GLMM and GLMM trees for school dropout modeling in Bengkulu Province

Herlin Fransiska[2], *Khairil Anwar* Notodiputro[1*], and *Bagus* Sartono[1]

[1]Statistics and Data Science Study Program, School of Data Science, Mathematics, and Informatics, IPB University, Meranti Wing Street, Dramaga Campus IPB, Bogor, West Java, 16680, Indonesia
[2]Statistics Study Program, Faculty of Mathematics and Natural Sciences, Bengkulu University, Kandang Limun Street, Bengkulu, 38371, Indonesia

Abstract. The complexity of machine learning models often makes interpretation difficult, despite their high accuracy. This study aimed to model the factors of school dropout among children aged 6-18 in Bengkulu Province in 2019 using a simpler and more interpretable approach. Generalized linear mixed models (GLMM) and GLMM trees were used in this study. GLMM is effective for data with both random and fixed components, whereas GLMM trees is a decision tree-based method that integrates the strengths of GLMM. The performance evaluation results showed no significant difference in accuracy between the GLMM and GLMM trees. The GLMM analysis identified gender, age, ownership of the Indonesia Smart Card (KIP), savings ownership, and social assistance as important variables. The GLMM trees model highlighted age, disability, savings ownership, and social assistance as important variables. In conclusion, both GLMM and GLMM trees are effective in modeling school dropout, but GLMM trees offer a model that can explain the relationships between variables in greater detail.

1 Introduction

Machine learning methods have magnetism across various domains because they can model complex, nonlinear relationships within data. Random forest is an ensemble technique that aggregates multiple decision trees to enhance predictive accuracy [1]. However, the random forest model is often considered a black box because of its complexity, which limits its interpretability. Consequently, simple models such as decision trees or generalized linear models (GLM) is often chosen when interpretability is crucial [2].

The generalized linear mixed model (GLMM) extends the GLM by incorporating random effects, making it suitable for grouped or hierarchical data. A GLMM provides flexibility in modeling nonlinear relationships and can handle various types of response variable distributions, making it appropriate for the complex data structures commonly found in social science research [3].

Recently, the GLMM method has been developed, such as GLMM Trees. GLMM trees combine decision trees and linear models. This method is specifically designed to analyze

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

data with mixed effects by recursively partitioning the data and fitting a GLMM to each subgroup. It can detect interaction effects and subgroup heterogeneity in complex data while maintaining transparency in the model interpretation [2].

In the context of education, school dropout refers to students who discontinue their education and leave school before completing the educational level in which they are enrolled without transferring to another institution [4]. School dropout is a persistent challenge that undermines human capital development and hinders socioeconomic progress, particularly in developing countries. In Indonesia, dropout among children aged 6–18 years remains a main concern, including in Bengkulu Province [5-7]. Previous research on dropout used machine learning, such as tree-based modeling [8]. This study did not consider the hierarchical structure of the data.

This study aimed to apply and compare the GLMM and GLMM Trees methods for modeling school dropout in Bengkulu (2019). These methods are considered appropriate given the presence of hierarchies or correlated data structures. This analysis focused on evaluating the performance of both approaches and identifying the most influential factors contributing to school dropout.

2 Materials and methods

In this study, GLMM and GLMM Trees were used to model school dropout cases in Bengkulu Province (2019).

2.1 Generalized Linear Mixed Model (GLMM)

Generalized Linear Mixed Models (GLMM) are an extension of Generalized Linear Models (GLM) that incorporate both fixed and random effects, making them especially suitable for analyzing hierarchical or clustered data structures [3, 9]. Although the GLM assumes that all observations are independent, this assumption often fails in real-world settings where data are grouped, such as individuals within geographic regions. By introducing random effects, the GLMM accounts for the correlation within clusters and allows the model parameters to vary across groups, thereby improving the accuracy and generalizability of the results.

2.2 Generalized Linear Mixed Model (GLMM) trees

The Generalized Linear Mixed Model Trees (GLMM Trees) method was introduced by Fokkema et al. and offers a solution to the challenges posed by hierarchical or clustered data in tree-based approaches. GLMM Trees utilize a recursive partitioned algorithm to divide data into internally homogeneous subgroups concerning predictor-response relationships while simultaneously incorporating random effects to model between-group variability [2]. This enables the tree to account for intracluster correlation, which is common in data where observations are nested within units such as schools, regions, or households, resulting in models that are both statistically robust and practically interpretable [2]. Empirical studies have demonstrated their effectiveness in various domains [2], [9-11].

2.3 Methods

The stages of this analysis are

a) The initial data were explored to obtain a general description of the data on children dropping out of school.
b) Data division into 75% training and 25% testing

c) Modeling using SMOTE with the GLMM and GLMM Trees methods. Because the number of children who dropped out of school was relatively small, the data were imbalanced. To overcome this problem, the synthetic minority oversampling technique (SMOTE) was used.
d) Conducting model evaluation using accuracy values.
e) Obtaining important variables and interpretation.

3 Result and discussion

This study was based on data from the National Socio-Economic Survey (SUSENAS)-BPS Statistics Indonesia in Bengkulu Province (2019). There were 4724 data. The data used in this study were dropout status (Y), gender (X1), age (X2), ownership of a Smart Indonesia Card (KIP) (X3), savings ownership (X4), disability status (X5), number of household members (X6), social assistance recipients (X7), and area type (rural/urban) (R) as random effects.

Table 1. Descriptive Statistics for Categorical Variables

Variable	**Category**	**Count**
Y (School Dropout Status)	Yes (Dropped Out)	192
	No (Still Enrolled)	4532
R (Area type)	Rural	3313
	Urban	1411
X1 (Gender)	Male	2399
	Female	2325
X3 (Ownership of a Smart Indonesia Card)	Owns	840
	Does Not Own	3884
X4 (Savings ownership)	Yes	846
	No	3478
X5 (Disability Status)	Yes	97
	No	4627
X7 (Social assistance recipients)	Yes	1179
	No	3545

Table 2. Descriptive Statistics for Numerical Variables

Variable	**Minimum**	**Maximum**	**Standard deviation**
X2 (Age)	6	18	12
X6 (Number of household members)	1	10	5

As shown in Table 1, 192 children dropped out, and Fig. 1 shows the distribution of student status (two categories) with "Yes (Dropped Out)" 4% and "No (Still Enrolled)" 96%. This indicates that the vast majority of students in the dataset are still enrolled, while only a small proportion have dropped out. The distribution was highly imbalanced, which is an important consideration for subsequent analyses (classification modeling). Without appropriate handling, such a class imbalance can lead to biased model performance, favoring the majority class and potentially reducing the predictive accuracy of the minority group.

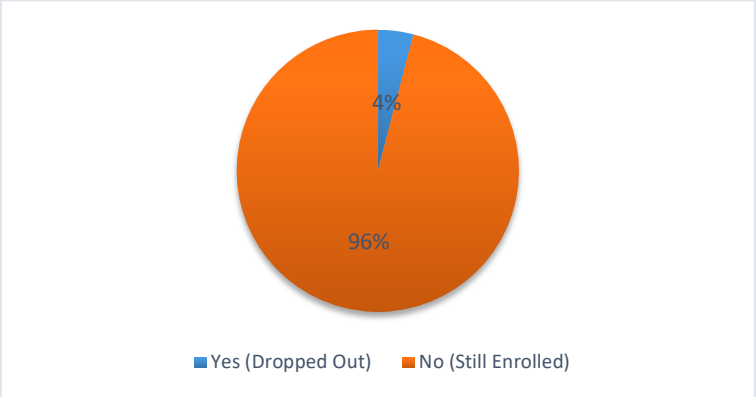

Fig. 1. Distribution of School Dropout Status (Y)

The data were divided into 75% training data and 25% testing data, where the modeling used a GLMM and GLMM trees with SMOTE first. The modeling was repeated ten times. The results of the GLMM analysis are displayed in the following summary:

Table 3. GLMM Estimation Results for School Dropout Status

Variable	Estimate	P-Value	Description
Intercept	-13.3148	< 2e-16	Significant
X1 (1)	0.54169	0.00095	Significant
X2	0.49074	< 2e-16	Significant
X3 (1)	1.24294	0.00011	Significant
X4 (1)	1.22261	0.00045	Significant
X5 (1)	0.64693	0.1249	Not significant
X6	0.08627	0.13966	Not significant
X7 (1)	0.75206	3.10E-05	Significant

Table 3. shows that the important variables based on the GLMM model that influence Y (school dropout status) are gender (X1), age (X2), ownership of a smart Indonesia card (KIP) (X3), savings ownership (X4), and social assistance recipients (X7). In this case, the random effect type (rural/urban) (R) was 0.02985.

Fig. 2. presents GLMM trees modeling showed that children with disabilities aged (X1) less than 14 years had a higher tendency to dropout of school than did children without disabilities (X5). Children who did not have savings come from families receiving social assistance (X7) and are aged (X1) for more than 14 years have a higher tendency to drop out of school. It can be said that physical and economic conditions affect whether they drop out of school. School dropout in elementary and junior high school is due to physical factors, whereas in senior high school is due to economic conditions. Urban random effects are smaller than rural effects.

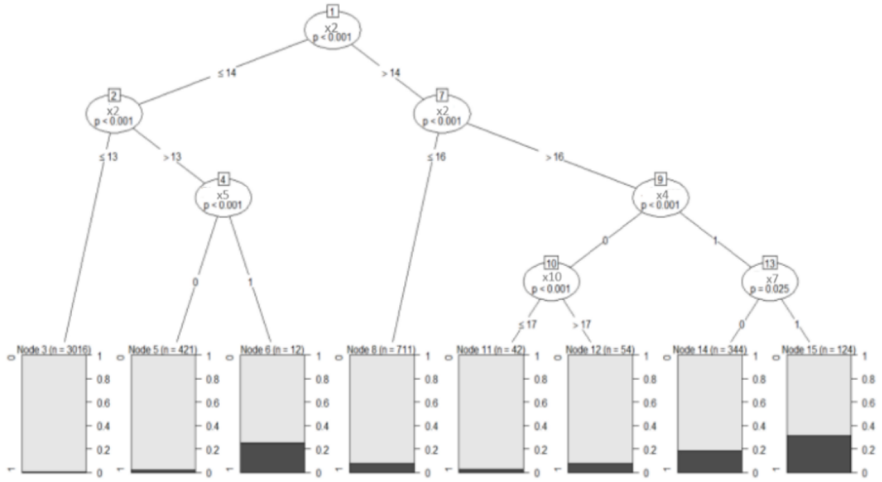

Fig. 2. Output of the Generalized Linear Mixed Model (GLMM) Trees for School Dropout Status

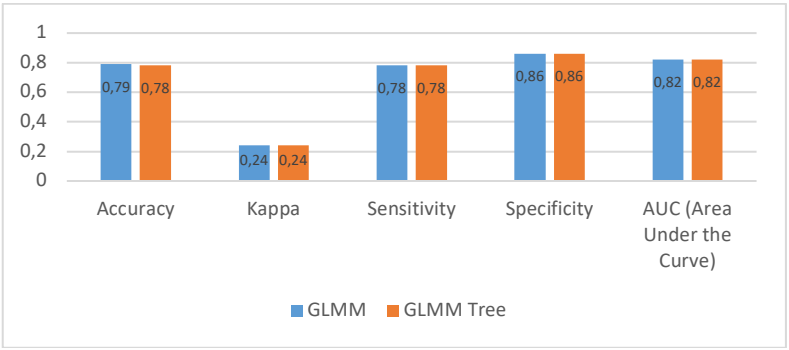

Fig. 3. Comparison of GLMM and GLMM_Trees Model Performance

Fig. 3. presents a comparison of five performance metrics between the generalized linear mixed model (GLMM) and its tree-based variant (GLMM Trees). Overall, both models demonstrated comparable performance across all metrics, with only minor differences in their accuracy. To clarify this, a paired sample difference test was conducted on the two methods on the five performance comparison measures, and the results showed that the p-value was greater than the 5% significance level in all criteria; therefore, it can be concluded that there is no difference in the performance of the two methods.

However, based on the outputs of the GLMM and GLMM tree models, the GLMM tree model can identify the relationship between factors that influence school dropout. Therefore, it can be said that GLMM trees can provide more interpretation in modeling school dropout. Compared with previous research on school dropout, this study accommodates the hierarchical conditions of data and can interpret relationships in a tree structure.

4 Conclusions

The application of GLMM to school dropout data showed that the model was influenced by gender, age, ownership of a Smart Indonesia Card, family savings ownership, and social assistance. The application of GLMM trees showed that the most important variables were age, disability, savings, ownership, and social assistance. The performance of the two methods was not significantly different; however, the GLMM trees were more interpretable than the GLMM.

References

1. H. Fransiska, A. M. Soleh, K. A. Notodiputro, and Erfiani, Evaluation of machine learning models based on household food insecurity data in Indonesia, In BIO Web of Conferences 171, p. 02011 (2025). https://doi.org/10.1051/bioconf/202517102011.
2. 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. **31**, 3, 329-341 (2021). https://doi.org/ 10.1080/10503307.2020.1785037.
3. C.E McCulloch and J.M. Neuhaus, Generalized Linear Mixed Models, (2015) https://doi.org/10.1016/B978-0-08-097086- 8.42017-9.
4. S.R. Ghazi, R. Ali, S. Shahzad, G. Shahzada, and K. Nawaz, Socio-economic factors as a cause of children dropout at primary level, Mediterranean Journal of Social Sciences. **2**, 3, 531-536 (2011). https://doi.org/10.5901/mjss.2011.v2n3p531.
5. I. R. Setyowati, K.A. Notodiputro, and A. Kurnia, A study of fixed-b asymptotic distribution models for analysing determinants of drop-out rates in Central Java. In AIP Conference Proceedings**, 2296**, 1, (2020). https://doi.org/10.1063/5.0030747.
6. D. Bianchi, E. Cavicchiolo, F. Lucidi, S. Manganelli, L. Girelli, A. Chirico, and F. Alivernini, School dropout intention and self-esteem in immigrant and native students living in poverty: The protective role of peer acceptance at school. *School Mental Health*, **13**, 2, 266-278 (2021). https://doi.org/10.1007/s12310-021-09410-4.
7. J. Niyogisubizo, L. Liao, E. Nziyumva, E. Murwanashyaka, and P.C. Nshimyumukiza, Predicting student's dropout in university classes using two-layer ensemble machine learning approach: A novel stacked generalization. *Comput. Educ.: Artif. Intell.* **3**, 100066 (2022). https://doi.org/10.1016/j.caeai.2022.100066.
8. Z. Song, S.H. Sung, D.M. Park, and B. K. Park, All-year dropout prediction modeling and analysis for university students. Applied Sciences, **13**, 2, 1143. 2023. https://doi.org/10.3390/app13021143.
9. C. M. Loan, Exploratory Data Analysis with Clustered Data: Simulation and Application with Oregon's Statewide Longitudinal Data System using Generalized Linear Mixed-Effects Model Trees, Doctoral dissertation, University of Oregon, United States (2024).
10. F. Asadi, R. Homayounfar, M. Farjam, Y. Mehrali, F. Masaebi, and F. Zayeri, Identifying Risk Indicators of Cardiovascular Disease in Fasa Cohort Study (FACS): An application of generalized linear mixed-model Tree. Arch Iran Med. **27**, 5, 239-247 (2024). https://doi.org/10.34172/aim.2024.35.
11. M. Fokkema, and A. Zeileis, Subgroup detection in linear growth curve models with generalized linear mixed model (GLMM) trees. *Behav Res.* **56**, 6759–6780 (2024). https://doi.org/10.3758/s13428-024-02389-1.