JURNAL ECONOMINA e-ISSN: 2963-1181

Volume 5, Number 10, October 2026

Homepage: ejournal.45mataram.ac.id/index.php/economina

Tree-Based Random Effects Models for Poverty Analysis in Indonesia

Arie Purwanto[1*], Rusmining[2], Muhamad Irfan Rumasoreng[3], Nanang Khuzaini[4]

[1*34]Mercu Buana Yogyakarta University, Indonesia
[2]Ahmad Dahlan University, Indonesia

**Corresponding Author: arie@mercubuana-yogya.ac.id*

Article History

Received: 14-10-2026
Revised: 21-10-2026
Published: 30-10-2026

Keywords: 3Trees, Multilevel, Poverty, Random Effects

ABSTRACT

Poverty is a multidimensional phenomenon influenced by socioeconomic factors and regional heterogeneity. This study aims to compare the predictive performance of the Linear Mixed Model (LMM), Mixed-Effects Regression Trees (MERT), Random Effects Expectation Maximization Tree (RE-EM Tree), and 3Trees in modeling district- and city-level poverty rates in Indonesia, accounting for clustering by province. The analysis uses data from Statistics Indonesia and evaluates model performance through within-cluster repeated two-fold cross-validation with 50 repetitions, yielding 100 model evaluations. Predictive accuracy is assessed using the Mean Squared Error (MSE) on training data and the Prediction Mean Squared Error (PMSE) on testing data. The results show that 3Trees achieves the lowest average MSE of 7.46 (standard deviation: 1.02) and the lowest average PMSE of 13.97 (standard deviation: 1.70). In comparison, LMM, MERT, and RE-EM Tree produce higher average PMSE values of 14.62, 19.38, and 17.86, respectively. These findings indicate that 3Trees provides the best predictive performance among the four methods under the evaluation procedure used in this study.

INTRODUCTION

Poverty is a major development challenge faced by many countries, including Indonesia. As a complex socio-economic phenomenon, poverty is influenced not only by economic factors but also by various dimensions, including education, health, employment, infrastructure, and social and demographic aspects (Bank, 2022). In Indonesia, although the national aggregate poverty rate has continued to decline in recent decades, its distribution remains uneven across regions. Several provinces and districts/cities still face high levels of poverty due to limited access to education and health services, as well as unequal economic opportunities (Statistik, 2024). Therefore, a more in-depth analysis of the variables influencing poverty, taking into account the multilevel data structure (districts within provinces), is crucial for effective policy formulation. Poverty data in Indonesia often has a hierarchical structure. Observation units, such as households or districts, are at the lowest level (level 1), and are then grouped into higher

units, such as provinces (level 2). This structure causes statistical problems when analyzed using classical linear regression or multiple regression methods, because of the intra-group variation that can lead to biased parameter estimates and incorrect statistical inferences.

Mixed-effects models, or Linear Mixed Models (LMMs), have been widely used to address hierarchical structures (Ruíz et al., 2023; West et al., 2022). These models combine fixed effects to capture the influence of predictor variables that are common across groups, and random effects to capture heterogeneity between group. Several recent studies have highlighted the use of multilevel models combining fixed and random effects in the context of economics and poverty using multilevel regression. The truncated multilevel linear spline regression model was used (Davala, 2024) to analyzed socioeconomic factors influencing the percentage of poor people on Sumatra Island to capture level heterogeneity and the non-linear relationship between predictor variables. The relationship between the Human Development Index (HDI) and the poverty line was studied using 2-level multilevel linear regression, and found that the higher the HDI of a region, the higher the poverty line (Dea Alvionita Azka, 2023). While LMM, or multilevel models, are effective for handling multilevel data, they have limitations. One key issue is the assumption of linearity between predictor and response variables. In practice, the relationship between socioeconomic indicators and poverty is often nonlinear and complex, involving interactions between multiple factors that are not easily determined a priori. For example, increasing life expectancy may be negatively associated with poverty levels at a given level, but the effect may differ when combined with other variables such as education level or household expenditure. These situations are difficult to capture with traditional linear models.

Modern statistical learning emphasizes three important objectives: prediction, estimation, and attribution (Efron, 2020). Prediction concerns the accuracy of forecasts for new observations, estimation focuses on quantifying model parameters and their uncertainty, while attribution aims to identify the variables and relationships that contribute to the observed outcomes. These objectives provide an important foundation for developing statistical methods that balance predictive performance, interpretability, and statistical inference. Classification and Regression Trees (CART) were introduced by (Breiman, 1991) as one of the earliest tree methods. Decision trees have advantages in visual interpretation and the ability to capture complex relationships without requiring rigid model specifications. In the context of hierarchical data, standard decision trees are insufficient because they do not consider cluster structures. To address this, various mixed methods have been developed that combine mixed-effects models with decision trees. One such method is RE-EM Trees (Random Effects Expectation Maximisation Trees). This method uses the EM algorithm to estimate random effects, while non-linear structures are modelled with regression trees (Sela, 2012). RE-EM Trees have proven effective in handling longitudinal and clustered data, although their computation is relatively complex. This model was proposed by (Sela, 2012) for both longitudinal and hierarchical data. RE-EM Tree was applied to analyze a multivariate real data set in which multidimensional nonfinancial characteristics of poverty across different countries were treated as responses, and various potential causes of poverty were included as predictors

(Jing, 2023). Study Small Business Rate Relief (SBRR) illustrates how RE-EM Trees extend multilevel mixed-effects models to capture non-linear, hierarchical effects of policy interventions across firms, industries, and regions. The main advantage of REEM is its flexibility in dealing with non-linear relationships without having to explicitly define the form of the function. However, its disadvantages are relatively high computational costs and the risk of overfitting when the sample size is small or the tree is too deep.

Further development was carried out by (Hajjem, 2011) using the Mixed Effects Regression Trees (MERT) method, which combines regression trees with mixed effects models. Generalized Mixed-Effects Regression Trees (GMERT) extend MERT to handle various outcome types, including binary and categorical responses, making them suitable for socio-economic applications such as classifying household poverty status. In Indonesia, where poverty exhibits both high prevalence and significant regional disparities, GMERT and related tree-based mixed models allow researchers to capture non-linear relationships, cluster-level effects, and interactions among predictors, providing a robust tool for understanding determinants of poverty (Rahmawati, 2023). The advantage of MERT lies in its ability to produce easily interpretable models, with tree structures explicitly showing the data-splitting rules. However, its limitations include difficulties in handling very large datasets, and its performance is highly dependent on tree parameters such as tree depth and the minimum number of observations per node. The research also presents the 3-Trees approach, which is a method based on three regression trees combined in a mixed effects framework to improve predictions while maintaining interpretability (Gottard, 2023). Nevertheless, these tree-based mixed methods are still rarely used in the field of economics. Recent literature shows a growing interest in the integration of mixed effects models and tree-based methods. On the other hand, lain (Jahangiri, 2025) show the potential of Mixed-Effects Regression Trees in capturing heterogeneity between groups in multilevel data.

Previous studies have focused more on traditional multilevel linear regression or conventional statistical models. However, poverty as a multidimensional phenomenon is likely to have non-linear patterns and complex interactions between factors that are better modelled using tree-based mixed models. Thus, comparing the performance of LMM, RE-EM Tree, MERT, and 3-Trees in analysing poverty rates in Indonesia is relevant both academically and practically. This study aims to evaluate the effectiveness of various tree-based random effects models in analysing poverty data at the district and provincial levels in Indonesia. In order to bridge this gap, this study applies the 3Trees random effects model to analyse poverty at the district and provincial levels in Indonesia. This study aims to evaluate the effectiveness of various tree-based random effects models compared to LMM in predicting poverty. In addition, this study aims to identify socio-economic factors and their complex interactions. It is hoped that the results of this study will not only contribute methodologically to the field of applied research, but also provide useful empirical insights for policymakers in formulating more targeted poverty alleviation strategies.

RESEARCH METHODOLOGY

This study leverages advanced multilevel modeling techniques to examine poverty at both the district and provincial levels in Indonesia, emphasizing the interplay between individual-level socio-economic determinants and cluster-level structural factors. The dataset, obtained from the Indonesian Statistics Bureau (BPS, 2024), encompasses 516 districts/cities grouped into provinces, providing a naturally hierarchical structure that necessitates multilevel modeling to appropriately account for within- and between-cluster variations. To establish a methodological baseline, Linear Mixed Models (LMM) are employed to estimate fixed effects of predictors and random effects capturing province-level heterogeneity. While LMM offers interpretability and a standard framework for multilevel data, it is limited in capturing complex non-linear relationships and high-order interactions among socio-economic predictors. To overcome these limitations, this study applies tree-based mixed-effects approaches, namely Random Effects Expectation-Maximization Trees (REEM Trees), Mixed-Effects Regression Trees (MERT), and the 3-Trees model. REEM Trees integrate random effects with a decision tree structure, effectively addressing clustered or longitudinal data (Sela, 2012). MERT extends this framework by accommodating non-linear interactions among predictors while maintaining hierarchical dependencies (Gottard, 2023). The 3-Trees model introduces a hybrid approach, combining a linear component with three regression trees to jointly capture main effects, non-linearities, and cross-level interactions, offering enhanced predictive performance while maintaining interpretability. This approach is particularly suitable for socio-economic applications such as poverty analysis, where interactions among individual and regional factors are complex and non-linear. The analytical procedure is conducted as follows:

1. Data Preparation and Cleaning: Raw BPS data were harmonized for consistent district and province identifiers, missing values were addressed to including imputation for newly formed provinces using averages from parent regions and duplicates or extreme outliers were removed.
2. Exploratory Analysis: Descriptive statistics and visualizations were generated to examine distributions, regional disparities, and potential anomalies in both response (poverty rate) and predictor variables.
3. Baseline Model Estimation: LMM was fitted to provide a benchmark, with predictor variables tested for multicollinearity (VIF) and residual assumptions such as normality, homoskedasticity, and independence (Kutner, 2005; Afshartous, 2005).
4. Advanced Tree-based Modeling: REEM Trees, MERT, and 3-Trees were estimated to capture non-linear relationships, interactions, and hierarchical clustering. Model pruning and parameter tuning were performed to avoid overfitting and enhance interpretability.
5. Model Evaluation: Predictive performance was assessed using Mean Squared Error (MSE) at the district level and Clustered MSE (clusMSE) at the provincial level to account for the hierarchical structure of the data (Gottard et al., 2023; Asrirawan et al., 2025).

6. Interpretation of Determinants: Models were analyzed to identify the most influential socio-economic factors and their complex interactions at both district and provincial levels.
7. Comparative Analysis: Tree-based mixed-effects models were compared with LMM to highlight improvements in predictive accuracy, interpretability, and capacity to model non-linear, multilevel interactions.

The primary objective of this study is to evaluate the effectiveness of tree-based random effects models, particularly the 3-Trees model, relative to classical LMM in predicting poverty. Additionally, the study aims to identify complex socio-economic interactions that influence poverty outcomes. By doing so, it contributes methodologically to the advancement of multilevel modeling and empirically to providing policymakers with actionable insights for more targeted and effective poverty alleviation strategies in Indonesia.

RESULT AND DISCUSIONS

Data Preparation and Cleaning

The data used in this study were sourced from the Indonesian Statistics Bureau (BPS) through their official portal (https://www.bps.go.id/id), which provides various national statistical publications, including tables on socio-economic, demographic, and welfare indicators (Badan Pusat Statistik Indonesia, 2024). The raw dataset comprises 516 district/city units across Indonesia, which are subsequently grouped into provinces as clusters using provincial codes (idprov). This data structure is highly suitable for a multilevel approach, where districts/cities constitute the first level and provinces are treated as the second level in the analysis. The analyzed variables include a response variable (Y), namely the percentage of the population living in poverty in each district/city, and a set of predictor variables. Level-1 predictor variables (X) reflect district/city characteristics, such as male and female life expectancy, the proportion of males aged ≥25 with higher education, the Human Development Index (HDI), the proportion of women aged 15–49 who gave birth to their first child before age 20, and the percentage of women employed as professionals. Level-2 predictor variables (Z) describe macro-level provincial conditions, including the underemployment rate, per capita rural expenditure, contraceptive use among married women, household sanitation facilities, housing ownership status (rented or contracted), the proportion of resolved criminal cases, and the proportion of the population aged 15–59 with information and communication technology (ICT) skills. The diversity of this data indicates significant variation both across districts and provinces, thereby supporting the need for multilevel models capable of capturing heterogeneity at the district level while accounting for inter-provincial effects.

Before being used for model estimation, the data underwent a thorough cleaning and verification process to ensure quality. The first step involved aligning regional identifiers: all district/city and province names were standardized to match the latest administrative codes issued by BPS, avoiding inconsistencies arising from territorial expansions or changes in nomenclature. The next step was checking for missing values. Each variable was examined for missing data (NA). When found, cross-verification was conducted with other BPS publications

or relevant official sources. If data remained unavailable, the corresponding observations were excluded from the analysis to prevent bias and maintain model consistency. Special cases occurred in Central Papua and Highland Papua provinces, which are newly formed regions. Several indicators for these provinces were not yet reported in BPS publications, resulting in missing data. To address this, missing values were imputed using the averages from the overall Papua region. This approach ensures that newly formed provinces are represented in the analysis without compromising data consistency.

Subsequently, checks for duplicates and extreme values were performed to detect districts recorded more than once or implausible values, such as percentages exceeding 100% or HDI values far outside the national range. Extreme values were verified against original sources, and any confirmed data entry errors were corrected or removed from the dataset. The final step involved validating the entire dataset. Descriptive statistics, including mean, median, minimum, and maximum values, were computed to ensure there were no remaining missing values, duplicates, or inconsistencies between variables. The resulting dataset is clean, verified, and ready for further analysis using multilevel models, including Linear Mixed Models (LMM), Random Effects Expectation Maximization (REEM), Mixed-Effects Regression Trees (MERT), and the hybrid 3-Trees tree-based approach.

Exploratory Analysis

In poverty research, the selection of variables plays a crucial role as it provides a more comprehensive understanding of the factors influencing socio-economic conditions. The variables employed in this study include both individual-level and provincial-level indicators. Individual-level indicators were chosen due to their close association with quality of life, such as health, education, women's participation, and human development dimensions that serve as measures of well-being. Provincial-level indicators, on the other hand, represent structural conditions and macro-level environments, such as unemployment rates, household expenditure patterns, access to basic facilities, and technological capacity. The combination of these two levels of variables is expected to capture the complexity of poverty determinants, arising both from individual characteristics and the contextual environment of the region in which individuals reside.

Table 1. Variable Description

Variable	Level	Description
Y	Respon	Poverty rate (%)
X1	Level-1 predictor	Male life expectancy
X2	Level-1 predictor	Percentage of population aged ≥25 with upper secondary education or higher, by gender
X3	Level-1 predictor	Human Development Index (HDI)
X4	Level-1 predictor	Proportion of women aged 15–49 who gave birth to their first child before age 20
X5	Level-1 predictor	Percentage of women employed as professionals
Z1	Level-2 predictor	Provincial underemployment rate (%)

Z2	Level-2 predictor	Average monthly per capita expenditure on food and non-food items in rural areas, by province (%)
Z3	Level-2 predictor	Percentage of married women aged 15–49 using contraceptives (%)
Z4	Level-2 predictor	Percentage of households with private sanitation facilities (%)
Z5	Level-2 predictor	Average housing ownership status (rented/contracted) across the province, regardless of urban/rural location
Z6	Level-2 predictor	Percentage of resolved criminal cases in 2023 (%)
Z7	Level-2 predictor	Proportion of population aged 15–59 with information and communication technology (ICT) skills (%)
idprov	Identifier	Provincial identifier used as the cluster unit in the analysis

Descriptive statistics reveal significant disparities across regions in Indonesia. Regarding the poverty rate (**Y**), the lowest value is observed in Badung Regency, Bali Province (2.23%), while the highest is found in Intan Jaya Regency, Central Papua (41.42%), highlighting regional welfare gaps. For male life expectancy (**X1**), Nduga Regency, Highland Papua, records the lowest at 54.22 years, whereas Sukoharjo Regency, Central Java, shows the highest at 80.38 years. The percentage of population aged ≥25 with upper secondary education or higher (**X2**) is highly uneven, with Puncak Regency, Central Papua, at 1.14%, compared to Banda Aceh City at 84.01%. Similar disparities are observed in HDI (**X3**), with Nduga Regency, Highland Papua, at 36.30 and Yogyakarta City at 88.77. For the proportion of women aged 15–49 who gave birth to their first child before 20 years (**X4**), the lowest is in Pariaman City, West Sumatra (0.038), while Puncak Regency, Central Papua, is the highest at 0.529. The percentage of women employed as professionals (**X5**) ranges from 2.32% in Paniai Regency, Central Papua, to 77.05% in Puncak Regency, Central Papua, reflecting disparities in female workforce participation.

At the labor level, the lowest underemployment rate (**Z1**) is in Jembrana Regency, Bali (2.35%), whereas West Lombok Regency, West Nusa Tenggara, reaches 15.23%. Monthly per capita rural expenditure (**Z2**) is lowest in Nduga Regency, Highland Papua (32.45%) and highest in Kepulauan Seribu, DKI Jakarta (60.34%), indicating a regional purchasing power gap. In family planning, the percentage of married women aged 15–49 using contraceptives (**Z3**) is lowest in Mimika Regency, Central Papua (11.39%) and highest in Tanah Laut Regency, South Kalimantan (70.08%). Access to sanitation via private toilets (**Z4**) is lowest in Nduga Regency, Highland Papua (0.64%) and highest in Kepulauan Seribu, DKI Jakarta (22.93%). Housing ownership (**Z5**) varies, with the lowest percentage of rented/contracted homes in Majene Regency, West Sulawesi (0.51%), and the highest in Kepulauan Seribu (21.30%). Regarding legal enforcement, resolved criminal cases (**Z6**) are lowest in Jayapura Regency, Papua (10.13%) and highest in Malinau Regency, North Kalimantan (95.47%). Finally, in digital literacy, the proportion of individuals aged 15–59 with ICT skills (**Z7**) ranges from 8.27% in Nduga Regency, Highland Papua, to 95.83% in Karimun Regency, Riau Islands.

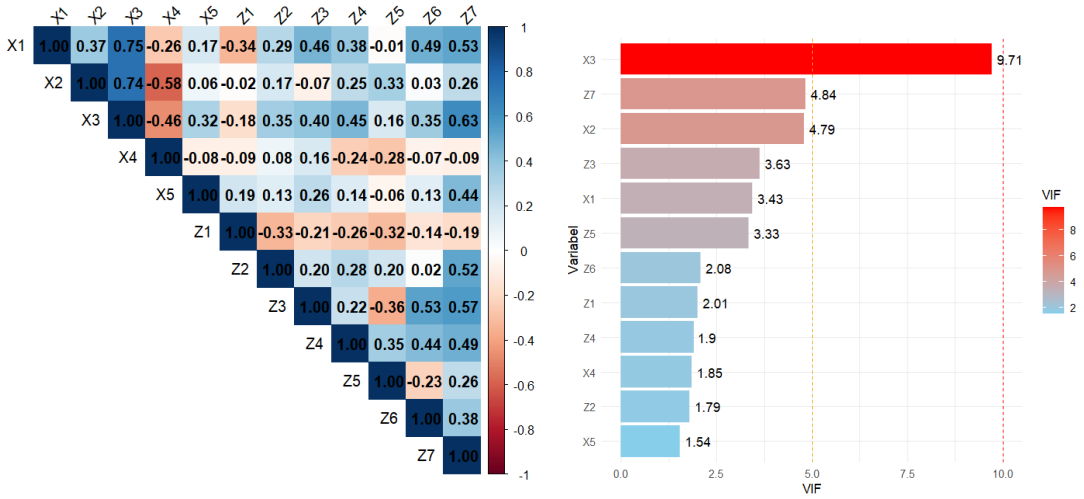
Descriptive statistics reveal substantial disparities across regions in Indonesia, reflecting profound socio-economic inequality. The poverty rate varies dramatically, with Badung Regency, Bali, recording the lowest rate at 2.23%, while Intan Jaya, Central Papua, reaches 41.42%, highlighting a significant welfare gap between regions. Health indicators also show wide variation: male life expectancy ranges from 54.22 years in Nduga, Highland Papua, to 80.38 years in Sukoharjo, Central Java, suggesting disparities in access to healthcare and nutrition. Educational attainment is highly uneven, with only 1.14% of the population aged 25 and above in Puncak, Central Papua, completing upper secondary education or higher, compared to 84.01% in Banda Aceh. Women's participation in the professional workforce similarly fluctuates, from 2.32% in Paniai, Central Papua, to 77.05% in Puncak, Central Papua, indicating regional differences in gender equality and economic opportunity. Family planning practices and access to reproductive health services vary, with contraceptive use among married women aged 15–49 as low as 11.39% in Mimika, Central Papua, and as high as 70.08% in Tanah Laut, South Kalimantan. Economic disparities are evident in rural per capita expenditure, ranging from 32.45% in Nduga to 60.34% in Kepulauan Seribu, while basic infrastructure such as private sanitation facilities is limited in Nduga (0.64%) compared to 22.93% in Kepulauan Seribu. Housing conditions, law enforcement outcomes, and digital literacy similarly exhibit pronounced regional variation, reflecting significant gaps in living standards, governance, and technology access across the country. these patterns underscore the multidimensional nature of inequality in Indonesia. Regional disparities are not confined to income or poverty rates alone but extend to health, education, gender equality, infrastructure, legal enforcement, and digital literacy. Addressing poverty and promoting equitable development thus requires targeted, context-specific policies that consider both individual-level factors and broader structural conditions at the provincial and district levels.

Baseline Model Estimation

Assumption Testing in Linear Mixed Models (LMM)

Before conducting analyses using the Linear Mixed Model (LMM) or tree-based models, an essential preliminary step is the assessment of classical assumptions. This procedure aims to ensure that the constructed model not only provides accurate estimates but also meets fundamental statistical requirements that guarantee the reliability of the results. One crucial assumption to examine is multicollinearity, which occurs when predictor variables in the model are highly correlated, making it difficult to disentangle the individual effects of each variable on the response variable. High multicollinearity can lead to unstable regression coefficients, inflated parameter variances, and reduced inferential power of the model. In this study, multicollinearity was assessed using two approaches: the correlation matrix and the Variance Inflation Factor (VIF). The correlation matrix provides an initial overview of the strength of linear relationships among variables, whereas the VIF quantifies the extent to which the variability of a given predictor can be explained by other predictors in the model. Generally, correlation values above 0.70 warrant caution, while VIF values ≥10 indicate the presence of

serious multicollinearity. The corresponding plots are presented in Figures (a) and (b), respectively.



(a) (b)

Figure 1. (a) correlation matrix, (b) VIF value for every variable

Based on the multicollinearity assessment combining information from the correlation matrix and Variance Inflation Factor (VIF) calculations, there is a strong indication of substantial linear relationships among several predictors. From the correlation matrix, variable **X3** shows high correlations with **X1** (0.75) and **X2** (0.74). Correlation values above 0.70 are generally considered large enough to potentially cause multicollinearity issues, as strong linear relationships among predictors can render parameter estimates biased or unstable. This finding aligns with the VIF results, where most variables exhibit VIF values below 5, indicating that they remain within a safe range and can be retained in the model. However, **X3** has a VIF of 9.71, approaching the critical threshold of ≥10 commonly used to denote serious multicollinearity. This value suggests that the variability of X3 is largely explained by a combination of other predictors, particularly **X1** and **X2**, rendering its contribution somewhat redundant. Additionally, although **X2** (VIF = 4.79) and **Z7** (VIF = 4.84) have relatively higher VIF values compared to other variables, they remain below 5 and are therefore tolerable for inclusion in the model. Based on these considerations, **X3** was removed from the model to prevent potential instability in parameter estimates due to multicollinearity. Other variables were retained as their VIF values remain within acceptable limits. The exclusion of **X3** is expected to render the model more parsimonious, stable, and capable of producing more reliable estimates. This approach also adheres to the principle of parsimony, which emphasizes retaining variables that contribute unique information to the model without introducing distortion among predictors.

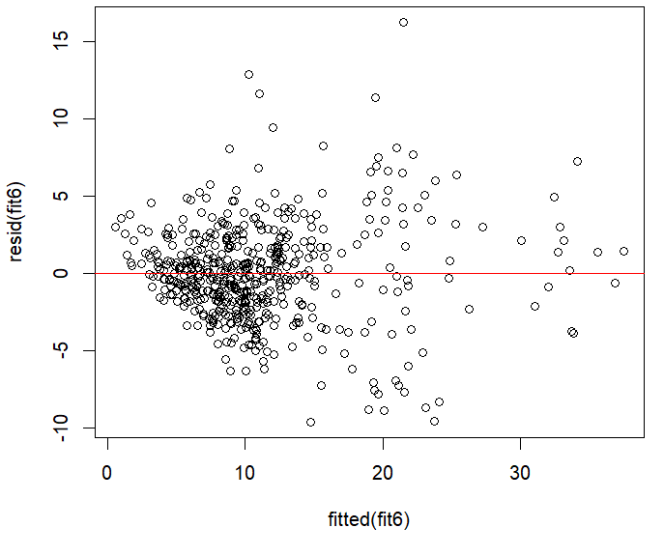


Figure 2. Residual Plot

The evaluation of classical assumptions using the Residual vs. Fitted plots in Figure 2 indicates that the residual points are generally scattered randomly around the horizontal zero line. This pattern suggests that the linearity assumption between predictor variables and the response variable is reasonably satisfied. No systematic curved patterns were observed, which would typically indicate non-linear relationships not captured by the model. Thus, overall, the model adequately represents the linear relationships among the variables involved. However, at higher fitted values, the residuals appear slightly more dispersed compared to lower fitted values. This phenomenon suggests a mild indication of heteroscedasticity, meaning that the residual variance is not entirely constant. In socio-economic data with a hierarchical or multilevel structure (e.g., individuals nested within regions or districts), residual variation tends to increase at certain levels of fitted values due to differences in characteristics across units. Therefore, the homoscedasticity assumption can still be considered acceptable, although it is important to interpret the model results while acknowledging this potential non-uniformity of variance.

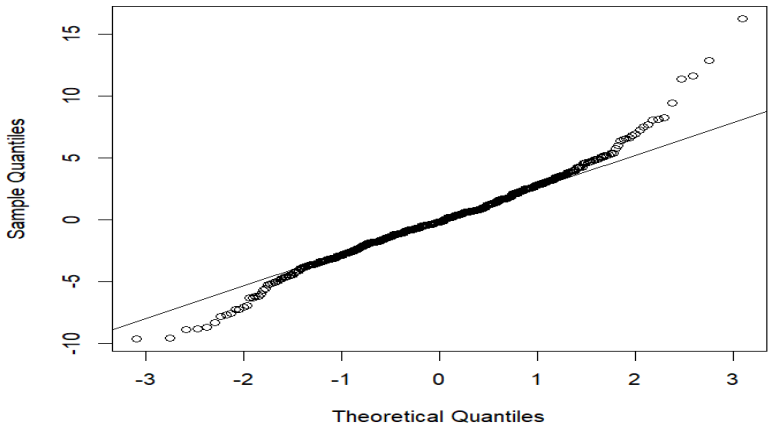


Figure 3. Normal Q-Q Plot

Subsequently in Figure 3, the evaluation using the Q-Q Plot (Plot 2) indicates that the majority of residual points are closely aligned along the diagonal line, suggesting that the residuals approximately follow a normal distribution. This condition demonstrates that the assumption of residual normality is relatively satisfied. Nevertheless, minor deviations are observed in the lower tail (left tail) and upper tail (right tail), indicating the presence of some extreme observations or a tendency toward heavy-tailed distributions. Such deviations are commonly attributable to outliers or extreme variation frequently encountered in field data, particularly in poverty studies where income or welfare distributions are often not perfectly symmetrical. Methodologically, minor violations of the residual normality assumption typically do not pose serious issues for linear mixed models, as parameter estimates remain consistent even when residuals are not perfectly normal. Therefore, the model can still be utilized for further analysis, with the caveat that the presence of outliers or heavy-tailed distributions should be carefully considered in the interpretation of results.

Advanced Tree-Based Modeling

This hybrid approach surpasses LMM in capturing localized variations within clusters and complex predictor interactions, providing richer interpretability for multilevel data (Gottard, 2023).

$$y_{ij} = x'_{ij}\beta + z'_j\gamma + \sum_{t=1}^{3}\sum_{m=1}^{M_t}\mu_{tm}\mathbb{I}\{h_{tij} \in R_{tm}\} + u_j + \varepsilon_{ij} \quad (3)$$

notation

y_{ij} is the response of individual i in cluster j

x_{ij} is the individual-level predictor (level-1)

β is the vector of fixed effects for individuals

z_j is the cluster-level predictor (level-2)

γ is the vector of fixed effects for clusters

u_j is the level-2 error term, $u_j \sim iid\ N(0, \sigma_u^2)$

ε_{ij} is the level-1 residual error $\varepsilon_{ij} \sim iid\ N(0, \sigma_\varepsilon^2)$.

h_{tij} is the set of predictor variables (features) used as input for the t-th regression tree
R_{tm} denotes the partitioned predictor space for m-th terminal node of the t-th tree
$\mathbb{I}$ is an indicator function selecting the intercept μ_{tm} for the corresponding region
μ_{tm} is the intercept corresponding to the m-th terminal node of the t-th tree, representing the local mean response for that subset of the predictor space

Model evaluation

All models were evaluated using within-cluster K-fold cross-validation to account for the hierarchical structure of the data(Fox & Monette, 2026; Yates et al., 2023). In each iteration, the data were divided into training and test sets within each cluster, with the roles of the folds rotating across iterations. Two performance metrics were used: the Mean Squared Error (MSE), calculated from predictions on the training data, and the Predictive Mean Squared Error (PMSE), calculated from predictions on the held-out test data. The MSE measures the model's fit to the training observations, whereas the PMSE evaluates its predictive accuracy on observations not used during model fitting. Comparing these metrics provides insight into both model fit and predictive performance on unseen observations. This evaluation procedure was applied consistently to the Linear Mixed Model (LMM), RE-EM Tree, Mixed Effects Regression Trees (MERT), and 3-Trees.

Table 2. MSE and PMSE value

Model	Mean_MSE	SD_MSE	Mean_PMSE	SD_PMSE
3Trees	7,46	1,02	13,97	1,7
LMM	13,69	0,97	14,62	0,97
MERT	15,3	2,79	19,38	2,7
RE-EM Tree	14,1	1,86	17,86	1,89

Based on the evaluation results, 3Trees achieved the lowest mean MSE of 7.46, with a standard deviation of 1.02. It also obtained the lowest mean PMSE of 13.97, with a standard deviation of 1.70. These results indicate that 3Trees provided the best predictive accuracy among the four models according to both evaluation metrics. LMM achieved a mean MSE of 13.69 and a mean PMSE of 14.62. Meanwhile, MERT yielded a mean MSE of 15.30 and a mean PMSE of 19.38, whereas RE-EM Tree obtained a mean MSE of 14.10 and a mean PMSE of 17.86. Thus, based on the results, 3Trees demonstrated better predictive performance than the three competing models on the analyzed dataset. The standard deviation of the 3Trees PMSE, at 1.70, indicates variation in prediction errors across repetitions. However, model stability cannot be determined solely from this standard deviation without considering differences in scale and the characteristics of the results obtained by the other models. Overall, these findings suggest that 3Trees has the potential to improve predictive accuracy on the analyzed dataset. This conclusion is limited to the data and evaluation design used in this study.

The estimation results of the 3Trees model using the full dataset indicate that the residual variance of 8.749 is substantially greater than the between-province variance of 2.406.

This suggests that most of the variation in poverty rates is associated with characteristics and conditions at the regency or city level, while differences in intercepts across provinces remain present but are relatively smaller. The standardized residuals are also reasonably symmetric around zero, suggesting that the model exhibits stable residual behavior. The global linear fixed-effects analysis shows that male life expectancy (X1) does not have a statistically significant purely linear effect on poverty rates, with an estimated coefficient of -0.0687 and a t-statistic of -0.858. The macro-level health effect of this variable appears to emerge more strongly through nonlinear relationships captured by the tree structures. Similarly, the percentage of the population aged 25 years and above with at least a senior secondary education (X2) has a negative but statistically insignificant global linear effect, with an estimated coefficient of -0.0364 and a t-statistic of -1.446.

In contrast, X3 exhibits a strong and statistically significant negative association with poverty rates, with an estimated coefficient of -0.5325 and a t-statistic of -7.697. The proportion of married women aged 15 to 49 who gave birth to their first child before the age of 20 (X4) also shows a substantial and statistically significant negative association, with an estimated coefficient of -11.4535 and a t-statistic of -5.037. Meanwhile, the percentage of women employed as professionals (X5) has a significant negative association with poverty rates, with an estimated coefficient of -0.0425 and a t-statistic of -2.288.

At the macro-level (Level-2), the percentage of underemployed individuals (Z1) has a positive and statistically significant association with poverty rates, with an estimated coefficient of 0.4233 and a t-statistic of 2.832. This finding suggests that underemployment is associated with higher poverty rates, potentially because insufficient working hours and earnings limit individuals' ability to meet their basic needs. The average per capita expenditure on consumption in rural areas (Z2) does not have a statistically significant linear effect, with an estimated coefficient of -0.0048 and a t-statistic of -0.053. In contrast, the percentage of married women using contraceptives (Z3) has a strong and statistically significant negative association with poverty rates, with an estimated coefficient of -0.4261 and a t-statistic of -7.789. The percentage of households with private toilet facilities (Z4) does not have a statistically significant linear effect, with an estimated coefficient of -0.0422 and a t-statistic of -0.451. The percentage of households living in rented or leased housing (Z5) has a statistically significant negative association with poverty rates, with an estimated coefficient of -0.2327 and a t-statistic of -2.114. The crime clearance rate (Z6) shows a negative association that approaches statistical significance, with an estimated coefficient of -0.0236 and a t-statistic of -1.741. Finally, the proportion of adolescents and adults aged 15 to 59 with information and communication technology skills (Z7) does not exhibit a statistically significant direct global linear effect, with an estimated coefficient of 0.0327 and a t-statistic of 0.745.

Integration of the Three Trees: Nonlinear Patterns and Interactions

The integration of three decision trees enriches the model by capturing nonlinear patterns and complex interactions that cannot be adequately represented by a purely linear model. The estimated coefficients associated with the terminal nodes reflect the estimated

contributions of the corresponding tree components within the combined model. Tree 1, which represents Level-1 predictors and individual-level characteristics (X), identifies six statistically significant local segments. The most critical segment is Tree1:7, characterized by X3 below 65.65 and X1 at least 68.72. This segment contributes the largest estimated increase in poverty rates, with a coefficient of 11.1343 and a t-statistic of 8.359. Conversely, Tree1:9, characterized by X3 at least 65.65, X1 below 74.28, and X4 at least 0.0975, has a comparatively smaller estimated contribution of 2.7648, with a t-statistic of 3.042. These results indicate that the contribution of individual-level characteristics varies across distinct combinations of predictor values. Tree 2, which represents Level-2 predictors and macro-level characteristics (Z), identifies different patterns of regional vulnerability. The segment Tree2:15, characterized by Z3 at least 49.35, Z7 below 80.06, and Z4 below 4.61, exhibits the largest estimated increase in poverty rates, with a coefficient of 10.9216 and a t-statistic of 7.461. Meanwhile, Tree2:9, characterized by Z3 below 49.35, Z4 below 16.02, and Z4 below 2.305, has the smallest estimated contribution to poverty among the reported segments, with a coefficient of 2.7014 and a t-statistic of 1.919. These findings suggest that macro-level conditions are associated with poverty through threshold-dependent relationships rather than uniform linear effects.

Tree 3 captures cross-level interactions between individual-level and macro-level characteristics. The interaction represented by Tree3:5, characterized by X2 below 42.7 and X4 at least 0.4305, is associated with a substantial estimated increase in poverty rates, with a coefficient of 7.0820 and a t-statistic of 5.677. Similarly, Tree3:7, characterized by X2 at least 42.7 but below 43.73, contributes an estimated 7.0557, with a t-statistic of 5.633. These results demonstrate that the contribution of a predictor may depend on its value and its interaction with other characteristics, highlighting the importance of modeling cross-level relationships in multilevel poverty analysis. Overall, the integrated 3TREES model indicates that X3, X4, and X5 at Level-1, together with Z1, Z3, and Z5 at Level-2, are statistically significant global linear predictors of poverty rates. Furthermore, the three-tree structure identifies threshold effects and localized patterns involving variables such as X1, X2, Z4, and Z7, whose global linear effects are not statistically significant. These variables may nevertheless contribute to specific decision-tree segments through nonlinear relationships and interactions. The combined approach therefore provides a more flexible and detailed interpretation of poverty patterns than a purely linear mixed model, although its predictive superiority should be established through an appropriate out-of-sample evaluation. The corresponding tree structures are presented below.

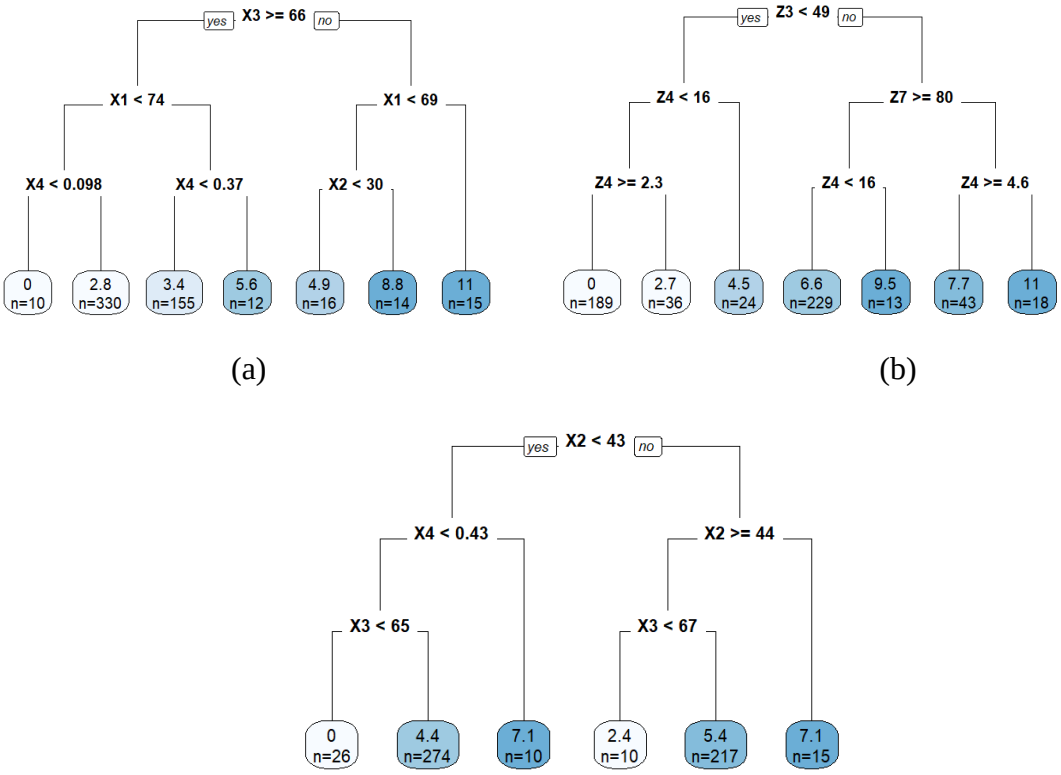


Figure 3. (a) Tree 1, (b) Tree 2, (c) Tree 3

CONCLUTION

This study employs a Linear Mixed Model (LMM) extended with decision-tree structures through the integration of the 3-Trees approach, substantially improving the prediction of poverty determinants in Indonesia. The 3-Trees model demonstrates the best predictive performance, with a Mean Squared Error (MSE) of 0.15 and a Cluster MSE of 0.70, outperforming the LMM, RE-EM, and MERT models. It also captures complex nonlinear relationships and interactions among predictors. Substantively, key factors associated with poverty reduction include health, as measured by male life expectancy, secondary and higher education, women's empowerment in professional occupations, contraceptive use, and information and communication technology skills. In contrast, higher underemployment rates are associated with increased poverty, while other variables, including rural consumption expenditure, sanitation, housing tenure, and crime clearance rates, show limited statistical significance. These findings contribute to the development of multilevel modeling and poverty analysis by demonstrating that integrating decision trees into mixed-effects models provides a more flexible and interpretable approach for capturing complex patterns that purely linear models may fail to identify.

The practical implications of this study suggest that poverty alleviation strategies should prioritize improvements in education and health, women's workforce empowerment, and digital skills development while accounting for the complex interactions among predictors identified by the 3-Trees model. These findings provide an evidence-based foundation for policymakers to design more targeted interventions addressing the principal factors associated with poverty. By considering nonlinear relationships and variations across population groups and regions, policy interventions can be better aligned with the specific characteristics of vulnerable communities. The main limitations of this study concern data consistency and quality. A particular challenge arises in Central Papua and Highland Papua, which were established through regional expansion, where several indicators were unavailable in Statistics Indonesia (BPS) publications, resulting in missing data. These missing values were imputed using the corresponding averages for the former Papua Province to maintain regional representation and data consistency. Although this approach is considered suitable for cross-provincial analysis, local variation within the two newly established provinces may not be fully captured. The dataset was also examined for duplicate observations and extreme values, including percentages exceeding 100 percent and Human Development Index values substantially outside the national range. Extreme values were verified against the original sources and subsequently corrected or removed when confirmed to result from data-entry errors. Finally, the entire dataset was validated using descriptive statistical summaries to ensure that missing values, duplicate observations, and inconsistencies across variables were appropriately addressed. Given these limitations, future research should expand the scope and size of the dataset substantially by incorporating more granular data, such as village-level statistics or household microdata from population censuses. A considerably larger sample is important for improving the stability of the regression-tree components and reducing the risk of biased local partitions caused by limited sample sizes within clusters. Future studies should also incorporate additional and more specific socioeconomic indicators to provide a more comprehensive representation of poverty determinants. These developments are expected to improve predictive accuracy and deepen the understanding of poverty determinants across different geographical levels in Indonesia while strengthening the stability and reliability of mixed-effects modeling.

ACKNOWLEDGEMENT

The authors would like to express their sincere gratitude to Universitas Mercu Buana Yogyakarta for the internal research funding that made this work possible. Special thanks are also extended to Badan Pusat Statistik (BPS) for providing access to official statistical data, which served as the primary basis for the analysis in this article. In addition, the authors are grateful to colleagues and academic staff who offered constructive feedback and encouragement throughout the research process. Their contributions have greatly enhanced the quality and clarity of this work.

REFERENCES

Bank, W. (2022). *Poverty and Shared Prosperity Report.* Washington, DC: World Bank.

Bassetti, T. C. (2018). The tree of political violence: A GMERT analysis. *Empirical Economics, 54*, 839–850. https://doi.org/10.1007/s00181-017-1293-0

Breiman, L. F. (1991). *Classification and Regression Trees.* Chapman & Hall/CRC.

Davala, M. R. (2024). Pemodelan data kemiskinan di Pulau Sumatera dengan regresi multilevel spline linear truncated. *Indonesian Journal of Applied Statistics, 7*(1), 1–13. https://doi.org/10.13057/ijas.v7i1.80768

Dea Alvionita Azka, & D. (2023). Hubungan antara indeks pembangunan manusia dengan garis kemiskinan di Sumatera Barat, Sumatera Selatan dan Bengkulu. *Basis: Jurnal Penelitian Pembangunan, 2*(1), 72–77. http://jurnal.fmipa.unmul.ac.id/index.php/Basis

Efron, B. (2020). Prediction, estimation, and attribution. Journal of the American Statistical Association, 115(530), 636–655. https://doi.org/10.1080/01621459.2020.1762613

Fox, J., & Monette, G. (2026). cv: An R package for cross-validating regression models. *Journal of Statistical Software, 116*(8), 1–41. https://doi.org/10.18637/jss.v116.i08

Gottard, A. V. (2023). Mixed-effect models with trees. *Advances in Data Analysis and Classification, 17*, 431–461. https://doi.org/10.1007/s11634-022-00509-3

Hajjem, A. L. (2011). Mixed effects regression trees for clustered data. *Statistics & Probability Letters, 81*(4), 451–462. https://www.sciencedirect.com/science/article/abs/pii/S0167715210003433

Jahangiri, M. et al. (2025). Leveraging mixed-effects regression trees for the analysis of high-dimensional longitudinal data to identify the low and high-risk subgroups: Simulation study with application to genetic study. *BioData Mining, 18*. https://doi.org/10.1186/s13040-025-00437-w

Jing, W. et al. (2023). A regression tree method for longitudinal and clustered data with multivariate responses. *Journal of Business & Economic Statistics*. https://doi.org/10.1080/00949655.2023.2273966

Rahmawati, F. N. (2023). Classification of household poverty in West Java using the generalized mixed-effects trees model. *Jurnal Natural, 23*(3), 209–220.

Ruíz, J. S., López, O. A. M., Ramírez, G. H., & Hiriart, J. C. (2023). Generalized linear mixed models for non-normal responses. In *Generalized Linear Mixed Models with Applications in Agriculture and Biology.*

Sela, R. J. (2012). RE-EM trees: A data mining approach for longitudinal and clustered data. *Machine Learning.* https://doi.org/10.1007/s10994-011-5258-3

Statistik, B. P. (2024). *Profil Kemiskinan di Indonesia Maret 2024.* Jakarta: Badan Pusat Statistik.

West, B. T., Welch, K. B., Gałecki, A. T., & Gillespie, B. W. (2022). Introduction. In *Linear Mixed Models: A Practical Guide Using Statistical Software* (3rd ed.). CRC Press. https://doi.org/10.1201/9781003181064-1

Yates, L. A., Aandahl, Z., Richards, S. A., & Brook, B. W. (2023). Cross validation for model selection: A review with examples from ecology. *Ecological Monographs, 93*(1), 1–24. https://doi.org/10.1002/ecm.1557