STATISTICS, OPTIMIZATION AND INFORMATION COMPUTING
Stat., Optim. Inf. Comput., Vol. **x**, Month 202x, pp 0–14.
Published online in International Academic Press (www.IAPress.org)

Boosting mixed-effects models with SMOTE: insights from Java's Human Development Index

Dimas Anggara [1,2,*], Anang Kurnia [1], Khairil Anwar Notodiputro [1], Indahwati [1]

[1] *School of Data Science, Mathematics and Informatics, IPB University, Bogor, Indonesia*
[2] *Directorate of Price Statistics, BPS Statistics Indonesia, Jakarta, Indonesia*

Abstract This study aims to evaluate the performance of various regression models in unbalanced and clustered data, using the 2018 Human Development Index (HDI) data for regencies in Java Island, Indonesia, as a case study. The models assessed include Linear Mixed Models (LMM), Generalized Estimating Equations (GEE), Mixed-Effects Regression Trees (MERT), and Gaussian Copula Marginal Regression (GCMR). These models share a common foundation in incorporating random effects, allowing for a fair and systematic comparison. The performance of the model was evaluated using two key metrics: The median absolute error (MedAE) and the root mean square error (RMSE), applied to both the original data set and an oversampled version generated using the Synthetic Minority Oversampling Technique (SMOTE). The results indicate that the application of SMOTE consistently improves the accuracy of the model. MERT achieved the lowest MedAE in both datasets, demonstrating superior capability in minimizing median prediction errors. Meanwhile, GCMR produced the best RMSE on the original data, highlighting its robustness in handling complex data structures without requiring oversampling. Residual analysis using boxplots further supports these findings, showing that SMOTE effectively reduces residual variability and improves model stability. Among the models evaluated, MERT exhibited the most consistent overall performance. These findings underscore the utility of oversampling techniques such as SMOTE in improving regression model performance on unbalanced and hierarchically structured data. Furthermore, both MERT and GCMR are identified as strong candidates for such analytical scenarios, contributing valuable insights toward developing more robust and accurate predictive models in data science and applied statistics.

Keywords copula, human development index, mixed models, oversampling, unbalanced

MSC 2020 subject classifications 62G09, 62J05, 62P25

DOI: 10.19139/soic-2310-5070-3011

1. Introduction

Clustered data refers to data that is organized into groups or clusters in which observations within the same group tend to be correlated or dependent on one another. This intra-cluster dependence often arises due to unobserved factors that similarly affect all observations within a group. Unlike independent data, where each observation is assumed to be unaffected by others, clustered data requires special analytical approaches. Classic examples of clustered data span various domains. In healthcare, for instance, patients treated in the same hospital may share similar characteristics due to environmental or institutional factors. In education, students within the same classroom may exhibit similarities influenced by teaching styles or shared social dynamics. In industrial and economic settings, branches of the same company operating in a specific geographic region may encounter similar market conditions, resulting in dependency across observations.

[*] Correspondence to: Dimas Anggara (Email: anggaradimas@apps.ipb.ac.id). School of Data Science, Mathematics and Informatics. IPB University. Meranti Road, Bogor, Jawa Barat Province, Indonesia (16680).

ISSN 2310-5070 (online) ISSN 2311-004X (print)

Traditional linear modeling approaches, such as simple or multiple linear regression, assume independence among observations. This assumption is violated in clustered data settings, leading to biased and inefficient estimates if not properly addressed. One widely adopted method to account for intra-cluster correlation is the Generalized Estimating Equations (GEE) approach [1]. GEE enables consistent estimation of regression parameters while accommodating correlated responses within clusters, without requiring a precise specification of the variance-covariance structure. This robustness has made GEE a popular choice in medical and social research, particularly for longitudinal and hierarchical data structures.

Recent advances in machine learning have introduced innovative methods for analyzing clustered data. One notable approach is the Mixed Effect Regression Tree (MERT), which integrates the interpretability of decision trees with the modeling power of linear mixed models [2]. While traditional decision trees partition data based on input variables to form homogeneous subgroups, MERTs extend this by incorporating random effects, allowing them to handle within-cluster correlation while retaining intuitive tree-based visualization [3]. Studies have demonstrated the effectiveness of MERT in medical domains, such as predicting disease risk while accounting for dependencies among patients in the same healthcare facility.

Another powerful framework for modeling dependencies in multivariate data is copula-based modeling. A copula is a function that links the marginal distributions of random variables to form a joint multivariate distribution. Introduced by Sklar in 1959, copulas offer a flexible mechanism to capture linear and non-linear dependencies without assuming specific marginal distributions [4]. Various copula families—such as Gaussian, t, and Archimedean copulas—enable the modeling of different types of dependency structures. In practice, copulas have found success in fields such as finance [5], where they are used to model the interdependence between asset prices, especially when traditional linear correlation fails to capture complex relationships. In clustered data contexts, copulas can be employed to model the joint distribution across dependent observations by combining marginal distributions into a comprehensive multivariate structure. This enables accurate risk assessment and prediction in domains such as economics, epidemiology, and engineering [6, 7, 8, 9]. The Gaussian Copula Marginal Regression (GCMR) method combines copula theory with marginal regression modeling to capture complex dependencies in clustered data [10]. GCMR estimates the marginal distributions of the dependent and independent variables separately and then links them using a Gaussian copula. This approach offers greater flexibility in handling non-normal distributions and capturing nonlinear dependencies compared to traditional linear models [11]. GCMR has shown improved predictive performance, particularly in cases where clustered data exhibits intricate dependency structures not well modeled by linear assumptions.

A common challenge in real-world clustered datasets is the imbalance in cluster sizes or class distributions, where one category may dominate the data, leading to biased models that perform poorly in minority classes [12]. This imbalance often skews predictive performance toward majority clusters or classes, diminishing the model's generalizability. Several strategies have been proposed to address data imbalance. Among them, oversampling techniques, such as the Synthetic Minority Oversampling Technique (SMOTE), artificially generate synthetic data for underrepresented classes to balance the dataset. Conversely, under-sampling reduces the size of majority classes, which, while effective, risks discarding valuable information [13]. In this study, only oversampling is applied due to the limited number of observations (119), as under-sampling could result in significant loss of information.

This study aims to compare the effectiveness of the Gaussian Copula Marginal Regression (GCMR) method in modeling unbalanced and clustered data with other established approaches such as Generalized Estimating Equations (GEE) and Mixed Effect Regression Tree (MERT). Through this comparison, the study seeks to provide more insight into the strengths and limitations of each approach in predicting outcomes from unbalanced and clustered datasets. By combining mixed-model-based regression techniques with the SMOTE oversampling method, the goal is to develop a model that is both flexible and accurate in handling intra-cluster dependencies and data imbalance. This, in turn, is expected to improve predictive performance and broaden applicability across various domains such as public health, economics, and finance. The proposed methods will be applied to Human Development Index (HDI) data from regencies on Java Island, Indonesia.

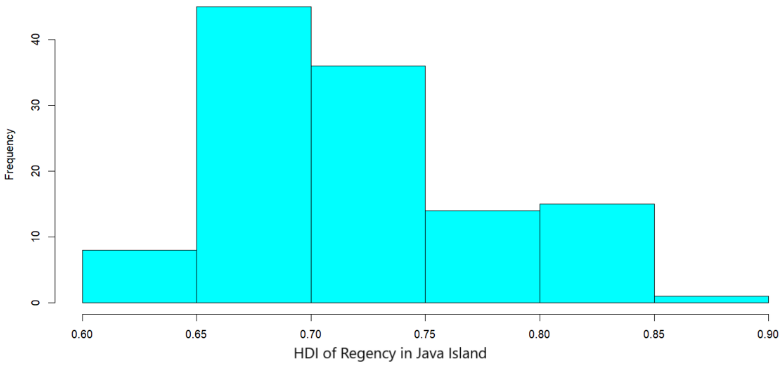


Figure 1. Histogram of Regency Human Development Index in Java Island 2018

2. Data and Methods

2.1. Data

This study focuses on the analysis of the Human Development Index (HDI) as the dependent variable, which is modeled in relation to a set of independent variables derived from the Village Potential Survey (PODES). The unit of analysis in this study is the regencies located on the island of Java, Indonesia. The data set includes administrative regions at the regency level, which exhibit various socioeconomic, cultural, and infrastructure characteristics that influence human development in the region.

The Human Development Index (HDI) serves as an indicator for measuring the quality of life in each area, based on three core dimensions of development: health, education, and standard of living. HDI is a key metric used to evaluate the extent to which a region can provide well-being for its population. The HDI data used in this study were obtained from Statistics Indonesia, BPS. Each Java regency has an HDI score derived from the measurement of the following three main components:

- Health: Measured by life expectancy at birth, reflecting the overall health conditions of the population.
- Education: Measured by the mean years of schooling and expected years of schooling, indicating the quality of and access to education in the region.
- Standard of Living: Measured by per capita income, representing the economic well-being of residents in each regency.

The HDI ranges from 0 to 1, where higher values indicate better human development outcomes. For instance, regions with higher HDI scores typically exhibit better access to healthcare and education services, as well as higher economic standards of living. The HDI values used in this study are derived from the 2018 National Socio-Economic Survey (SUSENAS). The histogram of the HDI variable for regencies in Java Island is presented in Figure 1.

The independent variables utilized in this study were derived from the Village Potential Survey (PODES), administered by Statistics Indonesia (BPS). As a comprehensive nationwide survey, PODES captures detailed information on village-level potential and socio-economic characteristics across Indonesia. The dataset encompasses a wide range of indicators that may affect the Human Development Index (HDI), including access to educational services, basic infrastructure such as clean water, electricity, and transportation networks, as well as the availability of healthcare facilities and other socio-economic determinants. The full list of independent variables employed in the analysis is presented in Table 1.

These data were initially collected at the village level and subsequently aggregated to the regency level. This study focuses on regencies located in Java Island, which comprises Jakarta, West Java, Central Java, Yogyakarta, East Java, and Banten. This coverage enables comparative analysis across regions with varying development levels and socio-economic characteristics. The HDI data used in this study refers to regency-level HDI in Java Island,

Table 1. Independent Variables Used in the Study

Name	Variables	Type
code	Province code in Java Island	categorical
X_1	Percentage of households without access to electricity (PLN and non-PLN)	numeric
X_2	Percentage of villages/sub-districts where most residents dispose of waste into rivers, irrigation channels, lakes, oceans, drains, ditches, or other places	numeric
X_3	Percentage of villages/sub-districts where most residents use open land, rivers, ponds, beaches, or ground pits as sanitation facilities	numeric
X_4	Percentage of villages/sub-districts with settlements along riverbanks	numeric
X_5	Percentage of villages/sub-districts with rivers contaminated by waste	numeric
X_6	Percentage of villages/sub-districts with slum settlements	numeric
X_7	Number of early childhood education centers and kindergartens per 1,000 residents	numeric
X_8	Number of elementary schools (SD/MI) per 1,000 residents	numeric
X_9	Number of secondary schools (SMP/MTs, SMA/MA, SMK) per 1,000 residents	numeric
X_{10}	Number of hospitals, health centers, polyclinics, and doctor practices per 1,000 residents	numeric
X_{11}	Number of maternity clinics, midwife practices, POSYANDU, and POLINDES per 1,000 residents	numeric
X_{12}	Number of pharmacies and drugstores per 1,000 residents	numeric
X_{13}	Number of medical doctors and dentists per 1,000 residents	numeric
X_{14}	Number of midwives per 1,000 residents	numeric
X_{15}	Number of malnutrition cases per 1,000 residents	numeric

Source: Village Potential Survey (PODES) 2018 by BPS, Statistics Indonesia

covering six provinces with a total of 119 regencies/cities. Table 2 presents the number of regencies per province. It is evident that the number of observations is unbalanced, with Jakarta, Yogyakarta, and Banten having fewer regencies compared to West Java, Central Java, and East Java. The modeling process will be conducted using two datasets: the original dataset and the dataset generated through the SMOTE, stratified by the provincial level region variable. These datasets are then further divided into training and testing sets. The training set is used for model development, with 70 percent of the data allocated for training and 30 percent for testing. The implemented models include Linear Mixed Models (LMM), Generalized Estimating Equations (GEE), Mixed Effect Regression Trees (MERT), and Gaussian Copula Marginal Regression (GCMR). The modeling is performed using the training dataset.

2.2. Methods

Several statistical methods are employed to analyze the data and identify relationships between variables, including Linear Mixed Model (LMM), Generalized Estimating Equations (GEE), Mixed Effect Regression Tree (MERT), and Gaussian Copula Marginal Regression (GCMR). Each method is selected based on its ability to handle dependency in clustered data and to capture potential non-linear relationships between variables. To address class imbalance in the data, the Synthetic Minority Oversampling Technique (SMOTE) is applied.

Table 2. Number of Regencies in Each Province in Java Island

Province	Number of Regencies
Jakarta	6
West Java	27
Central Java	35
Yogyakarta	5
East Java	38
Banten	8

Source: BPS, Statistics Indonesia

2.2.1. Synthetic Minority Over-sampling Technique (SMOTE). SMOTE is a widely adopted oversampling method that generates synthetic observations for the minority class, rather than duplicating existing samples. By introducing interpolated data points, SMOTE enhances the representation of the minority class, enabling machine learning algorithms to learn more robust and generalizable decision boundaries. SMOTE operates by synthesizing new samples through linear interpolation between existing minority class instances and their nearest neighbors. This process is conducted in the feature space, making it particularly suitable for datasets with continuous variables. The procedure involves the following steps: Selecting a minority class instance x_i as the reference point. Then, identifying k nearest neighbors of x_i within the minority class using the k-Nearest Neighbors (k-NN) algorithm. The last steps are generating synthetic samples by interpolating between x_i and one of its neighbors x_{nn}. To determine the nearest neighbors, SMOTE utilizes the k-NN algorithm with the Euclidean distance metric. The distance between two feature vectors x_i and x_j, each of dimension n, is computed as:

$$d(x_i, x_j) = \sqrt{\sum_{l=1}^{n} (x_{il} - x_{jl})^2} \tag{1}$$

In this study, the number of neighbors k is set to less than 5 (2, or 3 or 4). But the oversampling SMOTE results using k = 2, k = 3, or k = 4 did not show any significant differences. In this study, we used k = 3 for the discussion. For other provinces with more observations, the choice of k may be more flexible. This constraint represents one of the limitations of the dataset used in our study. Once a neighbor x_{nn} is selected, a synthetic sample x_{new} is created using the following interpolation formula:

$$x_{\text{new}} = x_i + \delta \cdot (x_{nn} - x_i) \tag{2}$$

where $\delta \in [0, 1]$ is a random scalar drawn from a uniform distribution. This ensures that the new sample lies along the line segment connecting x_i and x_{nn}, thereby expanding the minority class distribution in a controlled manner. One of the main advantages of SMOTE is its ability to balance the dataset without discarding samples from the majority class, thus preserving information. Another benefit is that it helps models better recognize patterns from the minority class [14]. However, SMOTE also has limitations. The generated synthetic samples may be less representative if the minority class data contain outliers. Additionally, when applied to datasets with categorical features, interpolation may produce less meaningful results.

This study was conducted using Human Development Index (HDI) data at the regency level across Java Island, where the number of regencies varies considerably among provinces. The Synthetic Minority Oversampling Technique (SMOTE), which is typically applied to categorical data involving class imbalance, was employed in this research by treating provinces as the classification variable. Some large provinces, such as the capital region, contain only a few regencies. However, the capital region has a high level of economic activity and strategic importance compared to other provinces that are also significant. Consequently, when performing prediction or parameter estimation with a limited number of observations, the predictive performance may be less accurate than in provinces with a greater number of regencies serving as data points. Therefore, this study can serve as a reference for future research that aims to conduct prediction or estimation under conditions of unequal sample sizes across provinces, particularly in cases where key regions such as the national capital have substantially fewer observations. The use of SMOTE in this study does not follow the conventional approach typically applied in classification regression, where the response variable is categorical and serves as the basis for class formation during oversampling. Instead, we adopt a different perspective: the class used for SMOTE is not the response variable, but rather the group variable (i.e., provinces in Java Island), which functions as a random effect in models such as LMM and MERT. Thus, SMOTE is employed to balance the distribution across categorical groups with unequal sample sizes, even though the response variable remains continuous. This approach aims to improve representation of minority groups within longitudinal or hierarchical data structures, without modifying the SMOTE algorithm theoretically. We acknowledge that applying SMOTE in mixed-effects regression is still relatively uncommon in the literature, and we include relevant references on oversampling in regression contexts [15, 16]. In this study, the ”UBL” package in the R programming language was used to perform the SMOTE oversampling technique.

2.2.2. Linear Mixed Models (LMM). LMM is a statistical framework designed to analyze data with hierarchical or clustered structures. LMM accommodates two types of effects: fixed effects and random effects [17]. Fixed effects capture the influence of covariates that are consistently observed across all groups (e.g., education level or income), whereas random effects account for group-specific variability. The LMM can be expressed in the following form:

$$Y_{ij} = X_{ij}\boldsymbol{\beta} + Z_{ij}\boldsymbol{\nu}_j + \epsilon_{ij} \tag{3}$$

where Y_{ij} denotes the HDI value for the j^{th} regency within the i^{th} group, X_{ij} is the covariate matrix for the observation, $\boldsymbol{\beta}$ represents the fixed effect coefficients, Z_{ij} is the design matrix for random effects, $\boldsymbol{\nu}_j$ represents the random effect coefficients, and ϵ_{ij} is the residual error term. This model captures variation in HDI across regencies while accounting for intra-group dependencies. The generalized form of this model is known as the Generalized Linear Mixed Model (GLMM). By integrating the link function used in Generalized Linear Models (GLMs) into the LMM framework, GLMM extends the applicability to response variables that follow distributions from the exponential family [18]. The equation for a Generalized Linear Model is given by [19]:

$$g(\mu_i) = X_i\boldsymbol{\beta} \tag{4}$$

When random effects are incorporated as in LMM, the GLMM takes the following form [20]:

$$g(\mu_{ij}) = X_{ij}\boldsymbol{\beta} + Z_{ij}\boldsymbol{\nu}_j \tag{5}$$

The function $g(\cdot)$ represents the link function, and in this study, the models employ the gamma distribution with a logarithmic link. The term μ_{ij} denotes the expected value of the response variable for the i^{th} observation within the j^{th} group. The generalized linear mixed model (GLMM) is implemented using the `lme4` package and the `glmer()` function in R. It is widely recognized that the Human Development Index (HDI) follows a non-symmetric distribution, typically modeled using the gamma distribution. Therefore, when analyzing datasets that violate the Gaussian assumption, it is recommended to select a link function that corresponds appropriately to the error structure or the presumed distribution of the response variable.

2.2.3. Generalized Estimating Equations (GEE). GEE are employed to account for dependence in clustered data structures, such as the HDI measurements across districts. GEE offers a more flexible alternative to traditional linear models by enabling parameter estimation without requiring explicit specification of the covariance structure among observations. It yields consistent estimates of regression parameters even in the presence of intra-cluster correlation. The general form of the GEE model is expressed as:

$$g(\mu_i) = X_i\boldsymbol{\beta} \tag{6}$$

where $g(\mu_i)$ represents the link function, and in this study, the models employ the gamma distribution with a logarithmic link, X_i denotes the covariate matrix, and $\boldsymbol{\beta}$ represents the vector of regression coefficients. Parameter estimation for $\boldsymbol{\beta}$ is performed by solving the following estimating equation iteratively, using the *Iteratively Reweighted Least Squares* (IRLS) method [21]:

$$U(\boldsymbol{\beta}) = \sum_{i=1}^{N} D_i^\top V_i^{-1}(Y_i - \mu_i) = 0 \tag{7}$$

In this formulation: - Y_i is the response vector for the i^{th} cluster, - μ_i is the expected value of Y_i based on the model, - $D_i = \frac{\partial \mu_i}{\partial \boldsymbol{\beta}}$ is the matrix of partial derivatives of μ_i with respect to $\boldsymbol{\beta}$, - V_i is the working variance-covariance matrix of Y_i, computed as:

$$V_i = A_i^{1/2} R(\alpha) A_i^{1/2} \tag{8}$$

where: - $A_i = \text{diag}(\text{Var}(Y_i))$ is a diagonal matrix of marginal variances, - $R(\alpha)$ is the working correlation matrix that characterizes the correlation among observations within a cluster. Several commonly used working correlation structures include:

- **Independent**: Assumes no correlation among observations, $R(\alpha) = I$.
- **Exchangeable (Compound Symmetry)**: Assumes a constant correlation among all observations.
- **Autoregressive (AR-1)**: Assumes that correlation decays exponentially with increasing time or distance between observations.
- **Unstructured**: Places no restrictions on the correlation pattern, allowing each pair of observations to have a distinct correlation.

In this study, the exchangeable (compound symmetry) working correlation matrix was employed. This choice is based on the assumption that all pairs of observations within the same area share a common correlation level. Compared to alternative working correlation structures, the exchangeable form is more appropriate for clustered data that lack a natural temporal or spatial ordering among units within the same area. For instance, using an independent correlation structure would imply that observations within a cluster are mutually uncorrelated, effectively reducing the model to an ordinary regression that ignores intra-cluster dependence. Similarly, the autoregressive (AR(1)) structure is unsuitable in this context because the correlation among observations does not decay according to distance or ordering, as typically observed in time-series data. The unstructured correlation matrix was also not adopted, as the number of observations in this study was insufficient to estimate all pairwise correlations—this structure generally requires a large sample size for stable estimation. All analyses in this study were conducted using the R software, specifically the "geepack" package, to fit the Generalized Estimating Equations (GEE) model with an exchangeable working correlation matrix.

2.2.4. Mixed Effects Regression Trees (MERT). MERT combine the flexibility of regression trees with the hierarchical modeling capabilities of Linear Mixed-Effects Models (LMM) [3]. This hybrid approach is designed to accommodate dependence in clustered data structures—such as HDI measurements across regencies—while maintaining interpretability through a tree-based structure. The general form of the MERT model is expressed as:

$$Y_{ij} = X_{ij}\boldsymbol{\beta} + Z_{ij}\boldsymbol{\nu}_j + \epsilon_{ij} \tag{9}$$

where Y_{ij} denotes the HDI value for the j^{th} regency within the i^{th} group, $X_{ij}\boldsymbol{\beta}$ represents the fixed effects modeled via regression tree partitions, $Z_{ij}\boldsymbol{\nu}_j$ captures the random effects associated with group-level variability, and ϵ_{ij} is the residual error term. While the model structure mirrors that of a standard LMM, MERT distinguishes itself by using a regression tree to estimate the fixed effects component $X_{ij}\boldsymbol{\beta}$. This allows the model to capture complex, nonlinear interactions among covariates, while the random effects $Z_{ij}\boldsymbol{\nu}_j$ account for intra-group correlation. In this study, the Mixed Effects Regression Tree (MERT) model was implemented using the `glmertree` package in R, which allows simultaneous estimation of tree-based fixed effects and group-level random effects. The modeling procedure consists of the following steps:

- **Hierarchical Data Structuring:** The dataset is organized to reflect its multilevel nature, with individual observations nested within categorical groups (e.g., provinces). The response variable is continuous, while the grouping variable serves as a random effect.

- **Tree-Based Partitioning of Fixed Effects:** The fixed effects component X_{ij} is modeled using a regression tree. The algorithm recursively partitions the data based on predictor values, forming terminal nodes that represent subgroups with distinct predictor-response relationships.
- **Estimation of Random Effects:** The random effects $Z_{ij}\nu_j$ account for within-group variability. These effects are estimated for each group, capturing deviations from the overall tree-based structure due to group-specific influences.
- **Model Fitting:** The `glmertree()` function is used to fit the model. It combines recursive partitioning for fixed effects with maximum likelihood estimation for random effects, producing a hybrid model that accommodates both nonlinearity and clustering.
- **Prediction Mechanism:** Predictions for a new observation Y_{ij} are obtained by summing the estimated mean response of the terminal node (from the regression tree) and the random effect associated with the group j. Formally, the prediction is given by:
$$\widehat{Y_{ij}} = \widehat{\mu_{\text{node}}} + \widehat{\nu_j}$$
where $\widehat{\mu_{\text{node}}}$ is the estimated mean of the terminal node, and $\widehat{\nu_j}$ is the estimated random effect for group j.
- **Subgroup Interpretation:** Although MERT does not yield interpretable regression coefficients, the tree structure provides intuitive decision rules that help identify meaningful subgroups. This is particularly useful for understanding heterogeneous effects across clusters and uncovering nonlinear interactions among predictors.

The main advantage of MERT is its ability to capture nonlinear interactions among variables while accounting for within-group dependence. This model enables the identification of subgroups with distinct characteristics that influence the Human Development Index (HDI).

2.2.5. Gaussian Copula Marginal Regression (GCMR). GCMR is a statistical framework that integrates marginal regression with the Gaussian copula to model dependence structures in clustered data. This approach enables the modeling of nonlinear relationships between dependent and independent variables and accommodates complex inter-cluster dependencies, such as those observed across districts. The Gaussian copula facilitates the combination of marginal distributions for each variable with a dependence structure defined via a copula function. This formulation provides enhanced flexibility in capturing nonlinear dependencies within the data. The marginal model for GCMR in clustered settings is given by:

$$g\left(\mathbb{E}[Y_i \mid X_i]\right) = X_i^\top \boldsymbol{\beta} \tag{10}$$

where $g(\cdot)$ is the link function, X_i is the covariate vector, and $\boldsymbol{\beta}$ denotes the regression parameters to be estimated. After specifying the marginal distribution $F_Y(y)$, GCMR employs the Gaussian copula to model the dependence among observations [10]. The Gaussian copula function with correlation matrix R is defined as:

$$C_R(u_1, u_2, \ldots, u_n) = \Phi_R\left(\Phi^{-1}(u_1), \Phi^{-1}(u_2), \ldots, \Phi^{-1}(u_n)\right) \tag{11}$$

where $u_i = F_Y(y_i)$ represents the marginal cumulative distribution value of the response variable, Φ^{-1} is the quantile function of the standard normal distribution, and Φ_R denotes the multivariate normal distribution with zero mean and correlation matrix R. The matrix R defines the correlation structure between clusters. Parameter estimation in the GCMR model is performed by maximizing the log-likelihood function of the marginal regression model combined with the Gaussian copula. The log-likelihood function is given by:

$$\ell(\boldsymbol{\beta}, R) = \sum_{i=1}^{n} \log f_Y(y_i \mid X_i, \boldsymbol{\beta}) + \log c_R(u_1, \ldots, u_n) \tag{12}$$

The first term, $\log f_Y(y_i \mid X_i, \boldsymbol{\beta})$, corresponds to the marginal regression component, while the second term, $\log c_R(u_1, \ldots, u_n)$, captures the contribution of the Gaussian copula in modeling the correlation structure. The parameters $\boldsymbol{\beta}$ and R are estimated jointly using the Maximum Likelihood Estimator (MLE). The modeling was conducted using the “gcmr” and ”MASS” package available in the R software.

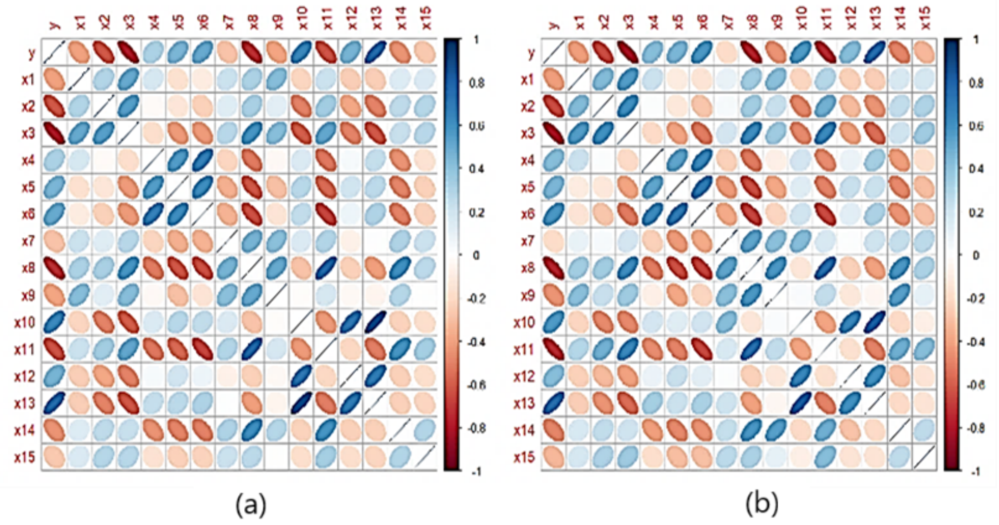


Figure 2. Pearson's correlation among variables: (a) original dataset, (b) dataset with SMOTE

3. Model Performance Evaluation

The performance of the models applied in this study is assessed using two quantitative indicators: *Root Mean Square Error* (RMSE) and *Median Absolute Error* (MedAE). RMSE evaluates the magnitude of prediction error by measuring the average squared differences between the predicted and actual values. It is particularly sensitive to large deviations, as larger errors contribute disproportionately to the overall score. The RMSE is computed using the following formula [22]:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{13}$$

where n is the number of observations, y_i is the actual value of the i^{th} observation, and $\hat{y}_i$ is the corresponding predicted value. In contrast, the *Median Absolute Error* (MedAE) measures prediction accuracy based on the median of the absolute differences between predicted and actual values. MedAE is more robust to outliers, as it relies on the median rather than the mean, making it less sensitive to extreme values in the data. The MedAE is defined as:

$$\text{MedAE} = \text{median}\left(|y_1 - \hat{y}_1|, |y_2 - \hat{y}_2|, \ldots, |y_n - \hat{y}_n|\right) \tag{11}$$

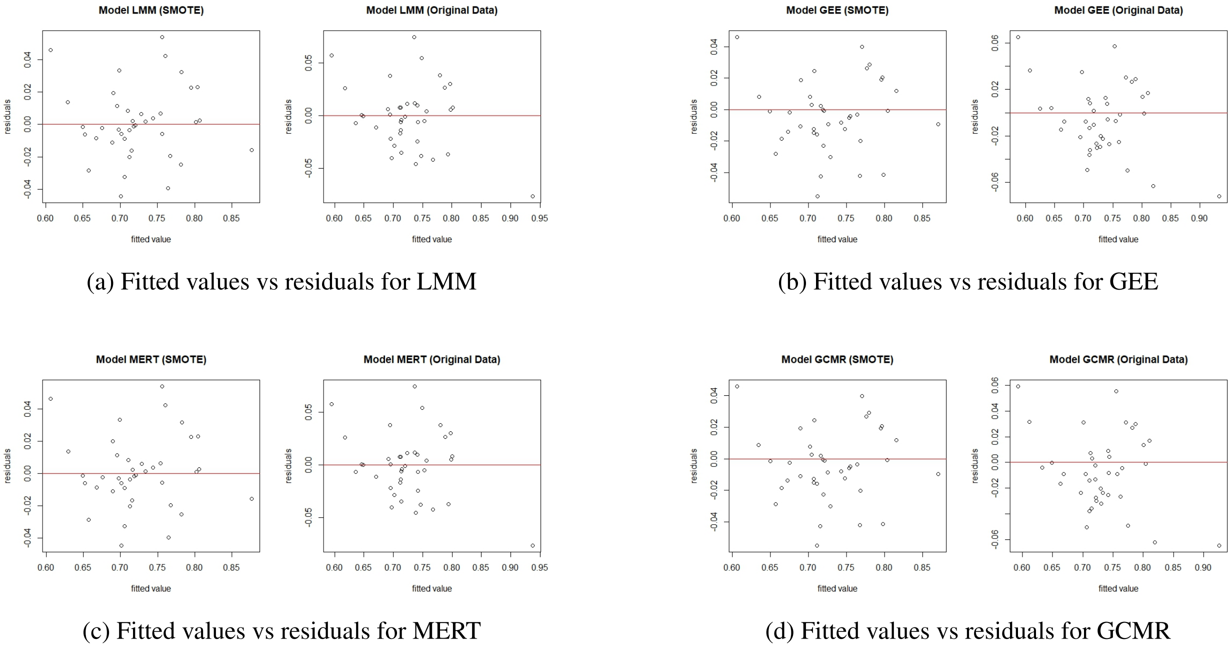
A model with strong predictive performance will exhibit lower values of both RMSE and MedAE, indicating minimal deviation between predicted and observed outcomes [23].

4. Results and Discussion

The relationships between variables are assessed using Pearson correlation coefficients, as the majority of the data used in the modeling process are numerical. As illustrated in Figure 2, negative correlations are represented by reddish hues, while positive correlations are indicated by bluish hues. Darker shades signify stronger correlations, whereas lighter shades correspond to weaker associations. Furthermore, Figure 2 shows that the response variable, namely the Human Development Index (HDI, y), is correlated with the predictor variables (x). This is evident from the correlation values between HDI (y) and all predictor variables (as detailed in Table 1), which consistently exhibit strong patterns—both positive and negative—without fading (refer to the first row or column in Figure 2).

The modeling process is conducted using two datasets: the original dataset and the SMOTE-generated dataset, stratified by provincial-level region variables. Each dataset is split into training and testing sets, with 70% allocated for training and 30% for testing. Model fitting is performed using the training data, and the estimated regression coefficients are presented in Table 3.

Model diagnostics were performed for LMM, GEE, MERT, and GCMR. The diagnostics can be observed through the plots of fitted values versus residuals, as shown in Figure 3. If the plots do not exhibit any discernible pattern, the linearity assumption of the model is considered to be satisfied.



(a) Fitted values vs residuals for LMM

(b) Fitted values vs residuals for GEE

(c) Fitted values vs residuals for MERT

(d) Fitted values vs residuals for GCMR

Figure 3. Residual diagnostics for each model: (a) LMM, (b) GEE, (c) MERT, and (d) GCMR

Model prediction and evaluation are performed using the testing data from the original dataset, which has not undergone SMOTE oversampling. The predictions generated by each model are shown in Figure 4. Figure 4(a) displays predictions from models trained on the original dataset, while Figure 4(b) shows predictions from models trained on the SMOTE dataset. The blue line represents the actual HDI values. In general, predictions from the SMOTE-trained models are closer to the actual values, although some exceptions exist.

The models are evaluated using Median Absolute Error (MedAE) and Root Mean Square Error (RMSE), as shown in Table 4. Lower values indicate better predictive accuracy. Upon inspection, the differences across models are relatively small. The MERT, GCMR, and LMM models demonstrate comparable performance, with MERT yielding the lowest MedAE values.

To visualize these metrics, a bar chart is presented in Figure 5, showing that models trained on the SMOTE dataset generally yield lower MedAE and RMSE values. This supports the effectiveness of SMOTE in improving model accuracy for unbalanced data. To assess the significance of differences between mixed models with and without SMOTE, paired t-tests were performed. The tests indicate that LMM, GEE, MERT, and GCMR each show statistically significant differences with p-values below 0.05. This demonstrates that applying the SMOTE technique to the regency-level HDI data on Java Island, Indonesia, yields predictions with significantly lower error.

Residual distributions are illustrated in Figure 6 using boxplots. Without SMOTE (Figure 6(a)), MERT and GCMR show smaller residual variation compared to LMM and GEE, with GCMR having the smallest residuals. Upon the application of SMOTE, all models show significant improvement in performance, as indicated by the more concentrated residuals.

Table 3. Estimated Regression Coefficients

Coef	**LMM**		**GEE**		**GCMR**	
	Origin	SMOTE	Origin	SMOTE	Origin	SMOTE
β_0	0.7530	0.7874	0.7276	0.7787	0.7534	0.7883
β_1	-2.6022	0.2831	0.8817	-4.9866	-2.6085	0.2995
β_2	-0.1046	-0.0768	-0.1889	0.2743	-0.1043	-0.0766
β_3	-0.0160	-0.0578	0.4106	-0.1644	-0.0167	-0.0586
β_4	0.0487	-0.0036	0.0907	0.1584	0.0489	-0.0034
β_5	-0.0120	0.0215	-0.2153	0.0014	-0.0127	0.0206
β_6	0.0223	-0.0466	0.4219	-0.0438	0.0228	-0.0447
β_7	0.0115	-0.0173	0.0292	0.0256	0.0120	-0.0175
β_8	-0.0829	-0.0673	0.0150	-0.0967	-0.0828	-0.0673
β_9	-0.0761	-0.0459	-0.9271	-0.0592	-0.0762	-0.0460
β_{10}	0.0692	0.0804	0.5145	0.2093	0.0684	0.0798
β_{11}	0.0290	0.0256	0.2361	0.0365	0.0282	0.0244
β_{12}	-0.0730	-0.0786	0.0489	-0.0594	-0.0736	-0.0794
β_{13}	0.1285	0.1268	-0.0652	-0.0778	0.1288	0.1260
β_{14}	0.0002	-0.0405	-0.2853	-0.0038	-0.0002	-0.0400
β_{15}	0.0162	-0.0002	-0.2959	-0.0761	0.0165	-0.0001

Note: MERT does not produce estimates for the regression parameters β.

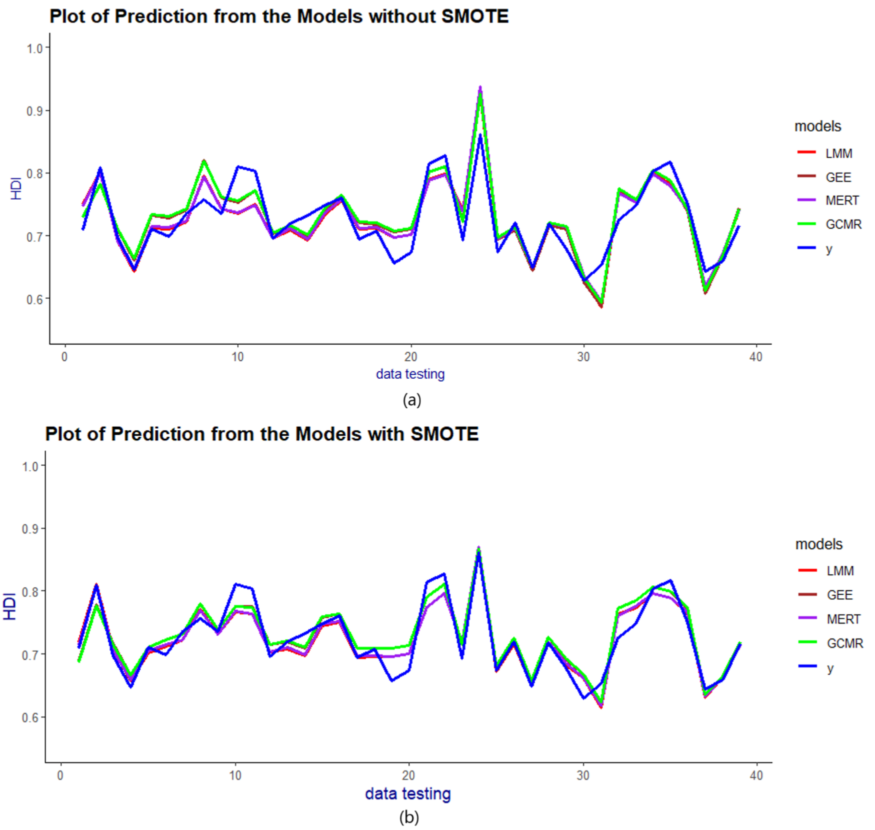
Table 4. Model Evaluation Metrics

Model	**MedAE**		**RMSE**	
	Origin	SMOTE	Origin	SMOTE
LMM	0.014110	0.011236	0.030400	0.021164
GEE	0.020975	0.017608	0.030253	0.022348
MERT	**0.014081**	**0.011222**	0.030415	0.021170
GCMR	0.023572	0.017841	**0.029573**	0.022414

The MERT model continues to demonstrate good consistency in performance, with residuals remaining concentrated both with and without SMOTE. The LMM and GEE models also show improved performance with SMOTE, as evidenced by the more concentrated residuals compared to the models without SMOTE. Overall, the SMOTE technique helps reduce residual variation and enhances model performance, with MERT demonstrating the most consistent and best performance. It can be concluded that the models using the SMOTE technique have better accuracy than those using the original data in the case of the HDI dataset from regencies in Java.

The trade-off between model interpretability and performance is indeed a central consideration in selecting statistical methods, particularly for applied researchers. Our analysis shows that MERT (Mixed Effect Regression Trees) achieves excellent predictive performance but operates as a “black box” without interpretable coefficients, making it more suitable in contexts where absolute accuracy is prioritized over model transparency. In contrast, models such as LMM (Linear Mixed Model) and GEE (Generalized Estimating Equations) provide parameters that are directly interpretable—either at the individual level (subject-specific, in LMM) or at the population level (population-averaged, in GEE).

Parameter interpretation in LMM allows researchers to understand both fixed and random effects on the response, as well as variation across groups or units, making it highly appropriate for longitudinal data with repeated or hierarchical structures. Similarly, in GEE, each coefficient represents the expected change in the population mean response when a predictor increases by one unit, yielding marginal inference that is relevant for policy or public health applications. Regarding model performance, it is important to recognize that models like GCMR



(a)

(b)

Figure 4. Fitted values from LMM, GEE, MERT, and GCMR models: (a) original dataset, (b) SMOTE dataset

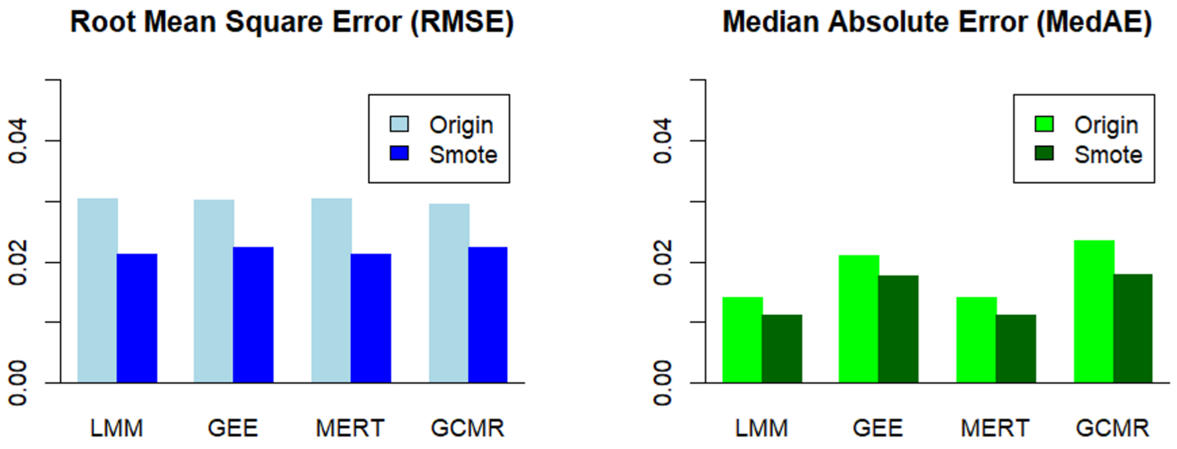


Figure 5. Comparison of MedAE and RMSE values for four models using original and SMOTE datasets

(Gaussian Copula Marginal Regression) may excel on original data due to their ability to capture complex non-linear dependencies in longitudinal data without relying on strict linearity or normality assumptions. GCMR's strength also lies in its capacity to separate the marginal model from the dependence structure via copulas, making it highly flexible and robust for real-world data that often deviate from normal distributions—features that standard linear models or GLM/GLMM may fail to fully capture.

However, after data engineering processes such as SMOTE (Synthetic Minority Oversampling Technique), especially when oversampling is performed through linear interpolation, the original non-linear relationships tend to be "smoothed out," making the dependencies appear more linear and potentially diminishing GCMR's advantage in detecting non-linear patterns. Theoretically, SMOTE expands the minority data region by interpolating between neighbors, resulting in smoothing or "regularization" of outliers and original non-linear dependencies—supporting the observation that complex models like GCMR are more effective on raw data with irregular dependency structures, while simpler or linear models may catch up in performance after SMOTE and even approach the effectiveness of non-linear models. Therefore, model selection should be guided by research objectives and data context: if accuracy is the absolute priority and coefficient interpretation is not the main focus, MERT is a viable choice; however, if understanding the mechanism of predictor effects is crucial for scientific justification or policy-making, LMM and GEE with interpretable parameters are more appropriate. On the other hand, GCMR is highly recommended for original data with complex dependency patterns that are difficult to capture using conventional models, although the smoothing effect of SMOTE may distort this advantage and bring its performance closer to linear models on engineered data. In conclusion, the trade-off between interpretability and performance is not solely determined by algorithm choice but is also heavily influenced by the characteristics and preprocessing of the data; thus, in real-world applications, method selection must be based on a balanced consideration of interpretability needs, performance demands, and the nature of the data being analyzed.

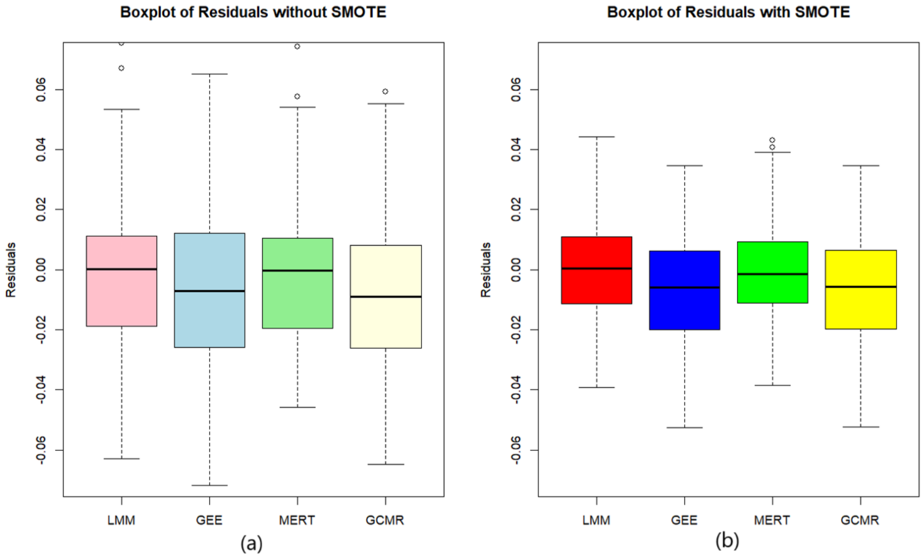


Figure 6. Boxplot of Residuals Models

5. Conclusion

Modeling using LMM, GEE, MERT, and GCMR is highly suitable for clustered data, as evidenced by the low MedAE and RMSE values during model evaluation. This indicates that all four models—LMM, GEE, MERT, and GCMR—demonstrate good predictive ability in the case of modeling the HDI of regencies in Java Island, Indonesia. However, the model with the smallest MedAE is the MERT (Mixed Effect Regression Tree) model, both for the original data (MedAE = 0.014081) and the oversampled data (MedAE = 0.011222). The MERT model shows good consistency, making it a recommended choice when the dataset undergoes an oversampling technique. On the other hand, the GCMR (Gaussian Copula Marginal Regression) model, although having a higher MedAE, performs well with the original dataset, with the smallest RMSE value (0.029573). This highlights the strength of the GCMR model in handling original data without the need for oversampling. MERT achieved superior evaluation metrics in terms of RMSE and MedAE, indicating strong predictive accuracy, but it is less interpretable because it does not produce regression coefficients. LMM and GEE showed lower predictive accuracy compared with MERT but offer greater interpretability through readily available regression coefficients and random-intercept estimates. GCMR effectively captures nonlinear patterns, which explains its high accuracy on the original (without-SMOTE) data.

The findings of this study suggest that the application of the SMOTE oversampling technique can improve the accuracy of models, as indicated by the lower MedAE and RMSE values. In other words, the use of the SMOTE technique helps reduce model errors, particularly for data with unbalanced observations across clusters, such as the HDI dataset for regencies in Java. Thus, clusters with fewer observations (minority clusters) are not overlooked and are given equal weight as other clusters. A key limitation of this study lies in the reliance on data drawn from regencies located on Java Island, Indonesia, which may possess unique regional attributes not necessarily representative of other areas. The modeling approach employed assumes intra-group dependence among regencies within the same province, while maintaining independence across provinces. A recommendation for future research is to use non-linear data to assess the performance of the best model when confronted with non-linearity between predictor variables and the response variable in making predictions.

REFERENCES

1. P. J. Diggle, K.-Y. Liang, and S. L. Zeger, *Analysis of Longitudinal Data*, 2nd ed., Oxford University Press, vol. 23, no. 21, 2002.
2. N. Sajithra and D. Ramyachitra, *Comparative analysis of various tree classifier algorithms for disease datasets*, International Journal of Engineering Trends and Technology, vol. 69, no. 6, pp. 8–13, 2021.
3. A. Hajjem, F. Bellavance, and D. Larocque, *Mixed effects regression trees for clustered data*, Statistics and Probability Letters, vol. 81, no. 4, pp. 451–459, 2011.
4. A. Sklar, *Fonctions De répartition à N Dimension Et Leurs Marges*, Publications de l'Institut de Statistique de l'Université de Paris, vol. 8, pp. 229–231, 1959.
5. G. De Luca, G. Rivieccio, and S. Corsaro, *Value-at-Risk dynamics: a copula-VAR approach*, European Journal of Finance, vol. 26, no. 2–3, pp. 223–237, 2020.
6. E. Ivanov, A. Min, and F. Ramsauer, *Copula-based factor models for multivariate asset returns*, Econometrics, vol. 5, no. 2, 2017.
7. K.-H. Chen and K. Khashanah, *Measuring systemic risk: Vine copula-GARCH Model*, Lecture Notes in Engineering and Computer Science, Newswood Limited, pp. 884–889, 2015.
8. Y. Guo, S. Huang, Q. Huang, H. Wang, L. Wang, and W. Fang, *Copulas-based bivariate socioeconomic drought dynamic risk assessment in a changing environment*, Journal of Hydrology, vol. 575, pp. 1052–1064, 2019.
9. D. She and J. Xia, *Copulas-Based Drought Characteristics Analysis and Risk Assessment across the Loess Plateau of China*, Water Resources Management, vol. 32, no. 2, pp. 547–564, 2018.
10. G. Masarotto and C. Varin, *Gaussian copula marginal regression*, Electronic Journal of Statistics, vol. 6, pp. 1517–1549, 2012.
11. P. Novianti, G. Gunardi, and D. Rosadi, *The application of Gaussian copula marginal regression for exploring the effect of weather to Covid-19 in Jakarta*, AIP Conference Proceedings, 2024.
12. C. Giusti, S. Marchetti, M. Pratesi, and N. Salvati, *Robust small area estimation and oversampling in the estimation of poverty indicators*, Survey Research Methods, vol. 6, no. 3, pp. 155–163, 2012.
13. R. Vaishali, B. Sarojini, and D. Sobya, *Prediction of COVID-19 Vaccine Side Effects using SMOTE and Ensemble Machine Learning Models*, International Journal of Engineering Trends and Technology, vol. 72, no. 4, pp. 324–332, 2024.
14. J. A. Benítez-Andrades et al., *Enhanced prediction of spine surgery outcomes using advanced machine learning techniques and oversampling methods*, Health Information Science and Systems, vol. 13, no. 1, p. 24, 2025.
15. L. Torgo, R. P. Ribeiro, B. Pfahringer, and P. Branco, *SMOTE for Regression*, In: Lecture Notes in Computer Science, pp. 378–389, 2013.
16. P. Branco, L. Torgo, and R. P. Ribeiro, *Pre-processing approaches for imbalanced distributions in regression*, Neurocomputing, vol. 343, pp. 76–99, 2019.
17. C. R. Henderson, *Estimation of Variance and Covariance Components*, Biometrics, vol. 9, no. 2, p. 226, 1953.
18. D. Anggara, K. A. Notodiputro, and B. Sartono, *Generalized Linear Mixed Models: Application for Consumer Price Index in Indonesia*, AIP Conference Proceedings, 2022.
19. D. Ferezagia, *Generalized linear model with bayes estimator: modeling the number of children in married couples*, Communications in Mathematical Biology and Neuroscience, vol. 2024, 2024.
20. M. E. Brooks et al., *glmmTMB balances speed and flexibility among packages for zero-inflated generalized linear mixed modeling*, R Journal, vol. 9, no. 2, pp. 378–400, 2017.
21. D. Anggara, Indahwati, and A. Kurnia, *Generalized linear mixed models approaches to modeling panel data: Application to poverty in east Nusa Tenggara*, Global Journal of Pure and Applied Mathematics, vol. 11, no. 5, pp. 2867–2875, 2015.
22. R. J. Hyndman and A. B. Koehler, *Another look at measures of forecast accuracy*, International Journal of Forecasting, vol. 22, no. 4, pp. 679–688, 2006.

23. H. Helaly, K. El-Rayes, E.-J. Ignacio, and H. J. Joan, *Comparison of Machine-Learning Algorithms for Estimating Cost of Conventional and Accelerated Bridge Construction Methods during Early Design Phase*, Journal of Construction Engineering and Management, vol. 151, no. 3, 2025.