Comparison of GMERF and GLMM Tree Models on Poverty Household Data with Imbalanced Categories

Ari Shobri Bukhari [1,2], Khairil Anwar Notodiputro [1*], Indahwati [1], and Anwar Fitrianto [1]

[1]Department of Statistics, IPB University, Bogor, Indonesia
[2]Badan Pusat Statistik (Statistics Indonesia)
*Corresponding author: khairil@apps.ipb.ac.id

Received: 22 Oktober 2024
Revised: 6 May 2025
Accepted: 16 June 2025

ABSTRACT – **Decision tree and forest methods have become popular approaches in data science and continue to evolve. One of these developments is the combination of decision trees with Generalized Linear Mixed Models (GLMM), resulting in the GLMM Tree, which is applicable to multilevel and longitudinal data. Another model, Generalized Mixed Effect Random Forest (GMERF), extends the concept of decision forests with GLMM, effectively handling complex data structures with non-linear interactions. This study compares the performance of GLMM Tree and GMERF models in classifying poor households in South Sulawesi Province, characterized by imbalanced categories. GLMM Tree provides a simple, interpretable classification through tree diagrams, while GMERF highlights variable importance. Initial tests show all three models (GLMM, GLMM Tree, and GMERF) achieve high accuracy and specificity but exhibit low sensitivity. By applying oversampling, sensitivity and AUC are significantly improved, though this is accompanied by a decline in accuracy and specificity, revealing a trade-off. The study concludes that while GLMM, GLMM Tree and GMERF have their strengths, using them together offers a more comprehensive understanding of poverty classification. Handling imbalanced data with oversampling is effective in increasing sensitivity, but careful consideration is needed due to its impact on overall accuracy.**

Keywords– **GLMM Tree, GLMM, GMERF, Poverty Household Classification, Oversampling**

I. INTRODUCTION

Decision tree and forest methods have become popular approaches in data science and continue to evolve. Decision-tree methods are widely used in data science due to their interpretability and flexibility. These machine learning techniques produce rule-based structures that are easy to visualize, making them suitable for decision-making applications. The integration of decision trees with Generalized Linear Mixed Models (GLMM) has led to the development of GLMM Tree, a single-tree method capable of handling hierarchical and longitudinal data [1]. This approach captures variations between clusters (e.g., individuals within groups), which is particularly relevant in social and medical research [2,3]. Compared to ensemble methods, GLMM Trees are more interpretable [4], although their predictive accuracy is often lower [5].

To overcome this limitation, this study also employs the Generalized Mixed Effects Random Forest (GMERF), an ensemble model that incorporates random effects into the random forest framework [6]. By generating multiple trees and aggregating their results, GMERF enhances predictive accuracy and captures complex non-linear interactions in hierarchical data [7,8]. Empirical studies show that GMERF outperforms traditional mixed-effects models in several domains, including education [7], and provides more robust predictions, particularly in multilevel settings [9, 10].

Fokkema *et al*. [1] found that the GLMM Tree offers a balance between interpretability and accuracy, performing slightly better than random forests but slightly worse than GLMM in their study. Random forests are also known to be more resistant to overfitting, especially when the number of predictors is large or when interactions are complex [9]. Consequently, GMERF is well-suited to address these modeling challenges.

This study aims to gain a broader understanding of the application of GLMM Tree and GMERF in the social field, particularly in classifying poor and non-poor households in South Sulawesi Province. The poverty line (GK) in South Sulawesi Province is among the lowest, set at Rp. 338,997 for urban areas and Rp. 322,223 for rural areas in 2019, which is below the national poverty line (BPS.go.id). This indicates that to meet basic needs, including 2,100 kilocalories per capita per day, less money is required compared to other provinces with higher poverty lines. The percentage of poor people in urban areas is relatively low, while it is significantly higher in rural areas. This condition reflects that the economic capacity of rural residents in the province remains relatively low.

Government intervention to address this issue begins with the classification of poor and non-poor households. The classification of poor households is necessary for the implementation of various policies, such as the distribution of The Non-Cash Food Assistance (BPNT), Healthy Indonesia Card (KIS), and Family Hope Program (PKH), as outlined in Presidential Regulation No. 63 of 2017. Errors in classification can lead to misdirected policies. The risk of such errors increases when considering the typically imbalanced nature of poor/non-poor household categories. As Sun *et al*. [11] pointed out, class imbalance often becomes a challenge in classification tasks, especially when the minority class has very low representation in the data. In such conditions, classification models tend to be biased toward the majority class and

may even fail to predict the minority class at all. Most classifiers, including decision trees and neural networks, perform optimally only when the response variable is well-balanced [12].

To address this issue, the study applies an oversampling approach (ROSE) to balance the class distribution and improve predictive accuracy for the minority class [13]. Oversampling artificially enlarges the minority class, thereby improving the model's ability to recognize poor households.

State of the art research shows that despite the widespread use of decision trees and forest models, the integration of GLMM with machine learning models like GLMM Tree and GMERF remains limited in socio-economic classification, especially in household poverty data. Moreover, the impact of oversampling (ROSE) on the performance of these models has not been systematically evaluated, particularly regarding the trade-off between sensitivity and specificity. This study addresses these gaps by exploring the combined use of GLMM, GLMM Tree and GMERF, along with a systematic evaluation of oversampling in poverty classification.

The objectives of this study are to:

i. Identify the influence of independent variables in the classification process of poor households using the GMERF, GLMM Tree, and GLMM models,
ii. Compare the performance of GMERF, GLMM Tree, and GLMM in making predictions,
iii. Demonstrate the impact of imbalanced data treatment on the dataset before applying decision-tree and random forest-based methods with random effects.

II. LITERATURE REVIEW

A. Generalized Linear Mixed Model (GLMM)

The addition of random components to a Generalized Linear Model (GLM), which originally only had fixed effects, transforms it into a Generalized Linear Mixed Model (GLMM) [14]. GLMM is capable of handling various response variable distributions and can address scenarios where observations are clustered, such as in multilevel and longitudinal data. The generalized linear mixed-effects model (GLMM) can be expressed with the following equation [15]:

$$\boldsymbol{\mu_i} = \boldsymbol{E}[\boldsymbol{Y_i}|\boldsymbol{b_i}] \quad \boldsymbol{i} = \boldsymbol{1}, \dots, \boldsymbol{l} \tag{1}$$
$$g(\boldsymbol{\mu}_i) = \eta_i$$
$$\eta_i = X_i\boldsymbol{\beta} + Z_i\boldsymbol{b}_i$$
$$\boldsymbol{b}_i \sim N_q(0, \boldsymbol{\Psi})\ ind$$

where:

i : group index, I is the total number of groups
n_i : number of observations in group i and $\sum_{i=n}^{I} n_i = j$
η_i : n_i- -dimensional linear predictor vector
X_i : $n_i \times (p+1)$ matrix of fixed-effect regressors for observations in group i
$\boldsymbol{\beta}$: $(p+1)$-dimensional vector of fixed-effect coefficients,
Z_i : $n_i \times q$ matrix of random-effect regressors,
$\boldsymbol{b}_i$: $(q+1)$ dimensional vector of random-effect coefficients
$\boldsymbol{\Psi}$: $q \times q$ within-group covariance matrix of random effects.

Fixed effects are identified by parameters that apply to the entire population, while random effects are identified by parameters specific to individual groups.

B. Generalized Linear Mixed Model Tree (GLMM Tree)

In the GLMM Tree framework, the fixed effects β_j are modeled as local parameters, meaning their values vary across terminal nodes of the tree, while the random effects remain global, shared across all groups in the data. To estimate this model, the fixed-effect component is replaced by a GLM Tree, where constant fits are assigned to terminal nodes, while random effects are estimated similarly to a standard GLMM [16].

The estimation process follows an EM-like iterative procedure, consisting of the following steps::

Step 0: Set the initial value of r and all values of $\hat{b}_{(r)}$ to 0.
Step 1: Set $r = r + 1$. Estimate the GLM Tree using $z_i^T \hat{b}(r-1)$ as the offset value.
Step 2: Fit a mixed-effects model

$$g(\boldsymbol{\mu}_{ij}) = X_i^T\boldsymbol{\beta}_j + z_i^T\boldsymbol{b} \tag{2}$$

with terminal nodes $j(r)$ from the GLM Tree estimated in Step 1. Calculate the posterior prediction $\hat{b}_{(r)}$ based on the estimated model.
Step 3: Repeat Steps 1 and 2 until convergence.

In this algorithm, random effects are estimated globally across all clusters and are not part of the partitioning process. Only the fixed effects are estimated locally within each terminal node. This structure allows GLMM Trees to combine the interpretability of decision trees with the flexibility of mixed models in handling clustered or hierarchical data [17,18].

C. Generalized Mixed Effect Random Forest (GMERF)

The GMERF algorithm, according to Pellagatti *et al.* [7] , integrates the strengths of random forests and generalized linear mixed models (GLMM) to handle hierarchical or clustered data. It estimates model parameters through the following key steps:

i. Initialization: The response vector (y) and covariates (cov) are used as inputs, along with the group variable (gr) as random effects. The covariates are initialized as both fixed and random effects.
ii. Initial Model Fitting: A Generalized Linear Model (GLM) is employed to predict the initial values of η_{ij} which represents the linear component of the model..
iii. Iteration:
   - At each iteration, the algorithm computes the target ($\eta - Z \times b$), where b denotes the random effect.
   - A random forest (RF) model is applied to predict the fixed effects $f(X)$.
   - The Generalized Linear Mixed Model (GLMM) is then estimated to obtain the random effects b_i using the results from the RF model and GLMM.
iv. Convergence: The iterative process continues until the parameter estimates stabilize or until the maximum number of iterations is reached.
v. Final Output: The final model produces the predictions for $\hat{\eta}_{ij} = f(X_{ij}) + Z_{ij}b_i$, and the response predictions are obtained by applying the inverse link function according to the distribution of the data.

This hybrid algorithm combines the nonparametric flexibility of random forests for modeling fixed effects with the structured estimation of random effects from GLMM, making it well-suited for complex hierarchical data with non-linear relationships [6,3,18,7].

D. Model Performance Evaluation

One of the standard methods for evaluating the performance of a classification system is the confusion matrix, which presents the number of correct and incorrect predictions compared to the actual class labels. It provides detailed insight into the types of classification errors made by the model. The structure of a binary classification confusion matrix is shown in Table 1.

Table 1 Confusion Matrix

Predicted Class	Actual Class	
	Class = 0	Class = 1
Class = 0	A	B
Class = 1	C	D

Accuracy in classification refers to the percentage of correctly classified data after testing the classification results. The higher the accuracy level, the more effective the classification model is considered to be [19]. The values for accuracy, sensitivity, and specificity can be calculated using the following formulas:

$$Accuracy = \frac{A+D}{A+B+C+D}; Sensitivity = \frac{A}{A+C}; \quad (3)$$
$$Specificity = \frac{D}{B+D}$$

The information in the confusion matrix can also be used to construct a Receiver Operator Characteristic (ROC) curve. The ROC curve is a probability curve that plots the True Positive Rate (TPR) against the False Positive Rate (FPR) at various threshold values, essentially separating 'signal' from 'noise'. Additionally, the Area Under the Curve (AUC) is a measure of the classifier's ability to distinguish between response variable categories and serves as a summary of the ROC curve.

$$TPR\ (True\ Positif\ Rate) = Sensitivity\ ; \quad (4)$$
$$FPR\ (False\ Positif\ Rate) = \frac{B}{B\ +\ D} = 1 - Specificity$$

The higher the AUC value for a classifier, the better its ability to distinguish between positive and negative classes [20,21]. When AUC = 1, the classifier perfectly distinguishes the classes. However, if AUC = 0, the classifier predicts all negatives as positives and all positives as negatives. An AUC = 0.5 indicates that the classifier cannot distinguish between positive and negative classes, implying that it predicts all data points randomly.

E. Handling Imbalanced Data

One of the most widely used approaches to handle imbalanced datasets is oversampling, in which synthetic data points are generated to augment the minority class. In this study, the Random Over-Sampling Examples (ROSE) method is selected over other oversampling techniques for several important reasons.

ROSE employs a smoothed bootstrap technique, generating synthetic examples by drawing from a kernel-based estimate of the predictor distribution conditional on the class label [22]. This allows ROSE to produce more diverse and realistic synthetic data that reflect the underlying structure of the minority class, even when the predictor variables are a mix of categorical and continuous types.

Compared to other popular techniques like SMOTE (Synthetic Minority Over-sampling Technique) and ADASYN (Adaptive Synthetic Sampling), ROSE offers greater flexibility and distributional realism. SMOTE generates new samples by interpolating between nearest-neighbor minority instances [23], while ADASYN focuses on generating more samples in harder-to-classify regions by weighting minority instances with higher classification difficulty [24]. However, both SMOTE and ADASYN are generally limited to numerical predictors, as interpolation is not well-defined for categorical data. ROSE, on the other hand, can accommodate mixed-type features, making it especially suitable for socio-economic

survey data.

In the context of this study—classifying poor and non-poor households using poverty survey data—ROSE was chosen due to the heterogeneous nature of the predictors, which include socio-economic and demographic variables with both numerical and categorical formats. ROSE provides greater flexibility in synthesizing realistic minority class samples without relying solely on instance pairings, and it helps the model learn more balanced patterns by preserving the overall distribution of the data.

Nonetheless, as with other oversampling techniques, ROSE is not without limitations. Since the generated samples are synthetic, they may introduce distributional distortion or overfitting if not validated properly. To mitigate this, the study employs cross-validation and evaluates performance on naturally imbalanced test data [25].

III. METHODOLOGY

A. Data

This study utilizes household sample data from the March 2019 National Socioeconomic Survey (Susenas) conducted by Statistics Indonesia (BPS) for the South Sulawesi Province. Households are classified as poor or non-poor by comparing their monthly per capita expenditure with the poverty line (PL) applicable to each district or city [26]. A household is categorized as poor if its per capita expenditure falls below the PL of its respective district.

The classification model uses the household's poverty status as the response variable, one random variable, and twelve candidate fixed-variable predictors. The district or city is modeled as a random variable to account for unobserved heterogeneity, given that households within the same area often share similar socio-economic characteristics due to local policy, environment, and access to public services. This approach aligns with the structure of multilevel or hierarchical data modeling [27].

The twelve candidate fixed-effect variables include characteristics related to housing conditions and the household head. These are termed "candidate variables" because not all may be included in the final model—some may be excluded based on statistical insignificance, model fit criteria, or convergence issues during estimation.

Table 2 Research Variables

Code	Variable (and Reference)	Scale	Factor
Y	Household Classification (Poor/Non-Poor)	Nominal	-
V	District/City	Nominal	Random
X1	Number of families living in the census building/house	Ratio	Fixed
X2	Ownership status of the occupied residence	Nominal	Fixed
X3	Floor area of the residential building exist	Ratio	Fixed
X4	Main material of the largest roof area	Nominal	Fixed
X5	Main material of the largest floor area	Nominal	Fixed
X6	Main water source used for drinking	Nominal	Fixed
X7	Main lighting source for the household	Nominal	Fixed
X8	Main fuel type used for cooking	Nominal	Fixed
X9	Land ownership	Nominal	Fixed
X10	Number of household member	Ratio	Fixed
X11	Highest diploma/certificate held by the household head	Ordinal	Fixed
X12	Type of occupation/industry sector	Nominal	Fixed

B. Data Analysis Procedures

The data analysis follows a structured series of steps comprising data preparation, modeling, and evaluation:

1) Data preprocessing, including household poverty classification based on per capita expenditure relative to the district/city poverty line, followed by data merging and case selection.
2) Descriptive analysis to provide a brief overview of poverty conditions in South Sulawesi Province.
3) Variable selection to identify relevant predictors through theoretical consideration and diagnostic checks (e.g., multicollinearity, convergence, and significance).
4) Model development using GMERF, GLMM Tree, and GLMM based on selected variables.
5) Data partitioning into 80% training and 20% testing sets using stratified sampling to preserve class distribution.
6) Prediction on the test set using the models trained in step 4.
7) Model evaluation based on accuracy, sensitivity, specificity, and AUC [20,21].
8) Performance comparison across the three models using the test results.
9) Class balancing by applying the ROSE oversampling method to the training set [22].
10) Re-evaluation by repeating steps 5 to 8 with the resampled training data.
11) Final comparison of model performance on datasets with and without oversampling.

IV. RESULTS AND DISCUSSIONS

A. Overview of Poverty Rate in South Sulawesi Province

Based on Figure 1, it can be observed that all regencies and cities in South Sulawesi Province have a poverty rate hovering around 10%. The lowest poverty rate is found in Makassar City at 4.28%, while the highest is in Jeneponto Regency at 14.88%. The average poverty rate across the province is 9.43%.

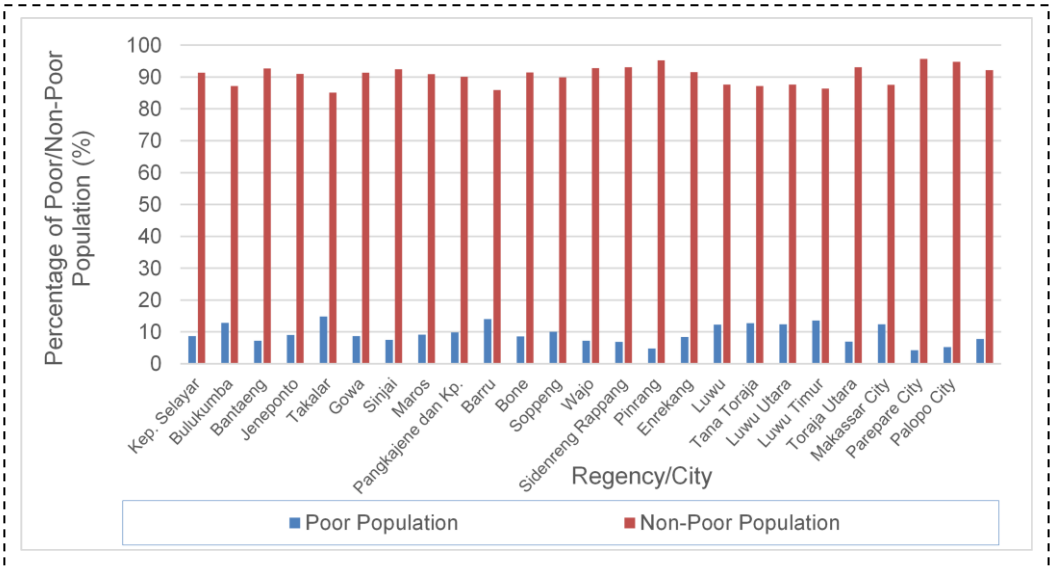

Figure 1 Percentage of Poor and Non-Poor Population by District in South Sulawesi Province, 2019

This distribution indicates a clear class imbalance between poor and non-poor households, which becomes a critical issue when developing classification models to identify poverty status based on socio-economic predictors. A dominant majority class (non-poor) may bias the learning process of many classification algorithms, potentially leading to poor detection performance for the minority class (poor households). Therefore, data imbalance handling is essential to improve model performance, particularly in terms of sensitivity or recall for the minority class [13,23].

In this study, the imbalance issue is addressed through an oversampling technique, aiming to enhance the classifier's ability to recognize poor households while maintaining an acceptable trade-off with overall model accuracy.

B. Independent Variable Selection

The data processing was conducted using the R programming language, employing the core functions glmerTree, glmer, and gmerf. Notably, the gmerf function is not currently distributed as a formal R package but was implemented using custom source code.

Initially, all independent variables were included in the GLMM, GLMM Tree, and GMERF models. However, none of these models produced convergent estimates, likely due to the inclusion of irrelevant or collinear predictors, as well as the complexity of interactions among variables. To address this issue, a variable selection process was performed. The first step involved fitting a Generalized Linear Model (GLM) to identify statistically significant predictors. This step serves as an efficient screening mechanism before fitting more complex models [28]. Variables found to be significant in the GLM were subsequently evaluated within the GLMM framework to confirm their contribution in the presence of random effects.

To further refine the model, Generalized Variance Inflation Factor (GVIF) diagnostics were applied to assess multicollinearity among the selected predictors. GVIF is an extension of the standard VIF that is suitable for categorical variables with multiple levels [29]. Only predictors with acceptable GVIF thresholds (typically GVIF < 5) were retained. As a result, the final set of independent variables used across all models included: the floor area of the residential building (X3), the main source of household lighting (X7), land ownership (X9), the number of household members (X10), the highest diploma/certificate held by the household head (X11), and the type of employment/industry of the household head (X12)..

C. GLMM Modeling

The initial Generalized Linear Mixed Model (GLMM) included all candidate predictors but resulted in a non-positive definite Hessian matrix, indicating instability in parameter estimation and potential multicollinearity or model overfitting [30]. To address this, non-significant variables were removed, and the model was refitted using only the significant predictors identified in the previous steps. The estimation results, presented in Table 3, show that the final model includes six significant predictors: X3 (floor area of the residence), X7 (main lighting source), X9 (land ownership), X10 (household size), X11 (education level of household head), and X12 (employment sector).

Table 3 Estimation of Parameters in the GLMM

Variable*	Estimate β	Std. Error	z value	Sig.	$Exp(\beta)$	$1/Exp(\beta)$
(Intercept)	2.253	0.209	10.770	< 2e-16	9.519	0.105
floor_area	0.842	0.065	12.984	< 2e-16	2.322	0.431
lighting2	-0.611	0.105	-5.798	6.70E-09	0.543	1.842
lighting3	-1.029	0.149	-6.914	4.71E-12	0.357	2.800
lighting4	-1.112	0.222	-5.004	5.61E-07	0.329	3.039
land_ownership5	-0.191	0.100	-1.909	0.0563	0.826	1.210
household_members	-0.781	0.033	-23.648	< 2e-16	0.458	2.185
education1	0.581	0.108	5.401	6.63E-08	1.789	0.559
education2	0.701	0.110	6.376	1.82E-10	2.017	0.496
education3	0.791	0.132	5.974	2.32E-09	2.206	0.453
education4	1.269	0.136	9.336	< 2e-16	3.556	0.281
education5	2.413	0.313	7.722	1.14E-14	11.170	0.090
employment_field1	0.177	0.097	1.827	0.0677	1.194	0.838
employment_field2	0.688	0.110	6.241	4.35E-10	1.990	0.502

Note: The main source of lighting consists of lighting1 (baseline) metered PLN electricity, lighting2 unmetered PLN electricity, lighting3 non-PLN electricity, and lighting4 non-electricity. Education consists of education0 (baseline) never attended school, education1 did not complete elementary school, education2 completed elementary school or equivalent, education3 completed junior high school or equivalent, education4 completed senior high school or equivalent, education5 college/university.

From the model output, it can be concluded that all selected predictors have statistically significant effects, with many variables exhibiting strong associations with poverty status. For categorical variables, the Exp(β) values facilitate interpretation in terms of odds ratios. For instance, households headed by individuals with a university-level education (education5) are 11.17 times more likely to be classified as non-poor compared to those who never attended school. Conversely, households using non-electric lighting (lighting4) are 3.04 times more likely to be poor compared to those using metered PLN electricity (lighting1). These findings underscore the role of educational attainment and infrastructure access in poverty classification.

D. GLMM Tree Modeling

During the initial modeling phase, the glmertree function failed to produce convergent results. This convergence issue was likely due to an overly complex tree structure, where the default splitting criteria (e.g., low minsplit and minbucket) led to terminal nodes with insufficient sample sizes. In such cases, the estimation of random effects within very small subsets becomes unstable or even undefined, especially when combined with non-linear link functions and hierarchical data structures. To address this issue, the optimization and tree-building settings were adjusted. Specifically, the optimizer was changed from the default (Nelder_Mead) to bobyqa, which is more robust for complex likelihood landscapes and has been recommended for mixed models in the lme4 package [30]. Additionally, the maxfun parameter was increased to 30,000 iterations, allowing the optimizer more opportunity to converge.

Moreover, the tree complexity was constrained by setting minsplit = 800, minbucket = 400, and maxdepth = 5. These adjustments limited the growth of the tree to ensure that each split occurred only when sufficient data were available, and that each terminal node retained enough observations for reliable estimation. These changes reduced model variance and improved convergence stability [16]. After implementing these changes, the model successfully converged, indicating that convergence issues in GLMM Trees can often be mitigated through careful tuning of both the optimizer parameters and tree growth constraints.

As illustrated in Figure 2, several terminal nodes represent groups with a notably higher proportion of poor households. In particular, nodes 14 and 6 exhibit poverty rates approaching 20%, substantially higher than other nodes. Node 14 is characterized by households with more than four members and household heads with junior high school education or less. Similarly, node 6 represents households with fewer than four members, residential floor area below 70 m², low education, and no access to metered electricity. A common thread between these nodes is low educational attainment, suggesting that this variable is a strong indicator of household poverty.

An insightful finding from this analysis is that the splitting variables in the GLMM Tree—which determine partitioning structure—correspond to the significant predictors in the GLMM. This alignment reinforces the validity of both modeling approaches and suggests that GLMM and GLMM Tree can be integrated to enrich interpretation, offering both inferential strength and intuitive subgroup analysis [16].

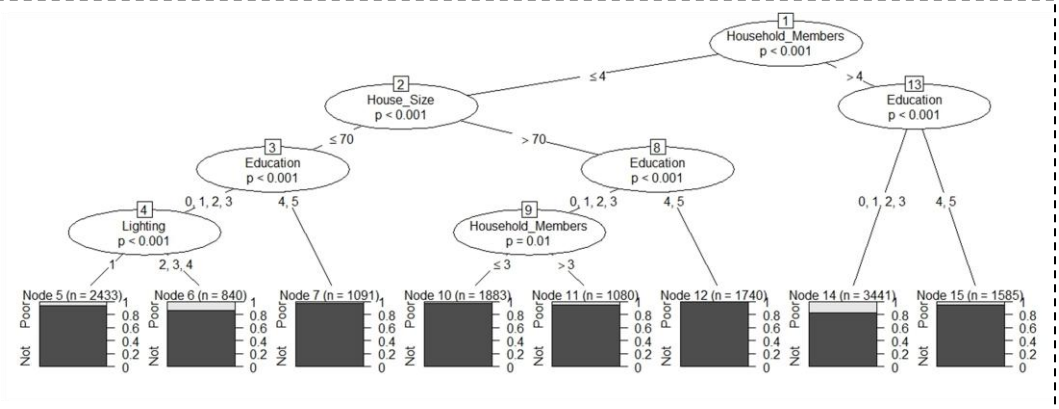

Figure 2 GLMM Tree for Classifying Poor/Non-Poor Households in South Sulawesi Province, 2019

Note.: The **lighting** is coded as (1) metered PLN electricity, (2) unmetered PLN electricity, (3) non-PLN electricity, and (4) non-electricity. **Education** is coded as (0) never attended school, (1) did not complete elementary school, (2) completed elementary school or equivalent, (3) completed junior high school or equivalent, (4) completed senior high school or equivalent, (5) college/university.

E. Modeling with GMERF

One of the distinctive outputs of the GMERF (Generalized Mixed Effect Random Forest) model is the variable importance metric, which is quantified using Increase in Node Purity (IncNodePurity). This metric indicates how much each variable contributes to reducing heterogeneity (impurity) in the model's terminal nodes. Variables with higher IncNodePurity values are considered more influential in predicting the response [31].

As shown in Table 4, the education level of the household head had the highest IncNodePurity value (4,770.82), suggesting that it is the most important predictor of household poverty in this model. This result is consistent with previous research, which shows that lower educational attainment is strongly associated with poverty risk due to its impact on employment opportunities and income-generating potential [32]. Households led by individuals with limited education often face structural barriers in accessing well-paying or formal sector jobs.

Table 4 Variable Importance in the GMERF Model

Variable	Increase in Node Purity
Floor Area of the Residential Building	4513.11
Main Source of Household Lighting	1535.28
Number of Household Members	3032.93
Highest Diploma/Certificate Held by the Household Head	4770.82
Type of Employment/Industry	1718.33

The floor area of the residential building also emerged as a strong distinguishing factor (4,513.11), reflecting material living conditions and indirectly capturing household wealth. Other important variables include the number of household members, type of employment, and main lighting source, all of which contribute meaningfully to classifying household poverty status within the GMERF framework.

F. Model Comparison on Imbalanced Data (before data treatment)

This section presents a comparative analysis of the predictive performance of the three models—GLMM, GMERF, and GLMM Tree—when applied to the original dataset with an imbalanced distribution between poor and non-poor households.

Initial modeling on the full dataset shows that the GLMM model yields the highest AUC value (0.5392), indicating a marginally better ability to distinguish between classes. However, all models exhibit poor classification performance for the minority class. To statistically evaluate the robustness of these models, a repeated random sub-sampling validation was conducted: 80% of the data were randomly selected as training sets and the remaining 20% as test sets, repeated over 30 iterations. For each iteration, accuracy, sensitivity, specificity, and AUC were computed, resulting in 30 observations per metric per model.

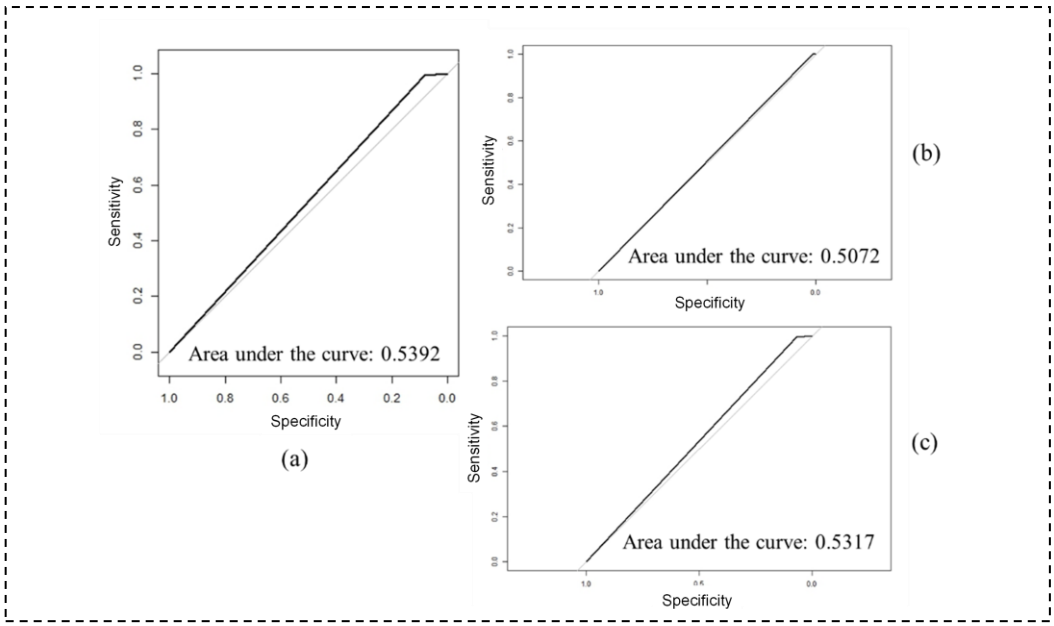

Figure 3 Comparison of ROC (Receiver Operating Characteristic) Curves between the Models (a) GLMM, (b) GMERF, and (c) GLMM Tree

The summary results show that all three models consistently produce high average accuracy (~0.92) and specificity (~0.99). However, sensitivity remains alarmingly low (< 0.1) across all models. This means that the models fail to correctly identify more than 90% of households actually classified as poor, instead misclassifying them as non-poor. Furthermore, the average AUC values hover around 0.5, indicating that the models are no better than random guessing in distinguishing between classes [20]. This confirms a key theoretical issue in imbalanced data classification: standard classifiers tend to favor the majority class, leading to misleading accuracy and specificity values that mask poor recall on the minority class [13].

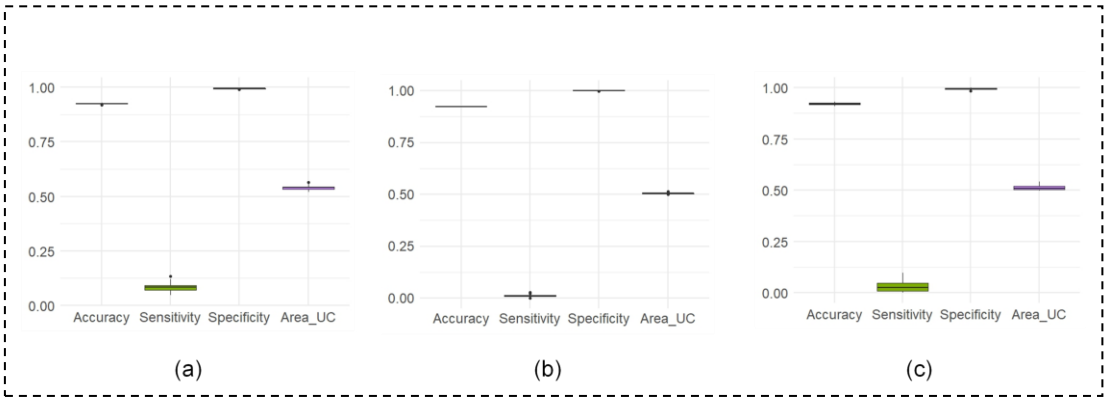

Figure 4 Boxplot diagram of accuracy, sensitivity, specificity, and AUC distribution on imbalanced data using the models (a) GLMM, (b) GMERF, and (c) GLMM Tree

To statistically assess model differences, an ANOVA test was conducted for each performance metric, and significant differences were observed in accuracy, sensitivity, specificity, and AUC across the three models. Follow-up Tukey post-hoc tests ($\alpha = 0.05$) revealed the nature of these differences, summarized in Table 5. The same column indicates that there is no significant difference based on the Tukey test.

Table 5 Results of the Tukey Mean Difference Test (alpha = 0.05)

Measure	Model	Mean and Comparison Order			Conclusion
		1	2	3	
Accuracy	GLMM		0.92479		GLMM Tree accuracy is significantly lower than the other models
	GMERF		0.92403		
	GLMMTree	0.92133			

Measure	Model	Mean and Comparison Order			Conclusion
		1	2	3	
Sensitivity	GLMM			0.08426	Sensitivity of all three models differs significantly, with GLMM the highest
	GMERF	0.01219			
	GLMMTree		0.03169		
Specificity	GLMM	0.99459			GMERF specificity is significantly higher than the other models
	GMERF		0.99976		
	GLMMTree	0.99459			
AUC	GLMM			0.53943	AUC differs significantly across all three models, with GLMM being the highest
	GMERF	0.50597			
	GLMMTree		0.51314		

From the Tukey test results, it can be concluded that:

- GLMM Tree has the lowest accuracy, likely due to over-partitioning and sensitivity to small node sizes in imbalanced datasets.
- GMERF achieves the highest specificity, suggesting it is more conservative in predicting poor households, which can be attributed to its ensemble-based averaging nature.
- GLMM demonstrates the highest sensitivity and AUC, making it the most reliable model for identifying poor households in an imbalanced context.

These findings align with existing literature which suggests that GLMMs perform well when the model structure captures key hierarchical effects, and that ensemble models such as GMERF may prioritize specificity due to aggregation effects [16]. However, the overall low sensitivity and AUC values across all models highlight a critical limitation: standard modeling approaches, without addressing class imbalance, fail to capture the minority class effectively. This reinforces the necessity of applying data balancing strategies—such as oversampling—to improve model fairness and minority-class prediction performance.

G. Model Comparison on Balanced Synthetic-Data (after data treatment)

To address class imbalance, the ROSE (Random Over-Sampling Examples) technique was applied to the training data. The resulting synthetic dataset was adjusted to reflect a 30:70 ratio between poor and non-poor households, effectively increasing the representation of the minority class by a factor of three. All models—GLMM, GMERF, and GLMM Tree—were then trained on this adjusted dataset, while predictions were evaluated using the untouched, naturally imbalanced test data.

As in the previous analysis, GLMM Tree initially failed to converge, which is consistent with known challenges in fitting complex mixed-effect models to small or unevenly distributed data subsets. Convergence was achieved only after adjusting key model parameters—specifically by increasing the minsplit, minbucket, and maxdepth values, and by using the bobyqa optimizer (see Section D). Possible causes of convergence failure include data structures that are not easily separable using linear boundaries, difficulties in estimating random effects in sparse cells [16], nd the optimizer's sensitivity to convergence tolerances or iteration limits [30].

As shown in Figure 5, the oversampling approach led to marked improvements in model sensitivity, with values increasing from below 0.1 to over 0.4 across all models. AUC values also improved significantly, rising from approximately 0.5 (before treatment) to above 0.6, indicating enhanced discriminatory ability. However, these improvements came with notable trade-offs: accuracy dropped from above 0.92 to around 0.88, and specificity declined from approximately 0.99 to 0.89.

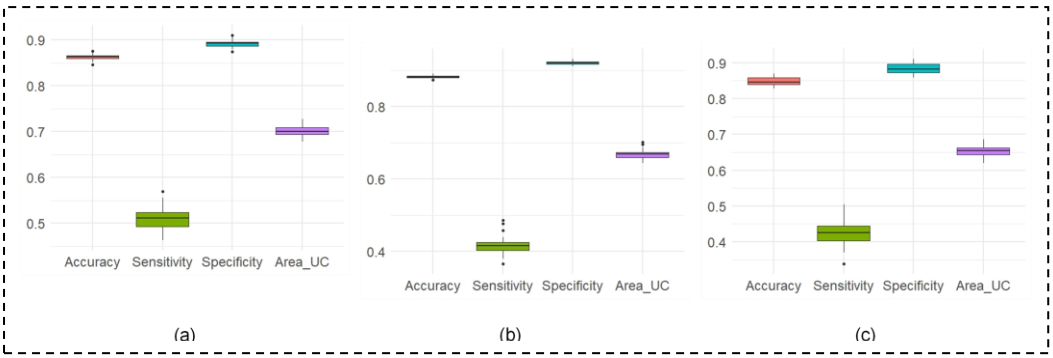

Figure 5 Boxplot diagram of the distribution of accuracy, sensitivity, specificity, and AUC values on balanced-synthetic data (after data treatment) using the models (a) GLMM Tree and (b) GLMM

These findings reflect the well-documented trade-off in imbalanced learning: as sensitivity improves due to better detection of minority class instances, the model tends to misclassify more majority class cases as positive, thereby reducing specificity and overall accuracy [11].

To statistically test performance differences across models, ANOVA followed by Tukey's HSD post-hoc tests ($\alpha = 0.05$) were conducted on the 30 repeated iterations. The results are summarized in Table 6. The same column indicates that there is no significant difference based on the Tukey test.

Table 6 Results of the Tukey Mean Difference Test (alpha = 0.05)

Measure	Model	Mean and Comparison Order			Conclusion
		1	2	3	
Accuracy	GLMM		0.8630		GMERF accuracy is significantly higher than the other models
	GMERF			0.8815	
	GLMMTree	0.8493			
Sensitivity	GLMM		0.5110		Sensitivity of GLMM is the highest
	GMERF	0.4196			
	GLMMTree	0.4230			
Specificity	GLMM		0.8922		GMERF specificity is significantly higher than the other models
	GMERF			0.9198	
	GLMMTree	0.8847			
AUC	GLMM			0.7016	AUC differs significantly across all three models, with GLMM being the highest
	GMERF		0.6697		
	GLMMTree	0.6538			

Based on these results, we conclude the following::

- GMERF yields the highest accuracy and specificity, reflecting its conservative nature in classifying positive (poor) cases—likely a result of its ensemble averaging mechanism.
- GLMM achieves the highest sensitivity and AUC, indicating superior performance in detecting poor households despite a higher rate of false positives.
- GLMM Tree performs lowest across most metrics, suggesting it is less robust—particularly when trained on synthetic data.

These results are consistent with prior literature, which suggests that GLMMs offer greater flexibility in modeling correlated hierarchical structures, whereas ensemble-based models such as GMERF tend to provide more stable and generalizable predictions[33, 16].

H. Oversampling Implications and Mitigation Strategies

Oversampling methods such as ROSE are effective in improving sensitivity on imbalanced datasets; however, they inherently modify the training data distribution and may introduce bias, overfitting, and distorted feature relationships [13, 22]. Specifically, ROSE employs smoothed bootstrap sampling to generate synthetic minority-class instances, which can deviate from the true joint distribution of predictors [25]. These alterations pose three major risks: (i) model overfitting to synthetic patterns, (ii) reduced generalizability to real-world data, and (iii) distortion of inter-feature correlations.

To mitigate these risks, this study implemented several safeguards. First, validation isolation was applied: oversampling was restricted to the training set, and model evaluation was conducted on a separate, unaltered test set to avoid contamination. Second, a repeated sub-sampling validation procedure (30 iterations) was used to reduce variance and sampling bias. Third, multi-metric evaluation—including accuracy, sensitivity, specificity, and AUC—was adopted to capture the nuanced trade-offs in model performance.

The results show that oversampling substantially improved sensitivity and AUC, particularly for GLMM. However, this came at the cost of reduced specificity and overall accuracy—highlighting a well-known trade-off in imbalanced learning [34]. Additionally, since this study did not apply probability calibration or class reweighting, the predicted probabilities should be interpreted cautiously, particularly in policy contexts.

Importantly, these safeguards were designed to ensure that enhanced detection of the poor (minority class) through synthetic oversampling does not lead to misleading conclusions or excessive compromise in generalizability. This is especially critical in poverty classification tasks, where false positives could result in misallocation of limited policy resources, while false negatives may lead to the exclusion of those most in need.

As a key limitation, this study did not incorporate post-hoc calibration or cost-sensitive learning strategies. Future research should explore probability calibration techniques (e.g., Platt scaling or isotonic regression), cost-sensitive loss functions, and hybrid resampling methods to achieve more balanced, robust, and policy-relevant model performance [33, 35].

V. CONCLUSIONS AND SUGGESTIONS

This study evaluated the performance of three classification models—GLMM, GLMM Tree, and GMERF—for identifying poor households, with a particular focus on addressing class imbalance in poverty data from South Sulawesi Province. The integration of machine learning approaches with Generalized Linear Mixed Models (i.e., GLMM Tree and GMERF) did not consistently outperform the conventional GLMM model in terms of classification metrics, under both imbalanced and balanced data conditions. Nonetheless, each model offered unique interpretative advantages that complement one another.

The GLMM Tree model provides an intuitive and interpretable tree structure that visually highlights key variables distinguishing poor households. In contrast, the GLMM model quantifies the statistical significance and effect size of independent variables, making it particularly well-suited for inferential analysis. GMERF, while also providing variable importance measures, demonstrated superior performance to GLMM Tree in several settings, particularly in terms of predictive accuracy and sensitivity after resampling.

Across all three models, the identified key predictors were generally consistent, indicating robustness in variable selection. Accordingly, employing these models in a complementary manner may enhance interpretability, providing both rigorous statistical inference and visually guided exploratory insights.

On the original imbalanced dataset, all models achieved high overall accuracy and specificity (exceeding 90%), but sensitivity remained critically low (below 10%), with area under the ROC curve (AUC) values approaching 0.5—reflecting limited effectiveness in detecting the minority class. Tukey's post-hoc tests confirmed statistically significant differences among the models, with GLMM yielding the highest sensitivity and AUC.

To mitigate the effects of class imbalance, the ROSE oversampling technique was applied. This led to a substantial improvement in sensitivity across all models, although it also resulted in marked reductions in both accuracy and specificity—highlighting the inherent trade-off in imbalanced learning scenarios. On the resampled data, GMERF achieved better accuracy and sensitivity than GLMM, suggesting its potential advantage under balanced conditions.

Future research should further investigate the convergence properties of the GLMM Tree model and extend its application to datasets characterized by nonlinear relationships to fully leverage the potential of machine learning-based approaches. Additionally, comprehensive evaluation of model stability, interpretability, and integration within the broader framework of mixed-effects modeling is recommended to inform policy-relevant decision-making and advance methodological development.

ACKNOWLEDGEMET

This research is funded by the Directorate General of Higher Education, Research, and Technology Ministry of Education, Culture, Research, and Technology in accordance research program implementation contract No: 027/E5/PG.02.00.PL/2024.

REFERENCES

[1] M. Fokkema, J. Edbrooke-Childs, and M. Wolpert, "Generalized linear mixed-model (GLMM) trees: A flexible decision-tree method for multilevel and longitudinal data," *Psychother. Res.*, vol. 31, no. 3, pp. 1–13, 2020, doi: 10.1080/10503307.2020.1785037.

[2] M. R. Segal, "Machine Learning Benchmarks and Random Forest Regression," *Biostatistics*, no. May, pp. 1–14, 2004, [Online]. Available: http://escholarship.org/uc/item/35x3v9t4.pdf

[3] G. Verbeke, G., & Molenberghs, "Linear Mixed Models for Longitudinal Data," *Linear Mixed Models for Longitudinal Data*. 2000. doi: 10.1007/b98969.

[4] J. R. Quinlan, "Induction of decision trees," *Mach. Learn.*, vol. 1, no. 1, pp. 81–106, 1986, doi: 10.1007/bf00116251.

[5] K. Topuz, A. Bajaj, and I. Abdulrashid, "Interpretable Machine Learning," *Proc. Annu. Hawaii Int. Conf. Syst. Sci.*, vol. 2023-Janua, pp. 1236–1237, 2023, doi: 10.1201/9780367816377-16.

[6] L. Breiman, "Random Forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.

[7] M. Pellagatti, C. Masci, F. Ieva, and A. M. Paganoni, "Generalized mixed-effects random forest: A flexible approach to predict university student dropout," *Stat. Anal. Data Min.*, vol. 14, no. 3, pp. 241–257, 2021, doi: 10.1002/sam.11505.

[8] D. R. Cutler *et al.*, "Random forests for classification in ecology," *Ecology*, vol. 88, no. 11, pp. 2783–2792, 2007, doi: 10.1890/07-0539.1.

[9] G. Biau and E. Scornet, "A random forest guided tour," *TEST*, vol. 25, no. 2, pp. 197–227, 2016, doi: 10.1007/s11749-016-0481-7.

[10] T. Hothorn, K. Hornik, and A. Zeileis, "Unbiased recursive partitioning: A conditional inference framework," *J. Comput. Graph. Stat.*, vol. 15, no. 3, pp. 651–674, 2006, doi: 10.1198/106186006X133933.

[11] Y. Sun, A. K. C. Wong, and M. S. Kamel, "Classification of imbalanced data: A review," *Int. J. Pattern Recognit. Artif. Intell.*, vol. 23, no. 4, pp. 687–719, 2009, doi: 10.1142/S0218001409007326.

[12] P. Kumar, R. Bhatnagar, K. Gaur, and A. Bhatnagar, "Classification of Imbalanced Data:Review of Methods and Applications," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 1099, no. 1, p. 012077, 2021, doi: 10.1088/1757-899x/1099/1/012077.

[13] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 9, pp. 1263–1284, 2009, doi: 10.1109/TKDE.2008.239.

[14] G. S. Datta and M. Ghosh, "Bayesian Prediction in Linear Models: Applications to Small Area Estimation," *Ann. Stat.*, vol. 19, pp. 1748–1770, 1991, doi: 10.1214/aos/1176348369.

[15] McCulloch, *Generalized, Linear, and Mixed Models*. Canada: John Wiley & Sons, Inc., 2001.

[16] M. Fokkema, N. Smits, A. Zeileis, T. Hothorn, and H. Kelderman, "Detecting treatment-subgroup interactions in clustered data with generalized linear mixed-effects model trees," *Behav. Res. Methods*, vol. 50, no. 5, pp. 2016–2034, 2018, doi: 10.3758/s13428-

017-0971-x.

[17] A. Zeileis, T. Hothorn, K. Hornik, A. Zeileis, T. Hothorn, and K. Hornik, "Interface Foundation of America Model-Based Recursive Partitioning Linked references are available on JSTOR for this article : Model-Based Recursive Partitioning," vol. 17, no. 2, pp. 492–514, 2016, doi: 10.1198/106186008X319331.

[18] A. Hajjem, F. Bellavance, and D. Larocque, "Mixed-effects random forest for clustered data," *J. Stat. Comput. Simul.*, vol. 84, no. 6, pp. 1313–1328, 2012, doi: 10.1080/00949655.2012.741599.

[19] P. and N. G. Mittal, "A comparative analysis of classification techniques on medical datasets," *IJRET Int. J. Res. Eng. Technol.*, vol. 3, no. 6, pp. 454–460, 2014, doi: 10.15623/ijret.2014.0306085.

[20] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognit. Lett.*, vol. 27, no. 8, pp. 861–874, 2006, doi: 10.1016/j.patrec.2005.10.010.

[21] J. Han, J., Kamber, M., & Pei, *Data Mining: Concepts and Techniques Third Edition*. San Francisco: Elsevier / Morgan Kaufmann, 2011. doi: 10.1016/C2009-0-61819-5.

[22] N. Lunardon, G. Menardi, and N. Torelli, "ROSE: A package for binary imbalanced learning," *R J.*, vol. 6, no. 1, pp. 79–89, 2014, doi: 10.32614/rj-2014-008.

[23] N. V Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE : Synthetic Minority Over-sampling Technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, 2002, doi: 10.1613/jair.953.

[24] H. He, Y. Bai, E. A. Garcia, and S. Li, "ADASYN: Adaptive Synthetic Sampling Approach for Imbalanced Learning," in *IEEE International Joint Conference on Neural Networks (IJCNN)*, Hong Kong, China: IEEE, 2008, pp. 1322–1328. doi: 10.1109/IJCNN.2008.4633969.

[25] S. M. A. Elrahman and A. Abraham, "A Review of Class Imbalance Problem," *J. Netw. Innov. Comput.*, vol. 1, pp. 332–340, 2013, [Online]. Available: https://cspub-jnic.org/index.php/jnic/article/view/42/33

[26] BPS, "Garis Kemiskinan dan Indikator Sosial Ekonomi Indonesia." [Online]. Available: https://www.bps.go.id

[27] Andrew Gelman and Jennifer Hill, *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge, UK: Cambridge University Press, 2007. doi: 10.1017/CBO9780511790942.

[28] D. W. Hosmer, S. Lemeshow, and R. X. Sturdivant, *Applied Logistic Regression: Third Edition*, Third. Hoboken, New Jersey: John Wiley & Sons, 2013. doi: 10.1002/9781118548387.

[29] J. Fox and G. Monette, "Generalized Collinearity Diagnostics," *J. Am. Stat. Assoc.*, vol. 87, no. 417, pp. 178–183, 2014, doi: 10.1080/01621459.1992.10475190.

[30] D. Bates, M. Mächler, B. M. Bolker, and S. C. Walker, "Fitting Linear Mixed-Effects Models Using lme4," *J. Stat. Softw.*, vol. 67, no. 1, 2015, doi: 10.18637/jss.v067.i01.

[31] A. Liaw and M. Wiener, "Classification and Regression by randomForest," *R News*, vol. 2, no. December, pp. 18–22, 2002, [Online]. Available: https://journal.r-project.org/articles/RN-2002-022/

[32] World Bank, "World Development Report 2018: Learning to Realize Education's Promise." The World Bank, Washington, DC, 2018. doi: 10.1596/978-1-4648-1096-1.

[33] B. Krawczyk, "Learning from imbalanced data: open challenges and future directions," *Prog. Artif. Intell.*, vol. 5, no. 4, pp. 221–232, 2016, doi: 10.1007/s13748-016-0094-0.

[34] M. Buda, A. Maki, and M. A. Mazurowski, "A systematic study of the class imbalance problem in convolutional neural networks," *Neural Networks*, vol. 106, pp. 249–259, 2018, doi: 10.1016/j.neunet.2018.07.011.

[35] A. Niculescu-Mizil and R. Caruana, "Predicting good probabilities with supervised learning," *ICML 2005 - Proc. 22nd Int. Conf. Mach. Learn.*, no. 1999, pp. 625–632, 2005, doi: 10.1145/1102351.1102430.