FaST - Jurnal Sains dan Teknologi, Vol.10, No. 1, Mei 2026
e-ISSN 2598-9596
DOI: https://doi.org/10.19166/fastjst.v10i1/10367
Available online at: https://ojs.uph.edu/index.php/FaSTJST/article/view/10367

Harnessing Predictive Modelling for Education Index: A Dual Approach with Random Forest and Multiple Linear Regression

Olivia Kristianti Kusuma[1)], Jessica[1)], Grace Felicia Christy Widjaya[1)],
Helena Margaretha[1)], Ferry Vincenttius Ferdinand[1)], Kie Van Ivanky Saputra[*1)]
[1]Department of Mathematics, Universitas Pelita Harapan, Jl MH Thamrin Boulevard, Lippo Karawaci, Tangerang, Banten, Indonesia, 15811

ABSTRACT

This study explores the predictive modeling of the Education Index (EI) using a dual approach using Random Forest and Multiple Linear Regression (MLR). The data, obtained from "Our World in Data" spanning 1990–2022, integrates socio-economic and infrastructure indicators, including GDP per capita, government spending on education, and access to electricity. This study includes 20 countries that are categorized by income level: Low-Income (Vietnam, Nepal, Myanmar, Pakistan, Zimbabwe), Lower-Middle-Income (Ghana, Bolivia, Cambodia, Egypt, Bangladesh), Upper-Middle-Income (Argentina, Brazil, Peru, Russia, Mexico) and High-Income (Germany, Italy, Portugal, Iceland, Greece). The analysis reveals that Random Forest outperforms MLR in terms of accuracy and lower error rates, while MLR provides better interpretability of variable relationships. With R2 of 99.34% by Random Forest Regression and 94% by Multiple Linear Regression (MLR). Key findings reveal that GDP per capita, primary and secondary completion rates, and internet usage significantly influence EI, underscoring the importance of economic conditions and infrastructure for educational outcomes. This study contributes to the field by offering comparative insights into machine learning and traditional statistical methods for educational analytics, providing a robust basis for policy development to enhance global education standards.

ARTICLE INFO

Keywords: education index; random forest; multiple linear regression; socioeconomic; infrastructure

***Corresponding author**:
kie.saputra@uph.edu

Article history:
Submitted 03 Oct 2025
Revised 25 May 2026
Accepted 26 May 2026
Online Available 26 May 2026
Published 26 May 2026

1. Introduction

Education plays a fundamental role in shaping the socio-economic development of nations. The Education Index, an integral component of the Human Development Index (HDI), serves as a critical measure of a country's educational performance and its contribution to overall development. Despite its importance, accurately predicting and understanding the factors influencing the Education Index remain challenging due to the complex interplay of socio-economic, technological, and demographic variables [1]. Addressing these challenges can provide policymakers with valuable insights for targeted interventions and resource allocation.

Studies have shown that the Education Index varies significantly across countries with different income levels. In low-income countries, limited government spending on education, inadequate infrastructure, and socio-economic challenges contribute to lower educational outcomes. For instance, a study on educational expenditures in low-income countries found that factors such as GDP per capita and population demographics significantly influence education spending, which in turn affects educational attainment [2].

In contrast, high-income countries typically exhibit higher Education Index scores due to substantial investments in education, better access to educational resources, and more developed infrastructures. The disparities in educational quality and access between low- and high-income countries highlight the impact of economic factors on educational outcomes. A comprehensive analysis of educational quality across countries emphasizes that higher income levels are associated with better educational outcomes, underscoring the importance of economic resources in achieving educational equity.

Middle-income countries often display a wide range of Education Index scores, reflecting diverse stages of development and varying policy priorities. Research on educational disparities in these countries indicates that while there have been improvements in educational access, significant challenges remain in ensuring quality and equity. For example, a study mapping educational disparities across low- and middle-income countries reveals substantial inequalities in educational attainment,

FaST - Jurnal Sains dan Teknologi, Vol.10, No. 1, Mei 2026
e-ISSN 2598-9596
DOI: https://doi.org/10.19166/fastjst.v10i1/10367
Available online at: https://ojs.uph.edu/index.php/FaSTJST/article/view/10367

suggesting that economic growth alone does not guarantee educational improvement [3]. These variations in the Education Index across different income levels underscore the complex interplay of economic, social, and policy factors that influence educational outcomes globally.

Several studies have explored the determinants of the Education Index. For instance, government spending on education, access to electricity, and internet usage have been identified as significant contributors to improving educational outcomes [4]. However, the variability in these factors across different countries, coupled with their interdependencies, necessitates robust predictive modeling approaches to derive actionable insights. Traditional statistical methods such as Multiple Linear Regression (MLR) have been widely employed for this purpose due to their interpretability and simplicity [5]. However, their effectiveness is often limited in capturing non-linear relationships between variables, highlighting the need for advanced machine learning techniques like Random Forest Regression [6].

Building on previous foundational analysis which has given spotlight to the predictive significance of female enrollment, child mortality as well as corruption on the Education Index [7], our study incorporates socioeconomic and infrastructure variables to capture a wide range of influences. This research expands the analysis towards the Educational Index (EI) by incorporating government spending on education, internet usage, primary and secondary completion rates, GDP per capita, population, and access to electricity. These variables were selected to bridge the gap in the previous study by integrating factors that influence access and quality in the education system. To illustrate, both government spending and electricity access provide structural support toward education, whereas internet usage underscores the role of technology in a modern learning environment. This multifaceted approach thus allows for a more comprehensive exploration of the dynamic factors shaping the global education outcome.

To further investigate these aspects, this study uses 20 countries categorized by their income level: Low-Income (Vietnam, Nepal, Myanmar, Pakistan, Zimbabwe), Lower-Middle-Income (Ghana, Bolivia, Cambodia, Egypt, Bangladesh), Upper-Middle-Income (Argentina, Brazil, Peru, Russia, Mexico), and High-Income (Germany, Italy, Portugal, Iceland, Greece). These countries were chosen to represent a balanced sample across different income levels as classified by the World Bank, ensuring a wide spectrum of economic diversity, educational infrastructure, and developmental stages. Additionally, the selection reflects diverse geographical and cultural contexts, enabling a broader understanding of how global and local factors influence educational outcomes. For example, low-income countries like Vietnam and Nepal demonstrate notable educational progress despite economic constraints, while high-income countries such as Germany and Iceland showcase well-established education systems. This curated sample enhances the relevance of the study by capturing both universal and income-specific factors affecting the Education Index, making the findings applicable to global audiences.

In this study, we aim to harness the potential of both Random Forest and Multiple Linear Regression (MLR) to predict the Education Index using a comprehensive dataset of socio-economic and demographic indicators. This dual approach leverages the interpretability of MLR and the predictive power of Random Forest. The primary objectives are to evaluate the predictive performance of both models, identify significant factors influencing the Education Index, and provide insights to inform policy decisions aimed at improving educational outcomes. While Random Forest excels in handling high-dimensional data and capturing non-linear interactions [8,9], its application in predicting education outcomes, compared to traditional methods like MLR, remains underexplored [10, 11]. Addressing this gap is crucial for understanding the relative strengths and weaknesses of these approaches and selecting the most effective predictive tools [12].

By addressing the identified research gap, this study is expected to contribute to the existing body of knowledge in two significant ways. First, it will offer a comparative analysis of traditional and machine learning methods in the context of educational analytics. Second, it will provide policymakers and educators with evidence-based insights into the drivers of the Education Index, facilitating more informed decision-making. The results of this research have the potential to guide resource allocation and policy formulation, ultimately contributing to the enhancement of global education standards.

FaST - Jurnal Sains dan Teknologi, Vol.10, No. 1, Mei 2026
e-ISSN 2598-9596
DOI: https://doi.org/10.19166/fastjst.v10i1/10367
Available online at: https://ojs.uph.edu/index.php/FaSTJST/article/view/10367

2. Methodology

This section outlines the research methodology used in forecasting the Education Index (EI), specifically using Random Forest alongside Multiple Linear Regression. The objective of this research is to compare the effectiveness of the two different models in predicting the educational index by utilizing a variety of socioeconomic and infrastructure variables.

Table 1 Data Sources

No	Variable	Source	Number of instances
1	Education index	**worldpopulationreview.com**	655
2	Government spending on education	**ourworldindata.org**	655
3	Share of population using internet	**ourworldindata.org**	655
4	Primary completion rate	**ourworldindata.org**	655
5	Secondary completion rate	**ourworldindata.org**	655
6	GDP per capita	**ourworldindata.org**	655
7	Population	**ourworldindata.org**	655
8	Access on electricity	**ourworldindata.org**	655

2.1 Data Collection

This research employs panel data derived from two primary databases (see **Table 1**), providing a cross-sectional time-series structure. The historical data spans from 1990 to 2022 and covers 20 countries. The total number of observations is 655, which is accounted for by the time dimension: 15 countries contribute the full 33-year time series (1990–2022), while the remaining 5 countries contribute 32 data points each, resulting in a total of (15×33) + (5×32) = 655 observations. The variables collected include government spending on education, access to electricity, primary and secondary completion rates, GDP per capita, and population. The dependent variable in this research is the Education Index (EI), which provides an overview of a country's education level.

2.2 Data Preprocessing

Before moving forward, we will undergo several data preprocessing steps. Linear Extrapolation, which was utilized in this study, estimates missing data points outside the range of observed values by fitting a linear trend line based on the closest available consecutive historical periods. Many variables such as GDP per Capita and Share Using Internet naturally exhibit strong upward or downward trends over time. Linear extrapolation preserves this trend per country, ensuring a more realistic and justifiable data structure for machine learning models.

However, some variables has structural missingness, which means certain macroeconomic variables are entirely missing across all years. In this situation linear extrapolation was mathematically impossible due to the absence of historical anchor points. We applied a Group-Based Mean Imputation, where the missing values were substituted using the annual average of countries within the same Country Classification.

Normalization of the numerical variable was conducted out using Gaussian Normalization, where data is transformed into a standard normal distribution with a mean of 0 and a standard deviation of 1 [13]. Lastly, the data is split into testing and training data, where 80% of the data was allocated to train the model, and the remaining 20% was utilized for testing and model evaluation.

2.3 Predictive Modelling Approaches

The model used in this research is the Random Forest and Multiple Linear Regression. These methods are chosen due to their ability to model the intricate relationship between the independent and dependent variables [14]. Random Forest provides flexibility in non-linear relationships and particularly effective in reducing overfitting whilst enhancing robustness. While, Multiple Linear Regression offers clear interpretation whilst upholding their simplicity through their defined coefficients [15].

FaST - Jurnal Sains dan Teknologi, Vol.10, No. 1, Mei 2026
e-ISSN 2598-9596
DOI: https://doi.org/10.19166/fastjst.v10i1/10367
Available online at: https://ojs.uph.edu/index.php/FaSTJST/article/view/10367

2.4 Model Evaluation

Both models were evaluated using multiple metrics, including Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Percentage Error (MAPE), and the coefficient of determination (R^2). These metrics provide a comprehensive assessment of accuracy, precision, and the ability of the models to explain variance in the Education Index (EI).

3.

3. Results and Discussion

3.1 Random Forest Model

The Random Forest model was created following pre-processing processes and the splitting of 80% of the data into the training dataset. The performance of the resulting model was evaluated, as shown in **Table 2**.

Table 2 Random Forest Performance Metrics

Metric	Value
MSE	0.000613
RMSE	0.024767
MAE	0.016496
MAPE (%)	2.7099
R2	0.975816

The Random Forest model exhibits excellent predictive result, with a low MSE (0.000613) and RMSE (0.024767) indicating a small deviation between the actual and predicted data. This confirms that the model has exceptional performance. This is reflected by the visualization of Actual and Predicted Education Index in **Figure 1**.

Table 3 Variable importance

Feature	Total increase in node purity
GDP per capita	0.767182
Share of population using internet	0.102819
Population	0.054212
Government spending on education	0.040617
Access to electricity	0.017306
Year	0. 010521
Country Classification_Upper Middle Income	0.004531
Country Classification_Lower Middle Income	0.001570

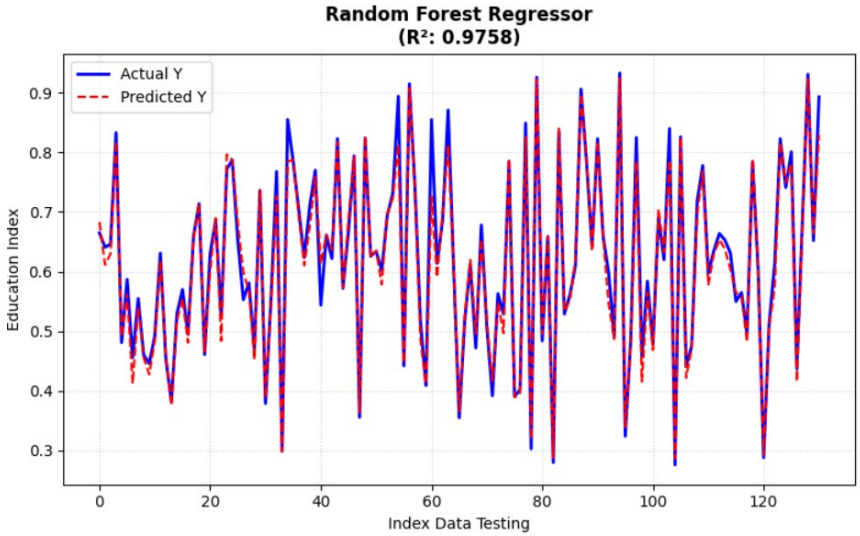

Figure 1 Actual and Predicted Education Index Random Forest

FaST - Jurnal Sains dan Teknologi, Vol.10, No. 1, Mei 2026
e-ISSN 2598-9596
DOI: https://doi.org/10.19166/fastjst.v10i1/10367
Available online at: https://ojs.uph.edu/index.php/FaSTJST/article/view/10367

We also have R^2=0.975816 indicates that this random forest model reflects that 97.6% of the variance observed can be explained in using this model, demonstrating its high predictive accuracy. The features of importance results was assessed using the total increase in Node Purity, providing an understanding of the contribution of each independent variable in **Table 3**. The feature importance analysis revealed that GDP per Capita serves as the most influential predictor, emphasizing a strong relationship between the economic and educational performance of a country. Economically, this is consistent with endogenous-growth theory: higher income levels expand fiscal space for educational spending while simultaneously elevating household capacity to invest in human capital. Partial-dependence profiles further indicate diminishing marginal returns beyond roughly, suggesting that policy leverage is greatest in low and lower middle income settings. Moreover, the use of internet and population variables also contribute significantly to the Education Index (EI)

Goverment spending on education and access to electricity emerges as the next predictors, highlighting the importance of investment and infrastructure..

3.2 Multiple Linear Regression

The multiple linear regression follows the equation below:

$$Y = \beta_0 + \beta_1 T + \beta_2 GS + \beta_3 SI + \beta_4 GDP + \beta_5 POP + \beta_6 AE + \beta_7 LI + \beta_8 LMI + \beta_9 UMI + \epsilon$$

where
Y: Dependent variable (Education Index)
β_0: Intercept
β_i: Coefficients
T: Year
GS: Government Spending on Education
SI: Share of Population Using the Internet
GDP: GDP per Capita
POP: Population
AE: Access to Electricity
LI: Country Classification: Lower Income
LMI: Country Classification: Lower Middle Income
UMI: Country Classification: Upper Middle Income
ε: Error term

Table 4 Multiple Linear Regression Performance Metrics

Metric	Value
MSE	0.004022
RMSE	0. 063423
MAE	0.050962
	1959024
MAPE (%)	8.8046
R2	0.841417

The model performance can be seen in **Table 4**. It can be observed that the regression model has a Mean Squared Error (MSE) notably higher than the Random Forest Regression model. The Mean Absolute Percentage Error (MAPE) of 8.8% also indicates a significant deviation in prediction from the actual values. This result suggests that the Random Forest Regression can capture relationships more effectively. This occurrence can be seen through **Figure 2**.

The R^2 value of 0.84 indicates that the Multiple Linear Regression model accounts for 84% of the variance in the Education Index, demonstrating a strong fit. Despite this, it remains notably lower than the R^2 achieved by the RF Model.

FaST - Jurnal Sains dan Teknologi, Vol.10, No. 1, Mei 2026
e-ISSN 2598-9596
DOI: https://doi.org/10.19166/fastjst.v10i1/10367
Available online at: https://ojs.uph.edu/index.php/FaSTJST/article/view/10367

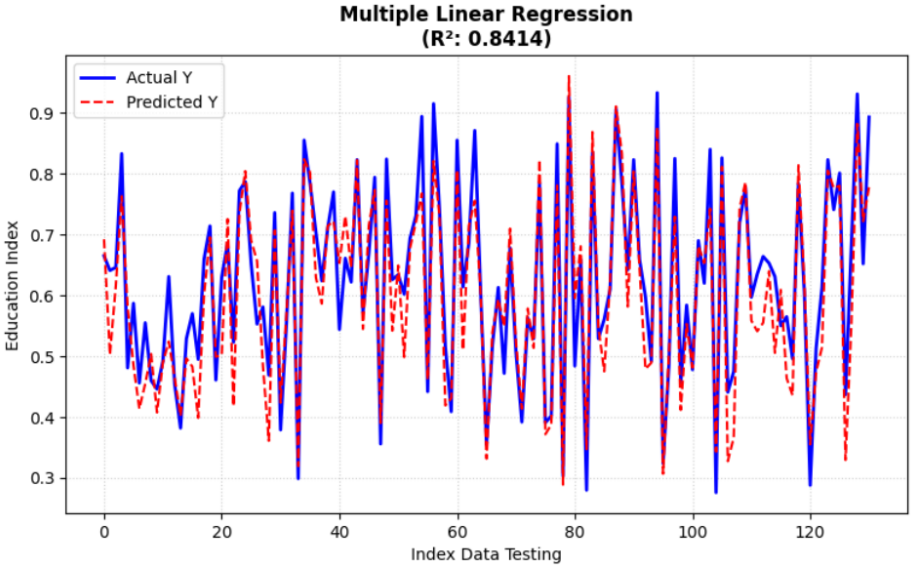

Figure 2 Actual and Predicted Education Index (Multiple Linear Regression)

Table 5 confirms that GDP per Capita, Population and Access of Electricity are the most influential predictor in the multiple linear regression (with all pvalues are less tan 2×10^{-16}). Variables that are likewise highly significant are Year and Government education spending. Variables that do not achieve significance levels are Share Using Internet. Country-income categorical variables behave as expected: relative to high-income economies (baseline), Upper-Middle groups has positive and significant offsets (pval < 0.01), whereas the Low-Income and Lower-Middle Income are not significant at the 5% level. These patterns mirror earlier cross-country evidence that economic capacity and universal basic education jointly explain most of the variation in education outcomes [16]. Collectively, both the statistical results and the external literature underscore economic growth and foundational schooling as complementary levers for improving national Education Index scores.

3.3 Comparison of Random Forest and Multiple Linear Regression

Several empirical studies in diverse applications corroborate the superiority of Random Forest over the Multiple Linear Regression in predictive tasks while simultaneously confirming the value of linear models for transparent inference.

Table 5 Multiple Linear Regression Coefficients

Coefficient	Estimate	Std. Error	p-value
Intercept	-8.5851	0.979	$<2\ 10^{-16}$
Year	0.0045	N/A	$<2\ 10^{-16}$
Gov's Spending (Education)	0.0016	0.001	0.016
Share Using Internet	$1.155\ 10^{-5}$	0.001	0.956
GDP per Capita	$9.149\ 10^{-6}$	$0.7.29\ 10^{-7}$	$<2\ 10^{-16}$
Population	$-7.488\ 10^{-10}$	$5.33\ 10^{-11}$	$<2\ 10^{-16}$
Access to Electricity	0.0014	0.001	$<2\ 10^{-16}$
Country Classification: Low Income	-0.0264	0.019	0.1724
Country Classification: Lower Middle Income	0.0085	0.019	0.654
Country Classification: Upper Middle Income	0.1123	0.014	$<2\ 10^{-16}$

Both models developed in this research provide valuable insights for the relationship of the predictors to the Education Index as well as their relative performance and interpretation. A comparative analysis has shown that it is evident that the Random Forest model outperforms the Multiple Linear Regression model across all the performance metrics, including with a higher R^2 value for the prediction model.

On the other hand, the MLR model provides critical interpretability through the regression coefficients of the model, enabling a better understanding of the relationship between the predictors and

the Education Index. Notably, in both the Random Forest and Multiple Linear Regression, the GDP per Capita as well as the Primary and Secondary Completion Rate emerge as statistically significant predictors highlighting their influence on education quality. Thus leaving a minor variation in other significant variables towards this model. This comparative analysis underlines a trade-off between predictive accuracy and interoperability. Together, whilst complementing each other, can offer a holistic framework to analyze the predictors of Education Index alongside providing meaningful insights.

4. Conclusion

This study underscores the effectiveness of predictive modelling approaches in understanding and forecasting the Education Index, with a comparative analysis of Random Forest and Multiple Linear Regression. The findings reveal that Random Forest significantly outperforms MLR in terms of accuracy, as evidenced by lower error rates and higher predictive precision. Conversely, linear regression model offers critical interpretability by elucidating the relationships between variables, highlighting a key trade-off between accuracy and insightfulness in model selection. This balance highlights the importance of selecting the appropriate model based on the specific goals of the analysis, whether prioritizing predictive precision or a deeper understanding of the relationship between variables.

The analysis identifies GDP per Capita, as well as Population and Access to Electricity as the most significant predictors of the Education Index. These factors emphasize the critical role of economic conditions and infrastructure development in shaping educational outcomes. For policymakers, this suggests that fostering economic growth, expanding access to quality education, and improving technological infrastructure should be strategic priorities to enhance global educational standards. By leveraging the strengths of both machine learning and traditional statistical methods, this research provides a robust framework for educational analytics. The primary limitation of this study stems from the structural missing data across certain countries and periods in the panel dataset [17, 18]. While minor missing data were addressed using linear extrapolation, severe missingness required group-based mean imputation using income-level classifications, which may smooth out country-specific anomalies. Additionally, to avoid data leakage and extreme multicollinearity, direct educational output proxies (Primary and Secondary Completion Rates) were deliberately excluded from the final model. Consequently, we focus strictly on macro-environmental and policy-driven inputs

As summary, our findings suggest two policy strategies: (i) macro-economic initiatives that lift per-capita income through productivity upgrades, inclusive growth, and financial-sector deepening and (ii) targeted investments that secure universal primary and secondary completion while expanding critical infrastructure such as electrification and broadband. Because the model's predictive dominance rests on GDP-per-Capita, scenario analysis using counterfactual GDP trajectories could provide actionable estimates of the educational gains achievable under alternative growth paths. Finally, while the Random Forest offers high accuracy, its black box nature warrants complementary linear or additive models for transparency, especially when communicating policy trade-offs to stakeholders

Acknowledgment

We would like to express our heartfelt gratitude to our colleagues, whose assistance and discussions provided both inspiration and motivation during the research and writing process. A We also thanks the Research and Community Service Unit (LPPM) of Universitas Pelita Harapan, whose financial support and institutional backing were essential for the successful completion of this study.

References

[1] S. Sukidin, W. Hartanto, R. N. Sedyati and S. Shofiyah, "Role of Education concerning the Gross Domestic Product. Human Development Index and Poverty Rate in East Java," *AL-ISHLAH: Jurnal Pendidikan*, vol. 15 no. 3, pp. 4140–4149, 2023. https://doi.org/10.35445/alishlah.v15i3.1716

[2] A. Hovhannisyan, R. Castillo-Ponce and R. Valdez, "The determinants of income inequality: The role of education," *Scientific Annals of Economics and Business*, vol. 66, no. 4, pp. 451–464, 2019. https://doi.org/10.47743/saeb-2019-0040

[3] Local Burden of Disease Educational Attainment Collaborators, "Mapping disparities in education across low- and middle-income countries," *Nature*, vol. 577, no. 7789, pp. 235–238, 2020. https://doi.org/10.1038/s41586-019-1872-1

[4] J. S. Jamal, M. Salam, A. N. Tenriawaru, D. Rukmana, M. H. Jamil and S. Saadah, "Determinant factors affecting the improvement of education index," *Jurnal Penelitian dan Evaluasi Pendidikan*, vol. 25, no. 1, pp. 88–96, 2021. https://doi.org/10.21831/pep.v25i1.40160

[5] F. Riandari, H. T. Sihotang and H. Husain, "Forecasting the Number of Students in Multiple Linear Regressions," *MATRIK: Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer*, vol. 21, no. 2, pp. 249–256, 2022. https://doi.org/10.30812/matrik.v21i2.1348

[6] H. S. Alim, N. Rohmah and M. Milawati, "Study of education leverage factors to improve sampang human development index," *Cendikia: Media Jurnal Ilmiah Pendidikan*, vol. 14, no. 3, pp. 366-374, 2024. https://doi.org/10.35335/cendikia.v14i4.4624

[7] O. Adeleke and P.E. McSharry, "Female enrollment, child mortality and corruption are good predictors of a country's UN Education Index," *International Journal of Educational Development*, vol. 90, pp. 102561, 2022. https://doi.org/10.1016/j.ijedudev.2022.102561

[8] G. Chairunisa, M. K. Najib, S. Nurdiati, S. F. Sanjaya, W. R. D. Andriani and D. Ekaputri, "Life Expectancy Prediction Using Decision Tree, Random Forest, Gradient Boosting, and XGBoost Regressions," *Jurnal Sintak*, vol. 2, no. 2, pp. 71-82, 2024. https://doi.org/10.62375/jsintak.v2i2.249

[9] A. Primajaya and B. N. Sari, "Random forest algorithm for prediction of precipitation," *Indonesian Journal of Artificial Intelligence and Data Mining*, vol. 1, no. 1, pp. 27–31, 2018 http://dx.doi.org/10.24014/ijaidm.v1i1.4562

[10] O. Dewi, G. E. Laukon, S. A. Sutresno and H. J. Christanto, "Modification of random forest method to predict student graduation data," *Jurnal Mantik*, vol. 7, no. 4, pp. 2949–2961, 2024. https://doi.org/10.35335/mantik.v7i4.4528

[11] S. N. Wahyuni, "Implementation of Multiple Linear Regression for Predicting Time Series Data in Infectious Diseases Using a Machine Learning Approach," *JATISI (Jurnal Teknik Informatika dan Sistem Informasi)*, vol. 11, no. 2, 2024. https://doi.org/10.35957/jatisi.v11i2.7878

[12] K. Spoon, J. Beemer, J. C. Whitmer, J. Fan, J. P. Frazee, J. Stronach, A.J. Bohonak and R. A. Levine, "Random Forests for Evaluating Pedagogy and Informing Personalized Learning," *Journal of Educational Data Mining*, vol. 8, no. 2, pp. 20–50, 2016. https://doi.org/10.5281/zenodo.3554595

[13] J. Raymaekers and P. J. Rousseeuw, "Transforming variables to central normality," *Machine Learning*, vol. 113, no. 8, pp. 4953–4975, 2024. https://doi.org/10.1007/s10994-021-05960-5

[14] S. Wijaya and Fauziah, "Analysis of the comparison between linear regression, random forest, and logistic regression methods in predicting Crude Palm Oil (CPO) price," *Brilliance: Jurnal Riset dan Konseptual*, vol. 3, no. 2, pp. 343–350, 2023. https://doi.org/10.47709/brilliance.v3i2.3334

[15] S. Obata, C. J. Cieszewski, R. C. Lowe III and P. Bettinger, "Random Forest regression model for estimation of the growing stock volumes in Georgia, USA, using dense Landsat time series and FIA dataset," *Remote Sensing*, vol. 13, no. 2, pp. 218, 2021. https://doi.org/10.3390/rs13020218

[16] R. J. Barro and J. W. Lee, "A new data set of educational attainment in the world, 1950–2010," *Journal of development economics*, vol. 104, pp. 184-198, 2013. https://doi.org/10.1016/j.jdeveco.2012.10.001

[17] J. Li, S. Guo, R. Ma, J. He, X. Zhang, D. Rui, Y. Ding, Y. Li, L. Jian, J. Cheng, and H. Guo, "Comparison of the effects of imputation methods for missing data in predictive modelling of cohort study datasets," *BMC Medical Research Methodology*, vol. 24, no. 1, pp. 41, 2024. https://doi.org/10.1186/s12874-024-02173-x

[18] S. M. Ribeiro and C. L. de Castro, "Missing data in time series: A review of imputation methods and case study," in *Learning and Nonlinear Models-Revista Da Sociedade Brasileira De Redes Neurais-Special Issue: Time Series Analysis and Forecasting Using Computational Intelligence*, vol. 19, no. 2, 2021. http://dx.doi.org/10.21528/lnlm-vol20-no1-art3