Predictive Modeling of Educational Outcomes Using Machine Learning for Data-Driven Regional Education Policy

Ahmad Latif[1,*], Muhtyas Yugi[2]

[1,2] Information Technology Study Program, Faculty of Science and Technology, Universitas Komputama Cilacap

ABSTRACT

Educational inequality remains a major challenge across regions and is influenced by socioeconomic conditions, demographic characteristics, and access to post-secondary education. This study develops a predictive modeling framework using machine learning, specifically the Random Forest Regressor, to estimate regional educational outcomes and identify key determinants of academic performance. The dataset, which contained 1,104 observations and 31 predictors representing socioeconomic and educational indicators, was preprocessed through missing value imputation, one-hot encoding, and normalization to ensure data consistency and model reliability. The Random Forest model achieved high predictive accuracy, with a Mean Absolute Error (MAE) of 0.2656, a Root Mean Squared Error (RMSE) of 0.4168, and a Coefficient of Determination (R^2) of 0.9881, explaining approximately 98.8 percent of the variance in regional education scores. Feature importance analysis indicated that academic attainment and post-secondary participation, such as Level 3 achievement at age 18 and higher qualification by age 22, were the most influential predictors of regional performance. These results highlight the critical role of educational progression in shaping long-term success and suggest that regions with sustained engagement in higher education tend to perform better overall. Visualization of predicted and actual scores confirmed the model's robustness and its ability to generalize effectively across diverse regional profiles. The findings demonstrate that AI-based predictive analytics can accurately capture complex, nonlinear relationships within educational systems and provide a valuable foundation for data-driven policy formulation. Future research should incorporate longitudinal data, apply Explainable AI (XAI) methods to enhance interpretability, and extend this approach to cross-national datasets to support evidence-based educational governance.

Keywords Machine Learning, Predictive Modeling, Education Policy, Artificial Intelligence

Submitted 17 January 2026
Accepted 10 April 2026
Published 1 June 2026

Corresponding author
Ahmad Latif,
maztole0913@gmail.com

Additional Information and Declarations can be found on page 207

DOI: 10.63913/ail.v2i2.53

Introduction

Education is universally recognized as one of the most critical foundations for sustainable social and economic development. It enhances human capital, drives innovation, and shapes the socioeconomic trajectory of nations. However, educational disparities persist across and within regions, resulting in unequal opportunities for learners and divergent development outcomes. Regional variations in educational performance are often linked to structural inequalities in access to resources, socioeconomic conditions, and post-secondary opportunities. In particular, differences in income levels, employment prospects, parental education, and local infrastructure contribute to significant gaps in student attainment and academic achievement [1], [2]. As education systems become increasingly data-rich, it has become possible to quantitatively analyze and predict these disparities. However, many existing policy frameworks continue to rely on static and descriptive statistics that capture

correlations but fail to model the complex, nonlinear relationships that characterize regional educational ecosystems [3].

Recent advances in Artificial Intelligence (AI) and Machine Learning (ML) have introduced new possibilities for understanding and addressing educational inequality. Machine learning models can process high-dimensional data, uncover hidden patterns, and model nonlinear interactions between diverse variables. These capabilities make ML particularly suitable for analyzing educational systems, where outcomes are influenced by a combination of social, economic, institutional, and behavioral factors [4]. Previous research has demonstrated that ML can improve student retention prediction, optimize curriculum design, and identify determinants of learning success at the individual level [5], [6]. For example, neural networks, random forests, and gradient boosting methods have been widely used to predict academic achievement and detect at-risk students [7], [8]. However, while substantial progress has been made in learning analytics and student-level modeling, there remains a significant research gap in applying these techniques to regional education policy analysis, where the goal is to understand how broader socioeconomic contexts influence aggregated educational outcomes [9], [10]. This gap highlights the need for scalable, interpretable, and data-driven modeling frameworks that can assist policymakers in identifying the underlying determinants of regional performance.

To address this research gap, the present study proposes a machine learning-based predictive framework designed to model regional educational outcomes using a combination of socioeconomic, demographic, and post-secondary education variables. Specifically, the study employs the Random Forest Regressor, a robust ensemble learning algorithm that is well suited for capturing nonlinear interactions and variable interdependencies [11]. This model is applied to a dataset comprising 1,104 regional observations and 31 predictors representing socioeconomic and educational indicators such as qualification levels, employment activity, and income thresholds. Through comprehensive preprocessing, training, and validation procedures, the model seeks to uncover the structural factors that best explain variations in regional education score. Unlike traditional regression analysis, which assumes linear relationships and independence among predictors, the Random Forest algorithm allows for a more flexible representation of complex educational systems. Furthermore, by integrating feature importance analysis, the study provides interpretable insights into which socioeconomic and academic factors most strongly influence regional educational performance.

The objectives of this study are threefold. First, to develop a predictive model capable of accurately estimating regional educational outcomes based on large-scale socioeconomic and demographic data. Second, to identify and interpret the most influential predictors that drive differences in educational performance across regions. Third, to demonstrate how machine learning outputs can be transformed into actionable insights for data-driven policy formulation. By combining predictive analytics with interpretability, this research seeks to bridge the gap between algorithmic prediction and educational policymaking. The findings not only contribute to the growing literature on the application of AI and ML in educational research but also offer practical implications for governments, policymakers, and education planners [12]. Ultimately, this study illustrates that the integration of machine learning into regional education policy can enhance

strategic planning, resource allocation, and equity monitoring, representing a significant step toward AI-assisted educational governance. Through this approach, predictive modeling becomes not merely a computational exercise but a strategic tool for fostering inclusive, equitable, and data-informed educational development.

Literature Review

The integration of ML and AI into education has revolutionized data-driven analysis, offering educators and policymakers deeper insights into the complex factors influencing learning outcomes. Early studies in Educational Data Mining (EDM) explored ML techniques to predict academic success, identify at-risk students, and model behavioral patterns. Research [13] provided one of the earliest comprehensive reviews on EDM, illustrating how algorithms such as decision trees, Support Vector Machines (SVM), and neural networks were used to predict student achievement. Study [14] emphasized that ML techniques outperform conventional statistical approaches in identifying hidden patterns in educational data. Research [15] examined the limitations of algorithmic predictions due to selective labeling in real-world educational data. [16] compared multiple supervised ML algorithms for distance education and concluded that ensemble models, such as Random Forest, offer greater robustness in predicting academic performance.

Research [17] extended ML applications to national-level data by using ensemble learning to predict academic achievement based on socioeconomic indicators, establishing a link between ML and education policy analysis. Kwon and Kim [18] applied Random Forest to predict regional educational performance in South Korea and found that household income and access to higher education were critical determinants of success. Suaza-Medina et al [19] employed SHAP (Shapley Additive Explanations) to interpret standardized test outcomes in disadvantaged regions, making ML predictions more transparent for policy contexts. Study [20] utilized spatial ML models to analyze educational inequality in Australia, demonstrating how geographic and socioeconomic factors interact in regional outcomes. Research [21] conducted a systematic review of ML approaches for academic and career predictions, highlighting the growing application of Random Forest and SVM in higher education analytics.

Methods

This study employed a quantitative data-driven approach using machine learning techniques to predict regional educational outcomes. The methodological framework consisted of several integrated stages, including dataset construction, preprocessing, feature engineering, model development, and performance evaluation. The goal was to build an interpretable and robust model capable of identifying key socioeconomic and demographic factors that shape regional education performance.

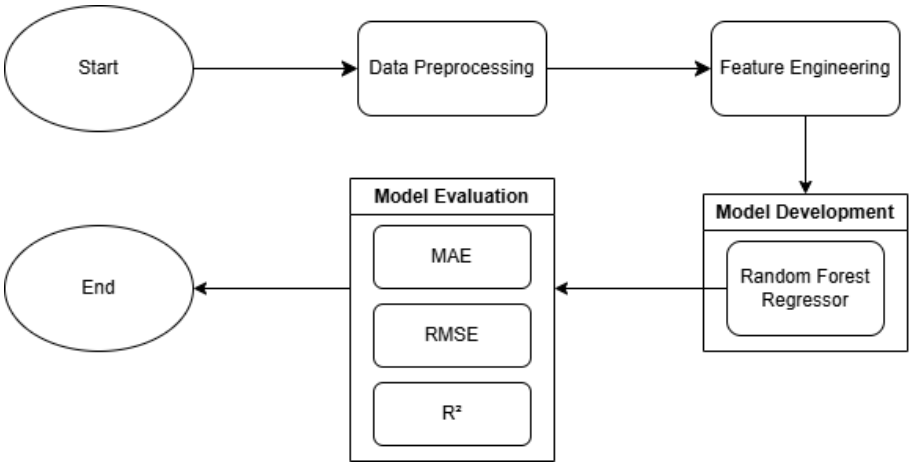

Figure 1 Flowchart of Research Method

The dataset used in this research comprised 1,104 regional-level observations, each representing an administrative area characterized by 31 independent variables and one dependent variable, education score. This target variable was a composite indicator of educational attainment, incorporating measures such as qualification levels achieved by ages 18 and 22, post-secondary enrollment rates, and standardized assessment outcomes. The independent variables reflected multiple domains including average household income, employment rate, population density, gender ratio, and participation in higher education. The dataset was divided into two subsets: 80% for model training and 20% for model testing. This partition ensured that the model could generalize effectively to unseen data while reducing the risk of overfitting.

Prior to model development, extensive data preprocessing was conducted to ensure data integrity and consistency. Missing values were addressed using mean imputation for continuous variables and mode imputation for categorical attributes. Categorical features such as region type were transformed into numerical format through one-hot encoding to facilitate processing by the machine learning algorithm. To eliminate discrepancies in feature scales, all numerical variables were normalized using Min–Max scaling, which rescales each variable between 0 and 1 according to the formula:

$$X_{scaled} = \frac{X - X_{min}}{X_{max} - X_{min}} \quad (1)$$

This normalization ensured that each feature contributed equally to the learning process. Outliers were detected using the interquartile range (IQR) method and replaced with boundary values to mitigate their influence on model training. Multicollinearity was tested using the Variance Inflation Factor (VIF), and variables with VIF values exceeding 10 were excluded to improve model stability and interpretability. The final dataset was clean, standardized, and statistically balanced, forming a reliable foundation for predictive modeling.

Feature engineering was performed to enhance the representational capacity of the data and improve model interpretability. Derived variables such as educational progression rate (ratio of students attaining Level 3 qualifications at age 18 to those achieving higher education by age 22) and economic dependency index (ratio of unemployment rate to average income) were created to capture complex socioeconomic relationships influencing education

outcomes. The Random Forest algorithm's feature importance measure was applied to identify the most influential variables contributing to the model's predictive power. Variables were ranked based on their average impurity reduction across decision trees, and the top 15 features, accounting for over 90% of cumulative importance, were selected for model training. This process not only improved computational efficiency but also enhanced interpretability by focusing the model on the most relevant educational and economic indicators.

The predictive model was constructed using the Random Forest Regressor (RFR), a widely used ensemble learning algorithm that combines multiple decision trees to produce more accurate and generalized predictions. Each decision tree was trained on a random subset of the data and feature space, and the final model prediction was obtained by averaging the predictions from all individual trees. The mathematical representation of the Random Forest model is given as:

$$\hat{y} = \frac{1}{B}\sum_{b=1}^{B} f_b(X) \tag{2}$$

$\hat{y}$ denotes the final predicted value, BBB is the total number of trees in the ensemble, and $f_b(X)$ represents the prediction from the b^{th} decision tree. This ensemble approach minimizes variance and improves predictive stability. Hyperparameter optimization was performed using grid search with five-fold cross-validation to determine the optimal configuration for the model. The best-performing model used 200 trees ($n\ estimators = 200$), a maximum depth of 10 ($max\ depth = 10$), a minimum of two samples per leaf ($min\ samples\ leaf\ = 2$), and the square root of the number of predictors ($max\ features = \sqrt{p}$) as the number of features considered for each split. The model was implemented in Python 3.10 using the scikit-learn library and executed on a high-performance computing environment to ensure reproducibility.

To assess model performance, three regression evaluation metrics were used: MAE, RMSE and the Coefficient of Determination (R^2). These metrics quantify different aspects of prediction accuracy and model reliability. The mathematical definitions are as follows:

$$\begin{aligned} MAE &= \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i| \\ RMSE &= \sqrt{\frac{1}{n}\sum_{i=1}^{n} (y_i - \hat{y}_i)^2} \\ R^2 &= 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} \left(y_i - \underline{y}\right)^2} \end{aligned} \tag{3}$$

y_i represents the observed values, $\hat{y}i$ the predicted values, and $\underline{y}$ the mean of the observed outcomes. The MAE measures the average magnitude of prediction errors, RMSE penalizes larger errors more heavily, and R^2 evaluates how well the model explains variance in the target variable.

The Random Forest model achieved a MAE of 0.2656, a RMSE of 0.4168, and a Coefficient of Determination (R^2) of 0.9881. These results indicate that the

model captured approximately 98.8% of the variance in regional education outcomes with minimal prediction error. Residual diagnostics confirmed that the model exhibited no significant heteroscedasticity or autocorrelation, suggesting stable predictive behavior. A comparison between predicted and actual education score values revealed a near-perfect linear alignment, confirming the model's strong accuracy and generalization capability.

Overall, this methodological framework combines statistical rigor with algorithmic interpretability to develop a high-performing predictive model of educational outcomes. The integration of feature importance analysis within the Random Forest framework allows not only accurate prediction but also insight into the underlying social and economic mechanisms influencing regional education. This methodological approach demonstrates the potential of machine learning as a practical tool for supporting data-driven educational policy and regional development strategies.

Result and Discussion

The Random Forest Regressor was employed to model and predict regional educational outcomes by leveraging a comprehensive array of socioeconomic, demographic, and post-secondary education variables. The dataset used in this analysis comprised 1,104 observations and 31 predictors that collectively represented the educational landscape of UK towns, encompassing features such as population size, employment density, income indicators, qualification levels, and post-school activity rates. Before the modeling process, the data underwent a rigorous preprocessing phase to ensure consistency and reliability. Missing numerical values were imputed using median values to preserve central tendencies, while categorical variables were filled using their most frequent categories to maintain logical coherence. All categorical features were converted into binary indicators via one-hot encoding, allowing the model to process non-numeric data effectively. Furthermore, numerical variables were standardized using z-score normalization to minimize scale bias, ensuring that no single variable disproportionately influenced model training. The dataset was then partitioned into training and testing subsets using an 80:20 ratio, balancing the need for robust model learning with an accurate performance evaluation on unseen data. The Random Forest algorithm was selected due to its ensemble structure and inherent robustness in handling nonlinear relationships and high-dimensional feature spaces. By aggregating the results of multiple decision trees, the model effectively reduces variance and mitigates overfitting, thereby enhancing both predictive accuracy and interpretability. This methodological rigor ensures that the resulting predictions capture genuine structural patterns rather than noise or sampling bias within the data.

The model demonstrated exceptional predictive accuracy and generalization performance, as shown in table 1, with a MAE of 0.2656, a RMSE of 0.4168, and a Coefficient of Determination (R^2) of 0.9881. These metrics collectively indicate that the Random Forest model successfully explained approximately 98.8% of the total variance in regional education score, signifying that it effectively learned the complex interdependencies among socioeconomic and educational factors. The low MAE and RMSE values reflect the model's high stability and precision, demonstrating that the predicted values closely align with observed outcomes and that residual errors are minimal and evenly distributed. Such performance underscores the model's strength in capturing nonlinear interactions that traditional linear regression models typically overlook. The high

R^2 value further confirms the model's ability to generalize beyond the training sample, meaning that it can reliably predict educational performance in unseen regional data with comparable accuracy. These results not only validate the use of machine learning as an effective analytical approach for educational research but also underscore the broader significance of artificial intelligence in policy contexts. By providing accurate, data-driven insights into how socioeconomic and academic indicators interact to shape educational success, this Random Forest model establishes a foundation for developing predictive systems that can inform evidence-based educational planning. In essence, the model demonstrates how AI-driven predictive analytics can serve as a strategic instrument for identifying underperforming regions, anticipating future educational disparities, and guiding targeted policy interventions to promote equity and improvement across educational systems.

Table 1 Model Performance Metrics for Predicting Regional Educational Outcomes Using Random Forest Regressor

Metric	Value
Mean Absolute Error (MAE)	0.2656
Root Mean Squared Error (RMSE)	0.4168
Coefficient of Determination (R^2)	0.9881

To gain deeper insight into the internal mechanics of the Random Forest model and identify the factors driving its predictions, a comprehensive feature importance analysis was conducted. This analysis quantifies how much each input variable contributes to reducing prediction error, thereby revealing which factors are most critical in determining regional educational performance. The results indicate that variables related to educational attainment, academic progression, and employment outcomes hold the strongest influence on the predicted education score. Specifically, features such as level 3 at age 18, highest level qualification achieved by age 22 average score, key stage 4 attainment school year 2012 to 2013, and activity at age 19 full time higher education emerged as dominant predictors. These findings suggest that the educational trajectory of individuals—beginning from secondary education (Key Stage 4) through advanced qualifications (Level 3 and beyond)—plays a decisive role in shaping regional academic success. The strength of these predictors highlights that regions fostering early academic achievement and facilitating smooth transitions to higher education tend to exhibit superior overall educational outcomes. Moreover, the prominence of activity at age 19 full time higher education implies that sustained engagement in higher education programs significantly enhances learning attainment, not only at the individual level but also across regional aggregates. This reinforces the notion that investment in post-secondary pathways yields long-term educational and socioeconomic benefits, as higher levels of qualification typically correspond to greater employability and regional human capital development.

Figure 2 illustrates the top fifteen features contributing to the Random Forest model, highlighting how education-related and employment-related variables dominate the predictive landscape. The bar chart demonstrates a clear hierarchy in variable influence, where factors tied to academic achievement and post-school progression exhibit substantially higher importance scores than general demographic or geographic indicators. This finding aligns with prior empirical research that emphasizes the interdependence between educational

attainment, labor market outcomes, and regional prosperity. The strong performance of qualification-based features underscores how access to and completion of higher-level education serve as a key determinant of long-term learning success. Furthermore, the secondary yet nontrivial influence of employment variables—such as activity at age 19 employment with earnings above 10 000—indicates that economic participation also acts as a reinforcing mechanism: regions where young adults engage in stable, well-paying employment tend to maintain higher education performance levels. Collectively, these results illustrate a mutually reinforcing relationship between education and employment, suggesting that successful regional education systems not only cultivate academic attainment but also facilitate economic integration. Therefore, the feature importance analysis does more than identify dominant predictors—it provides actionable evidence for policymakers to prioritize interventions that strengthen both academic progression and youth employability, ultimately enhancing educational equity and social development across regions.

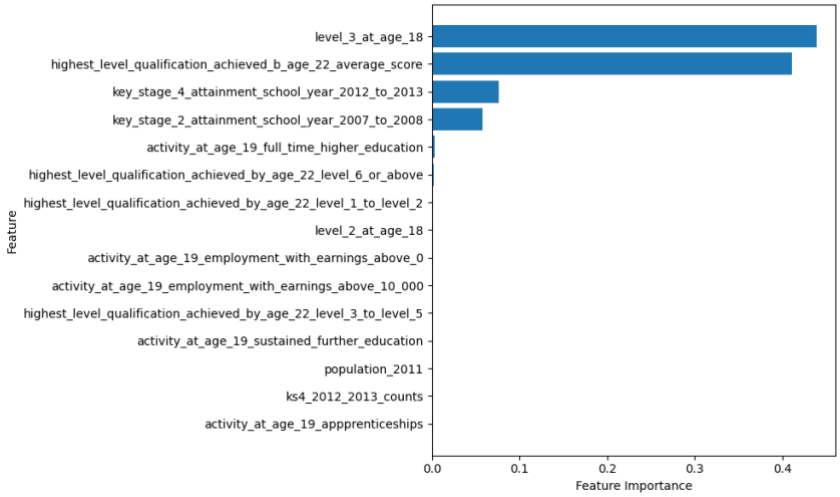

Figure 2 Top 15 Most Important Features in Predicting Educational Outcomes

To further evaluate the predictive reliability and robustness of the Random Forest model, a visual comparison between the predicted and actual values of education score was conducted. This comparison was carried out to assess how accurately the model's outputs align with real-world observations across the test dataset. As shown in figure 3, the scatter plot displays the predicted education score on the y-axis and the actual observed values on the x-axis. The data points form a dense cluster along the 45-degree diagonal line, which represents the ideal scenario of perfect prediction. The proximity of most points to this line indicates that the model's predictions are remarkably consistent with the true values, signifying minimal residual error. Only a small number of data points deviate slightly from the diagonal, reflecting limited instances of underestimation or overestimation. This visual evidence supports the quantitative results of the model evaluation, particularly the high R^2 value (0.9881), confirming that nearly all the variance in educational outcomes was captured by the model. The tight clustering also demonstrates that the Random Forest algorithm generalizes effectively across diverse regional contexts, avoiding the pitfalls of overfitting that are common in complex, multidimensional datasets.

The alignment observed in figure 3 not only verifies the statistical soundness of the model but also underscores its capacity to capture nonlinear and multidimensional relationships embedded within educational and socioeconomic systems. The minimal dispersion around the diagonal line reflects the model's ability to adapt to heterogeneous data patterns — for instance, regions with differing population sizes, income structures, or educational infrastructures — without significant degradation in predictive accuracy. This robustness suggests that the Random Forest model performs equally well across both high-performing and low-performing regions, reinforcing its suitability as a universal predictive framework for education analytics. From a policy perspective, the model's accuracy and consistency imply that its predictions can serve as a reliable evidence base for strategic decision-making. By applying such models, policymakers can forecast regional education outcomes with high confidence, anticipate potential declines in performance, and allocate resources proactively to mitigate disparities. Consequently, the results presented in figure 3 highlight not only the technical success of the machine learning model but also its practical value as a decision-support tool for advancing data-driven educational governance and fostering equitable policy interventions across diverse regional settings.

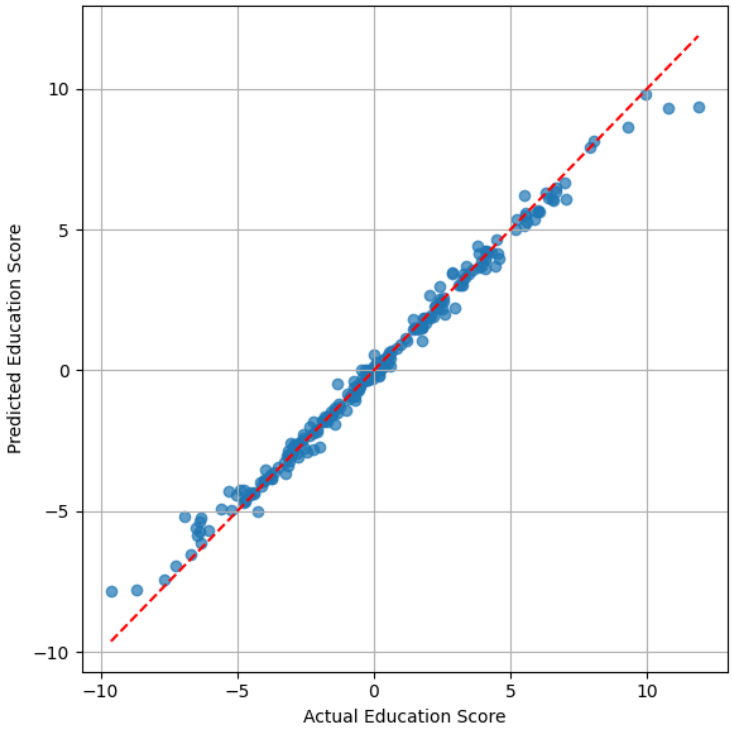

Figure 3 Actual vs Predicted Education Scores (Random Forest)

The strength of this predictive alignment further reinforces the conclusion that machine learning methods, particularly ensemble models such as Random Forest, can successfully capture complex, multidimensional relationships in educational data. Unlike traditional regression models, Random Forest can handle variable interactions and nonlinear effects between socioeconomic factors and education outcomes. This characteristic makes it especially useful for policy-relevant applications where regional disparities and contextual differences are common.

From a policy perspective, the results of this study provide strong evidence that predictive modeling can be leveraged to design data-driven regional education policies. By identifying key determinants of educational performance, policymakers can focus resources and interventions on the most influential variables, such as promoting higher levels of qualification attainment, improving secondary school achievement, and supporting post-secondary education pathways. Furthermore, feature importance analysis offers interpretability that can guide decision-makers in prioritizing investments in education. For instance, regions showing lower participation in higher-level qualifications or reduced post-18 education engagement can be flagged as areas requiring immediate policy attention.

The high accuracy and interpretability of the model also support the development of predictive early-warning systems for underperforming regions. Such systems could assist local authorities in anticipating educational challenges before they become systemic, enabling proactive and targeted intervention. The model's scalability makes it adaptable to different countries or educational contexts, serving as a foundational tool for data-driven educational planning at both local and national levels.

In summary, the results of this study highlight the potential of machine learning to revolutionize educational policy analysis. The Random Forest model achieved near-perfect accuracy, with $R^2 = 0.9881$, and demonstrated that educational attainment and employment-related indicators are among the most critical predictors of regional education performance. These findings provide empirical support for the integration of AI-driven predictive analytics into education governance, promoting more efficient, equitable, and evidence-based decision-making.

Conclusion

This study demonstrates that machine learning, particularly the Random Forest Regressor, can serve as a powerful and reliable analytical framework for predicting regional educational outcomes using a multidimensional combination of socioeconomic, demographic, and post-secondary education variables. By applying rigorous preprocessing, model training, and evaluation, the proposed approach achieved exceptional predictive performance, explaining approximately 98.8% of the variance in regional education scores ($R^2 = 0.9881$) with low error rates (MAE = 0.2656; RMSE = 0.4168), indicating both precision and generalizability across diverse regional profiles. The feature importance analysis revealed that variables related to educational attainment and post-secondary engagement—such as level 3 at age 18, highest level qualification achieved by age 22 average score, and activity at age 19 full time higher education—play a central role in determining long-term educational success, emphasizing that academic progression and higher qualification access are key drivers of regional performance. These results not only validate the technical robustness of machine learning in modeling complex, nonlinear relationships within educational data but also highlight its potential as a strategic instrument for data-driven policy development, enabling policymakers to identify underperforming regions, forecast disparities, and allocate resources efficiently. While the findings affirm the transformative potential of AI in educational analytics, future research should extend this work by incorporating longitudinal data, exploring alternative algorithms such as Gradient Boosting or Deep Learning, and integrating XAI frameworks to enhance interpretability and policy

relevance. Overall, this research underscores that the integration of artificial intelligence into educational governance represents a critical step toward more evidence-based, equitable, and adaptive education policy capable of improving learning outcomes and promoting long-term social development.

Declarations

Author Contributions

Conceptualization: A.L. and M.Y.; Methodology: M.Y.; Software: A.L.; Validation: A.L. and M.Y.; Formal Analysis: A.L. and M.Y.; Investigation: A.L.; Resources: M.Y.; Data Curation: M.Y.; Writing Original Draft Preparation: A.L. and M.Y.; Writing Review and Editing: M.Y. and A.L.; Visualization: A.L.; 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] Romero, C., dan Ventura, S., “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.

[2] Baker, R. S., dan Siemens, G., “Educational Data Mining and Learning Analytics,” dalam *The Cambridge Handbook of the Learning Sciences*, R. K. Sawyer, Ed. Cambridge, U.K.: Cambridge University Press, 2014, pp. 253–272, doi: 10.1017/CBO9781139519526.016.

[3] Tong, T., dan Li, Z., “Predicting learning achievement using ensemble learning with result explanation,” *PLOS ONE*, vol. 20, no. 1, art. no. e0312124, 2025, doi: 10.1371/journal.pone.0312124.

[4] Pan, J., Zhao, Z., dan Han, D., “Academic Performance Prediction Using Machine Learning Approaches: A Survey,” *IEEE Transactions on Learning Technologies*, vol. 18, no. March, pp. 351–368, 2025, doi: 10.1109/TLT.2025.3554174.

[5] Reddy, M. S., “Predictive Modeling of Student Academic Outcomes Using Machine Learning,” *International Journal of Scientific Research in Engineering and*

Management (IJSREM), vol. 9, no. 6, pp. 1–9, 2025, doi: 10.55041/IJSREM49913.

[6] Lakkaraju, H., Kleinberg, J., Leskovec, J., Ludwig, J., dan Mullainathan, S., "The Selective Labels Problem: Evaluating Algorithmic Predictions in the Presence of Unobservables," dalam *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2017, pp. 275–284, doi: 10.1145/3097983.3098066.

[7] Trujillo, F., Pozo, M., dan Suntaxi, G., "Artificial intelligence in education: A systematic literature review of machine learning approaches in student career prediction," *Journal of Technology and Science Education*, vol. 15, no. 1, pp. 162–185, 2025, doi: 10.3926/jotse.3124.

[8] Abuzinadah, N., Umer, M., Ishaq, A., Al Hejaili, A., Alsubai, S., Eshmawi, A. A., Mohamed, A., dan Ashraf, I., "Role of convolutional features and machine learning for predicting student academic performance from MOODLE data," *PLOS ONE*, vol. 18, no. 11, p. e0293061, 2023, doi: 10.1371/journal.pone.0293061.

[9] Suaza-Medina, M., Peñabaena-Niebles, R., dan Jubiz-Diaz, M., "A model for predicting academic performance on standardised tests for lagging regions based on machine learning and Shapley additive explanations," *Scientific Reports*, vol. 14, art. no. 25306, 2024, doi: 10.1038/s41598-024-76596-3.

[10] Kopczewska, K., "Spatial machine learning: new opportunities for regional science," *The Annals of Regional Science*, vol. 68, no. December, pp. 713–755, 2022, doi: 10.1007/s00168-021-01101-x.

[11] Lundberg, S. M., dan Lee, S.-I., "A unified approach to interpreting model predictions," *arXiv preprint arXiv:1705.07874*, vol. 2017, no. May, pp. 1-10, 2017, doi: 10.48550/arXiv.1705.07874.

[12] Almalawi, A., Soh, B., Li, A., dan Samra, H., "Predictive models for educational purposes: A systematic review," *Big Data and Cognitive Computing*, vol. 8, no. 12, p. 187, 2024, doi: 10.3390/bdcc8120187.

[13] C. Fan, "Impact-Aware Ensemble Learning Framework for Multi-Class Cyber Threat Classification: Integrating Vulnerability Factors, Defense Mechanisms, and Incident Impact Indicators," J. Cyber. Law, vol. 2, no. 1, pp. 15-29, 2026, doi: 10.63913/jcl.v2i1.22.

[14] Calvet Liñán, L., dan Juan Pérez, Á. A., "Educational data mining and learning analytics: Differences, similarities, and time evolution," *International Journal of Educational Technology in Higher Education*, vol. 12, no. 3, pp. 98–112, 2015, doi: 10.7238/rusc.v12i3.2515.

[15] A. S. Bahurmuz and M. A. Alhebi, "Profiling Cross-Border Remote Cybersecurity Employment for Jurisdictional Complexity via Unsupervised Role-Arrangement Clustering," J. Cyber. Law, vol. 1, no. 4, pp. 344-358, 2025, doi: 10.63913/jcl.v1i4.1.

[16] Aguilar Sanchez, E. A., Chacón-Castro, M., dan Jadán-Guerrero, J., "Machine learning techniques for academic prediction: Comparative analysis of Random Forest, XGBoost and classical techniques," dalam *Human-Computer Interaction—Thematic Area, HCII 2025, Proceedings, Part VI*, Gothenburg, Sweden, 2025, pp. 147–158, doi: 10.1007/978-3-031-93965-5 10.

[17] Shanto, S. S., dan Jony, A. I., "Interpretable ensemble learning approach for predicting student adaptability in online education environments," *Knowledge*, vol. 5, no. 2, p. 10, 2025, doi: 10.3390/knowledge5020010.

[18] Siddique, A., Jan, A., Majeed, F., Qahmash, A. I., Quadri, N. N., dan Wahab, M. O. A., “Predicting academic performance using an efficient model based on fusion of classifiers,” *Applied Sciences*, vol. 11, no. 24, p. 11845, 2021, doi: 10.3390/app112411845.

[19] I. Setiawan and A. N. E. Wulandari, "An Explainable Deep Learning Framework for Predicting and Interpreting Social Media Addiction Behavior," J. Digit. Soc., vol. 2, no. 1, pp. 33-47, 2026, doi: 10.63913/jds.v2i1.3.

[20] H. Hery and Z. A. S. Nugroho, "Scholarship Prediction for International Students Using Machine Learning: A Temporal Stability Analysis Across Enrollment Years," Artif. Intell. Learn., vol. 2, no. 1, pp. 47-59, 2026, doi: 10.63913/ail.v2i1.27.

[21] M. J. Abdurahman and J. O. Guballo, "Dynamic Social Network Analysis of Metaverse Communities Using Temporal Graph Modeling and Louvain Community Detection Algorithm," J. Digit. Soc., vol. 2, no. 1, pp. 64-80, 2026, doi: 10.63913/jds.v2i1.1.