JTAM (Jurnal Teori dan Aplikasi Matematika)
http://journal.ummat.ac.id/index.php/jtam
p-ISSN 2597-7512 | e-ISSN 2614-1175
Vol. 10, No. 2, April 2026, pp. 529-539

Random Forest-Based Modeling of Life Expectancy in Central Kalimantan

Mega Puspitorini[1*], Regina Wahyudyah Sonata Ayu[1], Dita Monita[1]
[1]Department of Mathematics, Universitas Palangka Raya, Indonesia
megapuspitorini@mipa.upr.ac.id

Article History:
Received : 04-10-2025
Revised : 12-01-2026
Accepted : 13-01-2026
Online : 01-04-2026

Keywords:
Life expectancy;
Random Forest regression;
Socioeconomic indicators;
Predictive modeling;
Central Kalimantan.

ABSTRACT

This study investigates key socioeconomic determinants of life expectancy (LE) and develops a regional-level predictive model using the Random Forest regression approach for regencies and municipalities in Central Kalimantan Province during 2016–2023. Although life expectancy is a core indicator of human development, empirical studies employing machine learning methods at the sub-provincial level in Indonesia remain limited. Using secondary data from Statistics Indonesia (BPS), this study examines the relationship between LE and selected indicators related to education, sanitation, health infrastructure, economic conditions, and demography. The Random Forest model exhibits robust predictive performance, achieving MAE values of approximately 0.29–0.30 and coefficients of determination (R^2) ranging from 0.71 to 0.74 across different evaluation schemes. Feature importance analysis identifies mean years of schooling as the most influential determinant of life expectancy, followed by access to proper sanitation and the availability of health facilities. These results highlight the prominent role of human capital and basic infrastructure in shaping regional health outcomes. By integrating machine learning techniques with regional socioeconomic data, this study extends existing life expectancy research in Indonesia through a data-driven modeling framework. Overall, this study supports evidence-based planning by highlighting priority intervention areas to improve life expectancy and human development in Central Kalimantan.

https://doi.org/10.31764/jtam.v10i2.35399

A. INTRODUCTION

The Human Development Index (HDI) is a composite measure used to assess the quality of life and development of a region by considering three main dimensions: health, education, and standard of living (Fauzi & Oxtavianus, 2014). One of the key indicators within the health dimension is life expectancy (LE), which measures the average number of years a person is expected to live based on population mortality data over a specific period (Leontyeva et al., 2025). In the context of Sustainable Development Goals (SDGs), particularly SDG 3, which targets good health and well-being, LE serves as a primary indicator of population health. Moreover, LE is closely linked to other SDG targets, including access to education (SDG 4), poverty reduction (SDG 1), and improved sanitation (SDG 6) (Maarip et al., 2024). Consequently, the analysis and prediction of life expectancy provide not only insights into public health conditions but also a broader evaluation of cross-sectoral development outcomes.

Central Kalimantan Province, with a population of approximately 2.8 million in 2024, recorded a life expectancy of 68.51 years for males and 72.44 years for females in 2024, show

a slight increase from the previous year (BPS-Statistics Indonesia, 2025). Despite this improvement, Indonesia's average life expectancy remains below the global average and displays substantial regional disparities. Similar patterns of inequality in life expectancy have been widely documented in regions characterized by heterogeneous socioeconomic and geographical conditions, including disparities across developed and developing countries as well as within economically diverse regions (Martín Cervantes et al., 2020; Meshram, 2020). Such disparities are often driven by unequal access to education, healthcare services, sanitation, and economic resources, which have been widely shown to influence life expectancy and to vary substantially across regions with different socioeconomic characteristics (Agarwal et al., 2019; Georgiev et al., 2024). In regions like Central Kalimantan, characterized by spatial inequality, demographic diversity, and uneven infrastructure development, identifying and predicting the key determinants of life expectancy requires analytical approaches capable of capturing complex and nonlinear relationships (Paramita et al., 2020). Therefore, developing a data-driven predictive framework is essential to support more targeted and effective regional health and development policies.

Research on life expectancy prediction has increasingly applied machine learning approaches that integrate socioeconomic and health-related variables. Early studies demonstrated that life expectancy is influenced not only by mortality patterns but also by broader economic, educational, and healthcare factors, and that machine learning models can improve predictive accuracy by incorporating these multidimensional determinants (Agarwal et al., 2019; Bali et al., 2021). Comparative analyses also highlighted differences between developed and developing countries, with healthcare expenditure and adult mortality emerging as key predictors (Meshram, 2020). In the European Union context, ensemble-based methods such as Random Forest have been shown to effectively identify the relative importance of socioeconomic and environmental factors, including income levels, education, and public policy interventions (Martín Cervantes et al., 2020, 2021).

More recent studies have extended machine learning applications to cross-country and global analyses, highlighting the multifactorial nature of life expectancy and confirming the relevance of education, health infrastructure, and social conditions across diverse settings (Aanegola et al., 2022; Chandirasekeran et al., 2022). Advances in ensemble and gradient boosting methods, such as Random Forest and XGBoost, have further demonstrated strong predictive performance in modeling life expectancy across large datasets (Lipesa et al., 2023; Ronmi et al., 2023). These findings collectively indicate that data-driven approaches are well suited to capture the complex and nonlinear relationships underlying life expectancy outcomes.

In the Indonesian context, existing studies on life expectancy prediction have largely relied on traditional statistical, spatial, or parametric approaches using official data from Statistics Indonesia (BPS). Previous works have applied methods such as spatial regression models, Bayesian Model Averaging, and causal decomposition techniques to examine regional disparities and determinants of life expectancy across provinces and at the national level, including studies in Central Java, East Java, and rural–urban comparisons in Indonesia (Al Azies & Vivi Mentari Dewi, 2021; Hakim et al., 2019; Sudharsanan & Ho, 2020). While traditional statistical and interpolation-based approaches remain commonly used in life expectancy studies in Indonesia and provide useful baseline insights (Maarip et al., 2024), such methods

are generally limited in capturing nonlinear relationships, complex interactions, and heterogeneous patterns across subnational units, as extensively discussed in the statistical learning literature (Hastie et al., 2009; James et al., 2021).

Despite the growing application of machine learning models for life expectancy prediction, empirical studies that integrate regional socioeconomic heterogeneity and apply ensemble-based methods at the sub-provincial level in Indonesia, particularly in Central Kalimantan, remain limited. This study addresses this gap by constructing a Random Forest Regression–based predictive model of life expectancy using panel data from 14 regencies and municipalities in Central Kalimantan during the period 2016–2023, while also evaluating the model's performance and assessing the contribution of each predictor variable.

B. METHODS

1. Data

This study adopts a quantitative research design with a predictive modeling approach, employing secondary data obtained from Statistics Indonesia (BPS) of Central Kalimantan Province. The dataset consists of socioeconomic and demographic indicators collected from 14 regencies and municipalities over the period 2016–2023, resulting in a total of 114 observations. It includes one dependent variable and seven independent variables, and all quantitative analyses were performed using Python-based computational tools to ensure transparency and reproducibility. All variables were selected based on data availability and their frequent use in empirical studies on life expectancy. The dependent variable is LE, measured in years, representing the average number of years a person is expected to live from birth. The independent variables are as follows:

a. Mean Years of Schooling (MYS): The average number of years of formal education completed by individuals aged 25 years and older, measured in years.
b. Number of Health Centers (NHC): The total number of primary healthcare facilities (*Puskesmas*) available in each regency or municipality, measured in units.
c. Access to Improved Sanitation (AIS): The percentage of households with access to sanitation facilities that meet national health standards (%).
d. Regional Minimum Wage (RMW): The official minimum monthly wage established by the provincial government, measured in Indonesian Rupiah (IDR).
e. Per Capita Expenditure (PCE): The average annual expenditure per individual, measured in IDR, serving as a proxy for household welfare.
f. Total Population (TP): The total number of residents in each regency or municipality, measured in persons.
g. Poverty Rate (PR): The percentage of the population living below the official poverty line (%).

2. Analytical Techniques

The data analysis was conducted through several sequential stages, including data preprocessing, model construction, validation, and interpretation.

a. Data Preprocessing

Prior to model development, the dataset was examined for outliers. Outlier detection was performed using the interquartile range (IQR) method, where observations lying below Q1 − 1.5×IQR or above Q3 + 1.5×IQR were excluded from the analysis. After outlier removal, all variables were standardized using the Z-score transformation to ensure comparability across different measurement scales and to improve model performance.

b. Exploratory Analysis

Initial exploratory analysis was conducted using scatter plots to visually examine the relationship between the independent variables and life expectancy. In addition, a correlation matrix was generated to identify the strength and direction of linear associations among variables and to assess potential multicollinearity issues.

c. Dataset Splitting

The dataset was divided into training and testing subsets using an 80:20 ratio. The training data were used to develop the Random Forest regression model, while the testing data were reserved exclusively for evaluating out-of-sample predictive performance.

d. Random Forest Regression Model

The predictive model was constructed using the Random Forest Regression algorithm, an ensemble learning method that aggregates predictions from multiple decision trees to improve accuracy and stability. The model was implemented using Python (version 3.13) with the scikit-learn library. The Random Forest model was built with 500 decision trees, and mean squared error was used as the splitting criterion. Other hyperparameters were set to maintain model stability given the relatively small sample size.

e. Model Validation and Performance Evaluation

Model performance was evaluated using four metrics: Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and the coefficient of determination (R^2). To ensure robustness, model validation was conducted using three complementary approaches:

1) a single 80:20 train–test split,
2) 10-fold cross-validation, and
3) repeated 10-fold cross-validation with 5 repetitions.

These validation strategies were employed to verify that model performance was consistent across different data partitions and not dependent on a single random split.

f. Feature Importance Analysis

After training the Random Forest model, feature importance analysis was conducted to assess the relative contribution of each independent variable to life expectancy prediction. Feature importance values were calculated using the mean decrease in impurity method derived from the trained Random Forest model. This analysis was used

to identify the most influential socioeconomic and demographic factors affecting life expectancy in Central Kalimantan.

C. RESULT AND DISCUSSION

1. Descriptive Statistics

Descriptive statistics were employed to summarize the characteristics of the variables used in this study. The summary measures include the minimum, maximum, mean, and standard deviation, which provide an overview of the central tendency, variability, and range of the data. This step serves as a preliminary stage to understand the dataset and to ensure that the information is suitable for further analysis, as shown in Table 1.

Table 1. Descriptive Statistics of the Data

Variable	Minimum	Maximum	Mean	Std. Deviation
LE	65.40	73.70	69.47	1.81
MYS	7.09	11.65	8.48	0.91
NHC	5	26	14.39	4.95
AIS (%)	22.16	94.99	62.70	19.68
RMW (million Rp)	2.06	3.59	2.81	0.40
PCE (Rp)	7792	14727	10858.71	1394.46
TP (thousand)	57.50	466.40	191.22	112.22
PR (%)	1.96	28.20	9.95	6.55

Table 1 presents the descriptive statistics of the dataset. The average LE in Central Kalimantan is 69.47 years, with moderate variation across regions. MYS reaches 8.48 years, reflecting relatively uniform levels of formal education. The NHC shows considerable disparity, ranging from 5 to 26 units per regency/municipality. AIS is highly uneven, with an average of 62.70% and values between 22.16% and 94.99%. The RMW averages 2.81 million rupiah, while PCE records 10.86 thousand rupiah on average. The TP varies substantially from 57.5 thousand to 466.4 thousand inhabitants, and the PR averages 9.95%, indicating substantial heterogeneity across areas. These findings reveal marked regional disparities in socio-economic development within Central Kalimantan.

2. Exploratory Data Analysis

Prior to predictive modelling, exploratory data analysis was conducted to examine the relationships between LEand the explanatory variables. Scatter plots were used to visualize potential association patterns, while a correlation matrix was employed to quantify linear relationships. As illustrated in Figure 1, most independent variables do not exhibit a clear linear relationship with LE. This suggests that the associations between these variables and LE are likely nonlinear or influenced by interactions among predictors. To obtain a more comprehensive understanding of the strength and direction of these associations, a correlation matrix was employed as an initial analytical tool. As illustrated in Figure 1, most independent variables do not exhibit a clear linear relationship with LE. This suggests that the associations are likely nonlinear or influenced by interactions among variables. The correlation matrix in Figure 2 confirms this observation, with correlation coefficients between LE and the independent variables ranging from −0.19 to 0.35. The strongest correlation is observed

between AIS and LE (r = 0.35), indicating a weak positive relationship. The generally low correlation values imply that simple linear relationships are insufficient to explain variations in life expectancy. This supports the use of non-parametric and ensemble-based approaches, such as Random Forest regression, which are capable of capturing complex nonlinear patterns and variable interactions.

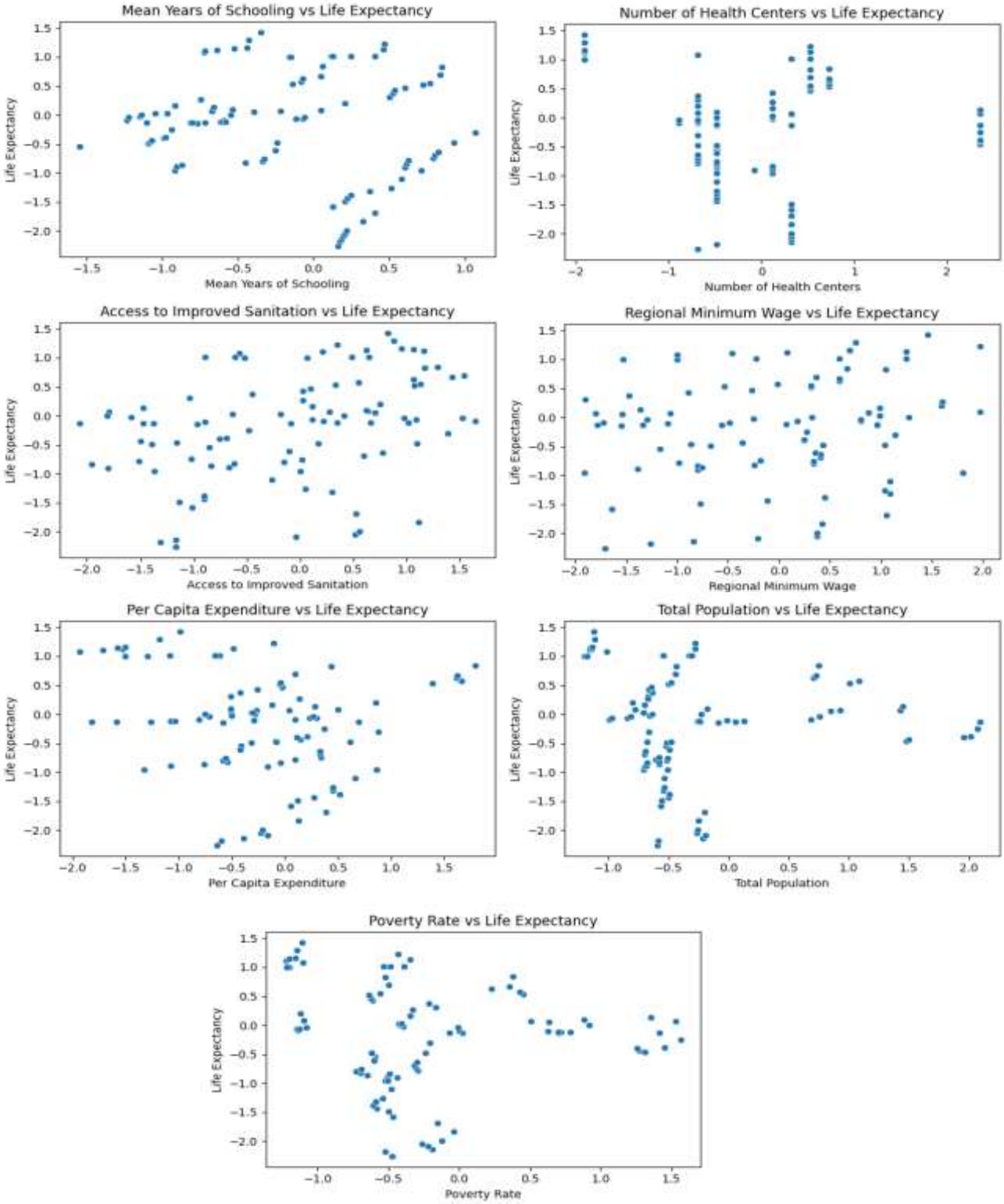

Figure 1. Scatter Plots of the Variables

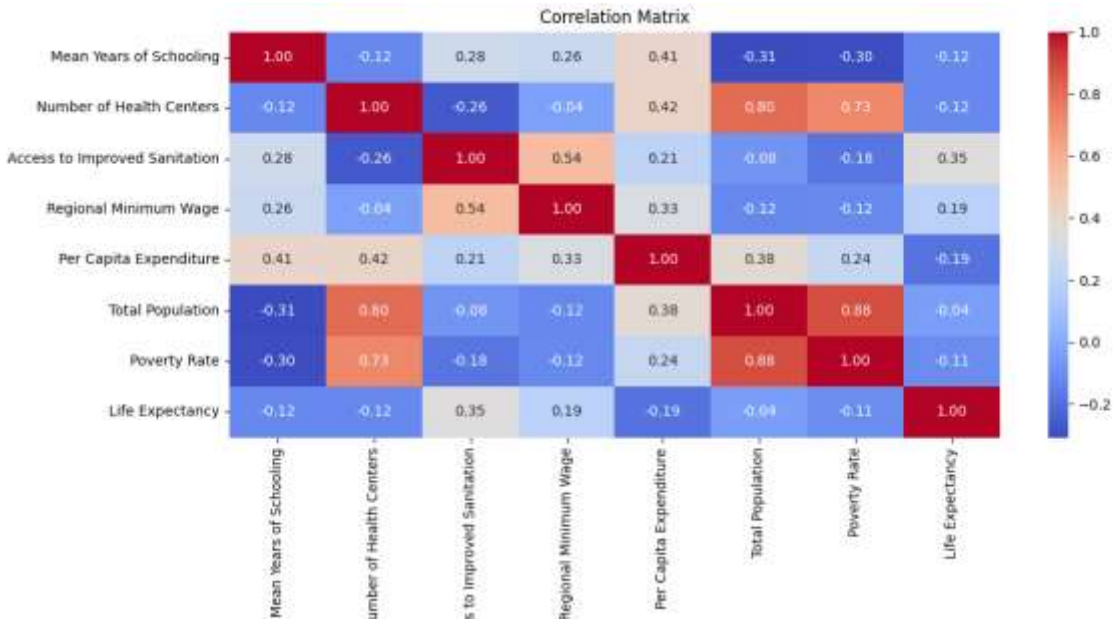

Figure 2. Correlation Matrix of Independent Variables and LE

3. Performance of the Random Forest Regression Model

The Random Forest Regression model demonstrated satisfactory predictive performance across different evaluation schemes as pesented in Table 2.

Table 2. Performance Metrics of Random Forest Regression Model

Evaluation Method	MSE	RMSE	MAE	R^2
Train-Test Split	0.15	0.39	0.30	0.74
10-Fold Cross Validation	0.23 ± 0.21	0.44 ± 0.20	0.29 ± 0.11	0.71 ± 0.18
Repeated 10-Fold CV (5 ×)	0.22 ± 0.21	0.43 ± 0.20	0.29 ± 0.12	0.71 ± 0.20

Table 2 summarizes the model performance based on train–test split, 10-fold cross-validation, and repeated 10-fold cross-validation. The Mean Absolute Error (MAE) remains stable across all evaluation methods, ranging between 0.29 and 0.30. This indicates that, on average, the predicted life expectancy deviates from the observed values by less than four months. Given that annual changes in regional life expectancy are typically small, this level of prediction error can be considered substantively acceptable.

The coefficient of determination (R^2) reaches 0.74 under the train–test split and averages approximately 0.71 under cross-validation procedures. The consistency of these values suggests that the model does not suffer from severe overfitting and maintains robust predictive performance across different data partitions. Furthermore, the proximity of the predictions to the identity line in the scatter plot indicates reliable estimation accuracy. To further evaluate the predictive performance, a scatter plot was generated to compare the predicted versus actual LE values, as shown in Figure 3.

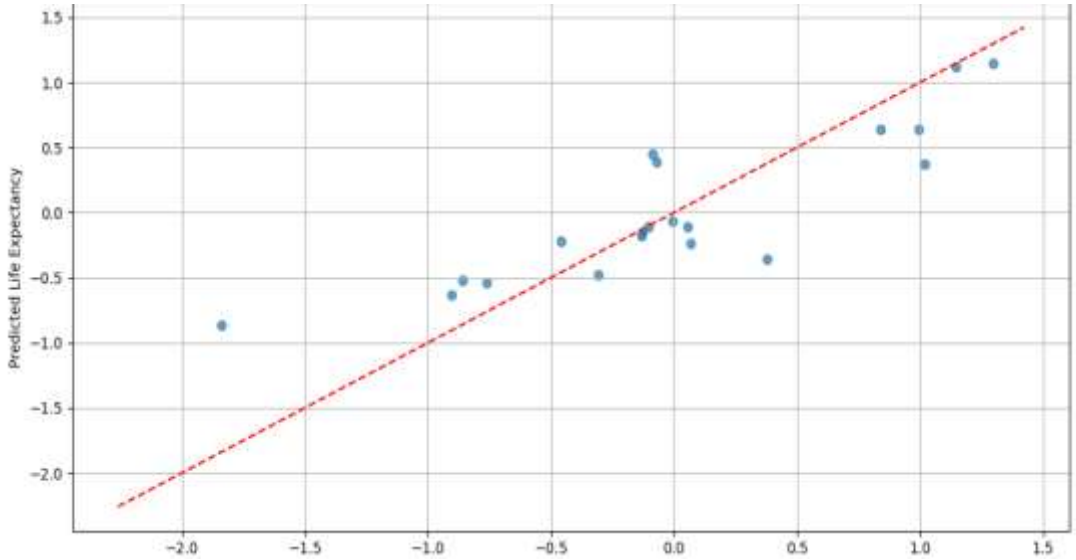

Figure 3. Scatter Plot of Predicted vs. Actual Life Expectancy

In the scatter plot, points that lie close to the diagonal (identity line) indicate accurate predictions, whereas greater deviations from the diagonal suggest larger prediction errors. In this study, most points were located near the identity line, demonstrating that the Random Forest Regression model provides reliable and accurate predictions of LE.

4. Feature Importance Analysis

Feature importance analysis was conducted to evaluate the relative contribution of each explanatory variable to life expectancy prediction, as shown in Figure 4.

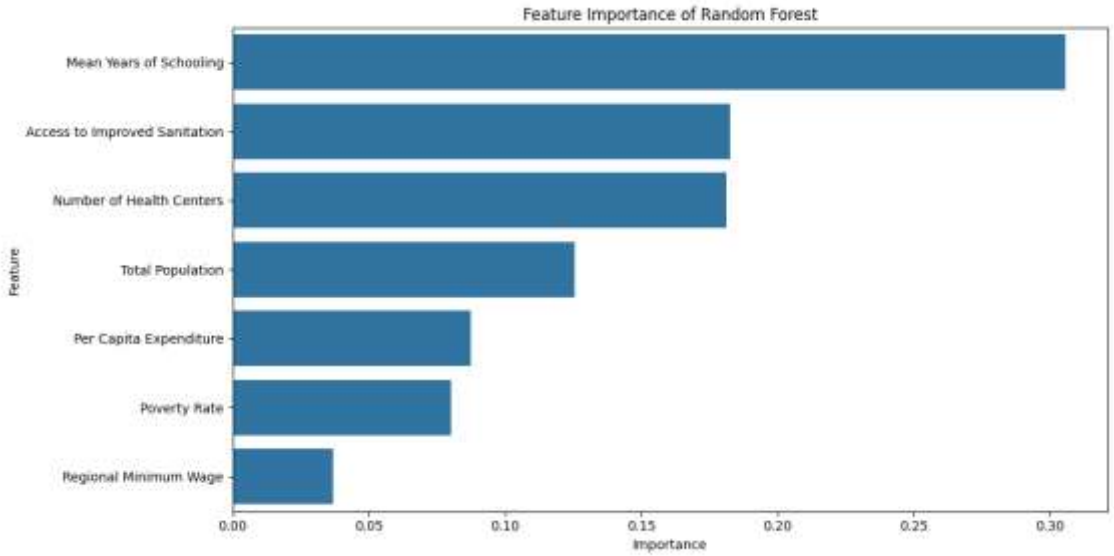

Figure 4. Feature Importance Bar Chart

As shown in Figure 4, MYS emerges as the most influential predictor, followed by AIS, NHC, TP, PCE, PR, and RMW. The findings of this study are broadly consistent with prior empirical research on life expectancy determinants. The dominant role of education aligns with studies highlighting Mean Years of Schooling as a key factor associated with longevity across both

developed and developing regions (Georgiev et al., 2024; Martín Cervantes et al., 2020). Likewise, the contributions of access to sanitation and healthcare infrastructure are supported by previous evidence emphasizing the importance of basic public health and healthcare availability in reducing mortality and improving life expectancy (Agarwal et al., 2019; Meshram, 2020). The relatively smaller influence of income-related variables, such as regional minimum wage, is consistent with findings suggesting that structural and social factors often exhibit stronger associations with life expectancy than purely economic indicators (Chandirasekeran et al., 2022). By focusing on sub-provincial data in Central Kalimantan, this study extends existing literature by demonstrating that these relationships hold within regions characterized by spatial inequality and heterogeneous development conditions.

D. CONCLUSION AND SUGGESTIONS

This study examined the socioeconomic determinants of LE across 14 regencies and municipalities in Central Kalimantan during the period 2016–2023 using a Random Forest Regression approach. The results demonstrate that the proposed model achieves robust predictive performance, explaining approximately 74% of the variation in LE, which confirms the suitability of machine learning methods for modeling complex and nonlinear relationships in regional socioeconomic data. From a scientific perspective, this study contributes to the existing literature by providing region-specific empirical evidence from Central Kalimantan, a context that has been relatively underexplored in previous life expectancy studies, particularly those employing ensemble learning techniques.

The feature importance analysis highlights Mean Years of Schooling as the most influential determinant of life expectancy, followed by access to improved sanitation and the availability of health centers. These findings reinforce the central role of human capital development, basic public health infrastructure, and healthcare accessibility in shaping population longevity at the regional level. The results also indicate that economic variables such as per capita expenditure and poverty rate play supporting, though less dominant, roles compared to education and sanitation-related factors.

Based on these findings, several policy implications can be drawn. First, regional governments should prioritize investments in education, particularly by improving access to and quality of schooling in rural and remote areas, as education emerges as the most critical factor influencing life expectancy. Second, accelerating the expansion of sanitation infrastructure and ensuring equitable access to basic healthcare facilities are essential to reduce health disparities across regions. Third, population growth and welfare-oriented policies should be integrated into regional development planning to ensure that public service capacity, especially in health and education sectors, keeps pace with demographic dynamics.

Despite these contributions, this study has several limitations. The analysis relies on secondary data at the regency and municipal levels, which may mask intra-regional disparities at more granular spatial scales. In addition, the set of explanatory variables is limited to available socioeconomic indicators and does not explicitly account for environmental quality or healthcare workforce characteristics. Future research is therefore encouraged to incorporate additional variables such as unemployment rates, physician-to-population ratios, clean water quality, and environmental indicators, as well as to apply spatial or spatiotemporal

modelling approaches using finer-scale data to support more targeted and effective policy interventions.

ACKNOWLEDGEMENT

The authors would like to express their sincere gratitude to the Faculty of Mathematics and Natural Sciences, Universitas Palangka Raya, for providing research funding that made this study possible. Appreciation is also extended to all parties who contributed directly or indirectly to the completion of this research.

REFERENCES

Aanegola, R., Nakamura Sakai, S., & Kumar, N. (2022). Longitudinal analysis of the determinants of life expectancy and healthy life expectancy: A causal approach. *Healthcare Analytics*, *2*, 100028. https://doi.org/10.1016/j.health.2022.100028

Agarwal, P., Shetty, N., Jhajharia, K., Aggarwal, G., & Sharma, N. V. (2019). Machine Learning For Prognosis of Life Expectancy and Diseases. *International Journal of Innovative Technology and Exploring Engineering*, *8*(10), 1765–1771. https://doi.org/10.35940/ijitee.J9156.0881019

Al Azies, H., & Vivi Mentari Dewi. (2021). Factors Affecting Life Expectancy in East Java: Predictions with A Bayesian Model Averaging Approach. *Jurnal Perencanaan Pembangunan: The Indonesian Journal of Development Planning*, *5*(2), 283–295. https://doi.org/10.36574/jpp.v5i2.214

Anselin, L. (2010). Thirty years of spatial econometrics. *Papers in Regional Science*, *89*(1), 3–26. https://doi.org/10.1111/j.1435-5957.2010.00279.x

Bali, V., Aggarwal, D., Singh, S., & Shukla, A. (2021). Life Expectancy: Prediction & Analysis using ML. *2021 9th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO)*, 1–8. https://doi.org/10.1109/ICRITO51393.2021.9596123

BPS-Statistics Indonesia. (2025). *Life Expectacy Rate by Province and Gender*. BPS-Statistics Indonesia. https://www.bps.go.id/en/statistics-table/2/NTAxIzI=/life-expectacy-rate-by-province-and-gender---year-.html

Chandirasekeran, P., Saravanan, S., Kannan, S., & Pattabiraman, V. (2022). Analyzing Implications of Various Social Factors on Life Expectancy. *National Academy Science Letters*, *45*(4), 311–316. https://doi.org/10.1007/s40009-022-01118-6

Fauzi, A., & Oxtavianus, A. (2014). Pengukuran Pembangunan Berkelanjutan di Indonesia. *MIMBAR, Jurnal Sosial Dan Pembangunan*, *30*(1), 42. https://doi.org/10.29313/mimbar.v30i1.445

Georgiev, V., Hadzhikoleva, S., & Hadzhikolev, E. (2024). Impact of Global Country Indicators on Life Expectancy. *Computer Science and Interdisciplinary Research Journal*, *1*(1). https://doi.org/10.70862/CSIR.2024.0101-04

Hakim, A. R., Yasin, H., & Rusgiyono, A. (2019). Modeling Life Expectancy in Central Java Using Spatial Durbin Model. *Media Statistika*, *12*(2), 152. https://doi.org/10.14710/medstat.12.2.152-163

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning*. Springer New York. https://doi.org/10.1007/978-0-387-84858-7

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An Introduction to Statistical Learning*. Springer US. https://doi.org/10.1007/978-1-0716-1418-1

Leontyeva, Y., Huang, Y., Cramb, S., Cameron, J., Baade, P., Mengersen, K., & Thompson, H. (2025). Bayesian Spatial Relative Survival Model to Estimate the Loss in Life Expectancy and Crude Probability of Death for Cancer Patients. *Statistics in Medicine*, *44*(3–4). https://doi.org/10.1002/sim.10287

Lipesa, B. A., Okango, E., Omolo, B. O., & Omondi, E. O. (2023). An application of a supervised machine learning model for predicting life expectancy. *SN Applied Sciences*, *5*(7), 189. https://doi.org/10.1007/s42452-023-05404-w

Maarip, S., Hermansyah, A., Hadraeni, S. N., Miqdad, S., Nuryadin, A. D., & Yuliyanti, S. (2024). Prediction of Life Expectancy in Indonesia by Implementing Website-Based Lagrange Polynomial

Interpolation. *International Journal of Applied Sciences and Smart Technologies*, *6*(2), 393–406. https://doi.org/10.24071/ijasst.v6i2.9167

Martín Cervantes, P. A., Rueda López, N., & Cruz Rambaud, S. (2020). Life Expectancy at Birth in Europe: An Econometric Approach Based on Random Forests Methodology. *Sustainability*, *12*(1), 413. https://doi.org/10.3390/su12010413

Martín Cervantes, P. A., Rueda López, N., & Cruz Rambaud, S. (2021). *Life Expectancy at Birth and Its Socioeconomic Determinants: An Application of Random Forest Algorithm* (pp. 383–406). https://doi.org/10.1007/978-3-030-61334-1_19

Meshram, S. S. (2020). Comparative Analysis of Life Expectancy between Developed and Developing Countries using Machine Learning. *2020 IEEE Bombay Section Signature Conference (IBSSC)*, 6–10. https://doi.org/10.1109/IBSSC51096.2020.9332159

Paramita, S. A., Yamazaki, C., & Koyama, H. (2020). Determinants of life expectancy and clustering of provinces to improve life expectancy: an ecological study in Indonesia. *BMC Public Health*, *20*(1), 351. https://doi.org/10.1186/s12889-020-8408-3

Ronmi, A. E., Prasad, R., & Raphael, B. A. (2023). How can artificial intelligence and data science algorithms predict life expectancy - An empirical investigation spanning 193 countries. *International Journal of Information Management Data Insights*, *3*(1), 100168. https://doi.org/10.1016/j.jjimei.2023.100168

Sudharsanan, N., & Ho, J. Y. (2020). Rural–Urban Differences in Adult Life Expectancy in Indonesia. *Epidemiology*, *31*(3), 393–401. https://doi.org/10.1097/EDE.0000000000001172