

GLMM AND GLMM TREE METHODS FOR MODELING STUNTING PREVALENCE IN WEST JAVA

Andi Tenriawaru¹, Gusti Arviana Rahman^{*2}, Arif Nur Alfiyan², Budi Wijaya Eauf¹, La Surimi¹, M. Ichsan Nawawi³

¹Program Studi Ilmu Komputer, Fakultas MIPA, Universitas Halu Oleo, Kendari, Indonesia

²Program Studi S-1 Statistika, Fakultas MIPA, Universitas Halu Oleo, Kendari, Indonesia

³Jurusan Matematika, UIN Alauddin, Makassar, Indonesia

Penulis korespondensi: Gusti Arviana Rahman (arviana.rahman@uho.ac.id)

This study compares the Generalized Linear Mixed Model (GLMM) and GLMM Tree methods for modeling stunting prevalence in West Java Province. The dataset consists of 216 district/city-year observations from 27 districts/cities observed during 2016-2023. A chronological hold-out validation scheme was used, with observations from 2016-2022 as training data (189 observations; 87.5%) and observations from 2023 as testing data (27 observations; 12.5%). Model performance was evaluated using Mean Squared Error (MSE) and Root Mean Squared Error (RMSE). On the training data, GLMM Tree produced lower errors (MSE = 8.6583; RMSE = 2.9425) than GLMM (MSE = 9.4121; RMSE = 3.0679). On the testing data, GLMM Tree also showed slightly better predictive accuracy (MSE = 22.0868; RMSE = 4.6997) than GLMM (MSE = 23.0027; RMSE = 4.7961). The GLMM results identified active Posyandu (X5) and pregnant women receiving blood supplement tablets (X6) as significant predictors with negative coefficients. The GLMM Tree further revealed subgroup heterogeneity through splits on low birth weight (X3 = 2.59%) and complete basic immunization coverage (X8 = 93%). These findings indicate that GLMM Tree provides a modest improvement in predictive performance and offers additional interpretability by identifying subgroup-specific patterns in stunting prevalence.

Kata Kunci — GLMM, GLMM Tree, Stunting Cases, West Java, Statistical Models.

I. INTRODUCTION

Generalized Linear Mixed Models (GLMMs) are robust statistical frameworks designed to analyze response variables that follow distributions from the exponential family, while simultaneously accommodating nonlinear relationships and accounting for correlated or hierarchical data structures [1]. Within GLMMs, explanatory variables can be modeled as either fixed effects, which represent population-level relationships, or random effects, which account for variability at the cluster or subject level. This ability to model both fixed

and random effects makes GLMMs particularly suited for complex data structures, such as longitudinal or clustered data, where observations are not independent.

An advanced extension of GLMMs is the Generalized Linear Mixed Model Tree (GLMM tree), which integrates recursive partitioning techniques with mixed-effects modeling. This method partitions the data based on predictor variables, allowing model parameters to vary within subgroups, which is particularly useful for datasets with complex interactions and heterogeneous structures [2]. The GLMM tree approach is advantageous when the data include a large number of predictor variables, and when these predictors interact in ways that vary across subgroups within the population. By combining model-based trees with GLMMs, the GLMM tree enhances the detection of parameter instability and allows for more nuanced modeling of complex interactions among variables [3].

Stunting remains a critical public health issue in Indonesia, serving as an indicator of chronic undernutrition and leading to significant long-term consequences for physical growth, cognitive development, and future human capital. According to the World Health Organization child growth standards, stunting is defined as a height-for-age z-score that is more than two standard deviations below the median of the reference population [4]. In Indonesia, efforts to reduce stunting are prioritized at the national level, as its determinants are multifactorial, involving maternal and child health services, socioeconomic conditions, sanitation, nutritional intake, and access to community-based health interventions [5], [6], [7]. Given its complex etiology, modeling stunting requires an analytical approach that accounts for both regional variations and intricate relationships among health, socioeconomic, and environmental factors.

The province of West Java, which consists of 27 districts and cities, provides a valuable context for studying stunting due to its socioeconomic diversity. The prevalence of stunting varies across these districts and cities, influenced by differences in poverty levels,

sanitation access, maternal health programs, infant feeding practices, immunization coverage, and the performance of community health services. The dataset, which spans from 2016 to 2023, reveals both temporal and spatial variability in stunting prevalence, making it an ideal case for analysis using GLMM and GLMM tree methodologies.

Although GLMMs have been widely applied to model clustered and longitudinal health data, they may be limited in capturing heterogeneous effects across different subgroups. The GLMM tree, by contrast, allows for flexible modeling of these subgroup differences by permitting the fixed-effect parameters to vary across terminal nodes while simultaneously accounting for random effects. Previous studies have demonstrated the utility of GLMM trees for detecting subgroup-specific interactions and handling complex data structures in various health-related contexts [3], [8], [9], [10]. However, the application of GLMM trees to stunting data, particularly in the Indonesian context, remains underexplored, highlighting a gap in both methodological and empirical research.

This study aims to compare the performance of GLMM and GLMM tree methodologies in modeling stunting prevalence in West Java Province. Specifically, it evaluates the predictive accuracy of both models using Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) as evaluation metrics, and explores the contribution of various health and socioeconomic predictors to stunting outcomes across regions. The findings of this study are expected to provide valuable insights into the application of GLMM tree models for hierarchical health data and inform more effective policies for stunting reduction in West Java.

The structure of this paper is as follows: Section 2 presents the data and methodologies employed in this study, including the GLMM and GLMM tree frameworks. Section 3 discusses the results and provides a comparative analysis of the models, while Section 4 concludes with a summary of the findings and recommendations for future research.

II. METHOD

A. Data

This study employed a longitudinal secondary dataset covering districts and cities in West Java Province over the period 2016–2023. The unit of observation was the district/city-year, consisting of 27 administrative regions observed repeatedly across eight years. The use of longitudinal data enables the analysis to capture both regional differences and temporal variations in stunting prevalence.

The response variable in this study was the percentage of stunting cases among children. The explanatory variables were selected based on their relevance to child nutrition, maternal and child health services, socioeconomic conditions, and environmental health indicators. These variables include the percentage of

households with access to decent sanitation, the percentage of poor population, the percentage of babies born with low birth weight, the percentage of infants under six months receiving exclusive breastfeeding, the percentage of active Posyandu, the percentage of pregnant women receiving blood supplement tablets, the coverage of vitamin A supplementation among toddlers aged 6–59 months, and the percentage of infants receiving complete basic immunization.

The data were obtained from official secondary sources, including Statistics Indonesia, reports from the Ministry of Health of the Republic of Indonesia, and district/city health profiles of West Java Province. In the model specification, districts/cities were treated as the grouping factor for the random intercept, while year was included as a random slope component to accommodate possible differences in temporal trends across districts/cities. The complete list of variables used in this study is presented in Table 1.

Table 1. Variables' List

Types of variables	Variables' names	Data types	Units	Data source
Response	Percentage of stunting cases (Y)	Numerical	%	BPS, West Java health profile
	Decent sanitation percentage (X1)	Numerical	%	BPS, West Java health profile
	Percentage of poor population (X2)	Numerical	%	BPS, West Java health profile
	Percentage of babies born with low birth weight (X3)	Numerical	%	BPS, West Java health profile
	Percentage of Giving Exclusive Breast Milk (ASI) to Babies <6 Months (X4)	Numerical	%	BPS, West Java health profile
	Percentage of Active Posyandu (X5)	Numerical	%	BPS, West Java health profile
	Percentage of Pregnant Women Receiving Blood Supplement Tablets (TTD) (X6)	Numerical	%	BPS, West Java health profile
	Percentage of Coverage for Toddlers (6-59 Months) Receiving Vitamin A (X7)	Numerical	%	BPS, West Java health profile
Fixed effect	Percentage of Complete Basic Immunization in Babies (X8)	Numerical	%	BPS, West Java health profile
	Years	Ordinal	-	BPS
Random slope	Districts/Cities	Categorical	-	BPS

B. Analysis

The analysis in this study was carried out using Microsoft Excel software, R version 4.3.2, and R Studio 2024.04.2 using package lme4, glmertree, ggplot2, and readxl. The processes of the analysis, as shown in Figure 1.

The dataset was divided using a chronological hold-out scheme based on year, not by random percentage splitting. The training data consisted of observations from 2016 to 2022, comprising 27 districts/cities observed over seven years, or 189 observations (87.5% of the total 216 observations). The testing data consisted of observations from 2023, comprising 27 observations (12.5% of the total observations). This year-based splitting strategy was selected to evaluate the ability of the models to predict stunting prevalence in a future period while preventing information leakage from the testing year into the training process. The predictor variables were selected based on prior studies and their substantive relevance to child nutrition, maternal health services, socioeconomic conditions, and environmental health.

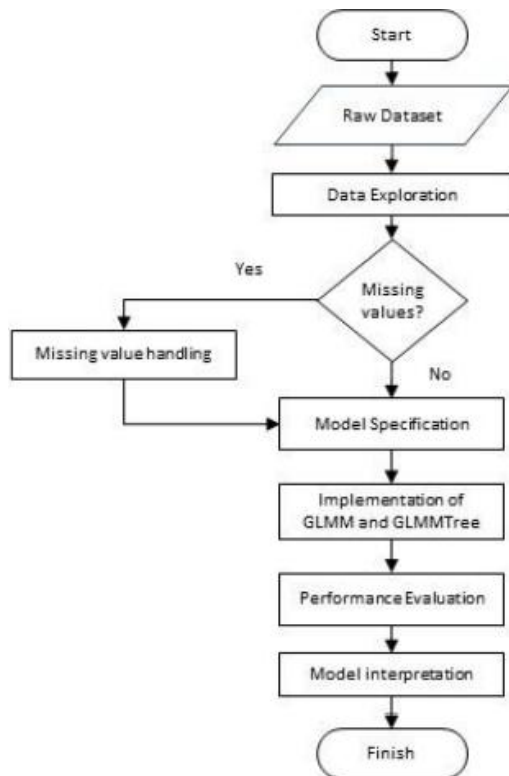


Figure 1. Research flow

Generalized Linear Mixed Models (GLMMs)

GLMMs provide a flexible framework for analysing data in which observations are not fully independent because they are grouped, clustered, or repeatedly measured over time. In this study, stunting prevalence is observed repeatedly for the same districts/cities across several years; therefore, the data have a longitudinal and hierarchical structure. A standard regression model may ignore this dependency, whereas a mixed model explicitly separates population-level effects from district/city-specific variations.

Let y_{it} denote the stunting prevalence in district or city i at year t , where $i = 1, 2, \dots, m$ and $t = 1, 2, \dots, T$. The conditional mean of the response variable is denoted by $\mu_{it} = E(y_{it}|b_i)$. In the GLMM framework, the relationship between μ_{it} and the explanatory variables is expressed through a link function as Equation 1.

$$g(\mu_{it}) = \eta_{it} \quad (1)$$

$$\eta_{it} = \beta_0 + \beta_1 X_{1it} + \beta_2 X_{2it} + \dots + \beta_p X_{pit} + b_{0i} + b_{1i}$$

Where $g(\cdot)$ is the link function, η_{it} is the linear predictor, $X_{1it}, X_{2it}, \dots, X_{pit}$ are predictor variables, and $\beta_0, \beta_2, \dots, \beta_p$ are fixed effect parameters representing the average effects across all districts/cities. The term b_{0i} and b_{1i} are random effects that allow each district/city to have its own baseline and temporal pattern of stunting prevalence.

The random effects are assumed as Equation 2 to follow a multivariate distribution:

$$b_i = \begin{pmatrix} b_{0i} \\ b_{1i} \end{pmatrix} \sim N \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \Sigma_b \right) \quad (2)$$

with covariance matrix

$$\Sigma_b = \begin{pmatrix} \sigma_{b0}^2 & \sigma_{b01} \\ \sigma_{b01} & \sigma_{b1}^2 \end{pmatrix}$$

where σ_{b0}^2 represents the variability of district/city-specific intercepts, σ_{b1}^2 represents the variability of district/city-specific slopes over time, and σ_{b01} represents the covariance between the random intercept and random slope.

In matrix notation, the GLMM can be written as Equation 3.

$$g(\mu_i) = X_i \beta + Z_i b_i \quad (3)$$

where X_i is the fixed-effect design matrix, Z_i is the random-effect design matrix, β is the vector of fixed-effect coefficients, and b_i is the vector of random effects for district/city i . This formulation allows the model to estimate both the general association between predictors and stunting prevalence and the unobserved heterogeneity among districts/cities.

Model parameters are commonly estimated using maximum likelihood. Since the likelihood function involves integration over the random effects, the marginal likelihood for district/city i can be expressed as Equation 4,

$$L_i(\beta, \Sigma_b, \phi) = \int f(y_i|b_i; \beta, \phi) f(b_i; \Sigma_b) db_i \quad (4)$$

and the total likelihood is written in Equation 5.

$$L(\beta, \Sigma_b, \phi) = \prod_{i=1}^m L_i(\beta, \Sigma_b, \phi) \quad (5)$$

In practice, the integral is often approximated numerically, for example by using the Laplace approximation. This estimation approach enables GLMMs to handle clustered data while maintaining a regression structure suitable for interpreting the effects of health, demographic, and socioeconomic predictors on stunting prevalence.

Generalized Linear Mixed Model Tree (GLMM tree)

Although GLMMs are useful for modelling clustered or longitudinal data, they generally assume that the fixed-effect coefficients are constant across the entire population. In public health data, this assumption may be too restrictive because the relationship between stunting prevalence and its determinants can differ across groups of districts/cities. For example, the effect of sanitation, poverty, immunization, or maternal health intervention may not be identical in all regional conditions. To address this issue, the Generalized Linear Mixed Model Tree combines the interpretability of mixed-effects models

with the flexibility of tree-based partitioning.

The GLMM Tree extends the GLMM by allowing fixed-effect parameters to vary across subgroups identified by recursive partitioning, while the random effects are still used to account for clustered observations. Suppose the data are divided into M terminal nodes. For an observation belonging to node m , the model can be written as Equation 6.

$$g(\mu_{it}) = X_{it}^T \beta_m + Z_{it}^T b_i \quad (6)$$

where β_m is the node-specific fixed-effect parameter vector, X_{it} is the vector of fixed-effect predictors, Z_{it} is the random-effect design vector, and b_i is the random effect for district/city i . This means that each terminal node may have a different regression structure, while the correlation caused by repeated observations within the same district/city remains accommodated through random effects.

The GLMM Tree algorithm can be summarized as follows:

- (1) A mixed-effects model is first fitted to account for the clustered structure of the data.
- (2) The fixed-effect component is examined for parameter instability with respect to candidate partitioning variables.
- (3) If parameter instability is detected, the algorithm selects the partitioning variable that provides the strongest evidence of heterogeneity.
- (4) The selected variable is split at an optimal cut-off point, creating more homogeneous subgroups.
- (5) The model is re-estimated within the resulting nodes, and the partitioning process continues until no statistically meaningful split is found or a stopping criterion is reached.
- (6) The final tree consists of terminal nodes, each representing a subgroup with its own fixed-effect pattern.

The advantage of the GLMM Tree is that it does not require the researcher to specify all possible interaction effects manually. Instead, subgroup structures are discovered from the data. This is particularly useful for stunting analysis because the influence of health service coverage, poverty, sanitation, and nutritional indicators may vary depending on regional characteristics.

Therefore, the GLMM Tree provides both predictive and explanatory value: it improves model flexibility while also revealing which variables define meaningful differences among districts/cities.

Model Evaluation

The predictive performance of the GLMM and GLMM Tree models was evaluated using Mean Squared Error (MSE) and Root Mean Squared Error (RMSE). These measures quantify the difference between the observed stunting prevalence and the predicted values produced by each model. The MSE is defined as:

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (7)$$

where y_i is the observed value, $\hat{y}_i$ is the predicted value, and n is the number of observations. The RMSE is obtained by taking the square root of MSE:

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (18)$$

A smaller MSE or RMSE indicates better predictive accuracy. In this study, these metrics were calculated for both training and testing data to assess not only how well each model fits the observed data but also how well it generalizes to unseen observations.

III. RESULT AND DISCUSSION

The dataset includes complete observations for all districts and cities, totaling 27 entities observed over the period from 2016 to 2023 (8 years). Figure 2 depicts the scatter plot of stunting cases as a percentage in West Java Province during this period with all variables' responses. The scatter plot shows that data illustrates different variations in trends. This variability suggests the need to account for varying intercepts and slopes in the model. Parameter estimations using the GLMM and GLMM tree models were conducted, employing the Laplace approximation and the GLMM tree algorithm for model specification.

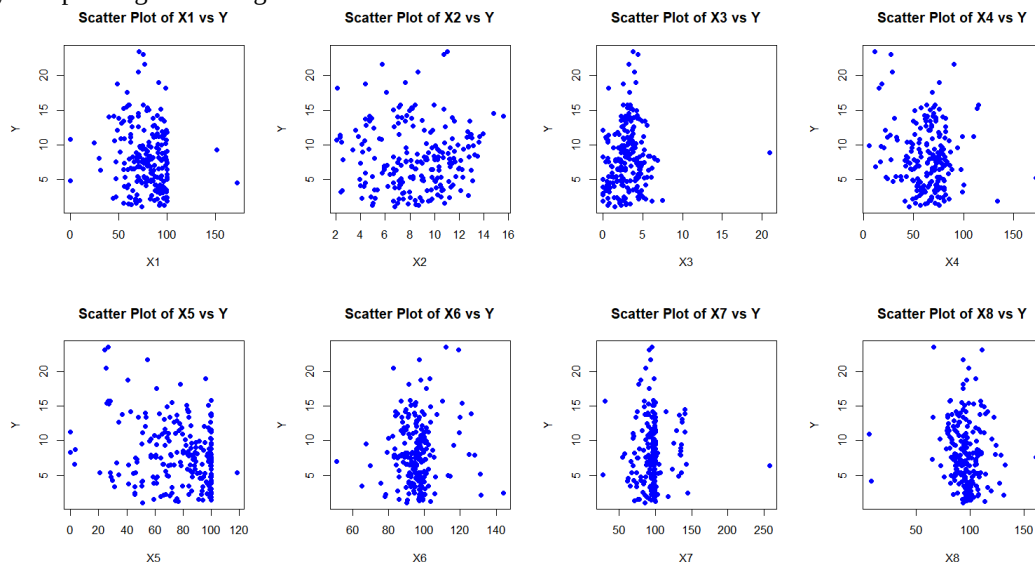


Figure 2. Scatter plots of stunting prevalence and predictor variables

The evaluation stage is carried out by using the MSE and RMSE value, as shown in Table 2.

Table 2. Estimation Method Evaluation using train data

Estimation Method	MSE	RMSE
GLMM	9.4121	3.0679
GLMM tree	8.6583	2.9425

According to Table 2, the GLMM tree shows notably superior performance in comparison to the GLMM, exhibiting lower MSE and RMSE values. This suggests that the GLMM tree more effectively captures patterns within the training data.

Table 3. Estimation Method Evaluation using testing data

Estimation Method	MSE	RMSE
GLMM	23.0027	4.7961
GLMM tree	22.0868	4.6997

Based on Table 3, it can be seen that the GLMM trees have lower MSE and RMSE compared to the GLMM, although the differences are not very large. This indicates that the GLMM tree is better at generalizing and more robust against unseen data. The relatively small performance difference between the two methods on the testing data suggests that both can be used, but the choice of method depends on whether priority is given to training performance or generalization ability.

The significant parameter estimates are presented in Table 4, while the GLMM Tree structure is shown in Figure 3. In the GLMM model, two predictors are significant at the 95% confidence level: the percentage of active Posyandu (X5) and the percentage of pregnant women receiving blood supplement tablets (TTD) (X6). The coefficient of X5 is -0.0054, indicating that a one-percentage-point increase in active Posyandu coverage is associated with a 0.0054 percentage-point decrease in stunting prevalence, holding other variables constant. The coefficient of X6 is -0.0072, indicating that a one-percentage-point increase in the percentage of pregnant women receiving blood supplement tablets is associated with a 0.0072 percentage-point decrease in stunting prevalence, holding other variables constant. Thus, the signs and interpretations of X5 and X6 are consistent with the negative coefficients shown in Table 4.

Table 4. The result of parameter estimation

Variables	GLMM tree			GLMM
	Node 3	Node 4	Node 5	
Intercept	5.969	17.425	6.255	2.2504
X2	0.326	-0.144	-0.196	-0.0013
X5	0.025	-0.044	-0.056	-0.0054 *
X6	-0.029	-0.073	0.088	-0.0072 *

Description: *Significant at 95% confidence level

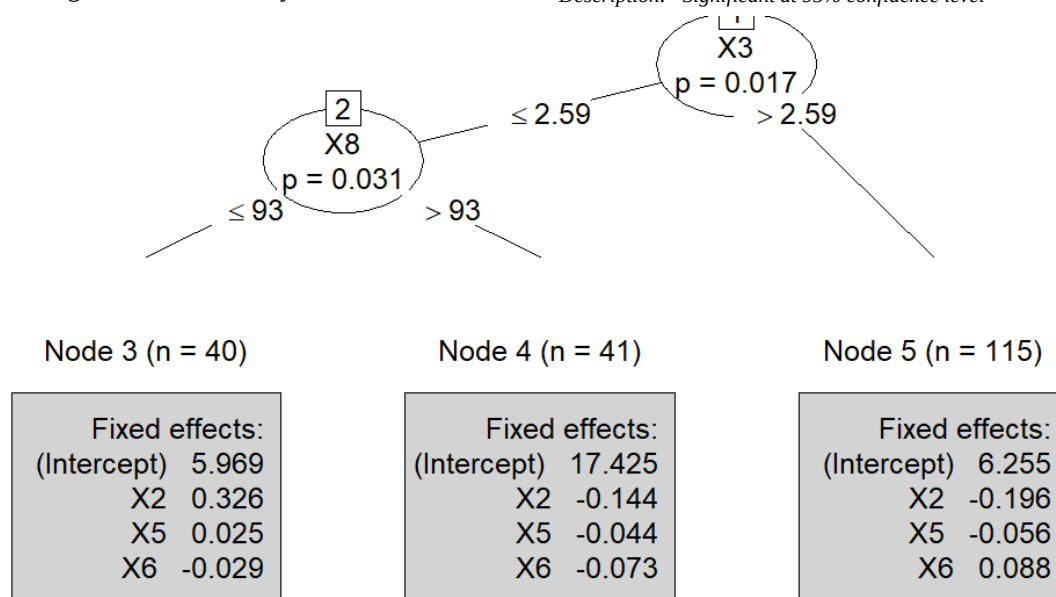


Figure 3. GLMM tree diagram

Based on Figure 3, the GLMM Tree separates the observations into subgroups using X3 and X8. The first split occurs at X3, the percentage of babies born with low birth weight, with a threshold of 2.59% ($p = 0.017$). This means that X3 provides the strongest evidence of parameter instability among the candidate splitting variables at the root node. Substantively, low birth weight is closely related to fetal growth, maternal nutrition, and early-life vulnerability; therefore, districts/cities with different levels of low birth weight may show different stunting patterns. For observations with $X3 \leq 2.59\%$, the model identifies a second split based on X8, namely complete basic immunization coverage, at the empirical threshold of 93% ($p < 0.031$). This threshold is data-driven and should be interpreted as the cut-off that best separates the remaining heterogeneous observations in this dataset, not necessarily as a universal policy

threshold. The split produces Node 3 ($n = 40$) for $X8 \leq 93\%$ and Node 4 ($n = 41$) for $X8 > 93\%$, while observations with $X3 > 2.59\%$ are assigned to Node 5 ($n = 115$). The terminal nodes indicate that the relationships involving X2, X5, and X6 differ across subgroup conditions defined by low birth weight and immunization coverage, supporting the use of GLMM Tree to capture heterogeneous structures in stunting prevalence.

IV. CONCLUSION

The conclusion from the above results is that the GLMM tree method demonstrates better performance in capturing patterns in the training data compared to the GLMM method, as indicated by lower MSE and RMSE values on the training data. According to the result of testing data, GLMM trees shows better generalization

ability with lower MSE and RMSE compared to the GLMM, although the differences are not very significant. This suggests that the GLMM trees are better to use in modeling the percentage of stunting cases in West Java. We can see that GLMM tree These results highlight the importance of specific health and socioeconomic factors in understanding the outcomes in different regions. The GLMM tree provides a nuanced view of how these variables interact in different subgroups.

DAFTAR PUSTAKA

- [1] J. M. McCulloch, Charles E.; Neuhaus, "Generalized linear mixed models," *Wiley StatsRef Stat. Ref. Online*, 2014, [Online]. Available: https://onlinelibrary.wiley.com/doi/book/10.1002/9781118445112?__cf_chl_tk=jire3CvJov0ibf4IR0aX_xOyTirTPaXHQCGJS_D2IYY-1777434839-1.0.1.1-Zm28pn2MHZEX1UNh6dQpH0SRQ8LhlJ4vdr4wNQ2Hug
- [2] H. Seibold, T. Hothorn, and A. Zeileis, "Generalised Linear Model Trees with Global Additive Effects," *Adv. Data Anal. Classif.*, no. 3, 2019, [Online]. Available: <https://www.springerprofessional.de/en/generalised-linear-model-trees-with-global-additive-effects/16178578>
- [3] 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.
- [4] "WHO Child Growth Standards based on length/height, weight and age.," *Acta Paediatr. Suppl.*, vol. 450, pp. 76–85, Apr. 2006, doi: 10.1111/j.1651-2227.2006.tb02378.x.
- [5] UNICEF and W. B. G. World Health Organization, *Levels and Trends in Child Malnutrition: UNICEF/WHO/World Bank Group Joint Child Malnutrition Estimates, Key Findings of the 2023 Edition*. 2023.
- [6] Kementerian Kesehatan Republik Indonesia, *Survei Kesehatan Indonesia (SKI) 2023 Dalam Angka*. 2023.
- [7] Pemerintah Provinsi Jawa Barat, *Prevalensi Balita Stunting Berdasarkan Kabupaten/Kota di Jawa Barat, 2019–2024*. 2025.
- [8] Y. Wei, L. Liu, X. Su, L. Zhao, and H. Jiang, "Precision medicine: Subgroup identification in longitudinal trajectories.," *Stat. Methods Med. Res.*, vol. 29, no. 9, pp. 2603–2616, Sep. 2020, doi: 10.1177/0962280220904114.
- [9] S. Bayu, K. A. Notodiputro, and B. Sartono, "GLMM and GLMMTree for Modelling Poverty in Indonesia," *Proc. Int. Conf. Data Sci. Off. Stat.*, vol. 2023, no. 1 SE-Data Science, pp. 121–131, Dec. 2023, doi: 10.34123/icdsos.v2023i1.333.
- [10] D. A. N. Sirodj, K. A. Notodiputro, and B. Sartono, "Perbandingan Kerja Binomial GLMM Tree dan BIMM Forest untuk Memodelkan Status Bekerja Penduduk," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 11, no. 1 SE-Ilmu Komputer, pp. 95–106, Feb. 2024,