Jurnal Matematika dan Statistika serta Aplikasinya, Vol. 13, No. 2, 90-97, 2025
DOI:https://doi.org/10.24252/msa.v13i2.62521

Generalized Linear Mixed Model Tree (GLMM-Tree) for The Classification of Direct Cash Transfer Recipients in West Java Province

M. Ichsan Nawawi*
Mathematics Study Program, Universitas Islam Negeri Alauddin Makassar, Indonesia
***Corresponding author**: ichsan.nawawi@uin-alauddin.ac.id

***Submission date**: 01 July 2025, **Revision**: 05 August 2025, **Accepted**: 07 November 2025

ABSTRACT
Generalized Linear Mixed Model Tree (GLMM-Tree) is a statistical method that combines the concepts of decision tree and Generalized linear mixed model (GLMM). Here are some key advantages including Flexibility in Handling Different Types of Data, Incorporation of Random Effects, Handling of Non-linear Relationships, Interpretability, Variable Selection, Robustness to Outliers, Capturing Interactions, No Need for Parametric Assumptions. The purpose of this study is to compare the GLMM and GLMM-tree methods for the classification of direct cash transfer recipients in West Java with 25890 observations using the GLMM-tree method. Looking at the MSE and RMSE values, GLMM-tree is superior to GLMM for both training and testing data.

KEYWORDS
GLMM, GLMM-tree, Cash Transfer Recipients

1. INTRODUCTION

Machine learning (ML) is the scientific study of algorithms and statistical models that computer systems use to perform specific tasks without being explicitly programmed. Learning algorithms in many applications that we use every day. Whenever web search engines like google are used to search the internet, one of the reasons why these search engines work so well is because of the learning algorithms that have learnt how to rank web pages. These algorithms are used for various purposes such as data mining, image processing, predictive analysis, and others. The main advantage of using machine learning is that, once the algorithm learns what to do with the data, it can do its job automatically. By using computer algorithms, Machine Learning (ML) allows machines to access data automatically and with better experience while learning. This has simplified life and become an indispensable instrument in several industries, such as agriculture, banking, optimisation, robotics, structural health monitoring, etc., to name a few [1], [2].

Generalized Linear Mixed Models (GLMM) fit a multilevel model on a binary response variable, but impose linear effects of covariates on the transformation of the response variable. In contrast, tree-based methods such as Classification and Regression Tree models (CART) learn the relationship between response and predictors by identifying dominant patterns in the training data. In addition, these methods allow a clear graphical representation of the results that is easy to interpret. The aim of this study is to create a new method, for non-Gaussian response variables, capable of maintaining the flexibility of the CART model and extending it to data structures, where multiple observations can be seen as samples in groups [3], [4].

Generalized Linear Mixed Model Tree (GLMM-Tree) is a statistical method that combines the concepts of decision tree and Generalized linear mixed model (GLMM). This hybrid approach is particularly useful for handling complex data structures where there are fixed and random effects, as well as interactions between variables that may change across different segments of the data. GLMM-Tree is a tree-based algorithm that allows for the detection of treatment-subgroup interactions while accounting

for the group structure of a data set. The algorithm uses model-based recursive partitioning to detect treatment-subgroup interactions and GLMMs to estimate the random effects parameters [5].

Several studies have been conducted using the Generalized Linear Mixed Model for factors that trigger student interest in continuing their studies at Syiah Kuala University (USK) in 2021, GLMM is suitable for this data because the response variable is Bernoulli distributed, and the random effects are assumed to be normally distributed [6]. GLMMs in the medical literature have increased to account for data correlation when modelling qualitative data or counts [7]. The study utilised a tree-based algorithm that can be used to detect interactions and non-linearities in GLMM tree type models, The GLMM tree algorithm builds on a model-based model that offers a flexible framework for subgroup detection [8], [9].

Cash Transfer is one of the government assistance programmes with the concept of providing cash or various other assistance, both conditional and unconditional cash transfers for people who are poor. The first country to implement the BLT programme was Brazil, and it was subsequently adopted by other countries. The funds provided and the mechanism of the BLT programme also varied depending on the characteristics of the community in that country. The BLT programme was first introduced under the administration of President Bambang Yudhoyono, and was first implemented in October 2005 - the same month, October 2005, that the government increased gas prices by 87.5%. The BLT programme was institutionalised through Presidential Instruction No. 12/2005. The second BLT programme was implemented in May 2008, the same month when the President again raised gas prices for the third time, this time by 33.3 per cent, which is still ongoing.

According to the Indonesian Ministry of Finance Directorate General of Treasury as of 20 April 2022, the Village Fund ceiling allocation is 544.358.095.000 which is distributed to 416 villages in the Bogor Regency area which includes Village BLT and Non BLT Village Fund.

The Direct Cash Transfer (BLT) programme in Indonesia is one of the largest targeted cash transfer programmes in the developing world. Targeted cash transfer programmes are increasingly used as a potentially important tool to help poor families. However, developing country governments often struggle to ensure that these programmes reach the people they are supposed to help. Targeting, or correctly identifying households eligible for assistance, can be challenging in the absence of reliable income data as a large proportion of the population works in the informal sector [1].

This research aims to classify Direct Cash Assistance (BLT) in West Java Province using Generalized Linear Mixed Model Tree (GLMM-Tree).

2. LITERATURE REVIEW

2.1 Generalized Linear Mixed Models (GLMM)

In situations where responses are not normal and correlated, we can use Generalized Linear Mixed Models for inference. The first introduction to this class of models was in McCullagh and Nelder's book Generalized Linear Models [10]. Models where they analysed salamander crossbreeding with GLMMs. Since its introduction, research on GLMMs has evolved from the introduction of penalised quasi-likelihood for approximate inference to more computationally complex Bayesian methods and non-Bayesian Monte Carlo methods.

Generalized Linear Mixed Model (GLMM) is a generalization of the linear model in which the independent variables contain random and fixed factors. These random effects are usually assumed to have a normal distribution. In general, GLMM can be written as follows:

$$E[y] = \beta_0 + \sum_{i=1}^{p} \beta_i x_i + zv \tag{1}$$

where y is a response variable of size $n \times l$, β_i is the estimated parameter, x is the independent variable, z is the standardized random vector and v is the random component.

In general, inference for GLMMs can be done either through likelihood estimation or by Bayesian methods. Likelihood estimation can focus on estimating the integrand (function being integrated) such as a penalised quasi-likelihood or trying to estimate the integral itself as in Laplace approximation or Gaussian-Hermite quadrature. Bayesian methods rely on estimating the posterior distribution of the parameters in the GLMM, which is usually done by Markov Chain Monte Carlo method (MCMC). Each of the given methods and how they estimate the parameters of the GLMM.

2.2 GLMM Tree

The GLMM Tree model is as follows

$$g(\mu_{ij}) = x_i^T \beta_j + z_i^T b \tag{2}$$

In the GLMM tree model, the fixed effect β_j is a local parameter, its value depends on the terminal node j, but the random effect b is global. Fokkema [5] developed a tree-based algorithm to detect interactions and non-linearity in GLMM models, which is called GLMM Tree. The GLMM Tree algorithm is built based on a recursive partitioning model.

The GLMM tree algorithm is based on the GLM tree algorithm, a specific case of the model based algorithm of the recursive partitioning algorithm of Zeileis et al [11]. The GLMM tree fits a recursive partitioning based on a (general) linear model: the nodes in the GLMM tree consist of GLMM-specific subgroups, which contain intercept terms and possibly effects of one or more predictor variables. The subgroups are described in terms of additional covariates: variables used to define partitions or subgroups, which are not included as predictors in the GLMM.

The GLMM tree algorithm extends the GLM tree algorithm by accounting for possible dependencies between observations in longitudinal or multilevel data sets. In such data sets, individual observations are nested in higher-level units: In multilevel data sets, individual observations may be nested in higher-level units, while in longitudinal data sets, measurements obtained on different occasions may be nested. Traditionally, such data sets are analysed with GLMM-type linear models, which take into account the correlated nature of the observations through the estimation of random effects. In (Generalized) linear models, this has been found to result in more accurate standard errors and lower type-I and -II errors. More recently decision tree methods have been developed that make it possible to analyse such correlated data structures.

Generalized Linear Mixed Model (GLMM) Trees combine the advantages of GLMMs with tree-based methods. Here are some key advantages:

1. Flexibility in Handling Different Types of Data: GLMM-Trees can handle various types of response variables, including continuous, binary, count, and other types of data.
2. Incorporation of Random Effects: GLMM-Trees allow for random effects, enabling the modeling of data with hierarchical or clustered structures, such as repeated measures or multi-level data.
3. Handling of Non-linear Relationships: Tree-based methods can capture complex, non-linear relationships between predictors and response variables, which may be difficult to model using traditional GLMMs.
4. Interpretability: Decision trees are often easier to interpret and visualize than other types of models. The tree structure provides a clear and straightforward way to understand how different variables affect the response.
5. Variable Selection: GLMM-Trees perform automatic variable selection, identifying the most important variables and interactions in the data without the need for manual intervention.
6. Robustness to Outliers: Tree-based methods are generally robust to outliers in the predictor variables, which can be a significant advantage over traditional GLMMs.
7. Capturing Interactions: GLMM-Trees naturally capture interactions between variables, which can be more challenging to specify and estimate in traditional GLMMs.
8. No Need for Parametric Assumptions: Unlike traditional GLMMs, which rely on parametric assumptions about the distribution of the response variable and the random effects, GLMM-Trees make fewer assumptions, providing more flexibility in modeling.

The algorithm of the Binomial GLMM Tree is as follows [5]:

1. Initial value estimation where the value of r and all values of $\hat{b}_{(r)} = 0$
2. For $r = r + 1$, perform GLM Tree modelling with $z_i^T \hat{b}_{(r-1)} = 0$ as the offset

3. Perform GLMM modelling $g(\mu_{ij}) = x_i^T \beta_j + z_i^T b$ with the final node $j(r)$ obtained from the GLM Tree obtained in Step 2. Then calculate the posterior predicted value $\hat{b}_{(r)}$
4. Repeat steps 2 and 3 until convergent [7], [8].

3. METHODOLOGY

The data used in this research is the National Socio-Economic Survey (SUSENAS) data in 2023 with a total of 25890 observations, where the districts and cities in West Java Province are 26 districts / cities consisting of 17 districts and 9 cities.

Table 1. Variables

Variables	Notation	Data Type
Status of Village Cash Assistance (BLT Village) Receipt	Y	Category
Worried about not having enough food	X_1	Category
Number of families living in this census building/home	X_2	Category
Ownership status of residential building	X_3	Category
Installed power at the meter	X_4	Category
Ownership of defecation facilities	X_5	Category
Land/land ownership status	X_6	Category
There is an art of owning a micro or small business	X_7	Category
Main building material of the roof of the widest house	X_8	Category
Main water source used for drinking	X_9	Category
Availability of soap, detergent or antiseptic liquid	X_{10}	Category

3.1 Data Analysis

The stages of data analysis are as follows:

1. Synchronisation of data from the 2023 Sakernas survey in West Java Province
2. Data pre-processing including: checking for missing or missing data.
3. Data exploration
4. Dividing training data and test data with a proportion of 80
5. Modelling using the binomial GLMM Tree method and GLMM
6. AIC Comparison of binomial GLMM Tree method and GLMM

4. RESULT & DISCUSSION

4.1 Data Exploration

The data used comes from the National Socio-Economic Survey (SUSENAS) data in 2023 with a total of 25890 observations in West Java Province. After data preprocessing, no missing values were found in the observations. The proportion of households that received the BLT programme was 6.94% (1798 observations) and households that did not receive BLT was 93.06% (24092 observations).

According to **Figure 1**, it can be seen that there is a difference in the proportion of households that received and did not receive BLT.

Figure 2 shows detailed data for variables X_1 to X_{10}. For variable X_1 the highest is not worried about not getting enough food, for variable X_2 the highest is the number of families who inhabit the building / house is 1 person, for variable X_3 the highest is the ownership status of the residential building is owned, for variable X_4 the highest is the installed power in the meter

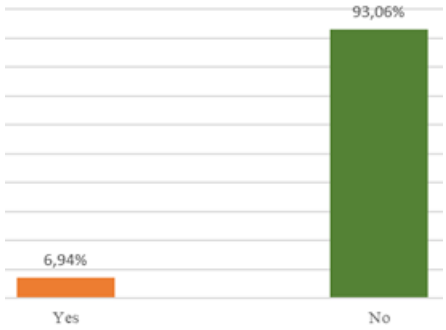

Figure 1. Diagram of BLT Receipt Status

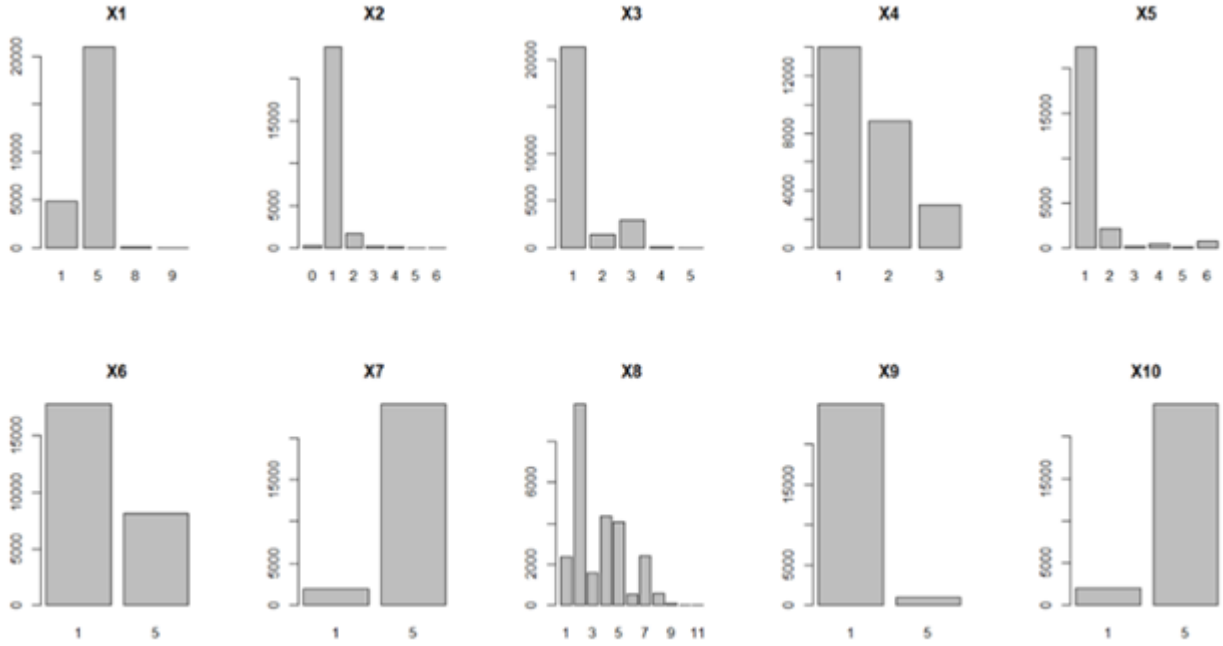

Figure 2. Diagram of X_1 to X_{10}

is 450 watts, for variable X_5 the highest is there is ownership of defecation facilities used by family members themselves, for variable X_6 the highest is the ownership of land / land by household members, for variable X_7 the highest is not having a micro or small business, for variable X_8 the highest is the main water consumed is refill water, for variable X_9 the highest is the main source of water used for drinking is drinking water, for variable X_10 the highest is available soap, detergent or antiseptic liquid.

4.2 Estimation Parameter of GLMM

Table 2 shows that only 3 out of 10 variables are significant to the status of receiving direct cash transfers. The variables that are influential are X_4 (power installed in the meter), X_5 (ownership of defecation facilities) and X_8 (source of water used for drinking).

From the aspect of installed power in the meter or X_4, households in West Java that received direct cash assistance with a 450 watt meter were 0.014 times compared to households with a 900 watt meter and 1300 watts or more.

From the aspect of ownership of defecation facilities or X_5 households in West Java who get cash transfers with ”available”, used only by members of their own household by 0.008 times compared to ”available”, used with certain households, ”available”, in communal washing latrines, ”available, in public washing latrines / whoever uses, ”available”, household members do not use, and no facilities.

Table 2. GLMM Parameter Estimation

Variable	Estimate	Std. Error	Z-Value	Pr(¿—z—)
(Intercept)	15.608133	0.0277310	56.284	$< 2.00 \times 10^{-16}$
X_1	0.0024412	0.0018757	1.301	0.19309
X_2	-0.0031748	0.0087679	-0.362	0.71728
X_3	-0.0002516	0.0049632	-0.051	0.95957
X_4	0.0143774	0.0049349	2.913	0.00357**
X_5	-0.0082731	0.0032890	-2.515	0.01189*
X_6	0.0019818	0.0018315	1.082	0.27921
X_7	0.0012039	0.0027708	0.434	0.66393
X_8	-0.0071906	0.0017799	-4.040	5.35×10^{-5}**
X_9	-0.0058141	0.0039503	-1.472	0.14107
X_{10}	-0.0011273	0.0027994	-0.403	0.68718

In terms of installed power at the meter or X_8, households in West Java receiving cash transfers with water sources used for drinking, branded bottled water are 0.0017 times higher than households with water sources used for drinking, refill water, tap water, borehole/pump, protected well, unprotected well, protected spring, unprotected spring, surface water, rainwater and other.

4.3 GLMM Tree

The GLMM tree for the unadjusted treatment outcome is presented in **Figure 3**. Electricity power installed in the meter was selected as the first predictor variable, with households using 450 watts of electricity faring worse than households using 900 and 1300 or more watts of electricity. In the group of households using 450 watts of electricity, the source of water used for drinking was branded bottled water, refill water, piped water, borehole/pump, unlined well and rainwater worse than tube well, surface water and other. At the lower level households using 450 watts of electricity, the source of water used for drinking is branded bottled water, refill water, piped water, borehole/pump, unlined well and rainwater with ownership status of residential building, owned and rent-free is worse than contract/rent, official and others, then for the lower level households using 450 watts of electricity, the source of water used for drinking is branded bottled water, ownership status of residential buildings, owned and rent-free with ownership of defecation facilities, ”available” used only by household members, ”available” in communal latrines, ”available” in public latrines/anyone uses, ”available” household members do not use worse than ”available” used with certain households and no facilities

The terminal nodes in **Figure 3** also present the standard errors for the subgroup mean estimates. These standard errors were calculated based on confirmatory mixed effects models, which account for variability between treatment centers, but not for tree structure searches. Thus, they provide a useful indication of variability, but may underestimate true variability. By considering standard errors, we can conclude that the unadjusted treatment results are not significantly different.

The predicted value of the random intercept is depicted in **Figure 4**, which shows a fairly symmetric distribution around 0. The estimated interclass correlation is 0.06. the poorest observed result was 9, and the best observed result was 78. Note that the error bars in **Figure 4** do not take into account tree search structures and therefore may be too small.

Table 3. Estimation Method Evaluation using Training and Testing Data

Model	MSE (Training)	RMSE (Training)	MSE (Testing)	RMSE (Testing)
GLMM	17.128	9.120	8.732	6.753
GLMM-Tree	16.210	9.001	7.811	5.901

Model evaluation was carried out by comparing the prediction accuracy of the GLMM tree with GLMM using training data and testing data. Results are presented. **Table 3** shows that the GLMM tree produces better accuracy results than GLMM for training data and testing data. This is supported by several studies including Fokkema et al[4]. Tree GLMMs produce slightly higher prediction accuracy than GLMMs [2]. GLMM-tree represents a lower computational load [5].

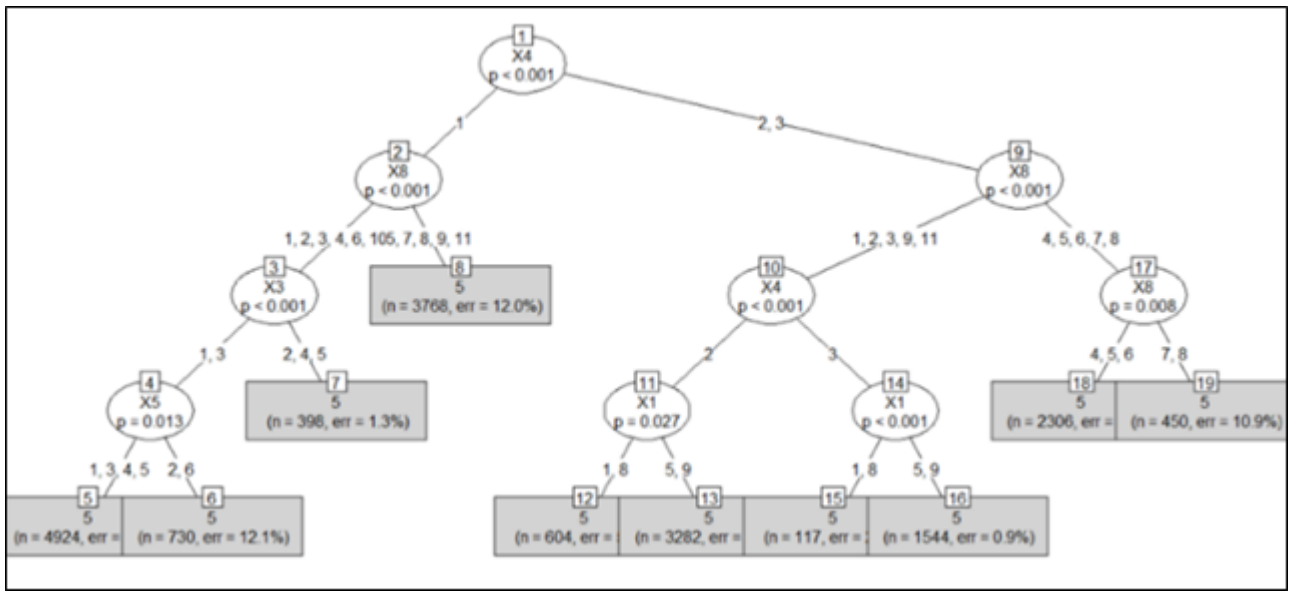

Figure 3. GLMM tree for the adjusted treatment outcome

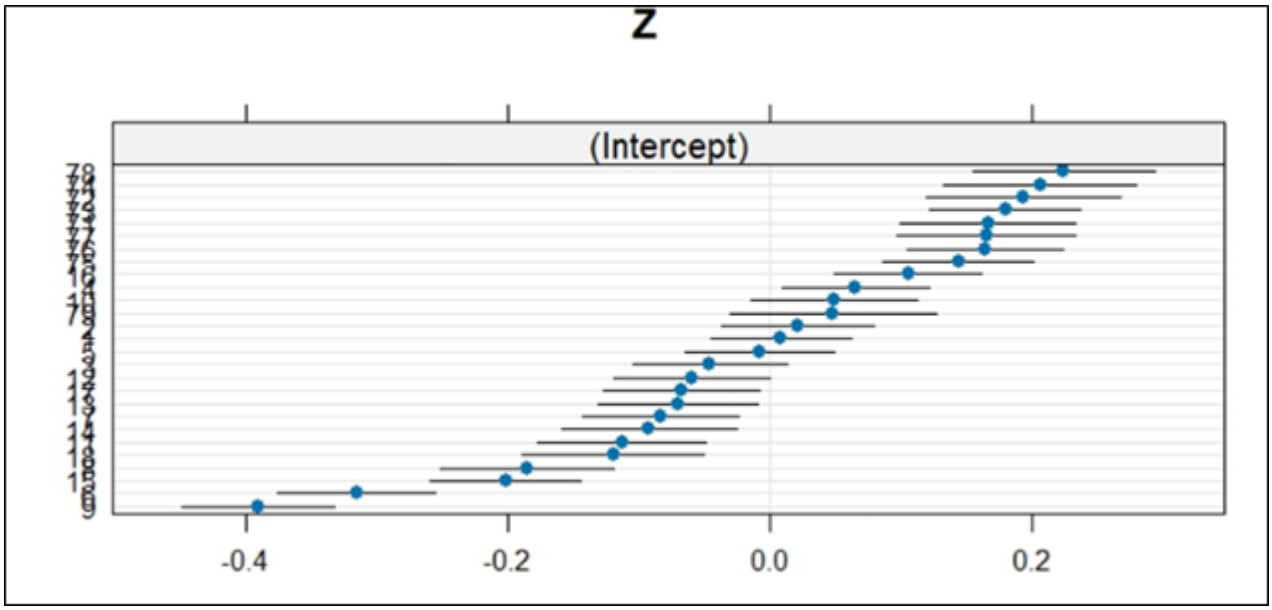

Figure 4. Random-effects predictions of the GLMM tree for the unadjusted treatment outcome

5. CONCLUSION

Based on the research results, the Generalized Linear Mixed Model Tree (GLMM-Tree) method proved superior to the Generalized Linear Mixed Model (GLMM) in classifying Direct Cash Assistance (BLT) recipients in West Java Province. Using 25,890 observations of the 2023 SUSENAS data, the evaluation results showed that the GLMM-Tree produced lower Mean Square Error (MSE) and Root Mean Square Error (RMSE) values in both training and testing data, indicating better predictive accuracy. Significant factors influencing BLT receipt include installed electricity capacity, ownership of toilet facilities, and drinking water sources. The advantages of GLMM-Tree in handling hierarchical data, non-linear relationships, and its ability to detect interactions between variables make it a more effective and flexible approach for large-scale socioeconomic data classification analysis.

REFERENCES

[1] B. Mahesh, "Machine learning algorithms – a review," *International Journal of Science and Research (IJSR)*, 2020, no. October.

[2] J. S. Ruíz, O. A. M. López, G. H. Ramírez, and J. C. Hiriart, *Generalized Linear Mixed Models for Categorical and Ordinal Responses*, 2023.

[3] M. M. Taye, "Understanding of machine learning with deep learning," *Computers (MDPI)*, vol. 12, no. 91, pp. 1–26, 2023.

[4] M. Fokkema, J. Edbrooke-Childs, and M. Wolpert, "Generalized linear mixed-model (glmm) trees: A flexible decision-tree method for multilevel and longitudinal data," *Psychotherapy Research*, pp. 1–13, 2020.

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

[6] A. Rusyana, K. A. Notodiputro, and B. Sartono, "A generalized linear mixed model for understanding determinant factors of student's interest in pursuing bachelor's degree at universitas syiah kuala," *Jurnal Natural*, vol. 21, no. 2, pp. 72–80, 2021.

[7] D. Agustin *et al.*, "Perbandingan kinerja binomial glmm tree dan bimm forest untuk pemodelan status bekerja penduduk," *Jurnal Teknologi Informasi dan Ilmu Komputer*, vol. 11, no. 1, pp. 95–106, 2024.

[8] M. Casals, M. Girabent-Farrés, and J. L. Carrasco, "Methodological quality and reporting of generalized linear mixed models in clinical medicine (2000–2012): A systematic review," *PLoS ONE*, vol. 9, no. 11, pp. 1–10, 2014.

[9] S. Bayu, K. A. Notodiputro, and B. Sartono, "Glmm and glmmtree for modelling poverty in indonesia," in *Proceedings of the International Conference on Data Science for Official Statistics*, vol. 2023, no. 1, 2023, pp. 121–131.

[10] P. McCullagh and J. Nelder, *Generalized Linear Models*, 2nd ed. Routledge, Jan. 2019. [Online]. Available: https://www.taylorfrancis.com/books/9781351445856

[11] A. Zeileis and K. Hornik, "Generalized m-fluctuation tests for parameter instability," *Statistica Neerlandica*, vol. 61, no. 4, pp. 488–508, 2007.