Multi-Level Student Performance Prediction Using Machine Learning Based on Socio-Demographic and Development Indicators

Dhedy Husada Fadjar[1,*], Aisyah Fortuna Nadapoetri Alhaq[2]

[1,2]Department of Computer Science and Electronics, Faculty of Mathematics and Natural Sciences, Universitas Gadjah Mada, Indonesia

ABSTRACT

This study investigates the application of machine learning techniques to predict student performance across multiple educational levels using socio-demographic and development indicators. A combined dataset covering Primary, Lower Secondary, and Upper Secondary levels was utilized to develop a multi-level predictive framework. A clean machine learning pipeline was implemented, including the removal of potential data leakage features, feature engineering through development index encoding, and standardized preprocessing. Two ensemble models, Random Forest and XGBoost, were evaluated using Mean Absolute Error, Root Mean Squared Error, and the coefficient of determination. The results show that Random Forest outperforms XGBoost, achieving the highest predictive performance with an R^2 score of 0.716, indicating strong explanatory power. Feature importance analysis reveals that education level is the most influential predictor, followed by temporal and macro-level variables such as year, regional classification, and development index, highlighting the significant role of socio-economic context in shaping student outcomes. Further analysis indicates that prediction performance varies across educational stages and that the model tends to underestimate high-performing students and overestimate low-performing students, suggesting limitations in capturing extreme values. Overall, the findings demonstrate that integrating machine learning with multi-level and macro-level analysis provides meaningful insights into student performance and supports data-driven educational policy and planning.

Keywords Student Performance Prediction, Machine Learning in Education, Multi-Level Analysis, Socio-Demographic Factors, Educational Data Mining

Introduction

Student performance prediction has become an increasingly important topic in the field of educational data mining and artificial intelligence in learning. Accurate prediction models can support early intervention, improve learning outcomes, and assist policymakers in designing effective educational strategies [1]. With the growing availability of large-scale educational datasets, machine learning techniques have been widely adopted to model complex relationships between student outcomes and various influencing factors [2]. Previous studies have demonstrated that factors such as socio-economic status, demographic characteristics, and institutional variables play a significant role in determining academic performance.

Recent advances in machine learning have enabled the development of more sophisticated predictive models, including ensemble methods such as Random Forest and gradient boosting algorithms [3]. These approaches have shown strong performance in handling non-linear relationships and heterogeneous

Submitted 25 May 2026
Accepted 5 August 2026
Published 17 November 2026

*Corresponding author
Dhedy Husada Fadjar,
dhedyhusadafadjar@mail.ugm.ac.id

Additional Information and Declarations can be found on page 265

DOI: 10.63913/ail.v2i4.73
© Copyright
2026 Fadjar and Alhaq

Distributed under
Creative Commons CC-BY 4.0

How to cite this article: Fadjar and Alhaq, "Multi-Level Student Performance Prediction Using Machine Learning Based on Socio-Demographic and Development Indicators," *Artif. Intell. Learn.*, vol. 2, no. 4, pp. 252-267, 2026.

data commonly found in educational settings. In addition, the integration of socio-demographic variables has improved the interpretability and predictive power of these models [4]. Some studies have also explored the use of temporal data and longitudinal analysis to capture changes in student performance over time. These developments represent the current state of the art in student performance prediction, where machine learning models are increasingly used to extract meaningful insights from complex educational data.

Despite these advancements, several important gaps remain in the existing literature. First, most studies focus on a single level of education, such as primary or secondary education, without considering differences across multiple educational stages. This limits the generalizability of the findings and overlooks structural variations in learning patterns. Second, many studies emphasize micro-level factors, such as individual or household characteristics, while giving less attention to macro-level indicators such as regional classification and development status. As a result, the broader socio-economic context influencing student performance is often underrepresented. Third, limited research has examined the generalization capability of machine learning models across different educational levels, which is essential for building robust and scalable predictive systems [5].

To address these gaps, this study proposes a multi-level machine learning framework for predicting student performance across Primary, Lower Secondary, and Upper Secondary education levels. The proposed approach integrates both micro-level socio-demographic variables and macro-level development indicators to provide a more comprehensive understanding of student outcomes. Furthermore, this study evaluates model performance across different educational stages and analyzes the relative importance of features to identify key factors influencing performance. By combining multi-level analysis with machine learning techniques, this study aims to contribute to the development of more generalizable and interpretable models for educational prediction.

The main contributions of this study are threefold. First, it introduces a multi-level predictive framework that captures structural differences across educational stages. Second, it incorporates macro-level socio-economic indicators into the modeling process, which are often neglected in previous studies. Third, it evaluates the generalization capability of machine learning models across multiple levels of education. These contributions provide valuable insights for both researchers and policymakers in understanding and improving student performance in diverse educational contexts.

Literature Review and Related Works

The application of machine learning in education has gained significant attention in recent years, particularly in the area of student performance prediction. Various studies have demonstrated that machine learning models can effectively capture complex relationships between student outcomes and multiple influencing factors. Traditional approaches often relied on statistical methods, but recent advancements have shifted toward data-driven models that provide higher predictive accuracy and flexibility [6], [7].

Several studies have explored the use of supervised learning algorithms such as decision trees, support vector machines, and neural networks for predicting academic performance. Among these, ensemble methods such as Random

Forest and gradient boosting have shown superior performance due to their ability to handle non-linear relationships and high-dimensional data [8], [9]. These models are particularly effective in educational datasets, which often contain heterogeneous and noisy features [10].

In addition to algorithm selection, the choice of input features plays a crucial role in model performance. Prior research has consistently highlighted the importance of socio-demographic variables, including family background, economic status, and geographic location, in influencing student outcomes [11], [12]. Incorporating such features into predictive models has been shown to significantly improve accuracy and interpretability [13].

Recent studies have also emphasized the importance of temporal and longitudinal data in capturing trends in student performance over time. By integrating time-related features, models can better reflect changes in educational systems and student behavior [14], [15]. However, most existing works focus primarily on micro-level factors and often overlook macro-level indicators such as regional development and national socio-economic conditions.

Another limitation in the current literature is the focus on single-level educational analysis. Many studies concentrate on a specific educational stage, such as primary or secondary education, without considering differences across multiple levels [16], [17]. This approach restricts the generalizability of findings and fails to capture structural variations in learning patterns across educational stages. Only a limited number of studies have attempted to incorporate multi-level analysis, and these often lack comprehensive evaluation across diverse contexts [18].

Furthermore, the issue of model generalization across different datasets and educational levels remains underexplored. While some studies report high predictive accuracy, they often do not evaluate how well the models perform when applied to different populations or educational stages [19]. This raises concerns about the robustness and scalability of existing approaches.

To address these limitations, recent research has started to integrate multi-level frameworks and macro-level indicators into machine learning models. These approaches aim to provide a more holistic understanding of student performance by combining individual-level and contextual factors [20]. However, there is still a need for comprehensive studies that systematically evaluate model performance across multiple educational levels while incorporating both socio-demographic and development indicators.

In summary, although significant progress has been made in applying machine learning to student performance prediction, gaps remain in terms of multi-level analysis, macro-level integration, and model generalization. These gaps highlight the need for more comprehensive frameworks that can capture the complexity of educational systems and provide more reliable and interpretable predictions.

Methodology

Research Framework

This study follows a structured machine learning workflow to ensure a systematic and reproducible analysis. The overall research steps are illustrated

in figure 1, which outlines the main stages including data collection, preprocessing, feature engineering, model development, evaluation, and multi-level analysis. The framework begins with integrating datasets from multiple educational levels, followed by data cleaning and transformation. Subsequently, machine learning models are trained and evaluated, and the results are further analyzed to extract insights regarding student performance across different educational stages.

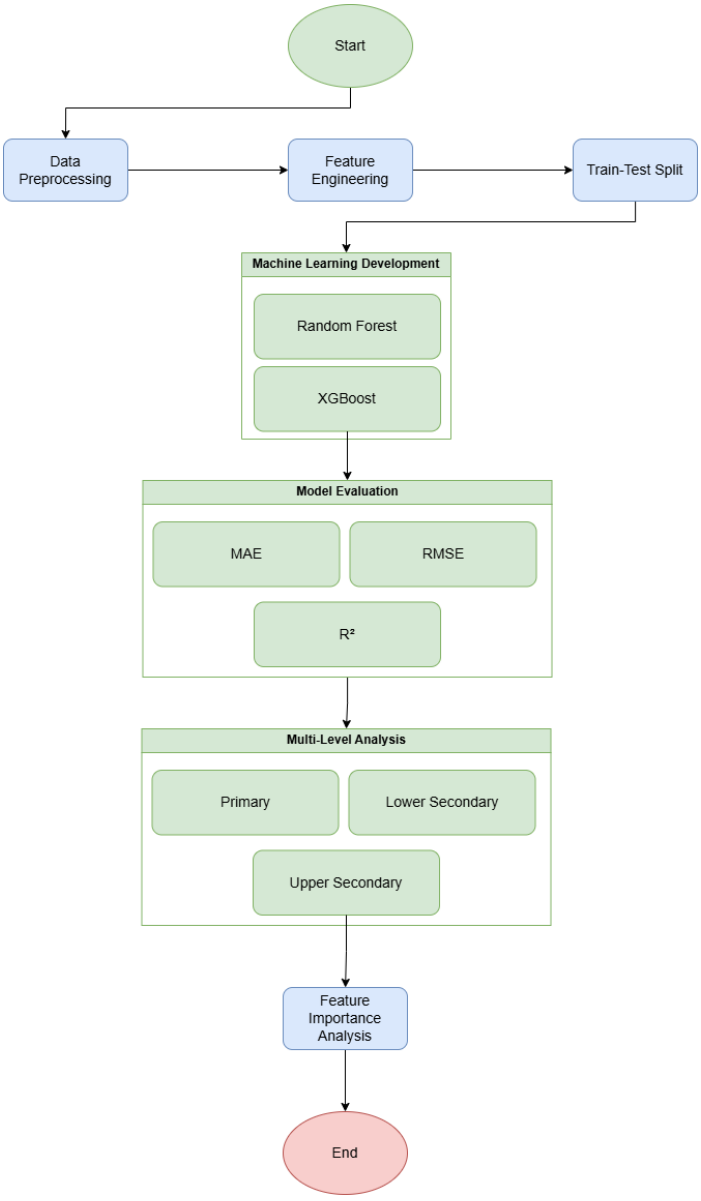


Figure 1 Research Framework of the Proposed Study

Dataset Description

This study utilizes a combined dataset consisting of three educational levels, namely Primary, Lower Secondary, and Upper Secondary. Each dataset contains student performance indicators along with socio-demographic and development-related variables collected across multiple countries. To enable multi-level analysis, a new categorical feature labeled Education Level was

introduced to distinguish between the three stages. The datasets were then merged into a single dataset to allow unified modeling and evaluation.

Non-informative attributes such as country identifiers, data source labels, and metadata were removed to ensure that only relevant predictive features were retained.

Data Preprocessing and Leakage Prevention

To ensure robust model performance, preprocessing steps were carefully applied. Potential data leakage features, such as gender distribution and wealth quintile indicators, were removed because they may implicitly encode the target variable and lead to biased predictions.

Missing values in numerical features were handled using median imputation, while categorical features were imputed using the most frequent value. Numerical features were standardized to ensure consistent scaling. These steps were implemented using a pipeline structure to guarantee consistency between training and testing phases.

Feature Engineering

Feature engineering was performed to improve model performance and capture more informative patterns within the dataset. Development-related categorical variables were transformed into numerical representations using ordinal encoding, where each category reflects a different level of socio-economic development. This transformation enables the model to interpret development status as a structured hierarchy rather than independent categories. By assigning ordered values to development levels, the model can better capture the relationship between socio-economic conditions and student performance. This is particularly important in a cross-country dataset, where variations in development levels often correspond to differences in educational infrastructure, access to resources, and overall learning environments. As a result, the encoded development index provides meaningful macro-level context that enhances the model's predictive capability.

In addition to socio-economic features, temporal information was incorporated to capture changes in student performance over time. The original time-related variable was transformed into a standardized Year feature, allowing the model to learn temporal trends and patterns across different periods. This transformation helps account for potential improvements or declines in educational outcomes due to policy changes, economic development, or external factors. By including temporal dynamics, the model is better equipped to generalize across different time frames and avoid biases related to specific periods. Overall, the integration of both socio-economic and temporal features strengthens the model's ability to capture complex, real-world relationships in the data and improves its overall predictive performance.

Model Development

Two ensemble machine learning models were used in this study, namely Random Forest and XGBoost, due to their strong capability in handling complex, non-linear relationships in tabular data. Random Forest operates based on a bagging strategy, where multiple decision trees are independently trained on randomly sampled subsets of the data and features. The final prediction is obtained by aggregating the outputs of all trees, which reduces

variance and improves model generalization. This approach makes Random Forest particularly robust to noise and overfitting, especially in datasets with heterogeneous feature distributions such as socio-demographic data. In contrast, XGBoost is a gradient boosting algorithm that constructs decision trees sequentially, where each new tree is trained to correct the errors of the previous ones. By optimizing a loss function through gradient descent and incorporating regularization techniques, XGBoost is able to achieve high predictive accuracy while controlling model complexity.

To ensure a consistent and reproducible modeling process, a unified machine learning pipeline was implemented. This pipeline integrates all preprocessing steps, including imputation, scaling, and categorical encoding, with the model training phase into a single workflow. By encapsulating these steps within a pipeline, the transformations learned from the training data are automatically and consistently applied to the testing data, thereby preventing inconsistencies and reducing the risk of data leakage. In addition, the pipeline structure simplifies the experimentation process, allowing different models to be evaluated under identical preprocessing conditions. This design improves the reliability of performance comparisons and ensures that the observed results are attributable to the models themselves rather than differences in data handling procedures.

Training Strategy and Evaluation Metrics

The dataset was split into training and testing sets using an 80:20 ratio. The models were trained on the training set and evaluated on the testing set to measure generalization performance.

Three evaluation metrics were used:

$$MAE = \frac{1}{n}\sum|y_i - \hat{y}_i|$$
$$RMSE = \sqrt{\frac{1}{n}\sum(y_i - \hat{y}_i)^2} \quad (1)$$
$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}$$

y_i represents the actual values, $\hat{y}_i$ represents the predicted values, and $\bar{y}$ is the mean of the actual values.

Multi-Level Analysis

To examine differences across educational stages, the dataset was divided into three subsets based on education level: Primary, Lower Secondary, and Upper Secondary. Separate models were trained and evaluated for each subset.

This approach allows the analysis to capture structural differences in student performance across educational levels and evaluate how well the models generalize within each level.

Feature Importance Analysis

Feature importance analysis was conducted using the Random Forest model to identify the most influential variables in predicting student performance. In Random Forest, feature importance scores are computed based on the average reduction in impurity, which reflects how much each feature contributes to

decreasing prediction error across all decision trees in the ensemble. Features that consistently provide better splits and reduce variance in the target variable are assigned higher importance scores. This approach allows the model to quantify the relative contribution of each variable in a data-driven manner, without requiring prior assumptions about feature relevance. As a result, the analysis provides a transparent and interpretable mechanism for understanding how different input variables influence the prediction outcomes.

The top-ranked features were further analyzed to examine the role of socio-demographic, temporal, and macro-level factors in shaping student performance. Socio-demographic variables capture individual and household characteristics, while temporal features reflect trends and changes over time. Macro-level indicators, such as regional classification and development index, provide broader contextual information related to economic and geographic conditions. By comparing the importance of these feature groups, the study is able to assess the relative impact of micro-level and macro-level factors on student outcomes. This analysis not only enhances the interpretability of the model but also provides meaningful insights for policymakers and educators by highlighting the key drivers of student performance across different contexts.

Algorithm 1: Multi-Level Student Performance Prediction

Input: Primary Dataset (D_p), Lower Dataset (D_l), Upper Dataset (D_u)
Output: Best model (M*), MAE, RMSE, R^2, Feature Importance
Steps:
Start
Data Merging:
Merge all datasets:

$$D = D_p \cup D_l \cup D_u$$

Label Addition:
Add the Education Level feature to the dataset:

$$D = D \cup \{Education_Level\}$$

Data Leakage Removal:
Remove features that may cause data leakage:

$$D = D \setminus L$$

Feature Engineering:
Create variables:
Dev_Index from Development_Regions
Year from Time_Period
Variable Definition:
Features:

$$X = D \setminus \{Y\}$$

Target:

$$y = Y$$

Preprocessing:
For each numerical feature $x_i \in X_{num}$:
Fill missing values with the median
Normalization:

$$x_i = \frac{x_i - \mu}{\sigma}$$

For each categorical feature $x_j \in X_{cat}$:
Fill missing values with the mode
Apply One-Hot Encoding

Data Splitting:
Split the data into training and testing sets:

$$(X_{train}, X_{test}, y_{train}, y_{test}) = split(X, y, 80{:}20)$$

Model Training:
For each model $m \in \{RF, XGB\}$:
Train the model:

$$m.fit(X_{train}, y_{train})$$

Prediction:

$$\hat{y} = m(X_{test})$$

Model Evaluation:
Compute evaluation metrics:
MAE:

$$MAE = \frac{1}{n}\sum | y_i - \hat{y}_i |$$

RMSE:

$$RMSE = \sqrt{\frac{1}{n}\sum(y_i - \hat{y}_i)^2}$$

R^2:

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}$$

Best Model Selection:
Select the model with the highest R^2:

$$M^* = \arg\max(R^2)$$

Multi-Level Analysis:
For each level $l \in \{Primary, Lower, Upper\}$:
Train the best model M^*on dataset D_l
Evaluate and record R_l^2
Feature Importance:
Extract feature importance from the best model:

$$F = importance(M^*)$$

Output:
Return M*, MAE, RMSE, R^2, and Feature Importance
End

Results

Model Performance

The performance of the machine learning models was evaluated using Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and the coefficient of determination (R^2), as presented in table 1. These metrics provide a comprehensive assessment of model accuracy by capturing both average prediction error and variance explanation. The results show that the Random Forest model achieved an R^2 score of 0.716, indicating that it explains approximately 71.6% of the variance in student performance. In comparison, the XGBoost model achieved a slightly lower R^2 score of 0.680, suggesting a marginally weaker ability to capture the underlying data patterns. This difference highlights the effectiveness of ensemble bagging methods in handling complex and heterogeneous educational data.

In addition to the R^2 metric, the error-based measures further confirm the superiority of the Random Forest model. Specifically, Random Forest obtained a lower MAE of 7.04 and RMSE of 10.77, compared to XGBoost with MAE of

7.58 and RMSE of 11.43. Lower MAE indicates that the average prediction error is smaller, while lower RMSE suggests fewer large deviations between predicted and actual values. Together, these results demonstrate that Random Forest provides more accurate and stable predictions. This performance advantage may be attributed to its robustness to noise and its ability to model non-linear relationships without extensive parameter tuning.

Table 1 Performance Metrics of Machine Learning Models

Model	MAE	RMSE	R^2
Random Forest	7.04	10.77	0.716
XGBoost	7.58	11.43	0.680

The comparative performance of the models is further illustrated in figure 2, which presents the R^2 scores for both Random Forest and XGBoost. As shown in the figure, Random Forest consistently achieves a higher R^2 value than XGBoost, indicating a stronger ability to explain the variance in student performance across the dataset. The visual difference between the two models highlights the relative stability and effectiveness of Random Forest in capturing complex relationships within the data. Although the performance gap is not extremely large, it is sufficiently consistent to suggest a meaningful advantage. Overall, both models demonstrate strong predictive capability, as reflected by their relatively high R^2 scores, reinforcing the relevance of socio-demographic and development indicators as informative features for modeling student performance.

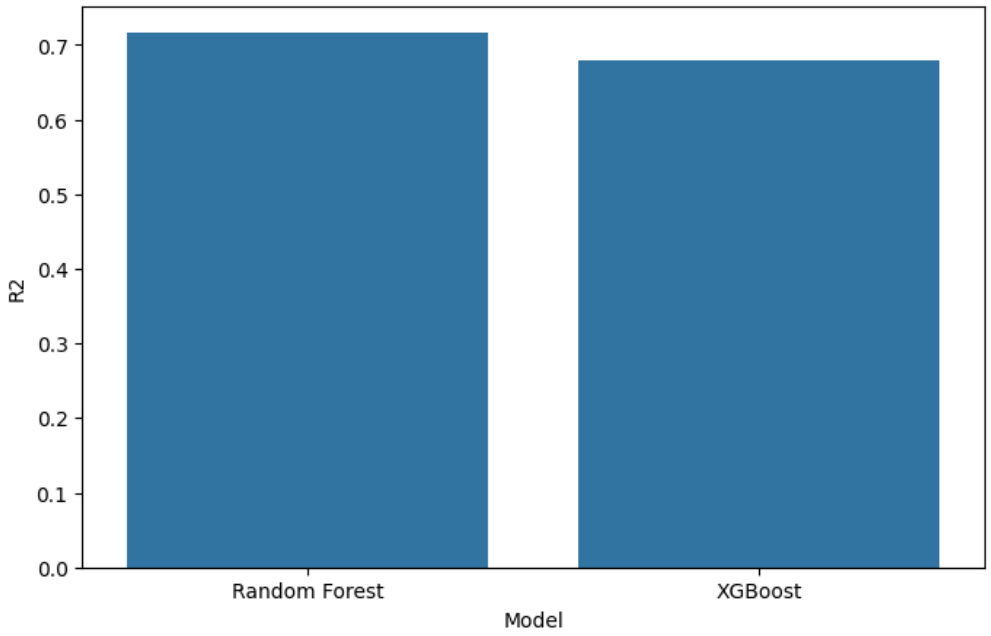


Figure 2 R^2 Performance of Machine Learning Models

Actual vs Predicted Analysis

The relationship between actual and predicted values is presented in figure 3, which provides a visual assessment of how well the model approximates the true student performance scores. As illustrated in the figure, the majority of data points are closely aligned with the diagonal reference line, indicating that the

model is able to capture the overall trend and general distribution of the data. This alignment suggests that the predictions are reasonably accurate for a large portion of the observations. However, a systematic pattern of deviation can be observed at the extremes of the distribution. For higher actual values, the model tends to underestimate the outcomes, while for lower values, it slightly overestimates them. This pattern indicates that the model has difficulty capturing extreme cases, which may be influenced by factors not fully represented in the dataset. Such behavior reflects a regression-to-the-mean tendency, where predictions are biased toward the central range of the data, resulting in reduced accuracy for outlier observations.

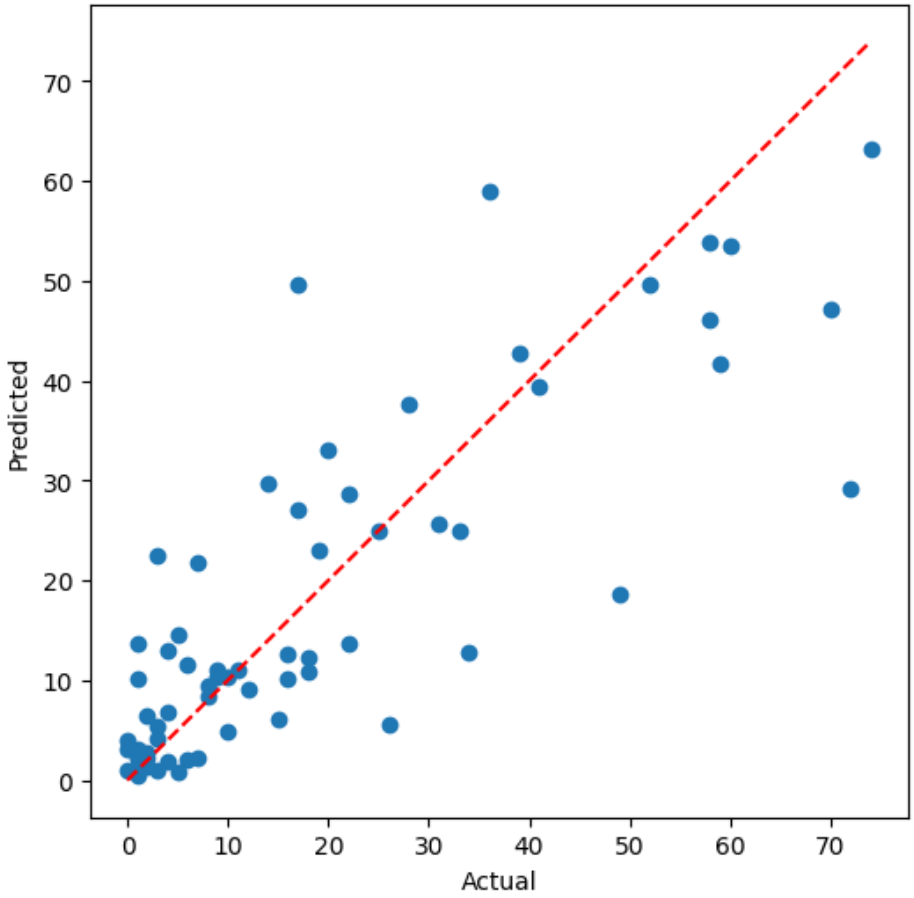


Figure 3 **Actual vs Predicted Values**

Error Distribution

The distribution of prediction errors is shown in figure 4, providing insight into the accuracy and reliability of the model across different observations. As indicated in the figure, the error values are approximately centered around zero, which suggests that the model does not produce consistent overestimation or underestimation and therefore does not exhibit significant systematic bias. The majority of errors are concentrated within a relatively narrow range, indicating that most predictions are close to the actual values and that the model performs consistently for typical cases. This concentration reflects stable predictive behavior and suggests that the model generalizes well across the dataset. However, the presence of several observations with larger error magnitudes

reveals that certain cases are more challenging to predict accurately. These larger deviations may be associated with complex patterns, extreme values, or latent variables that are not captured by the available features, highlighting potential limitations in the dataset and opportunities for further model improvement.

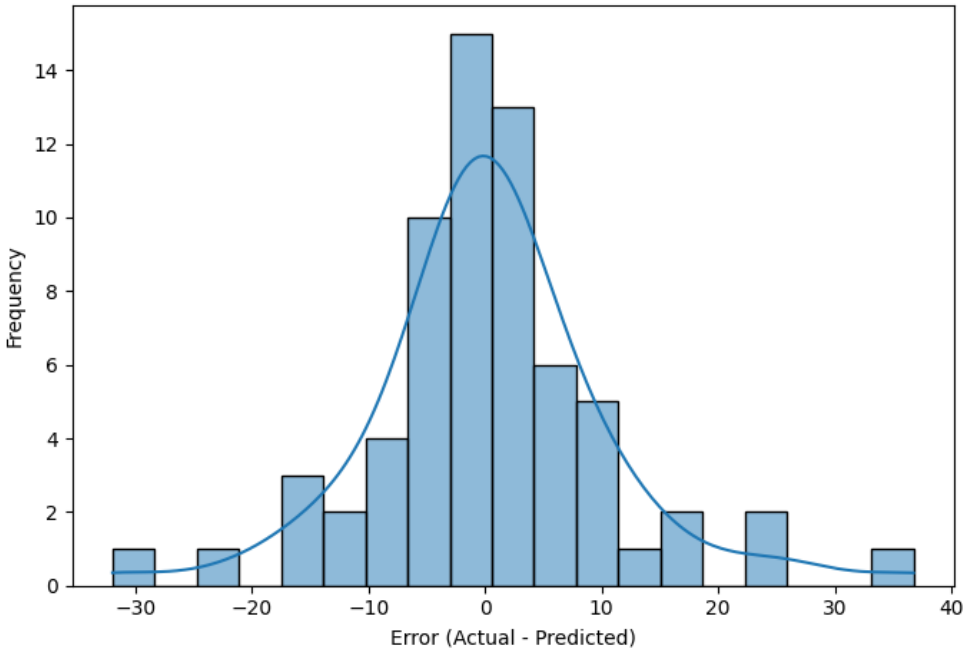


Figure 4 Prediction Error Distribution

Feature Importance Analysis

The importance of input features in the Random Forest model is presented in figure 5, which highlights the relative contribution of each variable in predicting student performance. As shown in the figure, the education level, particularly the Upper Secondary category, emerges as the most influential feature, indicating that differences across educational stages significantly affect model predictions. This finding suggests that student performance is not uniform across levels, but instead follows distinct structural patterns depending on the stage of education. In addition to education level, temporal features such as Year also play a prominent role, reflecting potential trends or changes in educational outcomes over time. Furthermore, macro-level indicators, including regional classification and development index, are among the top contributing features, emphasizing the impact of broader socio-economic and geographic conditions on student performance. Together, these results demonstrate that both temporal dynamics and socio-economic context are critical factors in shaping educational outcomes, and their inclusion enhances the model's ability to capture complex, real-world patterns.

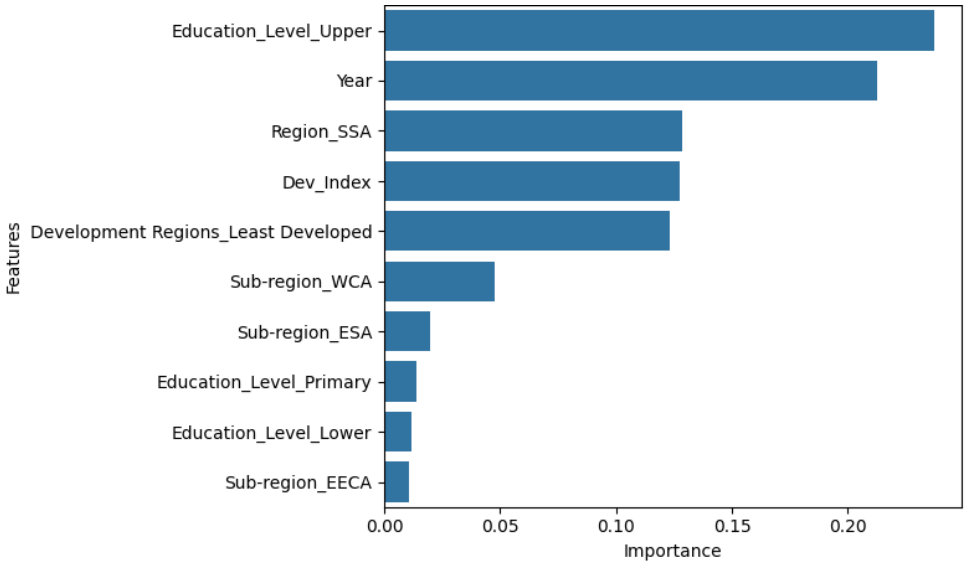


Figure 5 Top 10 Important Features

Multi-Level Insights

The dominance of education level as a key feature suggests that student performance prediction varies significantly across different educational stages. This finding indicates that each level of education, including Primary, Lower Secondary, and Upper Secondary, exhibits distinct learning characteristics and performance patterns. The strong influence of the Upper Secondary category, in particular, may reflect increased academic complexity, variation in curriculum, and broader socio-economic pressures that become more pronounced at higher levels of education. As a result, the model relies heavily on this feature to differentiate patterns in the data, highlighting that student performance cannot be treated as a homogeneous outcome across all educational stages.

These results reinforce the importance of adopting a multi-level modeling approach when analyzing student performance. A single generalized model that ignores educational levels may fail to capture the underlying structural differences and could lead to reduced predictive accuracy. By explicitly incorporating education level into the modeling framework, the analysis is better able to reflect real-world educational dynamics and heterogeneity. This approach not only improves model performance but also provides more meaningful insights into how different factors influence student outcomes at each stage, which is valuable for designing targeted educational policies and interventions.

Discussion

The findings of this study demonstrate that machine learning models, particularly Random Forest, are capable of effectively capturing the complex relationships between socio-demographic factors and student performance across multiple education levels. The superior performance of Random Forest compared to XGBoost suggests that ensemble bagging approaches are better suited for handling heterogeneous and potentially noisy educational data, especially in a cross-country context. This finding is consistent with previous studies showing that Random Forest performs strongly in educational

performance prediction tasks because of its robustness in handling diverse demographic, academic, and socio-economic attributes [21], [22]. The relatively high R^2 values achieved by both models indicate that the selected features provide strong explanatory power, confirming that student performance is closely linked to measurable socio-economic and contextual indicators. Similar findings have been reported in educational data mining studies, where demographic, socio-economic, academic, and contextual factors were found to significantly influence predictive accuracy and academic outcomes [23]. Furthermore, the feature importance analysis reveals that education level is the most dominant predictor, highlighting that performance patterns differ significantly across Primary, Lower Secondary, and Upper Secondary levels. This suggests that as students progress through the education system, the underlying factors influencing performance become more structured and potentially more complex, reinforcing the importance of considering educational stages explicitly in predictive modeling.

In addition to education level, the strong contribution of macro-level variables such as regional classification and development index emphasizes the critical role of broader socio-economic context in shaping student outcomes. These results indicate that student performance is not solely determined by individual-level characteristics, but is also influenced by structural inequalities and differences in educational resources across regions. Prior research using large-scale educational datasets has similarly shown that broader contextual variables and cross-background differences are important in explaining variations in student academic performance [24]. The observed prediction behavior, where the model tends to underestimate high-performing students and overestimate low-performing ones, suggests limitations in capturing extreme cases, possibly due to latent variables that are not included in the dataset. This highlights the need for incorporating additional factors such as school quality, teacher effectiveness, student behavior, engagement, and learning environment in future studies. Previous studies have also emphasized that behavioral, emotional, and school-related features can improve the prediction of student outcomes and provide more actionable insights for early intervention [25]. Overall, the results support the adoption of a multi-level modeling approach, as a single generalized model may not adequately represent the structural differences across educational stages, and a more nuanced framework can provide deeper insights for both research and policy development.

Conclusion

This study demonstrates that machine learning models can effectively predict student performance across multiple education levels using socio-demographic and development-related indicators. Among the evaluated models, Random Forest achieved the best performance, outperforming XGBoost in terms of MAE, RMSE, and R^2. This result indicates that Random Forest is more suitable for handling heterogeneous educational data, particularly in contexts involving cross-country variation and multi-level education structures. This finding is also consistent with previous studies showing that Random Forest and other ensemble-based models are effective for predicting student outcomes using demographic, socio-economic, and academic variables.

The results further show that education level is the most influential predictor of student performance, suggesting that performance patterns differ substantially across Primary, Lower Secondary, and Upper Secondary levels. In addition,

macro-level variables such as regional classification and development index contribute meaningfully to the prediction results, highlighting the importance of socio-economic context and structural inequality in shaping educational outcomes. These findings support prior research emphasizing that student performance is influenced not only by individual characteristics, but also by broader contextual, institutional, and socio-economic conditions.

Despite its promising results, this study has several limitations. The model still tends to underestimate high-performing students and overestimate low-performing students, indicating that extreme cases are not fully captured by the available features. This limitation may be caused by the absence of important latent variables such as school quality, teacher effectiveness, classroom environment, parental involvement, and student motivation. Therefore, future research should incorporate richer individual-, school-, and system-level variables, apply more advanced interpretability methods such as SHAP or LIME, and evaluate model robustness across different countries and education levels. Overall, this study confirms the value of a multi-level machine learning approach for understanding student performance and provides useful insights for data-driven educational planning and policy development.

Declarations

Author Contributions

Conceptualization: D.H.F., A.F.N.A.; Methodology: A.F.N.A.; Software: D.H.F.; Validation: D.H.F., A.F.N.A.; Formal Analysis: D.H.F., A.F.N.A.; Investigation: D.H.F.; Resources: A.F.N.A.; Data Curation: A.F.N.A.; Writing Original Draft Preparation: D.H.F., A.F.N.A.; Writing Review and Editing: A.F.N.A., D.H.F.; Visualization: D.H.F.; All authors have read and agreed to the published version of the manuscript.

Data Availability Statement

The data presented in this study are available on request from the corresponding author.

Funding

The authors received no financial support for the research, authorship, andor publication of this article.

Institutional Review Board Statement

Not applicable.

Informed Consent Statement

Not applicable.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] S. B. Kotsiantis, “Use of machine learning techniques for educational purposes: A decision support system for forecasting students’ grades,” *Artificial Intelligence*

Review, vol. 37, no. 4, pp. 331–344, 2012, doi: 10.1007/s10462-011-9234-x.

[2] C. Romero and S. Ventura, "Educational data mining: A review of the state of the art," *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, vol. 40, no. 6, pp. 601–618, 2010, doi: 10.1109/TSMCC.2010.2053532.

[3] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.

[4] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)*, San Francisco, CA, USA, 2016, vol. 2016, no. August, pp. 785–794, doi: 10.1145/2939672.2939785.

[5] R. S. J. d. Baker and K. Yacef, "The state of educational data mining in 2009: A review and future visions," *Journal of Educational Data Mining*, vol. 1, no. 1, pp. 3–17, 2009, doi: 10.5281/zenodo.3554657.

[6] J. Luan, "Data mining and its applications in higher education," *New Directions for Institutional Research*, no. 113, no. April, pp. 17–36, 2002, doi: 10.1002/ir.35.

[7] A. Peña-Ayala, "Educational data mining: A survey and a data mining-based analysis of recent works," *Expert Systems with Applications*, vol. 41, no. 4, pt. 1, pp. 1432–1462, 2014, doi: 10.1016/j.eswa.2013.08.042.

[8] P. Cortez and A. M. G. Silva, "Using data mining to predict secondary school student performance," in *Proceedings of the 5th Future Business Technology Conference (FUBUTEC 2008)*, Porto, Portugal, Apr. 2008, pp. 5–12. [Online]. Available: https://repositorium.uminho.pt/entities/publication/42d177d6-868b-4e6c-8a89-e1e7fa41c1d0

[9] S. Huang and N. Fang, "Predicting student academic performance in an engineering dynamics course: A comparison of four types of predictive mathematical models," *Computers & Education*, vol. 61, no. February, pp. 133–145, 2013, doi: 10.1016/j.compedu.2012.08.015.

[10] J. Xu, K. H. Moon, and M. van der Schaar, "A machine learning approach for tracking and predicting student performance in degree programs," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 5, pp. 742–753, Aug. 2017, doi: 10.1109/JSTSP.2017.2692560.

[11] S. R. Sirin, "Socioeconomic status and academic achievement: A meta-analytic review of research," *Review of Educational Research*, vol. 75, no. 3, pp. 417–453, 2005, doi: 10.3102/00346543075003417.

[12] T. R. Kumar, T. Vamsidhar, B. Harika, T. M. Kumar, and R. Nissy, "Students performance prediction using data mining techniques," in *Proceedings of the 2019 International Conference on Intelligent Sustainable Systems (ICISS)*, Palladam, India, 2019, vol. 2019, no. November, pp. 407–411, doi: 10.1109/ISS1.2019.8907945.

[13] A. Mueen, B. Zafar, and U. Manzoor, "Modeling and predicting students' academic performance using data mining techniques," *International Journal of Modern Education and Computer Science*, vol. 8, no. 11, pp. 36–42, 2016, doi: 10.5815/ijmecs.2016.11.05.

[14] S. Pati, K. Hashim, B. Brown, A. G. Fiks, and C. B. Forrest, "Early identification of young children at risk for poor academic achievement: Preliminary development of a parent-report prediction tool," *BMC Health Services Research*, vol. 11, no.

August, Art. no. 197, Aug. 2011, doi: 10.1186/1472-6963-11-197.

[15] G. W. Dekker, M. Pechenizkiy, and J. M. Vleeshouwers, “Predicting students drop out: A case study,” in *Proceedings of the 2nd International Conference on Educational Data Mining (EDM 2009)*, Cordoba, Spain, Jul. 1–3, 2009, pp. 41–50. [Online]. Available: https://educationaldatamining.org/EDM2009/uploads/proceedings/dekker.pdf

[16] O. W. Adejo and T. Connolly, “Predicting student academic performance using multi-model heterogeneous ensemble approach,” *Journal of Applied Research in Higher Education*, vol. 10, no. 1, pp. 61–75, 2018, doi: 10.1108/JARHE-09-2017-0113.

[17] M. Hussain, W. Zhu, W. Zhang, S. M. R. Abidi, and S. Ali, “Using machine learning to predict student difficulties from learning session data,” *Artificial Intelligence Review*, vol. 52, no. 1, pp. 381–407, 2019, doi: 10.1007/s10462-018-9620-8.

[18] S. K. Mohamad and Z. Tasir, “Educational data mining: A review,” *Procedia - Social and Behavioral Sciences*, vol. 97, no. November, pp. 320–324, 2013, doi: 10.1016/j.sbspro.2013.10.240.

[19] H. Agrawal and H. Mavani, “Student performance prediction using machine learning,” *International Journal of Engineering Research & Technology*, vol. 4, no. 3, vol. 111-113, Mar. 2015, doi: 10.17577/IJERTV4IS030127.

[20] A. Dutt, M. A. Ismail, and T. Herawan, “A systematic review on educational data mining,” *IEEE Access*, vol. 5, no. January, pp. 15991–16005, 2017, doi: 10.1109/ACCESS.2017.2654247.

[21] L. G. R. Putra, D. D. Prasetya, and Mayadi, “Student dropout prediction using Random Forest and XGBoost method,” *INTENSIF: Jurnal Ilmiah Penelitian dan Penerapan Teknologi Sistem Informasi*, vol. 9, no. 1, pp. 147–157, Feb. 2025, doi: 10.29407/intensif.v9i1.21191.

[22] U. P. Inyang and E. A. Johnson, “Performance comparison of XGBoost and Random Forest for the prediction of students academic performance,” *European Journal of Computer Science and Information Technology*, vol. 13, no. 2, pp. 1–21, 2025, doi: 10.37745/ejcsit.2013/vol13n2121.

[23] Y. R. Marín, L. Q. Huatangari, J. N. A. Tuesta, O. C. Caro, J. L. M. Guevara, E. S. Bardales, and R. C. Santos, “Analysis of factors affecting the academic performance of university students using machine learning,” *Scientific Reports*, vol. 15, no. November, Art. no. 44027, 2025, doi: 10.1038/s41598-025-28870-1.

[24] T. T. Tin, S. H. Lee, and O. M. Ikumapayi, “Educational big data mining: Comparison of multiple machine learning algorithms in predictive modelling of student academic performance,” *International Journal of Advanced Computer Science and Applications*, vol. 15, no. 6, pp. 1-13, 2024, doi: 10.14569/IJACSA.2024.0150664.

[25] S. Almutairi, H. Shaiba, and M. Bezbradica, “Predicting students’ academic performance and main behavioral features using data mining techniques,” in *Advances in Data Science, Cyber Security and IT Applications*, ser. *Communications in Computer and Information Science*, vol. 1097. Cham, Switzerland: Springer, 2019, vol. 2019, no. December, pp. 245–259, doi: 10.1007/978-3-030-36365-9_21.