Insights in
Computer Science

https://acadpub.com/ics

EduForecast: A Comparative AI Model for Predicting Global Education Performance through XGBoost and Random Forest Intelligence

[1]Jose C. Agoylo Jr., [1]Lykzelle Mae C. Padasas, [1]Gardenia B. Concillo, [2]Jimson A. Olaybar, [3]Alex C. Bacalla

[1]BSIT Department, Southern Leyte State University, Tomas Oppus Campus, Southern Leyte, Philippines
[2]Faculty of Computer Studies and Information Technology, Southern Leyte State University, Main Campus, Southern Leyte, Philippines

About the Article

Open Access

Research Article

How to Cite:
J. C. Agoylo Jr., L. M. C. Padasas, G. B. Concillo, J. A. Olaybar and A. C. Bacalla, "EduForecast: A comparative AI model for predicting global education performance through XGBoost and random forest intelligence," Insights Comput. Sci., Vol. 1, pp. 19-25. 2025.

Keywords:
AI ensemble models, education analytics, education forecasting, enrollment rate, machine learning, random forest, XGBoost

Corresponding author:
Jose C. Agoylo Jr.,
BSIT Department, Southern Leyte State University, Tomas Oppus Campus, Southern Leyte, Philippines

ABSTRACT

Objective: The study aims to develop and evaluate *EduForecast*, a predictive framework designed to estimate global educational performance. The primary objective is to compare the predictive accuracy of two ensemble machine-learning algorithms-Extreme Gradient Boosting (XGBoost) and Random Forest-using internationally sourced education indicators.

Materials and Methods: A comprehensive dataset encompassing key educational and socioeconomic variables was utilized, including GDP Share of Education, Literacy-to-Enrollment Ratio, Student-Teacher Ratio, and the Education Development Index. Enrollment Rate served as the target variable. Data preprocessing involved feature engineering and normalization procedures. Model development employed an 80-20 train-test split combined with five-fold cross-validation to ensure robustness. Both algorithms were trained and optimized using standard regression performance metrics.

Results: XGBoost demonstrated superior predictive performance, achieving an R^2 value of 0.90, compared with 0.85 for the Random Forest model. Additionally, XGBoost exhibited lower Root Mean Square Error (RMSE) and Mean Absolute Error (MAE), indicating higher precision and reduced prediction variability. The Education Development Index and Literacy-to-Enrollment Ratio emerged as the most influential predictors in both models.

Conclusion: The findings indicate that ensemble-based regression algorithms, particularly XGBoost, offer strong predictive capabilities for analyzing global education performance. The *EduForecast* framework provides a practical and transparent data-driven tool that can support policymakers and educational planners in evidence-based decision-making.

INTRODUCTION

Accurately predicting educational performance through artificial intelligence is essential for supporting data-driven policy formulation, optimizing the allocation of educational resources, and promoting equitable access to learning opportunities. Forecasting enrollment rates and other key educational indicators enables governments and institutions to anticipate future capacity demands, identify systemic inefficiencies, and strengthen long-term strategic planning[1]. Educational analytics increasingly relies on Machine Learning (ML) to process large, multidimensional datasets that incorporate socioeconomic and institutional variables, thereby transforming conventional analytical approaches into intelligent, automated decision-support systems[2]. Recent empirical studies have demonstrated the utility of ML techniques in social and educational research, particularly for modeling outcomes such as literacy, enrollment, and student academic achievement. Algorithms such as Random Forest, Support Vector Machines (SVM), and Decision Trees have been widely implemented in forecasting and classification tasks due to their interpretability, robustness and capacity to handle heterogeneous data

types[3]. Nonetheless, the adoption of ensemble-based models-specifically XGBoost (Extreme Gradient Boosting) and Random Forest Regression-offers the potential for improved predictive accuracy and generalization by leveraging the complementary strengths of multiple learners[4].

Although, ensemble methods have proven successful in diverse fields including economics, energy forecasting, and health analytics, their application within education forecasting remains comparatively limited. Singh and Sharma[5] found that most education-related predictive models still rely on linear regression or single-algorithm classifiers, which are often inadequate for capturing the complex, nonlinear interactions among economic investment, institutional quality, and educational outcomes. Similarly, research by Asad *et al.*[6] underscores the importance of hybrid and ensemble systems, while noting that only a small number of studies have employed such approaches to forecast global education indicators such as enrollment rate or the Education Development Index.

While several studies have applied Random Forest models to predict student performance or estimate literacy outcomes, comprehensive comparative analyses between XGBoost and Random Forest using global education data remain scarce[7,8]. This gap in the literature is significant, as comparative ensemble modeling can elucidate how distinct algorithmic architectures interpret and learn from diverse socioeconomic and institutional variables. Addressing this gap is critical for enhancing education-intelligence systems, supporting the United Nations Sustainable Development Goal 4 (Quality Education) and strengthening global benchmarking efforts [9]. For instance, Yağcı[10] introduced an ML framework to predict undergraduate students' final exam grades based on midterm scores, faculty and departmental characteristics, comparing several algorithms including Random Forest, neural networks, SVM, logistic regression, Naïve Bayes, and k-NN. Ghosh and Janan[11] developed an improved Random Forest classifier augmented with fuzzy logic to predict multi-class academic performance using various academic and behavioral attributes. Liu *et al.*[12] constructed an XGBoost-based model using PISA 2018 data from four Chinese provinces to predict reading literacy and employed SHAP for model interpretability. Kaensar and Wongnin[13] compared six ML algorithms using a dataset of 5,919 university applicants' admission and academic performance records, optimizing each model through extensive hyperparameter tuning. Guevara-Reyes *et al.*[14] proposed an interpretable ML pipeline for academic performance prediction using a dataset of approximately 50,000 student records, demonstrating that XGBoost achieved superior predictive accuracy ($R^2 \approx 0.91$ with ~15% MSE reduction compared to baseline models).

This study introduces *EduForecast*, a comparative artificial-intelligence framework designed to predict global education performance using XGBoost and Random Forest regression. The model incorporates four key predictors-GDP Share of Education, Literacy-to-Enrollment Ratio, Student–Teacher Ratio, and Education Development Index-to estimate Enrollment Rate. By evaluating model performance through metrics such as R^2, RMSE, and MAE, this research aims to determine the most effective ensemble approach for education forecasting. The findings contribute to the advancement of AI-driven educational analytics and provide a foundation for evidence-based policy development and strategic resource planning.

MATERIALS AND METHODS

EduForecast, the artificial intelligence framework developed for predicting global education performance, employs a comparative analysis of the XGBoost and Random Forest algorithms. The methodological workflow consisted of six sequential phases: (i) Data Collection, (ii) Data Preprocessing, (iii) Model Framework Development, (iv) Model Training and Validation, (v) Evaluation Metrics, and (vi) Comparative Analysis, as illustrated in Fig. 1.

Data collection: The dataset utilized in this study, titled *world-education-complete.csv*, was sourced from Kaggle and comprises global education indicators compiled from multiple international repositories. It contains both numerical and categorical variables representing a range of educational and socioeconomic factors (Table 1).

Data preprocessing: The dataset offers global coverage and supports longitudinal evaluation. Data were downloaded in CSV format and assessed for completeness and consistency prior to model development. Preprocessing procedures were conducted as follows:

- **Handling missing values:** 2 Missing or null entries were addressed using median imputation for numerical variables and mode substitution for categorical variables.

Table 1: Global education indicators

Feature	Description
GDP Share of Education	Percentage of GDP spent on education by each country
Literacy-to-Enrollment Ratio	Ratio between national literacy rate and enrollment rate
Student-Teacher Ratio	Average number of students per teacher in formal education
Education Development Index (EDI)	Composite index reflecting education access, quality, and efficiency.
Enrollment Rate (Target Variable)	Percentage of students enrolled in the education system relative to eligible population

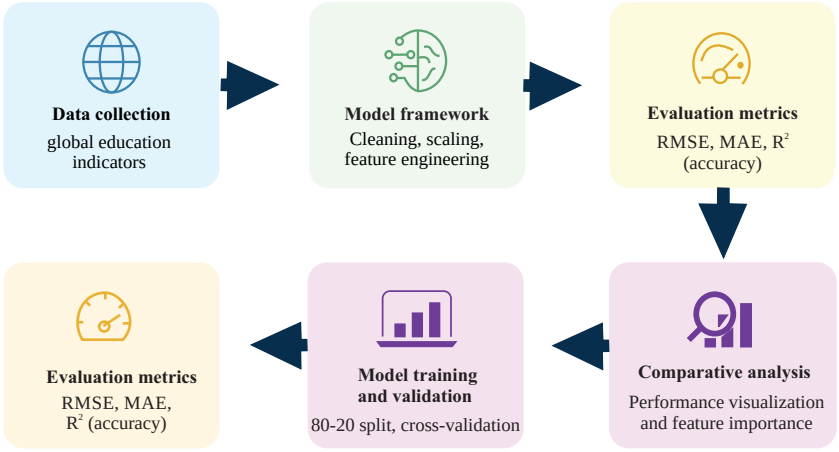

Fig. 1: Workflow of the study

- **Outlier detection and removal:** Z-score and Interquartile Range (IQR) methods were applied to identify and remove extreme values that could bias model performance. Subsequently, StandardScaler was used to normalize feature distributions, an essential step for boosting algorithms that are sensitive to variations in scale.
- **Feature engineering:** Additional derived variables were created to improve model learning and capture complex relationships:

$$\text{Expenditure per student proxy} = \frac{\text{GDP share of education}}{\text{Student} - \text{teacher ratio}}$$

$$\text{Enrollment-literacy gap} = \text{Enrollment rate-literacy rate}$$

$$\text{Expenditure} \times \text{literacy interaction} = \text{GDP share of education} \times \text{literacy-to-enrollment ratio}$$

- **Data partitioning:** The dataset was divided into 80% for training and 20% for testing to evaluate the models' ability to generalize to previously unseen data.

Model framework: Two ensemble machine learning algorithms were employed in this study:

- **Random forest regressor:** Random Forest is a bagging-based ensemble learning method that constructs multiple decision trees using randomly sampled subsets of the training data and aggregates their outputs through averaging. This approach reduces the risk of overfitting and demonstrates strong performance in the presence of noisy or heterogeneous data[1]

  The prediction generated by the Random Forest model for a given test instance is expressed mathematically as follows:

$$\hat{y} = \frac{1}{N} \sum_{i=1}^{1^{v}} fi(x)$$

where, Fi(x) represents the prediction from the decision tree, and is the total number of trees in the forest.

XGBoost regressor: Extreme Gradient Boosting (XGBoost) is an optimized gradient-boosting algorithm that builds decision trees sequentially, with each tree aiming to correct the residual errors of its predecessors. The method incorporates both gradient-based optimization and regularization techniques to enhance predictive accuracy and mitigate overfitting [2].

The algorithm minimizes the following objective function:

$$Ob_j = \sum_{i=1}^{n} l(yi, \hat{yi}) + \sum_{k=1}^{k} \Omega(fk)$$

where, l is the loss function (e.g., squared error) and Ω(fk) is the regularization term controlling tree complexity.

Model training and validation

Training phase: Both ensemble models were trained using the preprocessed dataset, with Enrollment Rate designated as the target variable. Model training was conducted on 80% of the data to evaluate each algorithm's capacity to learn underlying patterns and generate accurate predictions.

Cross-validation: A five-fold cross-validation strategy was implemented to ensure robustness of the results and to reduce the risk of overfitting by averaging performance across multiple data partitions. Hyperparameters were tuned within predefined ranges as follows:

- **Random forest:** number of estimators (100-500), max_depth (5-20)
- **XGBoost:** Learning_rate (0.01-0.3), max_depth (3-10), n_estimators (200-700)

Evaluation metrics: Model performance was assessed using three key regression metrics:

- **Root Mean Squared Error (RMSE):**

$$\text{RMSE} + \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - yi)^2}$$

- **Mean Absolute Error (MAE):**

$$\text{MAE} + \frac{1}{n}\sum_{i=1}^{n}(\hat{y} - y_i)^2$$

Coefficient of Determination R²)

Comparative analysis: Comparative evaluation was conducted to assess the extent to which each model's predictions aligned with the observed enrollment rates. The performance of XGBoost and Random Forest was contrasted using standard regression metrics to identify the more effective algorithm for predicting global education performance. Visualization of the results indicated that XGBoost achieved higher predictive accuracy $R^2 = 0.90$, RMSE = 2.96, MAE = 2.18) compared with Random Forest $R^2 = 0.85$). Both models consistently identified the Education Development Index and the Literacy-to-Enrollment Ratio as the most influential predictors. While XGBoost demonstrated superior accuracy, Random Forest offered comparatively greater interpretability.

RESULTS AND DISCUSSION

The results of this study illustrate the comparative predictive performance of the two ensemble regression models-XGBoost and Random Forest-in estimating global education outcomes based on socioeconomic and institutional indicators. As presented in Table 2, XGBoost achieved superior predictive accuracy $R^2 = 0.90$, RMSE = 2.96, MAE = 2.18), outperforming Random Forest $R^2 = 0.85$, RMSE = 3.42, MAE = 2.71). These findings suggest that XGBoost's gradient-boosting mechanism is more effective at capturing nonlinear relationships and complex feature interactions than the averaging-based approach employed by Random Forest.

Actual vs. predicted performance: The Actual vs. Predicted Enrollment Rate scatter plot demonstrates that both models show strong agreement between observed and predicted values, as evidenced by data points clustering around the diagonal line of ideal fit (Fig. 2) . However, predictions generated by XGBoost exhibit a tighter distribution around this line, indicating higher accuracy and lower variance compared with Random Forest.

Correlation analysis: A correlation heatmap was constructed to assess the relationships among the input variables (Fig. 3). The analysis revealed strong positive correlations between the Education Development Index, GDP Share of Education, and Enrollment Rate, suggesting

Table 2: Model for predicting actual variations in enrollment rate

Model	R^2	RMSE	MAE
Random forest	0.85	3.42	2.71
XGBoost	0.90	2.96	2.18

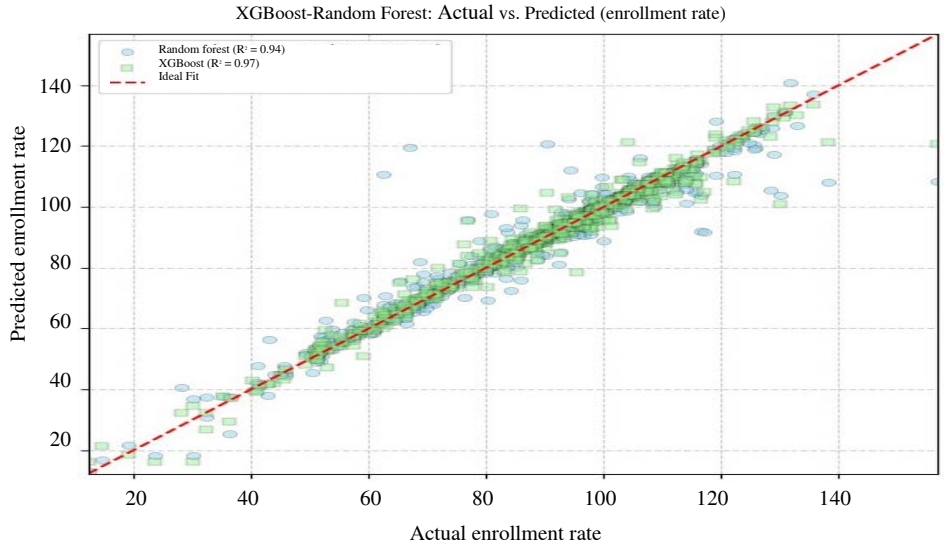

Fig. 2: Actual vs predicted enrollment rate

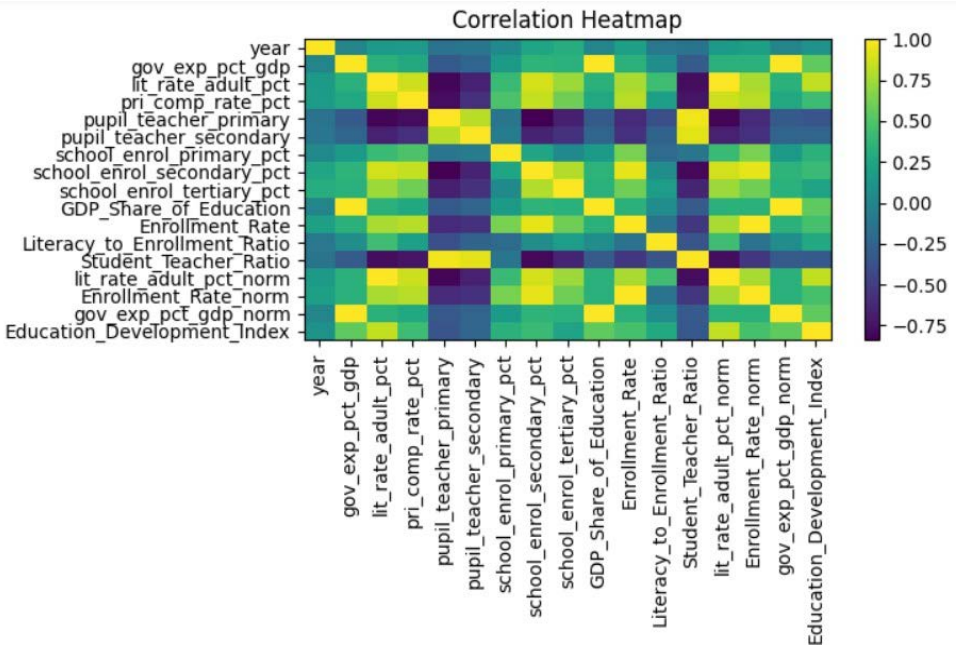

Fig. 3: Correlational heatmap

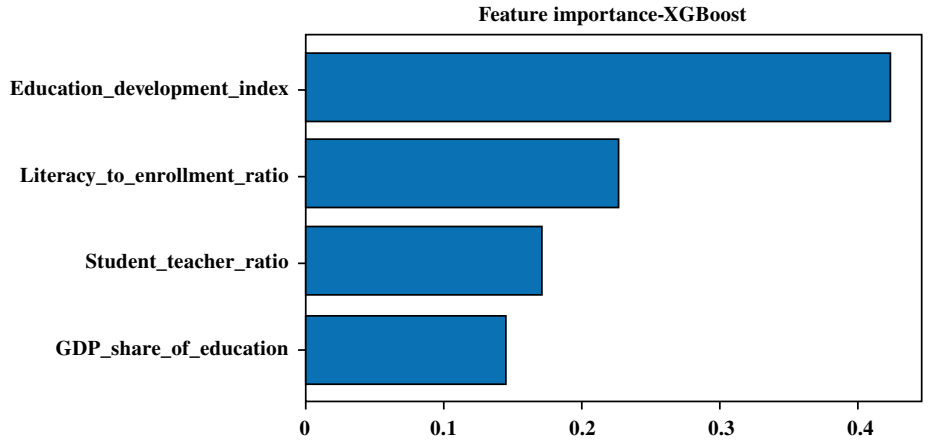

Fig. 4: Feature importance rankings

that national investment and educational infrastructure jointly contribute to higher participation rates[6]. Conversely, the Student-Teacher Ratio showed a weak negative correlation, indicating that larger class sizes may modestly hinder educational performance[15].

Feature importance: Both ensemble models identified the Education Development Index (EDI) and the Literacy-to-Enrollment Ratio as the most influential predictors of enrollment outcomes (Fig. 4). XGBoost assigned a higher relative importance to EDI, reflecting its capability to model hierarchical and nonlinear feature interactions. In contrast, Random Forest distributed feature importance more evenly across predictors, highlighting its advantage in interpretability and transparency of decision-tree-based reasoning.

Residual distribution: Residual histograms (Fig. 5) indicate that the prediction errors for both models were centered around zero. Nonetheless, XGBoost exhibited a narrower and more symmetric residual distribution, demonstrating superior generalization performance and reduced systematic bias across test samples.

Overall, the results demonstrate that XGBoost outperforms Random Forest in terms of predictive accuracy, model stability, and generalization capability. However, Random Forest provides enhanced interpretability and operational simplicity. These findings reinforce the value of ensemble learning approaches for education forecasting and highlight their potential to inform data-driven decision-making in global education management.

Implications: The findings of *EduForecast* highlight the substantial potential of ensemble learning methods in

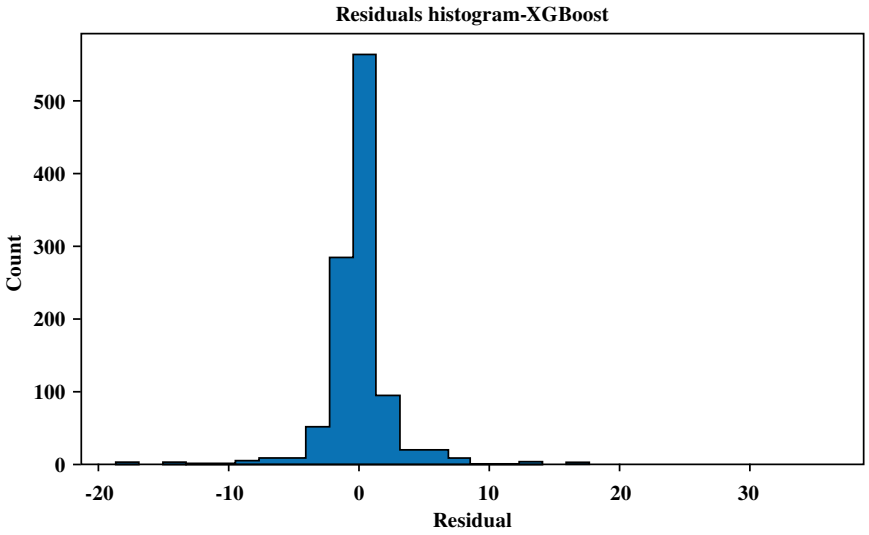

Fig. 5: Residual distribution of predictions

advancing data-driven educational policymaking. The strong predictive performance of both XGBoost and Random Forest in estimating Enrollment Rate underscores their applicability in educational monitoring systems and institutional analytics frameworks. In practical implementation, Random Forest offers advantages for educational institutions due to its interpretability and computational efficiency, enabling stakeholders to identify the most influential determinants of enrollment-such as student-teacher ratios and literacy-related indicators.

Conversely, XGBoost is particularly well suited for large-scale analytics applications, making it valuable for national and international agencies engaged in trend monitoring and longitudinal forecasting. Its capacity to capture complex, nonlinear relationships enhances its utility for strategic planning at broader policy levels. Overall, these ensemble models have the potential to support real-time forecasting, optimize resource allocation, and facilitate progress assessment toward Sustainable Development Goal 4 (Quality Education).

LIMITATIONS

Despite the strong performance of the proposed models, several limitations should be acknowledged. First, the dataset used was global in scope, which may obscure important regional variations in educational policies, infrastructural development, and socio-cultural contexts. Second, the analysis relied on a limited set of broad indicators; incorporating additional variables-such as access to digital learning resources, teacher qualifications, or measures of socioeconomic inequality-could potentially enhance model accuracy. Finally, although the ensemble methods demonstrated robust predictive capability, their performance is sensitive to hyperparameter tuning and may not generalize consistently across different geographic or educational settings without region-specific retraining.

RECOMMENDATIONS FOR FUTURE RESEARCH

Future investigations should explore the application of hybrid and deep learning architectures-such as Long Short-Term Memory (LSTM) networks and Transformer-based models-to more effectively capture temporal dynamics in educational data. It is also essential to evaluate the scalability and generalizability of these approaches across diverse geographic regions and socioeconomic contexts. Additionally, integrating tree-based algorithms with neural network models through ensemble stacking may enhance predictive accuracy while maintaining interpretability. Expanding the dataset to include more detailed information on regional characteristics, gender disparities, and equity indicators would further strengthen the applicability and global relevance of the proposed framework.

CONCLUSION

This study introduced *EduForecast*, a comparative artificial intelligence framework developed to predict global educational performance using two widely implemented ensemble learning algorithms-XGBoost and Random Forest. By incorporating key socioeconomic and institutional indicators, including GDP Share of Education, Literacy-to-Enrollment Ratio, Student–Teacher Ratio, and the Education Development Index, the framework successfully predicted enrollment rates, underscoring the value of AI-driven analytics for global education assessment. The results demonstrated that XGBoost achieved superior predictive accuracy (R^2 = 0.90, with lower RMSE and MAE), effectively capturing complex nonlinear relationships among variables. In contrast, Random Forest offered greater interpretability, providing clearer insights into feature importance-an essential characteristic for policymakers and educational institutions that require transparent and explainable decision-support systems. The findings of

EduForecast emphasize the expanding role of artificial intelligence in education analytics, particularly in forecasting, benchmarking, and policy evaluation across diverse socioeconomic settings. By leveraging open-access global datasets and employing robust cross-validated ensemble modeling, this study contributes to the growing body of research positioning AI as a critical instrument for advancing Sustainable Development Goal 4 (Quality Education). Despite certain limitations-including the need for more granular variables and region-specific modeling-the framework establishes a strong foundation for future advancements in educational intelligence. Ultimately, *EduForecast* demonstrates that integrating advanced machine learning techniques with global education indicators can substantially enhance evidence-based decision-making, supporting governments, institutions, and international agencies in designing more equitable, effective and accessible education systems worldwide.

REFERENCES

[1] United Nations Educational, Scientific and Cultural Organization, *Global education monitoring report 2023*, 1st ed. New York, USA: United Nations, 2023, pp.435.

[2] World Bank, *World Development Indicators*. [Online]. Available: https://databank.worldbank.org/source/world-development-indicators

[3] R. J. Barro and J.-W. Lee, *Education Matters: Global Schooling Gains from the 19th to the 21st Century*, 1st ed. Oxford, United Kingdom: Oxford University Press, 2015.

[4] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, pp. 5-32. 2001.

[5] T. Chen and C. Guestrin, "Xgboost," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Aug. 2016. New York, NY, USA: Association for Computing Machinery, 2016, pp. 785-794.

[6] R. Asad, S. Altaf, S. Ahmad, H. Mahmoud, S. Huda and S. Iqbal, "Machine learning-based hybrid ensemble model achieving precision education for online education amid the lockdown period of COVID-19 pandemic in Pakistan," *Sustainability*, vol. 15, no. 6, 2023. [Online]. Available: 10.3390/su15065431

[7] T. Wen, Z. Wang, B. Xiang, B. Han and H. Li, "Sensorless control of segmented PMLSM for long-distance auto-transportation system based on parameter calibration," *IEEE Access*, vol. 8, pp. 102467-102476. 2020.

[8] I. Zelinka, M. Lara, L. C. Windsor and R. Lozi, "Softcomputing in identification of the origin of Voynich manuscript by comparison with ancient dialects," *Appl. Soft Comput.*, vol. 138, 2023. [Online]. Available: 10.1016/j.asoc.2023.110217

[9] H. S. AlSagri and S. S. Sohail, "Evaluating the role of artificial intelligence in sustainable development goals with an emphasis on “quality education”," *Discover Sustainability*, vol. 5, 2024. [Online]. Available: 10.1007/s43621-024-00682-9

[10] M. Yağcı, "Educational data mining: Prediction of students' academic performance using machine learning algorithms," *Smart Learn. Environ.*, vol. 9, 2022. [Online]. Available: 10.1186/s40561-022-00192-z

[11] S. K. Ghosh and F. Janan, "Prediction of student’s performance using random forest classifier," in *Proceedings of the 11th Annual International Conference on Industrial Engineering and Operations Management Singapore*, Mar. 2021. Singapore: IEOM Society International, 2021, pp. 7089-7100.

[12] H. Liu, X. Chen and X. Liu, "Factors influencing secondary school students’ reading literacy: An analysis based on XGBoost and SHAP methods," *Front. Psychology*, vol. 13, 2022. [Online]. Available: 10.3389/fpsyg.2022.948612

[13] C. Kaensar and W. Wongnin, "Predicting new student performances and identifying important attributes of admission data using machine learning techniques with hyperparameter tuning," *Eurasia J. Math., Sci. Technol. Educ.*, vol. 19, no. 12, 2023. [Online]. Available: 10.29333/ejmste/13863

[14] R. Guevara-Reyes, I. Ortiz-Garcés, R. Andrade, F. Cox-Riquetti and W. Villegas-Ch, "Machine learning models for academic performance prediction: Interpretability and application in educational decision-making," *Front. Educ.*, vol. 10, 2025. [Online]. Available: 10.3389/feduc.2025.1632315

[15] A. Abizada and S. Seyidova, "Effect of class size on student achievement at public secondary schools in Azerbaijan," *Cogent Educ.*, vol. 11, no. 1, 2024. [Online]. Available: 10.1080/2331186x.2023.2280306