Jurnal Natural
Vol. 23, (3) 2023
DOI 10.24815/jn.v23i3.33079
pISSN 1411-8513
eISSN 2541-4062
ORIGINAL RESEARCH

Classification of household poverty in West Java using the generalized mixed-effects trees model

FARDILLA RAHMAWATI[1], KHAIRIL ANWAR NOTODIPUTRO[1*], KUSMAN SADIK[1]

[1]Department of Statistics, IPB University, Bogor, Indonesia

Abstract. Dealing with fixed effects and random effects can be accomplished by combining statistical modeling and machine learning techniques. This paper discusses the modeling of fixed effects and random effects using a statistical machine-learning approach. We used the generalized mixed-effects trees (GMET), a tree-based mixed-effect model for dealing with response variables that belong to the exponential family of distributions. In this study, both simulation and actual/empirical data utilized the GMET method to discover data conditions that were appropriate for employing this approach. The simulation data was generated using different response variable generations, as well as different values of the variance of random effect and fixed effect coefficients. The findings indicated that the GMET performs similarly for different response variable generation scenarios. However, it performed better when the fixed effect value and the variance of random effects were large. When applied to the empirical data, the GMET method describes fixed effects and random effects and classifies household poverty status quite well based on the area under curve (AUC) value. It has also revealed that important variables for poverty classification are the number of household members, owning land, the type of main fuel used for cooking, and the main source of water used for drinking. In order to address the socioeconomic disparity that leads to poverty, the government may become concerned about these factors. In addition to that information, the use of regional typology as a random effect in the model has also contributed to the variation of household poverty status. Based on research, the fixed effects in mixed models do not need to be linear and GMET may be employed in grouped data structures, giving the GMET technique the ability to compete with other approaches/methods.

Keywords: fixed effects, household per capita expenditure, machine learning, random effects, supervised learning

INTRODUCTION

Data mining can be applied to huge databases as an automatic search of patterns using computational techniques from statistics, machine learning, and pattern recognition. In machine learning, supervised learning uses training data to make decisions. Classification is one of the supervised machine learning techniques used as a predictive method and assumes that each category consists of a group of labeled groups [1]. Classification as a process involves the orderly and systematic assignment of each entity to one and only one class in a mutually exclusive and non-overlapping class system. This process is lawful (carried out according to a well-established set of principles governing class structure and class relationships) and systematic (mandating the consistent application of these principles within a prescribed framework of reality) [2].

*Corresponding Author:
khairil@apps.ipb.ac.id

Received: July 2023 | Revised: October 2023 | Accepted: October 2023

A decision tree is a classifier method consisting of nodes that form a “rooted tree”, where the tree is directed by a node called "root" [3]. Decision trees produce an efficient and easy-to-understand rule collection. All reputable data classifiers frequently employ enhanced tree pruning approaches (ID3, C4.5, CART, CHAID, and QUEST) [4]. Prediction accuracy can be improved using ensemble techniques such as bagging [5], boosting [6], and random forest [7]. However, complexity is the main disadvantage of ensemble trees; instead of a single decision tree, ensemble tree prediction

models are no longer visually understandable and have numerous trees [8].

Decision tree methods have been developed that allow for the analysis of correlated data structures. Correlated structures in decision tree analysis have been shown to the result in less complex trees and more accurate results [8]. Tree-based methods have been adopted to handle longitudinal and clustered data. One of them is done by Sela and Simonoff, who propose a regression tree method, also referred as a random effects expectation-maximization (RE-EM) tree [9]. However, the method can only handle Gaussian response variables and it's not suitable for classification problems [10]. To overcome this, Fontana et al. proposed a generalized mixed-effects trees (GMET) method to generalize the RE-EM tree approach by extending the use of the method to various classes of response variables belonging to exponential families. In addition, GMET can handle clustered data structures, similar to multilevel models [10].

In the mixed model, fixed effects and random effects can be described in the same model. Fixed effects are seen as constant in the population, and convey systematic and structural differences in response. Random effects are considered stochastic differences between groups or clusters [11]. Clusters are a series of observations for a particular subject in a study with repeated measurements. Mixed-effects have random effects term for each subject and treat random effects as effects that result independently of some probability distribution [12]. Fixed effects can be identified with sustainable covariates such as weights, baseline test scores, or socioeconomic status. The covariates are taken from continuous intervals (or sometimes multivalued ordinals) or factors, such as gender or treatment groups, that are categorical. Random effects are represented by unobserved random variables that are usually assumed to follow a normal distribution [13].

The GMET algorithm was developed to be able to describe the estimation of fixed effects and random effects. GMET is executed by initializing a random effect with zero, estimating response variables through a generalized linear model (GLM) (using appropriate link functions depending on the distribution of response variable families), constructing a regression tree using the estimated response variable (dependent variable), and fitting a mixed-effects model to estimate the part of the random effect using the presumed fixed effect part by trees. Flexibility and interpretation are the main advantages of the GMET algorithm. The method could model more complex functional forms by loosening the linear assumption of the fixed-effects part. The tree structure, as a result of summarizing complexity, is easy to interpret and communicate. GMET's performance could be comparable and outperformed when the data had a linear structure and a clear advantage in convergence time [10].

Other tree-based mixed model methods have been developed previously, such as the generalized mixed-effects regression tree (GMERT). GMERT is an extension of the mixed-effects regression tree (MERT) method designed for continuous outcomes to other types of outcomes (e.g., binary outcomes, counts data, ordered categorical outcomes, and multicategory nominal scale outcomes) [14]. Pellagatti et al. proposed a new statistical method, called generalized mixed-effects random forest (GMERF), that extends the use of random forest to the analysis of hierarchical data for any type of response variable in the exponential family, considering both continuous and discrete covariates and without assuming a closed form in the association between the response and the fixed-effects covariates [15]. Moreover, Gottard et al. proposed the mixed-effect model called the three-tree mixed-effect model (3Trees model) for jointly estimating an interpretable predictive mixed-effect model with two components: a linear part, capturing the main effects, and a non-parametric component consisting of three trees for capturing non-linearities and interactions among individual-level predictors, among cluster-level predictors, or cross-level [16].

In the economic field, actual problems with non-normal data involving random effects can be found, such as classifying household poverty status (binary) in regional groups. Poverty derives from the word "poor," which denotes having the ability to work or strive but not enough to meet basic requirements. Lack of basic necessities including food, clothes, shelter, and water to drink is a state known as poverty [17].The problem of poverty in Indonesia is not only in the high number or percentage of poverty but also in the disparity (difference) between regions (provinces or districts/cities), which is very high [18]. According to the Badan Pusat Statistik (BPS), West Java was the province in Indonesia with the third-largest number of poor people in March 2019.

Good poverty data can be used to evaluate government policies on poverty, compare poverty across time and regions, and determine

targets for poor people with the aim of improving their conditions [19]. One of the bases that can be used in policy determination for overcoming poverty in West Java is to pay attention to the factors of the Gross Regional Domestic Product (GRDP) and Human Development Index (HDI) [20]. The Human Development Index (HDI) is a summary assessment of average accomplishment in important areas of human development, including living a long and healthy life, being informed, and having a good quality of living [21]. In addition, isolation as a result of difficult topographical conditions is the main cause of high poverty. The pattern of population settlements that tend to be scattered in mountainous and valley areas is also a challenge. This results in limited access to basic services and assistance. People's lives remain underdeveloped, so it is difficult to get out of poverty [22]. Typological features demonstrate that the height of infrastructure does not necessarily influence economic development, dependent on geographical dynamics and regional resources [23]. Thus, regional groupings or typologies can be carried out based on GRDP, HDI, and the regional location (coastal and non-coastal) of districts/cities in West Java Province. The regional typology will be used as a random effect in GMET modeling in actual/empirical data study.

Based on this background, the study's objectives were to identify the data circumstances that were suitable for the GMET method by using simulation data (different kinds of data situations). In addition, the GMET algorithm is used to analyze actual/empirical data to identify variables or factors that can distinguish between households with low and high levels of poverty (poor and non-poor), as well as the impact of regional typology on the classification of household poverty status in West Java Province in 2019. The programming language used for all analyses is R, which uses RStudio as the Integrated Development Environment (IDE) software for R. The R code for the GMET algorithm can be found in the research of Fontana et al. [10].

METHODOLOGY

Model

The GMET model used to generate simulation data is given by equation:

$$\begin{aligned} &Y_{ij}\sim Bernoulli(p_{ij})\,;\; p_{ij}=E(Y_{ij}|b_i)\\ &logit(p_{ij})=\; f(x_{ij})+z_{ij}b_i\,;\; b_i\sim N(0,\Psi)\\ &i=1,\dots,7 \text{ dan } j=1,\dots,3400 \end{aligned} \quad (1)$$

Table 1. Response and fixed effects variables

Variable	Scale	Description
Y	Nominal	0: Non poor, 1: Poor
X_1	Ratio	Min: 1, Max: 10
X_2	Nominal	Category 1, 2, 3
X_3	Ordinal	Category Level 1, 2, 3

where $f(x_{ij})$ is the fixed effect part, $z_{ij}b_i$ is the random effect part, $x_{ij}=(x_{1ij},\dots,x_{10ij})^T$ is the fixed-effect covariate vector for each observation j in group i, and random intercept value: $z_{ij}=1\;\forall i,\forall j,\quad i=1,\dots,7,\quad j=1,\dots,3400$.

The GMET model used to actual/empirical data study is as follows:

$$\begin{aligned} &Y_{ij}\sim Bernoulli(p_{ij})\,;\; p_{ij}=E(Y_{ij}|b_i)\\ &g(p_{ij})=\; logit(p_{ij})=\ln\left(\frac{p_{ij}}{1-p_{ij}}\right)\\ &\qquad =\; f(x_{ij})+z_{ij}b_i\\ &b_i\sim N_q(0,\Psi)\,;\; \beta,b_i\; ind\\ &i=1,\dots,7 \text{ dan } j=1,\dots,n_i \end{aligned} \quad (2)$$

i is the group index, $I=7$ is the total number of groups or the number of regional typologies, n_i is the number of observations in the i-th group and $\sum_{i=n}^{I} n_i=J$. η_i is the n_i-dimensional linear predictor vector and $g(\cdot)$ is the logit link function. x_{ij} is the fixed effects covariate vector for each j-th observation in the i-th group, z_{ij} is the random effects covariate vector, b_i is the dimensional (1+1) vector of the random effects coefficient and Ψ is the 1x1 in-group covariance matrix of the random effects. Fixed effects are identified by a non-parametric CART tree model associated with the entire population, whereas random effects are identified by group-specific parameters. In the GMET model it is assumed that household j is nested within the regional typologies i.

Simulation Data

In generating simulation data, there are 1 (one) response variable (Y) and 3 (three) X_i fixed effects variables which can be seen in Table 1. Grouping will be used as a random effect variable.

The value of the random effects coefficient (b_i) is generated from the normal distribution with the mean zero and the variance value of the random effects coefficient are set to small (ψ_1) and large (ψ_2) where $\psi=\{\psi_1,\psi_2\}=\{4,36\}$. In addition, the scenarios that will be used in this simulation study are fixed effects coefficients/weights and response variable

generation. In the fixed-effects part, simulation scenarios are set to small and large fixed effects coefficients/weights. The scenarios of small and large fixed effects coefficients/weights for each response variable generation scenario are as follows:

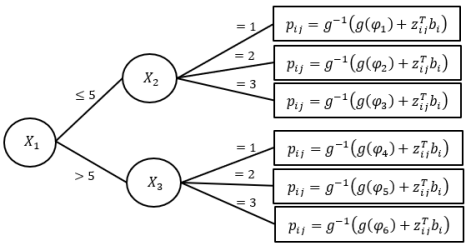


Figure 1. Mixed-effects tree structure used to generate p_{ij} values.

variables with nominal and ordinal scales for categories 1 and 2.

In general, the scenario of the simulation study data generation process can be seen in Table 3. The number of groups (I) simulated is $I = 7\ (i = 1.2, 3, \ldots, 7)$. The number of

Table 2. Fixed effects values scenario on data with response variable derived from mixed-effects tree structure

Fixed effects	φ^1	φ^2	φ^3	φ^4	φ^5	φ^6
Small	0.2	0.1	0.3	0.4	0.25	0.11
Large	0.99	0.9	0.87	0.79	0.8	0.95

$$f(x_{ij}) = \begin{cases} 0.3x_{1ij} - 0.11x_{21ij} - 0.12x_{22ij} - 0.21x_{31ij} - 0.22x_{32ij}\ ; \text{small} \\ 3x_{1ij} - 1.1x_{21ij} - 1.2x_{22ij} - 2.1x_{31ij} - 2.2x_{32ij}\ ; \text{large=10*small} \end{cases} \quad (3)$$

$$f(x_{ij}) = \begin{cases} -0.2\sqrt{x_{1ij}} - 0.11x_{21ij} - 0.12x_{22ij} + 0.31x_{31ij} + 0.32x_{32ij}; \text{small} \\ -2\sqrt{x_{1ij}} - 0.11x_{21ij} - 0.12x_{22ij} + 3.1x_{31ij} + 3.2x_{32ij}\ ; \text{large=10*small} \end{cases} \quad (4)$$

Table 3. Simulated data generation scenarios

Response variable generation (Y)	Fixed effects components (coefficients/weights)	Random effects components	
		$\psi_1 = 4$ (small)	$\psi_2 = 36$ (large)
Mixed- effects trees structure	Small	Data Scenarios 1	Data Scenarios 2
	Large	Data Scenarios 3	Data Scenarios 4
Linear fixed effect function	Small	Data Scenarios 5	Data Scenarios 6
	Large	Data Scenarios 7	Data Scenarios 8
Non-linear fixed effect function	Small	Data Scenarios 9	Data Scenarios 10
	Large	Data Scenarios 11	Data Scenarios 12

a. Response variable derived from mixed-effects tree structure. Each observation will be classified into 1 (one) of 6 (six) terminal nodes, which have a tree structure as shown in Figure 1, and will be calculated to obtain a p_{ij}value. The $g(\cdot)$ function is the logit link function. There are two specification scenarios for the value of the fixed effect (φ) on the data with response variables (Y) derived from mixed-effects tree structures, namely small and large, which can be seen in Table 2.
b. Response variable derived from linear fixed effect function. The function used can be seen in equation (3).
c. Response variable derived from non-linear fixed effect function. The function used can be seen in equation (4).

Each scenario of fixed effect coefficient values will be used to calculate the value of fixed effects and will be calculated with random effects to obtain p_{ij}values. In both functions, equations (3) and (4), each variable X_2 and X_3 variable use dummy variable to quantify

observations in each group are n_i = 3400, so the total observations simulated are n_{total}= $(I).(n)$ = 23800. The proportion of training data and testing data are 70% and 30%. The distribution of training data and testing data was carried out in each i-th group.

Actual/Empirical Data

The actual/empirical data used in this study is data from the Survei Sosial Ekonomi Nasional (Susenas) Kor 2019 (March) in West Java Province. The survey covered 23783 sample households. Households are the basic unit and household poverty status is a response variable (Y) in this study. In each district/city grouping is carried out based on regional typology, so each household will be nested in a regional typology. Regional typology as a random effect variable in this study grouped districts/cities based on three indicators: regional economic performance (high/low GRDP), human development index (high/low HDI), and regional location (has a coastlines/ does not have a coastlines). The fixed effects, random effect, and response variable used in this study can be seen in Table 4 and the regional typology

that become random effect with the number of households in each group can be seen in Table 5.

Procedures of Simulation Data Analysis

Stage 1: Data generation

1. Generating fixed effects variables X_1, X_2, X_3 as much as $n_{total} = 23800$.
2. Generating the coefficient of random effects variables b_i when $b_i \sim N(0, \Psi)$ with: $z_{ij} = 1\ \forall i, \forall j$ and $\Psi = \{\psi_1, \psi_2\} = \{4,36\}$.
3. Calculating p_{ij}.
   a. Response variable derived from linear fixed effect and non-linear fixed effect function:
      1) Calculating $logit(p_{ij}) = f(x_{ij}) + z_{ij}b_i$.
      2) Calculating $p_{ij} = inv\left(logit(p_{ij})\right)$.
   b. Response variable derived from mixed-effects tree structure:
      1) Identifying each observation by fixed effects variables to be classified into one terminal node in Figure 1.
      2) Calculating p_{ij} values in Figure 1 with fixed effects scenario in Table 2.
4. Generating response variable (Y) with $Y_{ij} \sim Bernoulli(p_{ij})$.
5. Divide data (12 data scenarios) into 70% training data and 30% testing data in each group.

Stage 2: Modeling

1. Performing the GMET algorithm using training data for each data scenario in Table 3.
2. Doing prediction on testing data based on the GMET model in stage 2 step 1.
3. Calculating the predictive misclassification rate (PMCR) from testing data. If the PMCR value is getting smaller, the GMET model performance is getting better.

$$PMCR = \frac{1}{7140}\sum_{i=1}^{7}\sum_{j=1}^{1020}\left|y_{ij} - \hat{y}_{ij}\right|$$

Stage 3: Model performance evaluation

1. Repeating stage 1 (steps 2-5) and stage 2 for 30 repetitions.
2. Calculating mean, standard deviation, and five-number summary of all PMCR values obtained.

Table 4. Variables used in actual/empirical data modeling

Code	Variable	Scale	Effect
Y	Household poverty status (0=Non Poor; 1=Poor)	Nominal	-
X_1	Regional Typology	Nominal	Random
X_2	Regional Type	Nominal	Fixed
X_3	Number of families living inside this census building/house	Ordinal	Fixed
X_4	Ownership status of occupied residential buildings	Nominal	Fixed
X_5	The main source of water used for drinking	Nominal	Fixed
X_6	Distance to nearest waste/fecal reservoir	Ordinal	Fixed
X_7	Owning land	Nominal	Fixed
X_8	The number of household members	Ratio	Fixed
X_9	The main sources of household lighting	Nominal	Fixed
X_{10}	The main types of fuel used for cooking	Nominal	Fixed
X_{11}	The largest source of financing in households	Nominal	Fixed

Table 5. Groups based on regional typology

Group named	Number of household samples	Description
HHY	2686	High GRDP, High HDI, Has a coastline
HHN	5849	High GRDP, High HDI, Does not have a coastline
HLY	0	High GRDP, Low HDI, Has a coastline
HLN	2309	High GRDP, Low HDI, Does not have a coastline
LHY	952	Low GRDP, High HDI, Has a coastline
LHN	1933	Low GDP, High HDI, Does not have a coastline
LLY	6534	Low GRDP, Low HDI, Has a coastline
LLN	3520	Low GDP, Low HDI, Does not have a coastline

Procedures of Actual/Empirical Data Analysis

1. Preprocessing data to tidy up data structures and prepare data to be analyzed.
   (1) Classifying each household into poverty status categories in accordance with the understanding of poor people according to the BPS "a person whose expenditure per capita per month is below the poverty line is considered to be poor" [24]. The poverty line is the limit of the rupiah

exchange rate needed to meet the minimum living needs [25].
   (2) Grouping districts/cities into regional typologies based on the characteristics of each district/city.
   (3) Combining explanatory variable categories in the category of variables that have a number of observations less than 10.
2. Doing data exploration to obtain an overview of actual/empirical data, such as using box plots and bar plots.
3. Divide the data into training data and testing data using stratified 10-fold cross-validation. The data will be divided into 10 equal parts, where each part becomes testing data and the other 9 parts become training data.
4. Doing the modeling and evaluation stages ten times, based on stratified 10-fold cross validation.
   (1) Performing the GMET algorithm using training data.
   (2) Doing prediction on testing data based on the GMET model in step 4(1).
   (3) Calculating the prediction evaluation value of testing data, namely PMCR, variance partition coefficient (VPC), area under curve (AUC), accuracy, sensitivity, specificity, precision, recall, F1-score, and Matthews correlation coefficient (MCC).
5. Calculating the mean of all testing data prediction evaluation values obtained.
6. Interpreting the results obtained from GMET modeling.

RESULTS AND DISCUSSION

Simulation Study

A simulation data generation process has been carried out using data generation scenarios: response variable generation scenarios, fixed effects' coefficient/weight scenarios, and various value scenarios of random effect coefficients. These scenarios were designed and applied to the data generation process to study the performance of the GMET algorithm in making predictions on simulated data with various data conditions (12 data scenarios). The number of groups (I) simulated was 7 to describe the number of groups in the actual/empirical data.

For each data scenario that has been modeled using training data and evaluated using testing data, statistics or a summary of the evaluation results are obtained, which can be seen in Table 6. The data scenario that has the smallest PMCR mean value is when the response variable generation comes from a linear fixed effect function with large fixed effect and random effect values. Data scenario 1 has the largest PMCR mean when the generation of response variables comes from a mixed-effects tree structure with small fixed and random effect values.

The PMCR mean value tended to be small when the variance of the random effect was large, whether the fixed effect values were large or small (see Table 7). Even when the fixed effect value was large, the PMCR mean value tended to be smaller than when the fixed effect value was small. Data scenario 7, which came from a linear fixed effect function, has quite interesting

Table 6. Summary of PMCR values (%)

Response variable generation (*Y*)	Fixed effects	Random effects	Data scenario	Mean	SD	Min	Q1	Q2	Q3	Max
Mixed-effects trees structure	Small	Small	1	23.82	6.44	13.36	18.54	23.83	28.08	37.72
		Large	2	9.27	5.32	0.25	5.21	8.73	13.17	20.98
	Large	Small	3	19.60	4.25	9.03	16.39	19.44	23.39	27.61
		Large	4	9.88	4.38	1.11	7.10	9.93	12.52	20.60
Linear fixed-effect function	Small	Small	5	20.32	4.33	12.62	17.61	19.83	23.46	29.45
		Large	6	9.88	4.39	0.52	6.70	10.18	12.61	21.22
	Large	Small	7	7.63	1.14	4.72	7.01	7.63	8.42	9.92
		Large	8	6.55	2.04	4.10	5.07	6.13	7.34	11.85
Non-linear fixed-effect function	Small	Small	9	22.28	6.37	11.43	17.71	21.66	24.93	36.16
		Large	10	9.24	4.50	0.14	6.13	9.03	12.14	18.36
	Large	Small	11	18.31	2.01	14.51	17.20	18.61	19.57	21.99
		Large	12	8.90	3.57	1.26	6.79	8.38	11.05	16.23

Table 7. PMCR mean values compare in each data scenario

Data scenarios	Fixed effects		Data scenarios	Random effects		Data scenarios	Response variable generation (*Y*)		
	Small	Large		Small	Large		Trees	Lin	Non-Lin
1vs3	23.82	19.6	1vs2	23.82	9.27	1vs5vs9	23.82	20.32	22.28
2vs4	9.27	9.88	3vs4	19.6	9.88	2vs6vs10	9.27	9.88	9.24
5vs7	20.32	7.63	5vs6	20.32	9.88	3vs7vs11	19.6	7.63	18.31
6vs8	9.88	6.55	7vs8	7.63	6.55	4vs8vs12	9.88	6.55	8.9
9vs11	22.28	18.31	9vs10	22.28	9.24				
10vs12	9.24	8.9	11vs12	18.31	8.9				

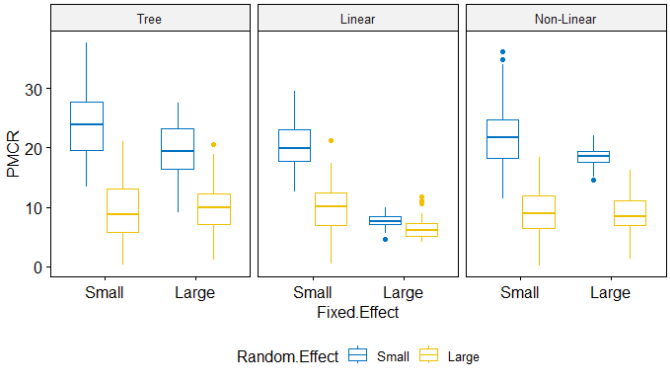
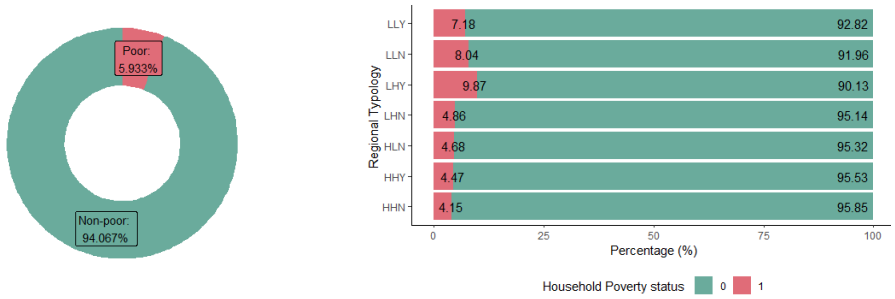


Figure 2. PMCR value boxplot.



(a) (b)

Figure 3. Pie chart of percentage of household poverty status (a) and stacked bar chart of percentage of household poverty status in each regional typology (b).

behavior where the PMCR mean value tends to be very smaller than data scenarios from mixed-effects tree structures and non-linear fixed effect functions (mean values tend to be large when fixed effects are large and random effects are small).

The PMCR value for each data scenario can be depicted to the distribution plot in Figure 2. The response variable data scenario group of the linear fixed effect function showed a different pattern than the other response variable data scenario group when the fixed effect was large (scenarios 7 and 8). Data scenarios 7 and 8 have the smallest variation in PMCR values compared to other data scenarios. In general, it can be seen that data scenarios with large random effects have a smaller distribution of PMCR values than scenarios with small random effects. PMCR values for each response variable generation group (mixed-effects trees, linear fixed effect functions, and non-linear fixed effect functions) tend to show relatively similar data distribution patterns.

Actual/Empirical Study

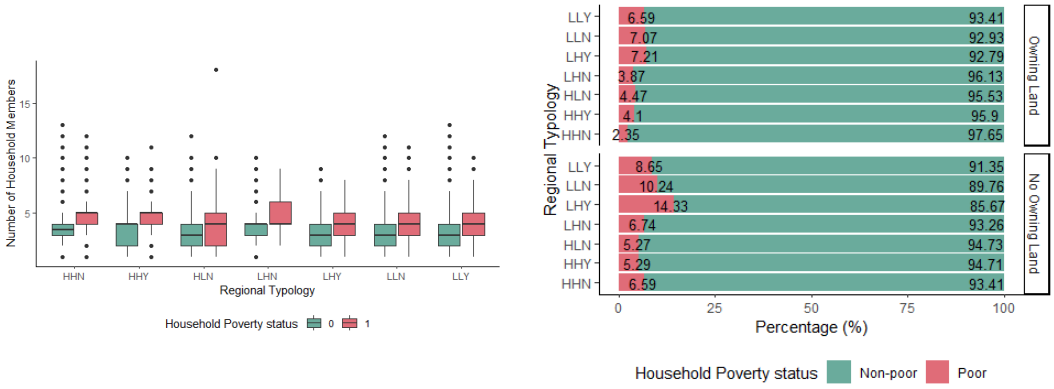
On the actual data that has been pre-processed, descriptive analysis is carried out by visualizing the data, as can be seen in Figures 3, 4, and 5. Figure 3(a) depicts the percentage of poor and non-poor households in poverty. The percentage of households with poor status is 5.933%, which is very small when we compared to the percentage of households with non-poor status, which has a value of 94.067%. The percentage of household poverty status in each regional typology can be seen in Figure 3(b). LHY areas (low GRDP, high HDI, has a coastline), LLN (low GRDP, low HDI, does not have a coastline), and LLY (low GRDP, low HDI, has a coastline) are the three regions that have the largest percentage of households with poor status compared to other regional typologies.

A stratified 10-fold cross-validation is used to evaluate the GMET model, so in the analysis process, ten models will be built using different testing data sets. Based on the GMET model used in modeling, it can be seen that the value of the importance level of the explanatory variable (variable importance) is known based on the estimated mixed-effects model tree. Over all, from ten GMET models, fixed effects explanatory variables that have the four highest variable importance values can be seen in Table 8.

The comparison distribution of the number of household members with poor and non-poor household status in each regional typology can be seen in Figure 4(a). Households with poor

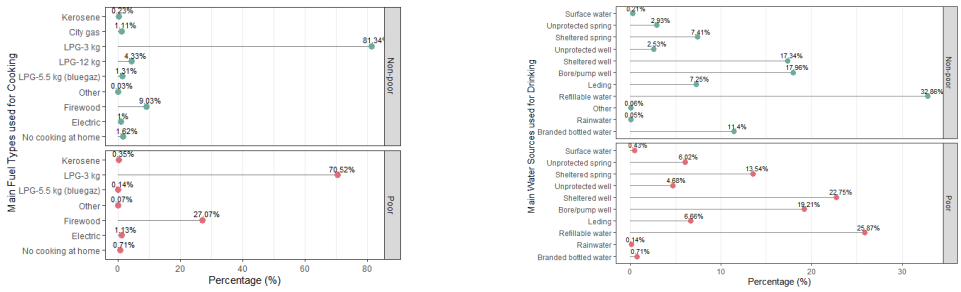
Table 8. Variable importance based on mixed-effects tree model estimates

Code	Explanatory variable	Average value of variable importance
X_8	The number of household members	2.66567
X_7	Owning land	1.87943
X_{10}	The main types of fuel used for cooking	1.28335
X_5	The main source of water used for drinking	1.20519



(a) (b)

Figure 4. Boxplot of the number of household members (a) and stacked bar chart of the percentage of poverty status of households in owning land and no owning land households (b) in each regional typology.



(a) (b)

Figure 5. Lollipop chart percentage of main fuel types used for cooking (a) and percentage of main water sources used for drinking (b) in each poverty status.

status tend to have a larger number of household members than those who are non-poor. Based on Figure 4(b), households that do not own land have a higher percentage of poor household status compared to households that own land in each regional typology.

The percentage of the main types of fuel used for cooking in households with poor and non-poor status can be observed in Figure 5(a). The types of fuel used by non-poor households are more diverse than those used by poor households, such as city gas and LPG-12 kg, which are only used by non-poor households. Meanwhile, firewood is still widely used by poor households compared to non-poor households. The main source of drinking water in poor and non-poor households mostly comes from refillable water. However, the percentage of main water sources used for drinking branded bottled water in non-poor households is much higher than in poor households (see Figure 5(b)).

The mean value of the intercept estimated from the random-effects GMET model for each regional typology can be seen in the confidence interval plot in Figure 6. Estimation points $\hat{b}_i$ for $i = 1,2,3,..,7$ are depicted with a blue point and the 95% confidence interval from the estimated point is depicted with a horizontal black line. The typology effect of the area is not significantly equal to zero when the confidence interval does not pass through point 0. It means that if the random effect value of the regional typology is not equal to zero, then the regional typology has an influence on poverty status.

As shown in Figure 7, the ROC curve of the 10 evaluation testing data sets shows a similar curve to each other, where the AUC means value is 0.753. The AUC means value is quite good based on the AUC value criteria

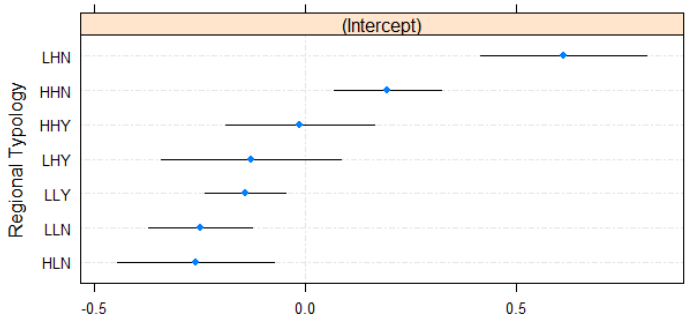


Figure 6. The average estimated random-effects intercept of the GMET model for each regional typology.

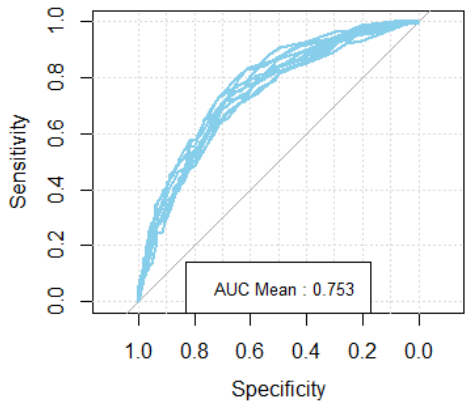


Figure 7. ROC Curve for 10-modeling.

Table 9. Performance evaluation of the GMET model

Evaluation	Mean
VPC	2.696
AUC	0.753
PMCR	32.108
Sensitivity	0.676
Specificity	0.727
Precision	0.975
Recall	0.676
F1-score	0.798
Accuracy	0.679
MCC	0.200

developed by Gorunescu [1]. Apart from evaluating the AUC value, other performance evaluations can be seen in Table 9. The average value of the variance partition coefficient (VPC) shows that an average of 2.696% of the variation in poverty status is caused by regional typologies in the GMET model.

Overall, the GMET algorithm can classify household poverty status well based on several performance evaluation scores mean tested (see Table 9). By using the GMET algorithm, the analysis results obtained become easy to understand. In addition, it can handle fixed effects and random effects at the same time in a tree-based model algorithm. This becomes useful in classifying household poverty status, where it can be seen that grouping based on regional typology as a random effect is one source of variation in the classification.

Performance of GMET

Based on the GMET algorithm for building a model and the type of response variables used in this research (binary response variable), the response variable originating from the Bernoulli family distribution are estimated using a generalized linear model (GLM) using the logit link function, which can be seen in equation (1). The GMET model, which consists of a fixed effects and a random effects part, uses the logit link function to obtain an estimated value from the estimated value of p_{ij}. The GLM algorithm that uses the logit link function will be the same as doing binary logistic regression modeling.

Testing classic assumptions such as autocorrelation, heteroscedasticity, normality, and linearity is usually carried out before carrying out regression modeling [26]. However, for a regression study that is not based on the OLS (Ordinary Least Square) example of logistic regression, the idea of the classical assumptions is not necessary [27]. Logistic regression differs from linear regression in that it does not maintain certain

Table 10. GVIF value of the actual/ empirical data dependent variables

Variable	GVIF	DF	$GVIF^{\frac{1}{2\cdot DF}}$	$\left(GVIF^{\frac{1}{2\cdot DF}}\right)^2$
X_1	1.729016	6	1.046686	1.095552
X_2	1.265348	1	1.124877	1.265348
X_3	1.218192	7	1.014198	1.028597
X_4	1.454002	4	1.047902	1.098098
X_5	2.038739	10	1.036258	1.073832
X_6	1.51458	1	1.230683	1.51458
X_7	1.340178	1	1.157661	1.340178
X_8	1.381201	1	1.175245	1.381201
X_9	1.052195	3	1.008516	1.017104
X_{10}	1.454706	9	1.021041	1.042524
X_{11}	1.146562	3	1.023056	1.046644

fundamental assumptions that linear and general linear models (as well as other ordinary least squares algorithm based models) do: (1) a linear relationship between the dependent and independent variables is not required in logistic regression; (2) the error terms (residuals) are not required to be normally distributed; (3) homoscedasticity is not required; and (4) the dependent variable in logistic regression is not measured on an interval or ratio scale [28]. Therefore, in this research that focuses on binary response variable, these general/classical assumptions are not tested.

Little or no multicollinearity among the independent variables is necessary for logistic regression. This implies that the independent variables should not be too correlated. Variance inflation factor (VIF) can be used to evaluate the assumption of multicollinearity, particularly in regression analysis. A value of VIF greater than 10 shows that multicollinearity exists and that the assumption is violated [28]. The square value of generalized VIF (GVIF), $\left(GVIF^{\frac{1}{2\cdot DF}}\right)^2 < 5$ is equivalent to a requirement that $VIF < 5$, thus that there is no evidence of multicollinearity in actual/empirical study (see Table 10).

As a statistical machine learning approach, the GMET method can model both fixed effects and random effects into a mixed model using a machine learning approach in its algorithmic or modeling stages. Although in this study both the simulation study and empirical study only focused on binary response variable, it was proven in the simulation study that the GMET method is good when the data structure comes from a mixed-effects tree structure, linear fixed-effects function, or non-linear fixed-effects function, GMET methods tend to have fairly good performance and are similar to one another. When the data structure is linear, the performance of the GMET method has the best value and this is in line with the results of a study conducted by Fontana et al. [10]. However, when faced with non-linear data structures, the GMET method still has good performance. This has been proven in actual/empirical data study, where the GMET method can handle cases of household poverty classification fairly well. Thus, this method has the potential to compete with other approaches or algorithms.

CONCLUSION

GMET models can perform supervised machine learning, especially classification with fixed effects derived from mixed-effects trees, linear functions, and non-linear functions. Based on the simulation study, the GMET model tends to have similar performance in each response variable generation and performs better when the variety of random effect components is large. In classifying the poverty status of households in West Java, the GMET algorithm showed a fairly good performance, with an average evaluation AUC value of 0.753. Variable importance based on mixed-effects model tree estimates are the number of household members, owning land, the type of main fuel used for cooking, and the main source of water used for drinking. Regional typology, as a random effect in modeling, contributed 2.696% of the variation in household poverty status.

ACKNOWLEDGMENT

The first author expresses her deepest gratitude to the supervisory committees, Prof. Dr. Ir. Khairil Anwar Notodiputro, M.S and Dr. Kusman Sadik, S.Sc., M.Sc., who have guided and provided very constructive support, correction, and direction. The authors also would like to thank the editors and reviewers who helped provide suggestions and corrections in this research manuscript.

REFERENCE

[1] Gorunescu, F. 2011. *Data mining: concepts, models and techniques* (Berlin: Springer). *DOI*: *10.1007/978-3-642-19721-5*.

[2] Jacob, E. K. 2004. Classification and categorization: a difference that makes a difference. *Libr. Trends.* **52** (3) 515–540.

[3] Rokach, L.; Maimon, O. 2005. Decision trees. *Data Mining and Knowledge Discovery Handbook*. 165–192. *DOI*: *10.1007/0-387-25465-X_9*.

[4] Jijo, B. T.; Abdulazeez, A. M. 2021. Classification based on decision tree algorithm for machine learning. *J. Appl. Sci. Technol. Trends*. **2** (1) 20–28. *DOI*: *10.38094/jastt20165*.

[5] Breiman, L. 1996. Bagging predictors. *Mach. Learn.* **24** (2) 123–140. *DOI*: *10.1007/BF00058655*.

[6] Freund, Y.; Schapire, R. E. 1997. A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. Syst. Sci.* **55** (1) 119–139. *DOI*: *10.1006/jcss.1997.1504*.

[7] Breiman, L. 2001. Random forests. *Mach. Learn.* **45** (1) 5–32. *DOI*: *10.1023/A:1010933404324*.

[8] Fokkema, M.; Edbrooke-Childs, J.; Wolpert, M. 2021. Generalized linear mixed-model (GLMM) trees: a flexible decision-tree method for multilevel and

longitudinal data. *Psychother. Res.* **31** (3) 329–341. *DOI*: *10.1080/10503307.2020.1785037*.

[9] Sela, R. J.; Simonoff, J. S. 2012. RE-EM trees: a data mining approach for longitudinal and clustered data. *Mach. Learn.* **86** (2) 169–207. *DOI*: *10.1007/s10994-011-5258-3*.

[10] Fontana, L.; Masci, C.; Ieva, F.; Paganoni, A. M. 2021. Performing learning analytics via generalised mixed-effects trees. *Data*. **6** (74) 1-31. *DOI*: *10.3390/data6070074.*

[11] Anderson, C.; Verkuilen, J.; Johnson, T. 2012. *Applied generalized linear mixed models: continuous and discrete data (for the social and behavioral sciences)* (Illinois: Springer).

[12] Agresti, A. 2019. *An introduction to categorical analysis, 3rd edition* (Hoboken: John Wiley & Sons, Inc.).

[13] West, B. T.; Welch, K. B.; Galecki, A. T. 2015. *Linear mixed models: a practical guide using statistical software, second edition* (New York: Chapman & Hall).

[14] Hajjem, A. 2010. *Mixed effects trees and forests for clustered data mixed effects trees and forests for clustered data*. (Dissertation, Canada: University of Montreal, 2010). [online] Available at: http://biblos.hec.ca/biblio/theses/003002.PDF [accessed 7 Jun. 2022].

[15] Pellagatti, M.; Masci, C.; Ieva, F.; Paganoni, A. M. 2021. Generalized mixed-effects random forest: a flexible approach to predict university student dropout. *Stat. Anal. Data Min. ASA Data Sci. J.* **14** (3) 241–257. *DOI*: *10.1002/sam.11505*.

[16] Gottard, A.; Vannucci, G.; Grilli, L.; Rampichini, C. 2022. Mixed-effect models with trees. *Adv. Data Anal. Classif.* **17** (2) 431–461. *DOI*: *10.1007/s11634-022-00509-3*.

[17] Arfiani, D. 2020. *Berantas kemiskinan* (Semarang: Alprin).

[18] Ritonga, H.; Surbakti, I.; Sodikin; Marsisno, W.; Suryaningsih, T.; Rizal, N.; Widya, C.; Ariewidayanti, D.; Taufiq, N. 2012. *Pendataan program perlindungan sosial (PPLS) 2011* (Jakarta: Badan Pusat Statistik).

[19] Avenzora, A.; Karyono, Y. 2008. *Analisis dan perhitungan tingkat kemiskinan tahun 2008* (Jakarta: Badan Pusat Statistik).

[20] Susanti, S. 2013. Pengaruh produk domestik regional bruto, pengangguran dan indeks pembangunan manusia terhadap kemiskinan di jawa barat dengan menggunakan analisis data panel. *J. Mat. Integr.* **9** (1) 1–18. *DOI*: *10.24198/jmi.v9.n1.9374.1-18*.

[21] [BPS] Badan Pusat Statistik. 2020. *Statistik Indonesia 2020* (Jakarta: Badan Pusat Statistik).

[22] Nanga, M.; HW, E. F.; Rahayuningsih, D.; Dinayanti, E.; Aulia, F. M.; Rismalasari, M.; Hafid, M.; Wahyu, R.; Putra, R.R.; Kartika, V.; Widaryatmo. 2018. *Analisis wilayah dengan kemiskinan tinggi* (Jakarta: Kedeputian Bidang Kependudukan dan Ketenagakerjaan Kementerian PPN/Bappenas).

[23] Kusuma, M. E.; Muta'ali, L. 2019. Hubungan pembangunan infrastruktur dan perkembangan ekonomi wilayah Indonesia. *J. Bumi Indones. DOI*: *oai:ojs.lib.geo.ugm.ac.id:article/1083*.

[24] [BPS] Badan Pusat Statistik. 2021. *Statistik Indonesia 2021* (Jakarta: Badan Pusat Statistik).

[25] Isdijoso, W.; Suryahadi, A.; Akhmadi. 2016. Penetapan kriteria dan variabel pendataan penduduk miskin yang komprehensif dalam rangka perlindungan penduduk miskin di kabupaten/kota. *The SMERU Research Institute.* 1-25. [online] Available at: http://www.smeru.or.id/sites/default/files/publication/cbms_criteria_ind.pdf [accessed 29 Sep. 2022].

[26] Flatt, C.; Jacobs, R. L. 2019. Principle assumptions of regression analysis: testing, techniques, and statistical reporting of imperfect data sets. *Adv. Dev. Hum. Resour.* **21** (4) 484–502. *DOI*: *10.1177/1523422319869915*.

[27] Ainiyah, N.; Deliar, A.; Virtriana, R. 2016. The classical assumption test to driving factors of land cover change in the development region of northern part of west java. *ISPRS - Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **XLI-B6** 205–210. *DOI*: *10.5194/isprsarchives-XLI-B6-205-2016*.

[28] Chreiber-Gregory, D.; Bader, K. 2018. Logistic and linear regression assumptions: violation recognition and control. *Southeast SAS Users Gr. Conf.* **47** (3) 1–22. [online] Available at: https://www.lexjansen.com/sesug/2018/SESUG2018_Paper-247_Final_PDF.pdf [accessed 13 Oct. 2023].