Jurnal Teknik Informatika (JUTIF)
P-ISSN: 2723-3863
E-ISSN: 2723-3871
Vol. 7, No. 2, April 2026, Page. 990-1006
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5407

A Comparative Study of Generalized Linear Mixed Model and Mixed Effects Random Forest for Analyzing Data with Outliers

Reza Arianti*[1], Khairil Anwar Notodiputro[2], Yenni Angraini[3]

[1,2,3]School of Data Science, Mathematics and Informatics, IPB University, Indonesia

Email: [1]rezaarianti@apps.ipb.ac.id

Received : Oct 29, 2025; Revised : Nov 14, 2025; Accepted : Nov 17, 2025; Published : Apr 15, 2026

Abstract

This study compares MERF and GLMM-NB in analyzing hierarchical data and focusing on the role of residual outliers and the application of winsorization. A two-stage analytical pipeline was implemented: (1) winsorization to reduce extreme residual values, and (2) model training using MERF and GLMM-NB. The dataset comes from the 2021 National Socio-Economic Survey (Susenas) in West Java Province, measuring tobacco consumption intensity. Two statistical approaches are compared, MERF and GLMM with a Negative Binomial distribution (GLMM-NB). Models were trained under two conditions: without winsorization (WIN0) and with two-sided 5% winsorization (WIN5). Winsorization was applied to the training data, and the test data were adjusted using thresholds from the training set. Model performance was assessed using Root Mean Squared Error (RMSE) and the train-test ratio. Under WIN0, GLMM recorded an RMSE of 49.65 for training and 42.27 for testing, while MERF achieved 35.96 and 39.94, respectively. After WIN5, GLMM showed a larger error reduction, with RMSE values of 34.90 (train) and 30.20 (test), while MERF dropped to 26.63 (train) and 28.64 (test). These results indicate that MERF provides higher predictive accuracy, whereas GLMM benefits more from winsorization. Household expenditure, employment status, age, and gender consistently emerged as key variables linked to tobacco consumption intensity. This study is the first to compare MERF and GLMM-NB with winsorization using Indonesia's hierarchical data. The analytical framework helps inform public health policies aligned with SDG 3: Good Health and Well-being, particularly in reducing tobacco-related health risks.

Keywords : *GLMM, Hierarchical Modeling, Data, MERF, Outliers, Tobacco Intensity, Winsorization.*

1. INTRODUCTION

Traditional linear regression is often unsuitable for high-dimensional datasets or for data that violate the assumption of independence. In recent years, mixed-effects (ME) models have emerged as an effective solution for analyzing high-dimensional data [1], [2], [3]. Their main advantage lies in the ability to combine fixed effects, which capture overall patterns, with random effects, which account for variability across groups. Among the available extensions, the Generalized Linear Mixed Model (GLMM) has been especially useful for count data, where Poisson and Negative Binomial distributions are most often applied [4], [5]. When data are overdispersed, which is common in social and health research, the Negative Binomial specification (GLMM-NB) is generally preferred [6], [7], [8].

Researchers have recently begun to apply machine learning techniques to hierarchical data. The Mixed-Effects Random Forest (MERF) has also gained attention as it modifies the Random Forest framework to account for variation between groups [9]. In practice, this adjustment allows the model to handle hierarchical structures effectively. Results so far suggest that MERF can handle complex prediction tasks effectively, and in several studies it even surpassed traditional models [10], [11], [12]. In Indonesia, for instance, applications in the education sector showed that MERF outperformed ordinary linear regression, pointing to its value in real-world contexts [13], [14], [15]. Recent studies

Vol. 7, No. 2, April 2026, Page. 990-1006
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5407

also show that MERF continues to be extended after the pandemic, such as MERF variants using PCA and MERoF for small area estimation [16] and MERF applications in post-pandemic digital learning analytics [17].

A persistent challenge in hierarchical data analysis is the influence of outliers. These values can distort statistical inference, and ME models are especially sensitive to them [18]. One way to deal with this problem is through winsorization, which replaces extreme observations with values set at chosen quantile thresholds. This adjustment improves the stability of mean and variance estimates while keeping all observations in the dataset [19]. Applying winsorization allows a fairer evaluation of outlier-sensitive models such as GLMM models and more robust approaches like MERF. Recent works also confirm Winsorization as an effective outlier treatment method in data mining and modeling [20], and post-pandemic studies highlight the urgency of handling outliers due to abnormal behavioral patterns during COVID-19 recovery periods [21], [22].

Comparative studies with PISA data have reported that GLMM and MERF perform competitively in multilevel contexts [23]. However, the role of residual outliers has not been fully examined, even though such values can affect predictive accuracy and the recognition of underlying patterns. This study seeks to bridge this gap by evaluating the performance of GLMM-NB and MERF when residual outliers are present. The focus is on how winsorization shapes model accuracy and stability, and whether it improves predictive performance. The study also examines the main factors linked to tobacco consumption, emphasizing consumption intensity rather than smoking status. This focus is important for designing tobacco control policies that are targeted, measurable, and sensitive to context. Treating outliers should help researchers identify the most influential variables with greater precision and, in turn, give policymakers a firmer empirical basis for decision-making. The explicit evaluation of winsorization on predictive improvement has also been recommended in recent literature as part of robust modeling workflows after the pandemic [24].

According to Badan Pusat Statistik (BPS-Statistics Indonesia) and the World Health Organization (WHO), smoking prevalence in Indonesia has remained consistently high [25]. This study relies on 2021 data, taken at a time when communities were beginning to recover from the pandemic.This period provides useful insight into how smoking patterns shifted as communities resumed daily activities. Evidence suggests that greater awareness of health risks during the COVID-19 outbreak may have influenced the way people smoked [26].

In selecting the study location, West Java stood out because it not only has the largest population in the country but also records one of the highest rates of smoking. These features make it a suitable context for examining differences in smoking across social and demographic groups. This study is also linked to the Sustainable Development Goals (SDGs), with particular attention to Goal 3, which emphasizes health and well-being through reductions in tobacco use. In this context, emphasis is placed on socioeconomic factors such as education and household purchasing power, as both have a strong influence on smoking behavior and its long-term health outcomes. Specifically, SDG Target 3.a emphasizes tobacco control through the WHO Framework Convention on Tobacco Control (WHO-FCTC).

To further clarify the research gap and justify the urgency of examining post-pandemic tobacco consumption using hierarchical modeling, a summary of relevant studies comparing GLMM, MERF, and outlier treatment approaches is presented in Table 1.

Based on the identified gap, this study explicitly aims to compare GLMM-NB and MERF while applying winsorization of residual outliers to improve predictive stability and identify the key socioeconomic determinants of tobacco consumption in Indonesia. The expected contribution of this study is to evaluate whether winsorization improves MERF predictive accuracy and robustness in hierarchical tobacco data.

Jurnal Teknik Informatika (JUTIF)
P-ISSN: 2723-3863
E-ISSN: 2723-3871
Vol. 7, No. 2, April 2026, Page. 990-1006
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5407

Table 1. Gap Study Table

Study / Reference	Method	Outliers handling	Hierarchical data	Gap in literature
[6]	GLMM-NB	✗	✓	Did not address residual outliers; no comparison with ML methods.
[11]	MERF	✗	✓	No winsorization; only focuses on prediction, not robustness.
[20]	Winsorization	✓	✗	No hierarchical modeling and not compared with GLMM/MERF.
[16]	MERF extension (PCA + Rotation Forest)	✗	✓	Did not evaluate residual outliers; no comparison with GLMM.
This study (2025)	GLMM-NB vs MERF + Winsorization	✓	✓	First empirical comparison assessing winsorization effect on GLMM-NB vs MERF for hierarchical count data.

2. METHOD

This study compares the performance of MERF and GLMM-NB on hierarchical data. The methods section explains the data used, the preprocessing carried out, how winsorization was applied, and the approach taken to evaluate predictive accuracy.

2.1. Generalized Linear Mixed Model (GLMM)

The Generalized Linear Mixed Model (GLMM) builds on the Generalized Linear Model (GLM) by incorporating both fixed and random effects [27]. With a Negative Binomial specification (GLMM-NB), it offers a flexible framework for modeling count data that reflects predictor relationships while also accommodating hierarchical structures and overdispersion [28].

This study analyzes the socioeconomic factors linked to tobacco consumption using data collected in 2021 from West Java Province. The dataset includes 21,290 individuals (k = 1, 2, …, 21,290) nested within 2,196 villages (l = 1, 2, …, 2,196). Because multiple individuals may reside in the same village, the data form a nested hierarchical structure in which observations within groups are not independent. Therefore, a random effect at the village level is included to accommodate between-group variation. The Generalized Linear Mixed Model with Negative Binomial distribution (GLMM-NB) is formulated as follows:

$$\boldsymbol{y}_{kl} = \boldsymbol{X}_{kl}^{T}\boldsymbol{\beta} + \mathbf{Z}_{kl}^{T}\boldsymbol{b}_{l} \quad \text{where} \quad \boldsymbol{y}_{kl} \sim \text{NB}(\mu_{kl}, \theta) \tag{1}$$

Notation:

$\boldsymbol{y}_{kl}$: Number of tobacco cigarettes consumed by individual k in village l
μ_{kl} : Conditional expected value of consumption, i.e., $\mu_{kl} = \text{E}(\boldsymbol{y}_{kl}|\boldsymbol{b}_{l})$
θ : Dispersion parameter controlling the degree of overdispersion
$\boldsymbol{X}_{kl}$: Matrix of fixed-effect predictors, where $\boldsymbol{X}_{kl} \in \mathbb{R}^{p\times 1}$, $p = 14$
$\boldsymbol{\beta}$: Vector of fixed-effect coefficients
$\boldsymbol{Z}_{kl}$: Design matrix for random effects (value = 1 for random intercept)
$\boldsymbol{b}_{l}$: Vector of level random effect, $\boldsymbol{b}_{l} \sim \mathcal{N}(0, \sigma_b^2\boldsymbol{I})$, $\boldsymbol{b}_{l} \in \mathbb{R}^{q\times 1}$

The number of tobacco cigarettes consumed represents discrete count data. In the Poisson model, it is assumed that the expected value equals the variance. However, an initial inspection of the data shows that the variance of the response is much greater than its mean, i.e., $Var(\boldsymbol{y}_{kl}) > \mathbb{E}[\boldsymbol{y}_{kl}]$, a

Vol. 7, No. 2, April 2026, Page. 990-1006
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5407

condition referred to as overdispersion. The Negative Binomial distribution addresses this by introducing a dispersion parameter θ, such that the variance is defined as

$$Var(\boldsymbol{y}_{kl}) = \mu_{kl} + \frac{\mu_{kl}^2}{\theta} \quad (2)$$

This parameterization, known as *nbinom2*, is available in the *glmmTMB* function in R (with family = *nbinom2*). Rather than eliminating overdispersion, the model directly incorporates it, making the analysis more robust to high variability in the data. The parameters $\boldsymbol{\beta}$, $\boldsymbol{b}_l$, and θ in the GLMM-NB are estimated using Maximum Likelihood Estimation (MLE), typically via Laplace approximation or adaptive Gauss–Hermite quadrature. both of which are available in the glmmTMB package for solving the likelihood function. The model is thus executed in R using glmmTMB, which supports the nbinom2 specification where the variance grows quadratically with the mean [29].

To ensure reproducibility of the statistical modeling process, the GLMM-Negative Binomial model was implemented using the glmmTMB() function in R, which estimates fixed and random effects simultaneously through maximum likelihood with Laplace approximation.

```
library(glmmTMB)
# GLMM-Negative Binomial (nbinom2) — Reproducibility Code
# Response variable:
#   Y = Number of tobacco cigarette consumption per individual (last 7 days)

# Predictor variables (X1 – X15):
# Random Effect:
# (1 | KODES) = Random intercept for each village (hierarchical structure)

model_nb <- glmmTMB(
  Y ~ X1 + X2 + X3 + X4 + X5 +
     X6 + X7 + X8 + X9 + X10 +
     X11 + X12 + X13 + X14 + X15 +
     (1 | KODES),              # random intercept per village/desa
  data   = train_data,
  family = nbinom2(link = "log")    # GLMM-NB (Negative Binomial Type 2)
)
summary(model_nb)
```

This code explicitly shows how the GLMM-NB model is fitted using glmmTMB(), ensuring transparency and reproducibility of the hierarchical model estimation.

2.2. Mixed Effect Random Forest (MERF)

The Mixed Effects Random Forest (MERF), introduced earlier, blends Random Forest with mixed effects modeling so that hierarchical or clustered structures can be properly handled. In Hajjem's work, MERF was applied in a way that allowed both fixed and random effects to be represented within the model, and this combination was shown to improve the accuracy of predictions [9]. The model can be expressed in its full vectorized form for cluster l as follows:

$$\boldsymbol{y}_l = f(\boldsymbol{X}_l) + \boldsymbol{Z}_l\boldsymbol{b}_l + \boldsymbol{\varepsilon}_l \quad (3)$$

$$\boldsymbol{b}_l \sim N(0, \boldsymbol{D}), \boldsymbol{\varepsilon}_l \sim N(0, \boldsymbol{\sigma^2}), \qquad l = 1, \ldots, 2.196$$

Notation:

$\boldsymbol{y}_l$: Number of tobacco cigarettes consumed by individual k in cluster l
$\boldsymbol{X}_l$: Matrix of fixed-effect predictors, $\boldsymbol{X}_l \in \mathbb{R}^{p\times 1}$,where $p = 14$
$\boldsymbol{Z}_l$: Design matrix for random effects (typically contains 1s for random intercepts),$q \times 1$
$\boldsymbol{b}_l$: Vektor of cluster-level random effects, $\boldsymbol{b}_l \in \mathbb{R}^{q\times 1}$
$f(\cdot)$: Nonparametric regression function based on Random Forest

let $\boldsymbol{y}_l = \left[y_{l1}, \ldots, y_{ln_l}\right]^T$ denote the response vector of size $n_l \times 1$, where n_l is the number of individuals (observations) in village l. The matrix $\boldsymbol{X}_l = \left[x_{l1}, \ldots, x_{ln_l}\right]^T$ represents the covariates for the fixed effects, with dimensions $n_l \times p$. Similarly, $\boldsymbol{Z}_l = \left[z_{l1}, \ldots, z_{ln_l}\right]^{[T]}$ is the covariate matrix for the random effects with dimensions $n_l \times p$, where each z_{lk} typically encodes the village identifier and $p = 14$ is the number of predictors. The cluster-level random effects are assumed to follow $\boldsymbol{b}_l \sim N(0, \boldsymbol{D})$ and the individual-level residuals within clusters follow $\boldsymbol{\varepsilon}_l \sim N(0, \sigma^2)$.

Let $r \in \{0,1,2, \ldots\}$ denote the iteration index in the MERF parameter estimation process. The parameters in MERF model are estimated iteratively, beginning with the following initial values: $\widehat{\boldsymbol{b}}_{l(0)} = 0$, the initial residual variance is set to $\hat{\sigma}^2_{(0)} = 1$, and the initial random effect covariance matrix is set to $\widehat{\boldsymbol{D}}_{(0)} = \boldsymbol{I}_q$. At each iteration r, the adjusted response is calculated as:

$$\boldsymbol{y}^*_{l(r)} = \boldsymbol{y}_l - \boldsymbol{Z}_l \widehat{\boldsymbol{b}}_{l(r-1)} \tag{4}$$

This step subtracts the contribution of the previous iteration's random effects from the original response. The Random Forest model is then trained using the input–output pairs $\boldsymbol{X}_l$ dan $\boldsymbol{y}^*_{l(r)}$. The fixed-effect prediction function $\hat{f}(\boldsymbol{X}_l)_{(r)}$ is estimated using only trees that exclude observations from cluster l, in order to prevent information leakage. Next, the residuals are computed as:

$$\hat{\boldsymbol{\varepsilon}}_{l(r)} = \boldsymbol{y}_l - \hat{f}(\boldsymbol{X}_l)_{(r)} - \boldsymbol{Z}_l \widehat{\boldsymbol{b}}_{l(r-1)} \tag{5}$$

These residuals are used to update the random effects $\widehat{\boldsymbol{b}}_{l(r)}$ and the total covariance $\widehat{\boldsymbol{V}}_{l(r)}$. The residual variance $\hat{\sigma}^2_{(r)}$ is updated using the mean of the squared residuals, with an adjustment for the trace penalty of the covariance. The random effect covariance matrix $\widehat{\boldsymbol{D}}_{(r)}$ is then updated from the current estimates of the random coefficients, taking into account the corrected covariance. This entire process is repeated until parameter convergence is achieved. conclusion. The pseudocode below illustrates the iterative MERF training process until convergence.

Input: X (fixed effects), Z (random effects group), y (response)
Initialize: u = 0 (random effects)
Repeat until convergence:
1. Compute pseudo-response: y* = y - Zu
2. Fit Random Forest on (X, y*)
3. Predict fixed component: ŷ_fixed = RF(X)
4. Estimate random effects using LME:

u = (Z'Z + λI)^(-1) Z' (y - ŷ_fixed)

Output: ŷ = ŷ_fixed + Z u

This pseudocode summarizes the iterative MERF estimation process and ensures clarity on how the fixed and random effects are updated until convergence.

Vol. 7, No. 2, April 2026, Page. 990-1006
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5407

2.3. Winsorization

The Winsor method was first introduced by Charles P. Winsor in 1946 as an alternative statistical technique to address the presence of outliers in observational data [30]. This approach estimates regression parameters by transforming the observed response values into Winsorized values. In this study, residual outliers residuals that deviate substantially from the majority distribution and can distort the data or mislead interpretation will be addressed. Winsorization in this study is applied externally to the GLMM-NB and MERF algorithms. The procedure replaces extreme residual values with predetermined thresholds, thereby limiting the effect of outliers without discarding relevant data. Winsorization, when applied at low levels of about 1% to 5%, has been shown to improve model robustness against outliers while keeping the overall structure of the dataset intact [31]. It is particularly useful for skewed distributions, since it reduces the influence of extreme values without noticeably changing the general shape of the data [32]. The definition of the winsorized observations is given below:

$$e_{kl}^{*} = \begin{cases} e_{kl} & ,\hat{\eta}\,(\alpha_1) \le e_{kl} \le \hat{\eta}\,(\alpha_2) \\ \hat{\eta}\,(\alpha_1) & ,e_{kl} < \hat{\eta}\,(\alpha_1) \\ \hat{\eta}\,(\alpha_2) & ,e_{kl} > \hat{\eta}\,(\alpha_2) \end{cases} \tag{6}$$

Equation (6) shows the two-sided winsorization applied to residuals. A residual e_{kl}^{*} remains unchanged when it falls between the quantile limits $\hat{\eta}\,(\alpha_1)$ and $\hat{\eta}\,(\alpha_2)$. If it lies outside this range, the value is replaced with the corresponding quantile threshold.

2.4. The Problem of Smoking Intensity

Since 2021, West Java Province has consistently ranked among the top five provinces with the highest smoking prevalence in Indonesia, with rates generally above the national average [33]. The high prevalence of smoking in this region is strongly linked to low levels of education and to geographic factors that make tobacco products widely available. Efforts to enforce smoke-free area policies (Kawasan Tanpa Rokok, KTR) in urban areas have also encountered persistent difficulties [34].

In Indonesia, smoking cannot be viewed purely as an individual habit; it is deeply embedded within wider social and economic circumstances. In West Java, for instance, cigarette use tends to rise as village income grows, yet it remains high even when income falls. This pattern indicates that tobacco is often treated as a spending priority, taking precedence over education and healthcare needs [35]. The role of advertising adds to this problem. Data from the 2019 Global Youth Tobacco Survey (GYTS) show that more than 64% of Indonesian adolescents are exposed to cigarette promotions through television and social media, and this exposure has a marked influence on smoking behavior [36]. While smoke-free regulations have helped reduce smoking in several provinces, in West Java their enforcement has so far failed to bring about a meaningful decline [37].

2.5. Data Description

The empirical data come from West Java Province and draw on the 2021 National Socio-Economic Survey (SUSENAS) [38] and the 2021 Village Potential Statistics (PODES) [39]. The dataset is hierarchical with two linked levels. Level 1 comprises villages or administrative communities identified by the KODES identifier. Level 2 contains individuals (KODIN) nested within those villages. In total, the data include 21,290 individuals from 2,196 villages. The target variable (Y) in this study is the number of tobacco cigarettes consumed per week (in sticks) recorded for each individual. The explanatory variables analyzed are listed in full detail in Table 2.

Vol. 7, No. 2, April 2026, Page. 990-1006
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5407

Table 2. Description of Predictor Variables

No	Variable Code	Variable Description	Scale
1	X1	Presence of other household members who smoke	Nominal
2	X2	Highest educational attainment of the individual	Ordinal
3	X3	Marital status: never married	Nominal
4	X4	Marital status: divorced	Nominal
5	X5	Marital status: widowed	Nominal
6	X6	Employment status: working or not	Nominal
7	X7	School status: currently enrolled or not	Nominal
8	X8	Respondent's gender	Nominal
9	X9	Respondent's age at the time of the survey	Ratio
10	X10	Respondent's age at first marriage	Ratio
11	X11	Access to cell phone usage	Nominal
12	X12	Individual-level expenditure (in thousand rupiah)	Ratio
13	X13	Total Number of Health Insurance Owned	Ratio
14	X14	Reported Health Complaints	Nominal
15	X15	Urban or rural residence	Nominal

This study uses 15 predictors covering individual attributes and village-level context. Several categorical variables, including marital status and main activity, were recoded as dummy variables for use in the regression models. To avoid multicollinearity from the dummy-variable trap, one category per variable was set as the reference group. For marital status, three dummy variables were created: X3 (never married), X4 (divorced), and X5 (widowed). Each takes the value 1 if the condition applied to the respondent and 0 otherwise. Married individuals were used as the reference group. The main activity status was measured using X6, which equals 1 if the respondent was working at the time of the survey. X7 which equals 1 if the respondent was currently enrolled in school.

The variable X13 is a numerical variable representing the total number of health insurance policies owned, whereas X14 is a categorical (binary) variable coded as 1 for the presence of health complaints and 0 for the absence of complaints. Other variables are defined as follows: X1 indicates whether other household members (excluding the respondent) smoked, coded 1 for yes and 0 for no. X8 records the respondent's gender, coded 1 for male and 0 for female. X9 is the respondent's age in years at the time of the survey, and X10 is the age at first marriage. X11 denotes mobile-phone access, coded 1 if the respondent had access and 0 otherwise. Finally, X12 measures individual-level expenditure, expressed in thousands of rupiah. X2 representing the respondent's highest level of education attained, was recoded into an ordinal scale for analytical purposes. Originally consisting of 25 categories, education levels were grouped into four: code 1 for no formal education, code 2 for elementary and junior high school (or equivalent), code 3 for senior high school, diploma, and undergraduate degrees, and code 4 for postgraduate education, including master's, doctoral, and professional programs. The last one, X15 captures residential classification, with 1 indicating urban areas and 0 for rural areas.

2.6. Model Performance Metrics

Model performance was evaluated with Root Mean Square Error (RMSE). This statistic is the square root of the mean squared difference between predicted and observed values, so large errors weigh more heavily in the result. Because it is sensitive to outliers, RMSE is particularly appropriate for datasets with high variability or skewed distributions [40], [41]. In this study, RMSE was computed separately for the training and testing sets to evaluate accuracy at each stage. To further examine model

Vol. 7, No. 2, April 2026, Page. 990-1006
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5407

fit, the RMSE ratio (train/test) was calculated, with values close to one indicating stability and good generalization.

Figure 1 illustrates the step-by-step procedure implemented in this study, starting from data preprocessing (standardization and winsorization), model training (GLMM-NB and MERF), and ending with model evaluation using RMSE, MAE, and prediction ratio.

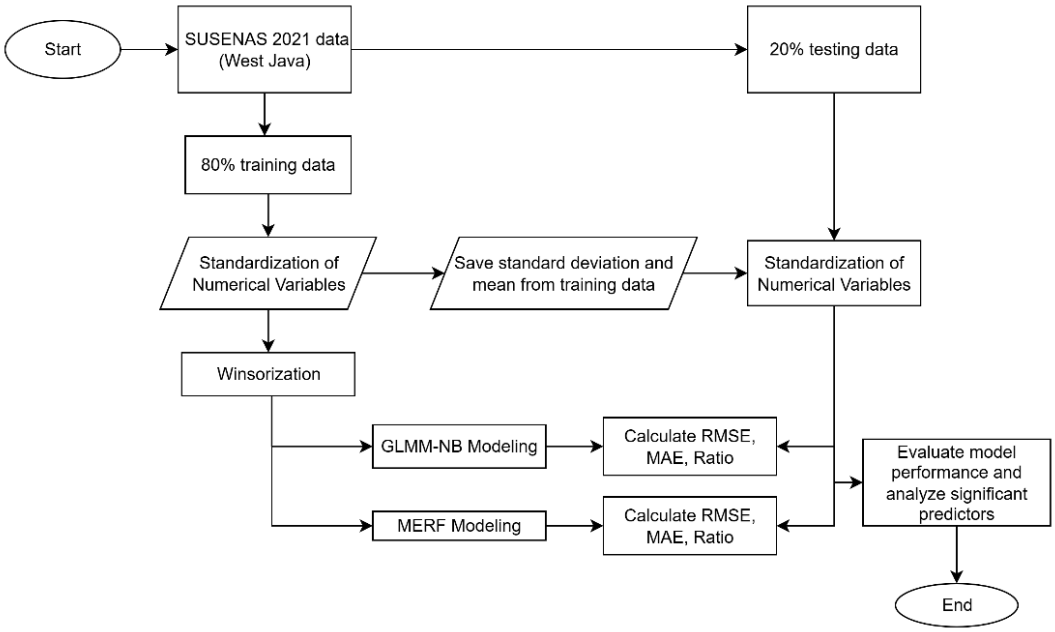
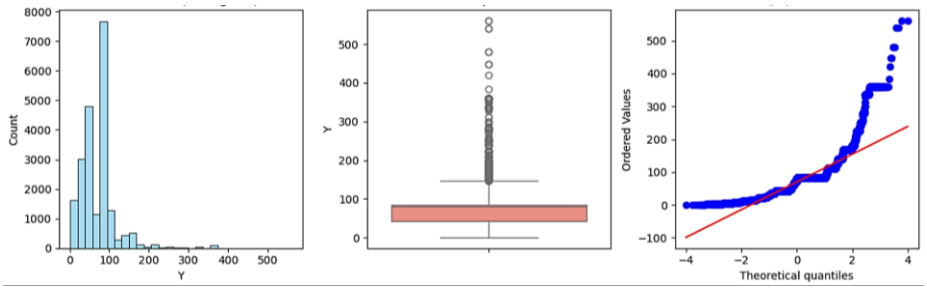


Figure 1. Flowchart of the research stages

3. RESULT

3.1. Exploratory Analysis

The analysis began with an examination of the target variable (Y) to describe its distribution, identify possible outliers, and assess features such as skewness and overdispersion. This step offered an initial overview of the data and guided subsequent decisions about model specification and potential transformations.



(a) (b) (c)

Figure 2. Histogram of the response variable Y (a), Boxplot of Y (b), and Q-Q plot of Y (c)

As illustrated in Figure 2, the response variable *Y* (number of cigarettes smoked) is not symmetrically distributed. The histogram and boxplot show a right-skew with several outliers in the upper tail, while the Q-Q plot departs clearly from the diagonal, especially at higher values. Normality tests using Shapiro-Wilk and D'Agostino confirm this impression, both giving p-values below 0.001.

Vol. 7, No. 2, April 2026, Page. 990-1006
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5407

The Intraclass Correlation Coefficient (ICC) is 0.2429, which means that about 24.29% of the variation in Y comes from differences between KODES groups. This points to a strong hierarchical structure in the data and justifies the use of mixed-effects models such as the GLMM.

Table 3. Likelihood Ratio Test (LRT)

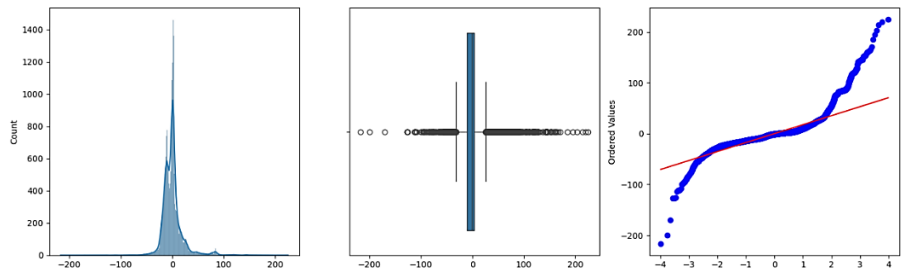
Model	logLik	AIC	Deviance
model_fixed	-85766	171569	171531
model_full	-84896	169833	169793

The Likelihood Ratio Test (LRT) was carried out to compare a fixed-effects-only model with a full specification that also included random effects , as summarized in Table 3. The difference in log-likelihoods produced an LRT value of 1,738.6 with 1 degree of freedom, yielding a p-value smaller than 2.2e-16. This shows that the improvement from adding random effects is statistically significant. The full specification provides a noticeably better fit than the fixed-effects model, which is consistent with the ICC result reported earlier. On this basis, the use of a GLMM with random effects at the village level (KODES) is well supported, capturing the hierarchical nature of the data and strengthening model validity.

An examination of the response variable, weekly tobacco consumption, gave a mean of 70.20 and a variance of 2,220.58. From these values, the dispersion ratio was calculated as:

$$\text{Dispersion Ratio} = \frac{Var(Y)}{Mean(Y)} = \frac{2220.58}{70.20} \approx 31{,}65$$

This result indicates that the variance far exceeds the mean, which is a statistical indication of overdispersion in the data. Therefore, the use of the Negative Binomial distribution, particularly with the nbinom2 parameterization, is more appropriate, as it allows greater flexibility in modeling overdispersed count data through the dispersion parameter [20]. This confirms the methodological justification for employing the GLMM-NB model as the primary modeling approach.



(a) (b) (C)

Figure 3. Histogram of residuals (a), Boxplot of residuals (b), and Q-Q plot of residuals (c)

The residuals did not follow a normal distribution. Both the histogram and the boxplot showed an uneven spread, and the Q-Q plot bent away from the diagonal line, especially in the tails (Figure 3a–c). Formal tests of normality, Shapiro-Wilk and D'Agostino, both produced p-values below 0.001, confirming that the residuals are not normally distributed. This further reinforces the presence of a strong hierarchical structure, thereby supporting the use of mixed-effects models.

3.2. Data Preprocessing

The preprocessing was carried out in two main stages, namely data splitting and numerical standardization. In the preprocessing stage, the dataset was first divided into training (80%) and testing (20%) subsets using stratified sampling. This method guaranteed that all groups represented in the test

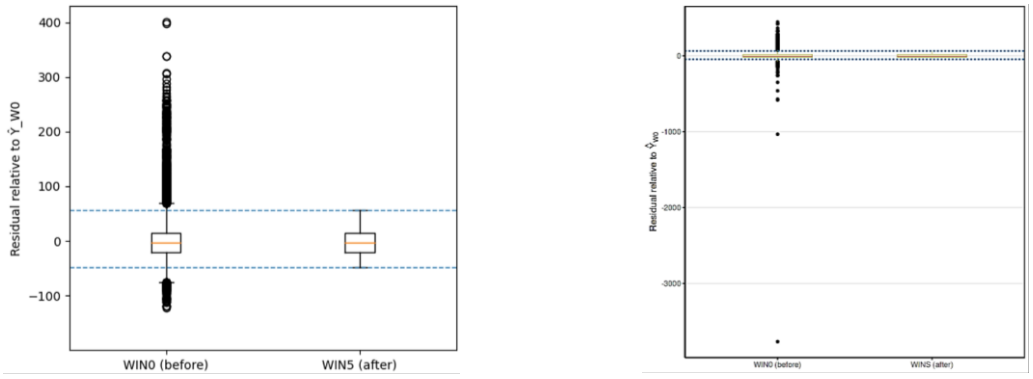
Vol. 7, No. 2, April 2026, Page. 990-1006
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5407

data were also present in the training data, meaning that the separation took place only at the individual level. By doing so, the hierarchical nature of the dataset was maintained and the risk of clusters appearing only in the test set was avoided.

The next stage addressed the standardization of numerical variables. The variables age (X9), age at first marriage (X10), household expenditure (X12), and number of insurance policies (X13) were scaled with the StandardScaler. The mean and standard deviation were derived from the training data and subsequently applied to the test data. This procedure kept the scales aligned across both datasets and prevented any leakage of information from the test set into the training process.

3.3. Winsorization Results

A separate analysis was conducted to evaluate the effect of winsorization on residual distributions. The residuals were first obtained from the initial models without winsorization (WIN0), and then the winsorization procedure was applied, as residuals can only be calculated after model estimation. In this process, the residuals from the test set were capped using the thresholds determined from the training set, thereby preventing information leakage.



(a) (b)

Figure 4. Boxplots of Residuals Before and After Winsorization: (a) MERF, (b) GLMM-NB

The results show that winsorization reduced the impact of extreme residuals in both MERF and GLMM-NB. In Figure 4a, MERF residuals that were widely spread at WIN0 became more concentrated after WIN5, with most extreme values at both ends diminished. A similar pattern appears in Figure 4b, where GLMM-NB residuals, which had shown large deviations and several points far from the median line, were compressed into a narrower range after winsorization. These changes show that applying WIN5 reduced the influence of outliers, resulting in residual distributions that are more stable and easier to interpret.

3.4. Descriptive Performance of the Models

Table 4. Descriptive Performance Metrics of MERF and GLMM-NB Across Winsorization Levels

Perlakuan	RMSE *Train*		RMSE *Test*		RMSE Ratio	
	GLMM	MERF	GLMM	MERF	GLMM	MERF
WIN0	49.6500	35.9600	42.2700	39.9400	0.8510	1.1110
WIN5	34.9000	26.6300	30.2000	28.6400	0.8650	1.0750

During training, MERF consistently produced more accurate results, with lower RMSE values than GLMM-NB across both winsorization levels (Table 4). At WIN0, MERF recorded a training RMSE of 35.96, compared with 49.65 for GLMM-NB. A similar pattern was found at WIN5, where MERF

Vol. 7, No. 2, April 2026, Page. 990-1006
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5407

declined to 26.63, while GLMM-NB dropped to 34.90. Although the relative reduction was greater for GLMM-NB, MERF still produced lower absolute errors, highlighting its stronger fit to the training data. At the same time, this outcome suggests that MERF tends to adhere more closely to the training set, raising the possibility of mild overfitting.

On the test data, both models showed improvement after winsorization, with errors decreasing overall. For GLMM-NB, the RMSE declined from 42.27 at WIN0 to 30.20 at WIN5, while MERF dropped from 39.94 to 28.64 over the same range. MERF remained slightly ahead at each level, but the gap between the two was narrow, implying that its advantage is relative rather than absolute.

Looking at the RMSE Test/Train ratio, GLMM-NB displayed greater balance, remaining close to 1 across conditions, which reflects its stronger ability to maintain equilibrium between training and testing. By contrast, MERF consistently showed ratios above 1, ranging from 1.11 (WIN0) to 1.08 (WIN5). Although the ratios improved with stronger winsorization, they still revealed a gap between the very low training error and the higher test error.

3.5. Identification of Important Variables

In the MERF framework, variable importance was evaluated using the Mean Decrease in Impurity (MDI). This measure reflects the average drop in impurity, such as variance in regression, each time a variable is used to split the data at a decision node in the forest. From a statistical perspective, it can be understood as a breakdown of the output variance, offering a clear and interpretable gauge of how much a variable contributes [36]. A higher MDI value signals that the variable plays a stronger role in improving model performance. This metric enables relative interpretation of feature influence in non-parametric models such as MERF, where conventional coefficient-based interpretation is not applicable [42].

Table 5. Important Variables in MERF

Ranked Value	Variable kode	MDI Score
1	X12	0.5871
2	X9	01417
3	X6	0.1268
4	X10	0.0389
5	X8	0.0378
6	X2	0.0162
7	X1	0.0161
8	X13	0.0089
9	X11	0.0059
10	X15	0.0052
11	X14	0.0049
12	X5	0.0041
13	X4	0.0034
14	X7	0.0021
15	X3	0.0009

The MERF model with 5% winsorization (WIN5) revealed that prediction was dominated by a few variables, as shown in Table 5. Household expenditure (X12, 0.5871) was by far the strongest predictor, followed by age (X9, 0.1417) and employment status (X6, 0.1268), both of which

Vol. 7, No. 2, April 2026, Page. 990-1006
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5407

substantially influenced smoking intensity. Among the additional predictors, age at first marriage (X10, 0.0389) and gender (X8, 0.0378) also contributed, though their influence was modest. Variables such as X2, X1, X13, X11, and X15 had smaller contributions, all below 0.02, while X14, X5, X4, X7, and X3 recorded very low values, at or under 0.005. This pattern suggests that, beyond the top five predictors, most variables had only a weak connection to smoking intensity and functioned largely as supporting factors.

In the GLMM, variable importance was determined from the fixed-effect coefficients. These coefficients describe both the direction and magnitude of each predictor's effect on the outcome and are evaluated for significance using p-values or confidence intervals. A variable is regarded as important if its effect is statistically significant, which makes this approach particularly useful for hierarchical data that incorporate both fixed and random effects.

Table 6. ANOVA Fixed Effect GLMM

Variable code	estimate	p-value
X15	-0.0601	<0.0000
X8	0.3861	<0.0000
X9	-0.0389	<0.0000
X10	-0.0163	0.1240
X11	0.0938	<0.0000
X13	-0.0071	0.1070
X14	-0.0355	<0.0000
X1	0.1058	<0.0000
X2	-0.0824	0.00114
X3	-0.2127	<0.0000
X4	-0.0971	<0.0000
X5	-0.1122	<0.0000
X6	0.2146	<0.0000
X7	-0.3449	<0.0000
X12	0.1600	<0.0000

Table 6 shows that for the GLMM estimated with 5 percent winsorization (WIN5), four of the five leading predictors identified by MERF were also significant. Household expenditure (X12, $p < 0.0000$) showed a positive effect, age (X9, $p < 0.0000$) had a negative effect, and both employment status (X6, $p < 0.0000$) and gender (X8, $p < 0.0000$) showed positive effects. In contrast, age at first marriage (X10) was not significant ($p = 0.1240$), suggesting a weaker role in explaining smoking intensity.

4. DISCUSSIONS

The findings of this study provide a nuanced perspective on modeling tobacco consumption using hierarchical health survey data. Across all levels of winsorization, the Mixed Effects Random Forest (MERF) consistently achieved higher predictive accuracy, as reflected in lower RMSE values for both training and test datasets. However, the Test/Train RMSE ratio indicated that GLMM-NB maintained a more stable balance between training and test performance, whereas MERF exhibited early signs of mild overfitting. Overall, these results emphasize a trade-off: MERF is preferable when predictive precision is the primary goal, while GLMM-NB is more suitable for ensuring generalization stability across datasets.

Vol. 7, No. 2, April 2026, Page. 990-1006
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5407

Winsorization proved effective in reducing prediction errors for both models, but its impact was particularly pronounced for GLMM-NB. This suggests that GLMM-NB is more sensitive to extreme residual values, while MERF, due to its tree-based structure, is inherently robust to outliers. This pattern aligns with prior studies demonstrating the resilience of Random Forest–based models to extreme values[13], [16], [42]. The negative correlation observed between the percentage of winsorization and evaluation metrics reinforces the practical utility of winsorization, particularly for parametric approaches in hierarchical health data [20].

Variable importance analyses further illustrate the complementary insights offered by the two modeling approaches. As shown in Table 7, both MERF and GLMM-NB consistently identified household expenditure (X12), employment status (X6), gender (X8), and age (X9) as the strongest predictors of tobacco consumption intensity. In MERF, these predictors carried the highest Mean Decrease in Impurity (MDI), whereas in GLMM-NB they were statistically significant, with positive effects for expenditure, gender, and employment status, and a negative effect for age. The alignment between these methods reinforces the conclusion that economic and demographic factors play a dominant role in shaping smoking behavior, while other variables have a more secondary influence. These findings are consistent with previous studies on tobacco consumption in Indonesia, which highlight the influence of economic capacity, age, gender, and employment on smoking behavior [25], [33], [34], [35], [36], [37]

Table 7. Synthesis of Important Predictors Across Models

Rank	Variable	MERF (MDI)	GLMM-NB (Estimate / p-value)
1	X12 (Household expenditure)	0.5871	0.1600 / <0.0001
2	X9 (Age)	0.1417	-0.0389 / <0.0001
3	X6 (Employment status)	0.1268	0.2146 / <0.0001
4	X8 (Gender)	0.0378	0.3861 / <0.0001

Note: Only the four top predictors significant in both models are included for clarity.

A synthesis of recent literature on hierarchical and mixed-effects modeling is summarized in Table 8, highlighting both methodological advances and alignment with health-related findings.

Table 8. SOTA Comparison of Hierarchical and Mixed Effects Modeling Studies

Year	Study & Context	Key Findings / Relevance
2001	Breiman,L. “Random Forests”	Introduced Random Forest; robust to outliers, non-parametric, foundational for MERF.
2021	Mayapada et al.	MERF outperformed RF in hierarchical educational data; emphasizes accuracy vs. interpretability.
2023	Lee et al.	GLMM-NB sensitive to outliers; winsorization improved performance for zero-inflated count data.
2024	Heiling et al.	Efficient high-dimensional penalized GLMM estimation; highlights scalability for complex datasets.
2024	Ananda et al.	Modified MERF with PCA and rotation forest; improved robust prediction in hierarchical data.
2024	Abuzaid & Alkrunz	Evaluated winsorization methods; supports preprocessing for robust modeling in MERF and GLMM.

Vol. 7, No. 2, April 2026, Page. 990-1006
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5407

Year	Study & Context	Key Findings / Relevance
2025	Olaniran et al.	Mixed effect gradient boosting for high-dimensional longitudinal data; modern hybrid ML-statistical approach.
2021–2022	Baktiar & Utiayarsih [33], Fahmi [34], Diniyati & Achmad [35]	Household expenditure, gender, age, and employment significantly influence smoking prevalence in West Java and forest households. Supports predictors identified by MERF and GLMM-NB.

An important limitation of this study is its reliance on a single dataset and the use of only two-sided winsorization. Future research could explore alternative strategies for handling outliers, such as trimming, robust regression, or adaptive winsorization. Expanding the analysis to other provinces and multiple survey years would further validate generalizability and provide stronger guidance for policy and health interventions [32], [33].

4.1 Implications for Computer Science

From an informatics perspective, this study demonstrates the practical value of integrating machine learning and statistical approaches in analyzing complex social and health data. By combining these paradigms, analysts can effectively leverage hierarchical information while maintaining strong predictive performance. MERF is particularly suitable when prediction accuracy is the primary goal, whereas GLMM-NB offers interpretability and stable generalization, which are crucial for decision-making in public health. Furthermore, preprocessing strategies such as winsorization enhance the robustness and reliability of both models, especially when dealing with datasets that contain outliers. Overall, these findings emphasize that model selection should be guided not only by predictive objectives but also by practical considerations in big data informatics, including scalability, robustness to outliers, and interpretability [3], [20], [43]. In conclusion, model selection for hierarchical health data should consider both research priorities and practical constraints: MERF for precision in prediction, GLMM-NB for stability in generalization, and careful preprocessing to mitigate the influence of extreme values. Integrating these approaches provides a bridge between statistical rigor and computational scalability, with clear implications for policy analysis and health informatics.

5. CONCLUSION

MERF consistently achieved lower RMSE values than GLMM-NB across all winsorization levels (e.g., 28.64 vs. 30.20 under WIN5), confirming its superior predictive accuracy. Winsorization effectively reduced prediction errors in both models, with a more pronounced effect in GLMM-NB, indicating this model's higher sensitivity to residual outliers. Overall, MERF is recommended when maximizing predictive precision is the main goal, whereas GLMM-NB is more suitable for achieving robust performance when data irregularities are present. Both models identified household expenditure (X12), age (X9), employment status (X6), and gender (X8) as consistent and dominant predictors of smoking intensity, aligning with existing tobacco-related literature. These results contribute to the development of an analytical framework that supports SDG 3 (Good Health and Well-Being) through evidence-based policy design in developing countries. Future research may extend this framework by incorporating deep mixed models and robust regression techniques for multi-provincial or longitudinal datasets, thereby strengthening computational approaches to overdispersed health data in informatics and public health analytics.

Vol. 7, No. 2, April 2026, Page. 990-1006
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5407

CONFLICT OF INTEREST

The authors declare that they have no conflict of interest regarding the authorship or the subject matter of this paper.

REFERENCES

[1] Y. Angraini, K. A. Notodiputro, A. Saefuddin, dan T. Toharudin, "Latent factor linear mixed model (LFLMM) for modelling flanders data," *Communications in Mathematical Biology and Neuroscience*, vol. 2020, hlm. 1–14, Mei 2020, doi: 10.28919/cmbn/4610.

[2] B. Suseno, K. A. Notodiputro, dan B. Sartono, "GLMMTree for Modelling Poverty in Indonesia."

[3] H. M. Heiling, N. U. Rashid, Q. Li, X. L. Peng, J. J. Yeh, dan J. G. Ibrahim, "Efficient Computation of High-Dimensional Penalized Generalized Linear Mixed Models by Latent Factor Modeling of the Random Effects," Apr 2024, [Daring]. Tersedia pada: http://arxiv.org/abs/2305.08201

[4] W. W. Stroup, *Generalized Linear Mixed Models Modern Concepts, Methods and Applications*, 1st edition. New York: CRC Press Taylor & Francis Group, 2016. doi: https://doi.org/10.1201/b13151.

[5] J. Salinas, R. Osval, A. Montesinos López, G. Hernández, R. Jose, dan C. Hiriart, "Generalized Linear Mixed Models with Applications in Agriculture and Biology."

[6] K. H. Lee, C. Pedroza, E. B. C. Avritscher, R. A. Mosquera, dan J. E. Tyson, "Evaluation of negative binomial and zero-inflated negative binomial models for the analysis of zero-inflated count data: application to the telemedicine for children with medical complexity trial," *Trials*, vol. 24, no. 1, Des 2023, doi: 10.1186/s13063-023-07648-8.

[7] D. A. N. Sirodj, K. Sadik, dan A. Kurnia, "Modeling The Incidence of Malnutrition in Bogor Regency using Zero-Inflated Negative Binomial Mixed Effect Model," *BAREKENG: Jurnal Ilmu Matematika dan Terapan*, vol. 18, no. 2, hlm. 0961–0972, Mei 2024, doi: 10.30598/barekengvol18iss2pp0961-0972.

[8] P. R. Sihombing, K. A. Notodiputro, dan B. Sartono, "Comparison of GEE and GLMM Methods for Longitudinal Data (Case Study: Determinants of the Percentage of Poor People in Indonesia, 2015-2019)," dalam *AIP Conference Proceedings*, American Institute of Physics Inc., Okt 2022. doi: 10.1063/5.0103254.

[9] A. Hajjem, F. Bellavance, dan D. Larocque, "Mixed-effects random forest for clustered data," *J Stat Comput Simul*, vol. 84, no. 6, hlm. 1313–1328, 2014, doi: 10.1080/00949655.2012.741599.

[10] P. Krennmair dan T. Schmid, "Flexible domain prediction using mixed effects random forests," *J R Stat Soc Ser C Appl Stat*, vol. 71, no. 5, hlm. 1865–1894, 2022, doi: 10.1111/rssc.12600.

[11] R. A. Lewis, A. Ghandeharioun, S. Fedor, P. Pedrelli, R. Picard, dan D. Mischoulon, "Mixed Effects Random Forests for Personalised Predictions of Clinical Depression Severity," 2023, [Daring]. Tersedia pada: http://arxiv.org/abs/2301.09815

[12] A. Fakhrurrozi, "On The use of Mixed Effects Machine Learning Regression Models to Capture Spatial Patterns: A Case Study on Crime," Master Thesis, University of Twente, Enschede, 2019.

[13] R. Mayapada, B. Susetyo, dan B. Sartono, "A Comparison between Random Forest and Mixed Effects Random Forest to Predict Students ' Math Performance in Indonesia," *International Journal of Sciences: Basic and Applied Research (IJSBAR)*, vol. 57, hlm. 1–8, Mar 2021.

[14] R. Anisa, A. Kurnia, dan I. Indahwati, "Cluster Information of Non-Sampled Area In Small Area Estimation," *IOSR Journal of Mathematics*, vol. 10, no. 1, hlm. 15–19, 2014, doi: 10.9790/5728-10121519.

[15] F. Zubedi, B. Sartono, dan K. Anwar, "Jurnal Natural," vol. 22, no. 2, hlm. 108–116, 2022, doi: 10.24815/jn.v22i2.25499.

[16] R. Ananda, K. A. Notodiputro, dan M. N. Aidi, "Modified Mixed Effects Random Forest in Small Area Estimation Using PCA and Rotation Forest with Correlated Auxiliary Variables," *Scientific Journal of Informatics*, vol. 11, no. 3, hlm. 705–720, Agu 2024, doi: 10.15294/sji.v11i3.10633.

Vol. 7, No. 2, April 2026, Page. 990-1006
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5407

[17] J. Shi, "Investigating Mixed Effects Random Forest Models in Predicting Investigating Mixed Effects Random Forest Models in Predicting Satisfaction with Online Learning in Higher Education Satisfaction with Online Learning in Higher Education." [Daring]. Tersedia pada: https://digitalcommons.du.edu/etd

[18] P. J. Huber dan J. Wiley, "Robust statistics," *Data Handling in Science and Technology*, vol. 20, no. PART A, hlm. 339–377, 1981, doi: 10.1016/S0922-3487(97)80042-1.

[19] R. R. Wilcox, *Introduction to robust estimation and hypothesis testing*, 3rd ed. Amsterdam: Academic Press, 2012.

[20] A. Abuzaid dan I. Alkrunz, "A COMPARATIVE STUDY ON UNIVARIATE OUTLIER WINSORIZATION METHODS IN DATA SCIENCE CONTEXT," *Statistica Applicata*, vol. 36, no. 1, hlm. 85–99, Jul 2024, doi: 10.26398/IJAS.0036-004.

[21] M. Atif, M. Farooq, M. Shafiq, T. Alballa, S. Abdualziz Alhabeeb, dan H. Abd El-Wahed Khalifa, "Uncovering the impact of outliers on clusters' evolution in temporal data-sets: an empirical analysis," *Sci Rep*, vol. 14, no. 1, Des 2024, doi: 10.1038/s41598-024-75928-7.

[22] C. Lartey, J. Liu, R. K. Asamoah, C. Greet, M. Zanin, dan W. Skinner, "Effective Outlier Detection for Ensuring Data Quality in Flotation Data Modelling Using Machine Learning (ML) Algorithms," *Minerals*, vol. 14, no. 9, Sep 2024, doi: 10.3390/min14090925.

[23] A. A. Mangino, J. H. Bolin, dan W. H. Finch, "Fixed Effects or Mixed Effects Classifiers? Evidence From Simulated and Archival Data," *Educ Psychol Meas*, vol. 83, no. 4, hlm. 710–739, Agu 2023, doi: 10.1177/00131644221108180.

[24] O. R. Olaniran, S. F. Olaniran, J. Allohibi, A. A. Alharbi, dan N. M. S. Alharbi, "Mixed effect gradient boosting for high-dimensional longitudinal data," *Sci Rep*, vol. 15, no. 1, Des 2025, doi: 10.1038/s41598-025-16526-z.

[25] *WHO global report on trends in prevalence of tobacco use 2000-2025 Fourth edition WHO global report on trends in prevalence of tobacco use 2000-2025, fourth edition ISBN 978-92-4-003932-2 (electronic version)*. 2021. [Daring]. Tersedia pada: http://apps.who.int/bookorders.

[26] A. E. Yunianto, N. A. Q. A'yunin, E. Emy Yuliantini, M. Haya, dan A. Faridi, "Knowledge and Healthy Behavior of the West Java People Related to COVID-19 Pandemic," *Annals of Tropical Medicine and Public Health* ·, vol. Volume 7, no. Issue 6, 2021, doi: 10.36295/AOTMPH.2021.7602.

[27] C. E. McCulloch, S. R. Searle, dan J. M. Neuhaus, *Generalized, Linear, and Mixed Models*, 2 ed. New Jersey: John Wiley & Sons, 2008.

[28] J. M. . Hilbe, *Modeling count data. Hilbe (Arizona State University and Jet Propulsion Laboratory, California Institute of Technology)*. Cambridge University Press, 2014.

[29] M. E. Brooks *dkk.*, "glmmTMB balances speed and flexibility among packages for zero-inflated generalized linear mixed modeling," *R Journal*, vol. 9, no. 2, hlm. 378–400, Des 2017, doi: 10.32614/rj-2017-066.

[30] A. H. Welsh, *Approaches to the robust estimation of mixed models,* vol. 13. Amsterdam: Elsevier Science, 1997.

[31] V. Barnett dan Lewis T., *Outliers in Statistical Data*, 3rd ed., vol. XVII. Boston: J. Wiley & Sons, 1994.

[32] M. Hubert dan S. Van Der Veeken, "Outlier detection for skewed data," dalam *Journal of Chemometrics*, John Wiley and Sons Ltd, 2008, hlm. 235–246. doi: 10.1002/cem.1123.

[33] A. F. Baktiar dan T. S. Utiayarsih, "Identification of Factors Affecting Smoking Prevalence in West Java using Spatial Modeling," *Indonesian Journal of Statistics and Its Applications*, vol. 6, no. 1, hlm. 114–131, 2022, doi: 10.29244/ijsa.v6i1p114-131.

[34] M. A. Fahmi, "Correlation Between Smoke-Free Areas and Smoking Behavior in Indonesia," *Jurnal Berkala Epidemiologi*, vol. 8, no. 2, hlm. 117, 2020, doi: 10.20473/jbe.v8i22020.117-124.

[35] Dian Diniyati dan Budiman Achmad, "Tobacco use and its impact on poverty among forest households: The cases of Indonesia," *World Journal of Biology Pharmacy and Health Sciences*, vol. 11, no. 3, hlm. 060–066, 2022, doi: 10.30574/wjbphs.2022.11.3.0139.

Vol. 7, No. 2, April 2026, Page. 990-1006
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.2.5407

[36] R. Fauzi, I. Arumsari, M. A. Maruf, dan A. Ahsan, "Association of Tobacco Advertising, Promotion, and Sponsorship (TAPS) exposure on smoking intention and current smoking behavior among youth in Indonesia," *J Subst Use*, vol. 29, no. 1, hlm. 54–60, Sep 2022.
[37] W. Septiono, M. A. G. Kuipers, N. Ng, dan A. E. Kunst, "The impact of local smoke-free policies on smoking behaviour among adults in Indonesia: a quasi-experimental national study," *Addiction*, vol. 115, no. 12, hlm. 2382–2392, 2020, doi: 10.1111/add.15110.
[38] Badan Pusat Statistik (BPS), "Data Mikro Survei Sosial Ekonomi Nasional (SUSENAS) Jawa Barat 2021," Jawa Barat , 2021.
[39] Badan Pusat Statistik (BPS), "Data Potensi Desa (PODES) 2021 Provinsi Jawa Barat," Jawa Barat, 2021.
[40] T. Chai dan R. R. Draxler, "Root mean square error (RMSE) or mean absolute error (MAE)?," 28 Februari 2014. doi: 10.5194/gmdd-7-1525-2014.
[41] T. O. Hodson, "Root-mean-square error (RMSE) or mean absolute error (MAE): when to use them or not," 19 Juli 2022, *Copernicus GmbH*. doi: 10.5194/gmd-15-5481-2022.
[42] L. Breiman, "Random Forests," *Mach Learn*, vol. 45, hlm. 5–32, 2001, doi: 10.1023/A:1010933404324.
[43] Q. Zou, B. Chen, Y. Zhang, X. Wu, Y. Wan, dan C. Chen, "Mixed-effects neural network modelling to predict longitudinal trends in fasting plasma glucose," *BMC Med Res Methodol*, vol. 24, no. 1, Des 2024, doi: 10.1186/s12874-024-02442-9.