yield better performance but would introduce additional complexity and tuning requirements, which may reduce interpretability and increase the risk of overfitting under limited sample sizes. Third, while the reported improvements are consistent across evaluation metrics and cross-validation, the study does not claim that complex models are ineffective in educational prediction broadly. Instead, the results indicate that for structured datasets with strongly predictive prior-performance features, simpler models may be sufficient and preferable.

Overall, this work contributes a measured empirical perspective: rather than assuming that more sophisticated machine learning models will always outperform simpler alternatives, the findings suggest that model selection in educational analytics should be driven by data characteristics and deployment requirements, with interpretability and practicality considered alongside predictive accuracy.

## Conclusion

Using the UCI Student Performance dataset, students' final grades in Mathematics and Portuguese was predicted by comparing three different machine learning models, linear regression, random forest, and integrated models. The findings indicate that Linear Regression outperforms the other models, achieving the highest $R^2$ values and the lowest MSE, RMSE, and MAE values in both subjects. This suggests that Linear Regression is particularly effective at capturing key features to develop accurate predictive models for student performance. However, the results also highlight the challenges in creating a universally accurate model for different subject modules. The comparative analysis between Mathematics and Portuguese datasets suggests that distinct features may be necessary for predicting performance in STEM versus non-STEM subjects. Across both subjects, prior academic performance indicators (G1 and G2) are the most influential predictors, highlighting the central role of early-term achievement in forecasting final outcomes.

Future research can extend this work in three directions. First, to support truly early-stage intervention, models should be evaluated using predictors available at the start of the term and compared under the same interpretability–accuracy lens. Second, validation on more diverse datasets and educational contexts would strengthen generalisability and improve relevance to broader regional practices. Third, while advanced ensembles and deep learning methods may be beneficial for larger and more complex data sources, their added value should be assessed against the practical costs of deployment, transparency, and maintenance, particularly in educational settings where interpretability and accountability are essential.

## Acknowledgement

The authors would like to express their sincere gratitude to the International Conference on Higher Education Learning and Teaching (ICHELT2024) Organizing Committee for providing the platform for this research. The authors also thank Xi'an Jiaotong-Liverpool University for supporting the completion of this study. No external funding was received for this work. The source code used in this research has been made publicly available at:

https://github.com/JiajunGuo1027/AI-Driven-Student-Performance-Prediction-and-Analysis-of-Influencing-Factors.git

## Conflict of Interest

The authors declare that there is no conflict of interest associated with the publication of this manuscript.

## References

Chen, L., Chen, P., & Lin, Z. (2020). Artificial Intelligence in Education: A Review. *IEEE Access*, *8*, 75264–75278. https://doi.org/10.1109/ACCESS.2020.2988510

Cortez, P. (2008). *Student Performance* [Data set]. UCI Machine Learning Repository. https://doi.org/10.24432/C5TG7T

Kuleto, V., Ilić, M., Dumangiu, M., Ranković, M., Martins, O. M. D., Păun, D., & Mihoreanu, L. (2021). Exploring Opportunities and Challenges of Artificial Intelligence and Machine Learning in Higher Education Institutions. *Sustainability*, *13*(18), 10424. https://doi.org/10.3390/su131810424

Luan, H., Geczy, P., Lai, H., Gobert, J., Yang, S. J. H., Ogata, H., Baltes, J., Guerra, R., Li, P., & Tsai, C.-C. (2020). Challenges and Future Directions of Big Data and Artificial Intelligence in Education. *Frontiers in Psychology*, *11*. https://doi.org/10.3389/fpsyg.2020.580820

Luan, H., & Tsai, C.-C. (2021). A Review of Using Machine Learning Approaches for Precision Education. *Educational Technology & Society*, *24*(1), 250–266.

Shafique, R., Aljedaani, W., Rustam, F., Lee, E., Mehmood, A., & Choi, G. S. (2023). Role of Artificial Intelligence in Online Education: A Systematic Mapping Study. *IEEE Access*, *11*, 52570–52584. https://doi.org/10.1109/ACCESS.2023.3278590

Vakhabova, S. A., Valery V. Kosulin, & Ana Zizaeva. (2019). *Artificial intelligence in education: Challenges and opportunities for sustainable development*. https://www.semanticscholar.org/paper/Artificial-intelligence-in-education-%3A-challenges/697ba06bfcabbbde6292d979b87b2642115f1099

Yağcı, M. (2022). Educational data mining: Prediction of students' academic performance using machine learning algorithms. *Smart Learning Environments*, *9*(1), 11. https://doi.org/10.1186/s40561-022-00192-z

Zafari, M., Bazargani, J. S., Sadeghi-Niaraki, A., & Choi, S.-M. (2022). Artificial Intelligence Applications in K-12 Education: A Systematic Literature Review. *IEEE Access*, *10*, 61905–61921. IEEE Access. https://doi.org/10.1109/ACCESS.2022.3179356