BAREKENG: Journal of Mathematics and Its Applications
March 2026 Volume 20 Issue 1 Page 0815–0828
P-ISSN: 1978-7227 E-ISSN: 2615-3017
https://doi.org/10.30598/barekengvol20iss1pp0815-0828

POISSON MIXED MODELS WITH A BOOSTING APPROACH FOR THE ANALYSIS OF COUNT DATA

Ita Wulandari[1*], Khairil Anwar Notodiputro[2], Bagus Sartono[3], Anwar Fitrianto[4], Anang Kurnia[5]

[1]*Departmen of Applied Statistics, Politeknik Statistika STIS*
Jln. Otto Iskandardinata 64C, 13330, Indonesia

[1,2,3,4,5]*School of Data Science, Mathematics, and Informatics, IPB University*
Jln. Meranti, Kampus IPB, Dramaga-Bogor, 16680, Indonesia

Corresponding author's e-mail: * *ita.wulandari@stis.ac.id*

Article Info

Article History:

Received: 4th June 2025
Revised: 27th June 2025
Accepted: 29th July 2025
Available online: 24th November 2025

Keywords:

Boosting;
Count data;
High-dimensional data;
Poisson mixed models.

ABSTRACT

Boosting is a powerful technique for enhancing predictive accuracy by iteratively reweighting observations, and is particularly effective in high-dimensional settings and for variable selection. While previous studies have demonstrated the advantages of integrating boosting with generalized linear mixed models (GLMMs) for binary outcomes, its application to count data within hierarchical frameworks remains limited. This study addresses that gap by extending boosting methods to count data through the development of a boosted Poisson mixed model (bPMM), a novel approach for small area estimation and variable selection in complex survey designs. The proposed model is applied to fertility data in the Indonesian provinces of Bali and East Nusa Tenggara, where the response variable is the number of live births and the predictors include twenty-eight socio-demographic covariates. Using the Akaike Information Criterion (AIC) for model selection, three significant variables were identified in Bali (Model 1), and one in East Nusa Tenggara (Model 2). The results demonstrate that bPMM not only improves variable selection in high-dimensional settings but also accommodates hierarchical structure in count data.

How to cite this article:

I. Wulandari, K. A. Notodiputro, B. Sartono, A. Fitrianto, and A. Kurnia, "POISSON MIXED MODELS WITH A BOOSTING APPROACH FOR THE ANALYSIS OF COUNT DATA", *BAREKENG: J. Math. & App.*, vol. 20, iss. 1, pp. 0815-0828, Mar, 2026.

Journal homepage: *https://ojs3.unpatti.ac.id/index.php/barekeng/*
Journal e-mail: *barekeng.math@yahoo.com; barekeng.journal@mail.unpatti.ac.id*

Research Article · Open Access

1. INTRODUCTION

One of the main problems in statistical research is the development of algorithms for building models and selecting variables. This problem can arise when data sets are not high-dimensional. For example, continuous variables can be entered into statistical models in the form of linear, nonlinear, or interaction with other predictor variables. To overcome this problem, various regression techniques have been developed in recent years. The progress in statistical methodology is mainly due to the fact that classic techniques for building models in the selection of variables (such as generalized linear modeling with stepwise selection) are known to be unreliable or even biased [1].

A different and more appropriate method for variable selection has been developed in machine learning-one of them is the boosting approach. According to [1]-[4], boosting is one of the most powerful learning methods, introduced about twenty years ago. Although originally designed for a classification problem, the method can also be applied to regression. A general description of the boosting method can be found in [4]. This concept can be seen as a breakthrough for several methods called ensemble schemes that rely on the principle of repetitive prediction by repeating or resampling original data sets and finally averaging the results of individual classifications.

The boosting algorithm is fairly well-known, particularly AdaBoost, which is used for classification of data [5], [6]. The success of AdaBoost [7], [8] is attributed to its ability to decompose classification prediction errors by reducing both bias and variance. Another important aspect is its association with determining the optimal number of boosting steps.

Research [9] further improved the concept of boosting as a gradient descent optimization method and boosting expansion method to deal with problems in the regression method. This research succeeded in proving an exponential dependency between bias and variants of the boosting method, which explains that the boosting algorithm is resistant to overfitting. This discovery represents some of the most important results regarding the theoretical nature of the boosting algorithm. Similar studies that are related to these results are [10], [11].

Generalized Linear Mixed Models (GLMM) are right for modeling dependency structures from longitudinal data and designs with repeated measurements. However, its use is usually limited to several variables, because the presence of many variables produces unstable estimates. Research [12] proposes a GLMM boosting approach (GLMM) that can be used in high-dimensional research when many variables are influential in the model by implicitly selecting variables. For the resulting estimator complexity, the information criteria are used; otherwise, it can combine random slope on linear effects that produce a flexible and suitable GLMM in cases where simple random intercepts cannot capture all variations of the effects of all subjects. This method has been investigated in simulation studies with the Poisson and Bernoulli connecting functions and the application of real data in cases of AIDS (Acquired Immune Deficiency Syndrome). Research [13] using gradient boosting to predict ‘‘at-fault’’ accidents on car loss costs. Another case was developed by researchers [13]-[16].

This research will be applied to fertility cases. Fertility (birth) can be measured by the number of live births a mother has given. The fertility rate of an area is measured by the Total Fertility Rate (TFR) [16]. This is one indicator to compare the success between regions in implementing the Family Planning (KB) program. Based on the IDHS data, Indonesia’s TFR has decreased from 2.6 in 2012 to 2.4 in 2017. However, this figure has not yet reached the Strategic Plan, which sets 2.3 for 2017. That is because there is still a TFR gap between provinces. In 2017, the highest TFR was in the province of East Nusa Tenggara (NTT) of 3.4, and the lowest was in the province of Bali of 2.1 [17].

These regional disparities highlight the critical need for localized, data-driven analyses of fertility patterns that account for both demographic heterogeneity and contextual factors. Despite this, most prior studies have predominantly employed binary or Gaussian response models, which may be ill-suited for modeling count-based fertility outcomes. Moreover, the application of boosted Poisson mixed models (bPMM) within the context of fertility research remains notably underexplored—particularly when addressing high-dimensional explanatory variables and hierarchical data structures commonly encountered in large-scale survey designs. To bridge this methodological gap, the present study applies the bPMM framework to model fertility data in Indonesia. This approach offers improved variable

selection capabilities, accommodates random effects to capture cluster-level variation, and provides more nuanced insights into the socio-demographic determinants of fertility.

2. RESEARCH METHODS

2.1 Generalized Linear Mixed Models

Generalized Linear Mixed Models (GLMM) are the result of the development of two models, namely Linear Mixed Models (LMM) and Generalized Linear Models (GLM). In its development, GLMM can answer more complex problems, especially those related to random effects, variance components, and the shape of the distribution of response variable data that does not have to be normally distributed. The model with random effects is expected to be more efficient in identifying the distribution of random components, to be able to explain more precisely the effects of these random components.

Let y_{it} is an observation of t in i clusters $i = 1,\dots,n, t = 1,\dots,T_i$, collected in $\mathbf{y}_i^T = (y_{i1},\dots,y_{i\,p})$, $\mathbf{x}_{it}^T = (x_{it\,1},\dots,x_{it\,p})$ are covariate vectors associated with fixed effects, $\boldsymbol{z}_{it}^T = (z_{it\,1},\dots,z_{it\,p})$ is a covariate vector associated with random effects. It is assumed that the observations of y_{it} are conditionally independent of the middle-value $\mu_{it} = E(\mathbf{y}_{it}|\mathbf{b}_i, \mathbf{x}_{it}, \mathbf{z}_{it})$ and the variance of $var(\mathbf{y}_{it}|\mathbf{b}_i) = \phi v(\mu_{it})$, where $\upsilon(.)$ is a known function of variance and ϕ is a variant of the parameter scale [18]. The GLMM model, which has a continuous and monotonous interface, can be derived, namely:

$$\begin{aligned}\boldsymbol{\eta} &= g(\mu_{it}) = \beta_0 + x_{it}^T\beta + z_{it}^T b_i \\ &= \beta_0 + \eta_{it}^{par} + \eta_{it}^{rand}.\end{aligned} \tag{1}$$

The connecting function that can be used is log for $\boldsymbol{y}_{ij}|\mathbf{b}_i \sim Poisson(\mu_{it})$, Therefore, Eq. (1) becomes bPMM [19], [20]:

$$\begin{aligned}\boldsymbol{\eta} &= \log(\mu_{it}) = \beta_0 + x_{it}^T\beta + z_{it}^T b_i \\ &= \beta_0 + \eta_{it}^{par} + \eta_{it}^{rand}.\end{aligned} \tag{2}$$

2.2 Boosted Poisson Mixed Models

Boosting machine learning is proposed to improve classification procedures by combining estimates with reweighted observations. Reweighting is related to minimizing the loss function iteratively [21]. Boosting has been extended to the regression problem in L2-estimation [22]. The procedure is very similar to the gradient descent method by using the specific loss function [23]. The initial idea of boosting is the urgent need for estimation problems for high-dimensional models.

The boosting algorithm presented in this study is based on the likelihood function and operates by repeatedly fitting residuals using 'weak learners', which are simple models (e.g., single-variable regressions) that individually have limited predictive power but can be combined to form a strong overall model. This process of sequentially updating the model by fitting one variable at a time is known as component-wise boosting, allowing for efficient variable selection in high-dimensional settings. That step means that only one predictor component, in this case, is fitted at a time. More precisely, the model contains intercepts and only one $x_r\beta_r$ in one iteration step. This study uses $\mathbf{x}_{i.r}^T = (x_{i1r,},\dots,x_{iT_ir,})$ notation for covariate vectors on the r-linear effect on the 1st cluster, $r = 1,\dots,p$. Therefore, the corresponding r-th matrix contains only intercepts, and the r-th covariate vector is $\boldsymbol{X}_{i.r} = [\mathbf{1}, \boldsymbol{x}_{i.r}]$ and $\boldsymbol{X}_{.r} = [\mathbf{1}, \boldsymbol{x}_{.r}]$ for each i-th cluster and the entire sample. For the i-cluster of predictors that only contain the r-th covariate has the following form of Eq. (3):

$$\boldsymbol{\eta}_{ir} = \boldsymbol{X}_{i.r}\breve{\boldsymbol{\beta}}_r + \boldsymbol{Z}_i\boldsymbol{b}_i \tag{3}$$

where $\breve{\boldsymbol{\beta}}_r^T = (\beta_0, \beta_r)$ and for all samples are:

$$\boldsymbol{\eta} = \boldsymbol{X}_r\breve{\boldsymbol{\beta}}_r + \boldsymbol{Z}\boldsymbol{b} \tag{4}$$

where $\boldsymbol{\eta}$ is the vector of linear predictors, $\boldsymbol{X}_r$ is the design matrix for the r-th covariate and intercept, $\breve{\boldsymbol{\beta}}_r^T = (\beta_0, \beta_r)$ is the coefficient vector for the intercept and the r-th covariate, $\boldsymbol{Z}$ is the random effect design matrix, and $\boldsymbol{b} \sim \boldsymbol{N}(\boldsymbol{0}, \sigma_b^2 \boldsymbol{I})$ the random effect vector.

A more complete description of the single step of the bPMM algorithm is the calculation of Fisher's score and matrix functions, calculation of variance-covariance components, and determining initial values, stopping criteria, and selection in bPMM. The calculation of the variance-covariance matrix $\widehat{\mathbf{Q}}^{(l)}$ is performed using the Restricted Maximum Likelihood (REML) estimation approach or other alternative approaches. This approach is an alternative form of the Maximum Likelihood (ML) estimator, which is often used to reduce bias in estimating ML. REML estimation as a method of estimating variance-covariance components in an Unbalanced Incomplete Block Design based on optimization of the log-likelihood function. Another algorithm is Expectation-Maximization (EM). However, this algorithm has the main disadvantage of a slow process of convergence. Therefore, the EM algorithm is rarely used except to provide the initial value of another algorithm.

2.3 Case Study

Fertility refers to the number of children born alive, with the understanding that children who have been born in living conditions are showing signs of life. Fertility is usually measured by the frequency of births occurring in a given population and is more accurately represented as the number of live births per person or partner during their fertile period. One measure of fertility is the Total Fertility Rate (TFR), which is defined as the number of live births of men and women per 1000 women living until the end of their reproductive period. In 2017, the lowest Total Fertility Rate (TFR) occurred in Bali, while the highest was in NTT (Nusa Tenggara Timur). A high TFR value can often reflect several underlying conditions, such as the low average age of first marriage or the number of early marriages, low education levels, and low socio-economic levels. Therefore, various efforts are needed to suppress the TFR, one of which is by re-evaluating the family planning program. Seeing these conditions, this case study aims to analyze the factors influencing the number of live births in the two provinces, Bali and NTT (Nusa Tenggara Timur), to assess the success of their respective family planning programs in both regions. The factors influencing the high or low live birth rates can be divided into two categories: demographic and non-demographic. This study uses twenty-eight predictor variables representing both demographic and non-demographic factors. The response variable is the number of live births in women of reproductive age (15-49 years). The method used to model the number of live births in Bali and NTT is bPMM. While research [18] used bGLMM (EM) and bGLMM (REML), this study uses the REML algorithm, considering its model goodness and stability. The model for the case in Bali will be referred to as Model 1, and the model for the case in NTT will be referred to as Model 2.

2.3.1 Data

This study uses secondary data from the 2017 IDHS. The IDHS data is part of the international Demographic and Health Survey (DHS) program designed to provide information on birth rates, deaths, family planning, and health. Something similar to the IDHS is also carried out in Latin American, Asian, African, and Middle Eastern countries. In general, the IDHS questions are the same as the DHS (Demographic and Health Surveys) pattern. The response variable in the study used was the number of live children born to women within their reproductive age (15-49 years). The permanent effects used in this study consist of twenty-eight variables, which are presented in Table 1. The random effect used was the cluster, represented by the Census Block used in the 2017 IDHS. The number of Census Blocks used in Indonesia was 1.950, covering 49.250 households, and with a total of 59.100 women of childbearing (reproductive) age. In this study, sample coverage in Bali was 32 Census Blocks with a total of 500 women of childbearing age, while the sample from NTT included 86 Census Blocks comprising 1.327 women of childbearing age.

Table 1. The Explanatory Variables Used in This Study

Variables	Initial Notation	Description
Maternal Age	X_1	Continuous variable in years (15–49)
Age of first marriage	X_2	Continuous variable in years
Husband/partner's age	X_3	Continuous variable in years
Residential area	X_4	0: Rural; 1: Urban
Pregnancy history	X_5	0: Complications; 1: No complications
Marital status	X_6	0: No married; 1: Married/Never married
Education Level	X_7	0: No education; 1: Primary; 2: Secondary; 3: Higher
Working status	X_8	0: No; 1: Yes
Decision making: health problems	X_9	0: Doesn't involve a wife; 1: Involve wife
Decision making: Large household purchases	X_{10}	0: Excludes wife; 1: Involve wife
Decision making: Visits to family or relatives	X_{11}	0: Excludes wife; 1: Involve wife
Decision making: The money the husband/partner earns	X_{12}	0: Excludes wife; 1: Involve wife
The household has electricity	X_{13}	0: No; 1: Yes
The household has: radio	X_{14}	0: No; 1: Yes
The household has: television	X_{15}	0: No; 1: Yes
The household has: refrigerator	X_{16}	0: No; 1: Yes
The household has: bicycle	X_{17}	0: No; 1: Yes
The household has: motorcycle/scooter	X_{18}	0: No; 1: Yes
The household has: car/truck	X_{19}	0: No; 1: Yes
Has an account in a bank /financial institution	X_{20}	0: No; 1: Yes
Wealth Indkes	X_{21}	0: Poor; 1: Middle; 2: Rich
Current contraceptive method	X_{22}	0: No; 1: Yes
Desire for more children	X_{23}	0: Wants; 1: Undecided; 2: No more children desired
Husband's desire for children	X_{24}	0: Husband wants more; 1: Same preference (both); 2: Husband wants fewer
Ideal number of children	X_{25}	0: > 2 children ; 1: <= 2 children
Family Planning: radio	X_{26}	0: No; 1: Yes
Family Planning: television	X_{27}	0: No; 1: Yes
Family Planning: newspaper/magazine	X_{28}	0: No; 1: Yes

2.3.2 Data Analysis Procedure

The steps to get a bPMM for Model 1 and Model 2 are as follows:

1. Conduct data exploration processes on all variables used in research.
2. Initialization by calculating initial values $\widehat{\boldsymbol{\mu}}^{(0)}, \mathbf{Q}^{(0)}, \mathbf{b}^{(0)}$ with $\boldsymbol{\eta}^{(0)} = \mathbf{X}\widehat{\widehat{\boldsymbol{\beta}}}^{(0)} + \mathbf{Z}\mathbf{b}^{(0)}$.
3. Iteration for all l = 1, 2, ..., max with refitting residuals and calculating the covariance-variance component by estimating from $\widehat{\mathbf{Q}}^{(l)}$ is obtained by the REML estimation approach.
4. The criteria stop by finding the appropriate complexity of the model by using effective degrees of freedom, which are given by the trace of the hat matrix [1].

5. Selection in bPMM, the given hat matrix shows the complexity of the model determined by the information criteria used in the boosting step that minimizes AIC or BIC.

$$AIC_r^{(l)} = -2l\left(\hat{\boldsymbol{\mu}}_r^{(l)}\right) + 2\,\text{trace}\left(\hat{\mathbf{H}}_r^{(l)}\right),$$

$$BIC_r^{(l)} = -2l\left(\hat{\boldsymbol{\mu}}_r^{(l)}\right) + 2\,\text{trace}\left(\hat{\mathbf{H}}_r^{(l)}\right)\log(n)$$

with

$$l\left(\hat{\boldsymbol{\mu}}_r^{(l)}\right) = \sum_{i=1}^{n} \log f\left(\mathbf{y}_i \left| \hat{\boldsymbol{\mu}}_{ir}^{(l)}\right.\right)$$

For example, chosen $L := \{1,2,\ldots,l_{max}\}$the lopt component, where $AIC^{(l)}$ or $BIC^{(l)}$ is the smallest, i.e.

$$l_{opt} = \arg\min_{l\in\mathcal{L}} AIC^{(l)}$$

$$l_{opt} = \arg\min_{l\in\mathcal{L}} BIC^{(l)}$$

Finally, the estimation of the parameters $\hat{\boldsymbol{\delta}}^{(l_{opt})}, \hat{\mathbf{Q}}^{(l_{opt})}$ is obtained and corresponds to the model $\hat{\boldsymbol{\mu}}^{(l_{opt})}$, for $r \in \{1, \ldots, p\}$ component j which produces the smallest $AIC_r^{(l)}$ or $BIC_r^{(l)}$ with $\left(\hat{\boldsymbol{\delta}}_j^{(l)}\right)^T = \left(\hat{\beta}_0^*, \hat{\beta}_j^*, (\hat{\mathbf{b}}^*)^T\right)$.

6. Interpretation of both models.

3. RESULTS AND DISCUSSION

Fertility refers to the actual production capacity of a population (actual reproductive performance) or the number of live births experienced by one or a group of women. As outlined in the previous chapter, one of the objectives of this research is to find out the success of the family planning program. For this purpose, the number of live births is categorized into two groups: (1) a maximum of two live-born children, and (2) more than two live-born children. Based on Table 4, it can be seen that in Bali, the majority of women who have a maximum number of children born alive tend to reside in rural areas, have no history of pregnancy complications, are currently married, have completed primary education, and demonstrate good autonomy status as indicated in variables X_7 to X_{12}. These women also generally have better economic status, as indicated by variables X_{13} to X_{21}, and demonstrate active participation in family planning programs, as reflected in variables X_{22} to X_{28}. Overall, these variables show higher proportions in the first category. Similar patterns are also observed among women in NTT, with the main differences found in the characteristics of residence area, pregnancy history, marital status, and wealth index.

NTT remains the region with the highest TFR in Indonesia. Fig. 1 shows that in NTT, there are still women who have more than 10 live-born children, while women in Bali province only have a maximum of 10 live-born children. Across both provinces, most women have 2 to 3 live-born children.

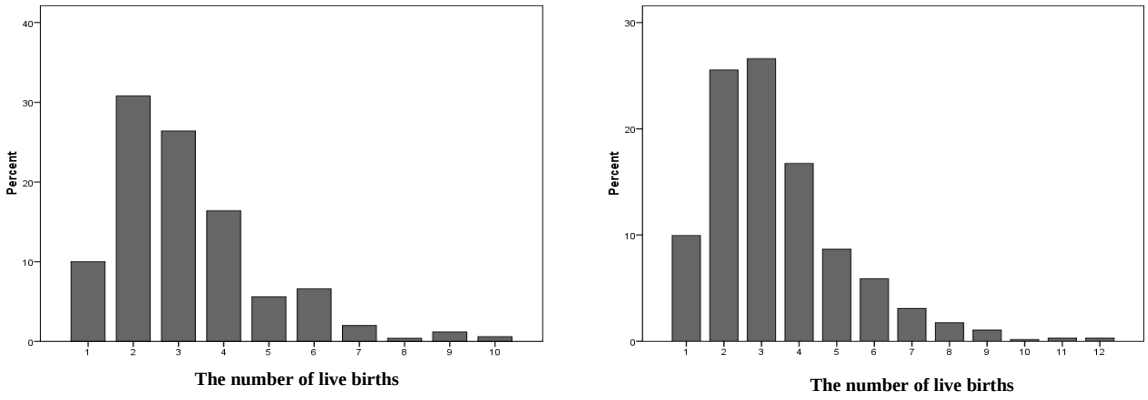

Figure 1. Live Births Percentage Among WRA in (a) Bali and (b) East Nusa Tenggara
(Source: RStudio)

Data plots on the variables of wives' age, age of first marriage, and husbands' age in the provinces of Bali and NTT have similar patterns (Fig. 2). The highest number of live births in Bali province is reached by women at the age of 30-40 years, with the age of first marriage around 25 years, and the age of the husband between 30-40 years. NTT Province has the highest number of live births, with most women aged 25-30 years, had their first marriage at around 20 years old, and the age of the husband is between 30-40 years. When viewed from the six data plots, there is no correlation.

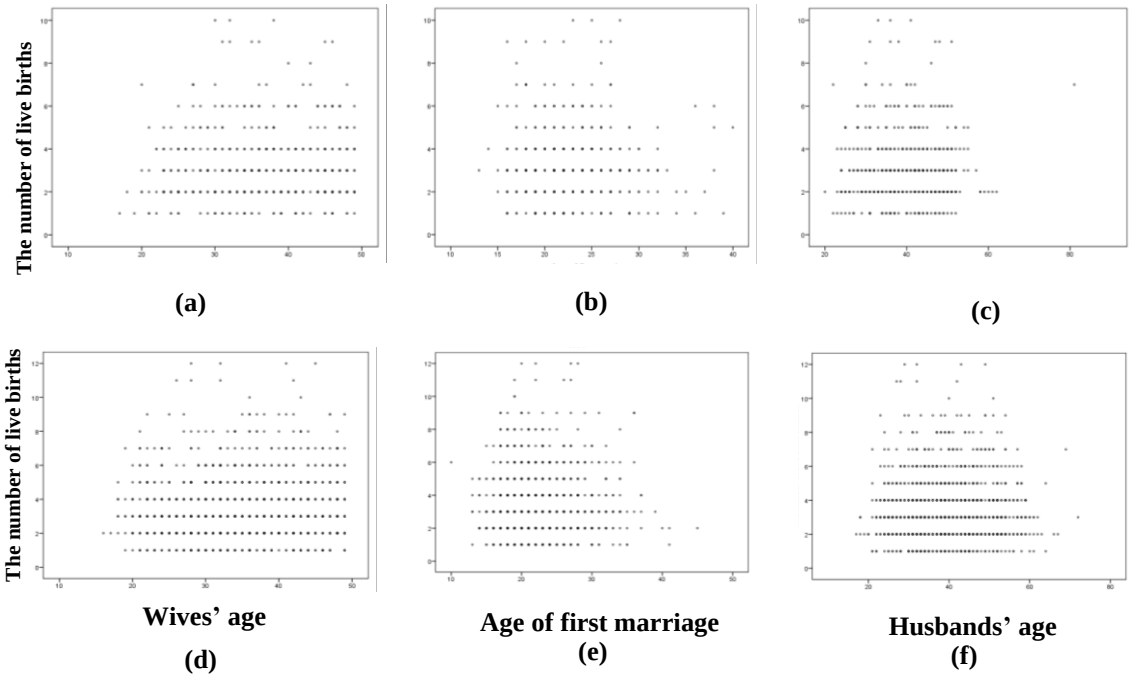

Figure 2. Scatter Plot Variable: The Number of Live Births with (a) Wives' Age in Bali, (b) Age of First Marriage in Bali, (c) Husbands' Age in Bali, (d) Wives' Age in NTT, (e) Age of First Marriage in NTT, and (f) Husbands' Age in NTT
(Source: RStudio)

3.1 Boosted Poisson Mixed Models for Live Births in Bali

An illustration of how model 1 works using the REML algorithm for the coefficients that enter the model is shown in Fig. 3 (a). In this study, the maximum step (l_{max}) used was 100 steps. The selected variables model 2 are the residential area (X_4), households with motorcycles (X_{18}), receiving family planning information from radio (X_{26}), and the source of family planning information comes from

television (X_{27}). Estimation of the coefficients is carried out in stages by setting the initial coefficients to zero.

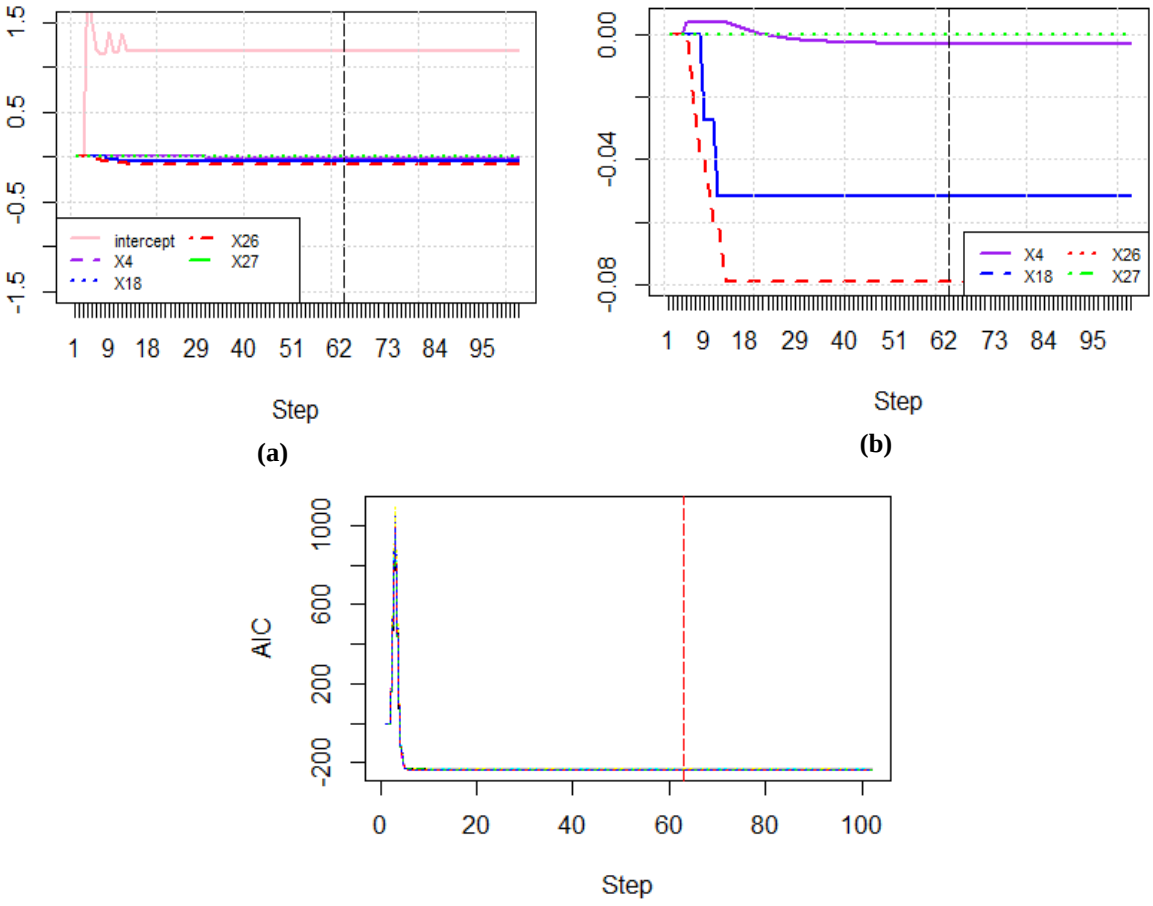

Figure 3. The Plot of Predictor (a) Model with Intercept (b) Model without Intercept, and (c) AIC Values in Model 1
(Source: RStudio)

In Fig. 3 (b), a plot was generated without including the intercept term in order to identify significant variables entered into model 1. It can be seen that from twenty-eight variables, only three variables significantly affect the number of live births in Bali; these variables are X_4, X_{18} and X_{26}. The first variable entered into the model is X_4, followed successively by the variable X_{26}, and the last is X_{18}. It can be seen in Fig. 3 (b) that the optimum step (l_{max}) selected in model 1 are at step 63, even after step 48, all coefficients have stabilized, even though the variables X_{18} and X_{26} have begun to stabilize at step 14. This is under research by [18], which states that the optimum step in bPMM is determined by the minimum AIC and BIC values at the boosting step. If we look at the AIC value, before the 10th step, the AIC value of all variables has started to stabilize, and in the previous step, the AIC value has experienced a very sharp spike Fig. 3 (c).
Model 1 has form:

$$\log(\mu_{ij}) = 1.18420 - 0.00286X_{4_{ij}} - 0.05174X_{18_{ij}} - 0.07910X_{26_{ij}} + 0.00005X_{27_{ij}} + b_i$$

with μ_{ij} is the average number of live births of women of reproductive age in the i Census Block and b_i is a random intercept in the Census Block.

Table 2. Estimator Coefficient, Standard Error, and p-value on Model 1

Description	Coefficient	Standard Error	p-value	Description	Coefficient	Standard Error	p-value
Fixed Effect				**Fixed Effect**			
(Intercept)	1.1842	0.1424	0	X_{16}	0	0	-
X_1	0	0	-	X_{17}	0	0	-
X_2	0	0	-	$X_{18(1)}$	-0.0517	0.0203	0.0100
X_3	0	0	-	X_{19}	0	0	-
$X_{4(1)}$	-0.0029	0.0027	0.0283	X_{20}	0	0	-
X_5	0	0	-	X_{21}	0	0	-
X_6	0	0	-	X_{22}	0	0	-
X_7	0	0	-	X_{23}	0	0	-
X_8	0	0	-	X_{24}	0	0	-
X_9	0	0	-	X_{25}	0	0	-
X_{10}	0	0	-	$X_{26(1)}$	-0.0791	0.0273	0.0030
$X_{11(1)}$	0	0	-	X_{27}	0.0001	0.0131	0.9971
X_{12}	0	0	-	X_{28}	0	0	-
X_{13}	0	0	-	**Random Effects**			
X_{14}	0	0	-		Variance		
X_{15}	0	0	-	Cluster (BS)	0.0340		

As seen in Table 2, the area of residence (X_4), households with motorcycles (X_{18}), and receiving family planning information from the radio (X_{26}) significantly influenced the number of live births in Bali with different real levels. Women who live in rural areas are estimated to have approximately 0.3% more children compared to those in urban areas. Similarly, women who do not own a motorcycle are expected to have about 5.3% more live births than those who do. Moreover, women who are not exposed to family planning information through the radio tend to have approximately 8.2% more live births than women who receive such information. These results suggest that limited economic resources and restricted access to reproductive health information are associated with higher fertility. The findings highlight the importance of targeted interventions to improve both economic empowerment and access to family planning outreach, particularly in rural communities.

These findings suggest that women living in rural areas, with low economic status, and with no family planning, tend to have higher fertility rates. This condition is further supported by the percentage of women in Bali who have live births of more than 2 people and live in rural areas is 51 percent, where 76.2 percent do not possess a motorcycle, and 60.6 percent do not receive family planning information from the radio. However, this trend appears to contrast with contraceptive usage data, where 96 percent of women who have family members who live in rural areas, are reporting to use contraception-highlighting the complexity of factors influencing fertility. Despite these factors, Bali continues to report the lowest fertility rate among provinces.

The model 1 uses the Census Block as a random effect. In the model, a different coefficient will be obtained for each j-woman in the i Census Block. Table 2 shows that the diversity of the Census Block random variables is 0.0340.

3.2 Boosted Poisson Mixed Models for Modeling Live Births in NTT

Based on Fig. 4 (a) and (b), it can be seen that the variable residential area (X_4) and the household has a refrigerator (X_{16}) enter model 2. The smallest AIC value was obtained in step 6 with a value of -890.739. It can be seen in Fig. 4 (c) that the AIC value for all coincident variables indicates the AIC value obtained tends to be almost the same for each variable. The highest AIC value was obtained in the first step with a positive AIC value, then down in the second step with a negative AIC value, and began to stabilize in the sixth step.

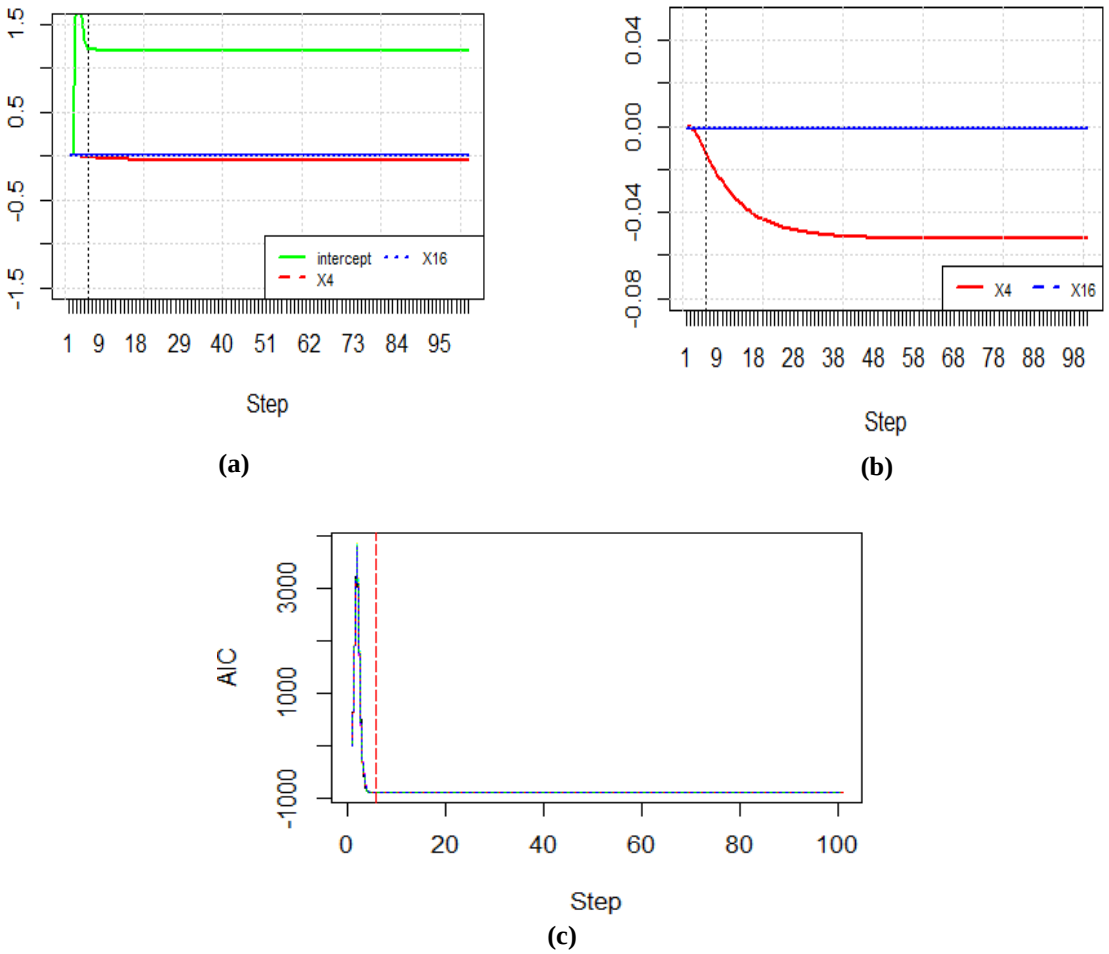

Figure 4. The Plot of Predictor (a) Model with Intercept, (b) Model without Intercept, and (c) AIC Values in the NTT Model
(Source: RStudio)

Regional variables of residence significantly affect the number of live births in NTT with a significant level of 10 percent (Table 3). This can also be observed from Fig. 4 (b), where the coefficient value of the variable does not coincide with zero, unlike the household has refrigerator variable, which tends to coincide with zero.

Table 3. Estimator Coefficient, Standard Error, and *p*-value on Model 2

Description	Coefficient	Standard Error	*p*-value	Description	Coefficient	Standard Error	*p*-value
Fixed Effect				**Fixed Effect**			
(Intercept)	1.2113	0.0254	0	X_{19}	0	0	-
X_1	0	0	-	$X_{20(1)}$	0	0	-

Description	Coefficient	Standard Error	p-value	Description	Coefficient	Standard Error	p-value
X_2	0	0	-	X_{21}	0	0	-
X_3	0	0	-	X_{17}	0	0	-
X_4	-0.0162	0.0090	0.0730	X_{18}	0	0	-
X_5	0	0	-	X_{22}	0	0	-
$X_{6(1)}$	0	0	-	X_{23}	0	0	-
X_7	0	0	-	X_{24}	0	0	-
X_8	0	0	-	$X25$	0	0	-
X_9	0	0	-	X_{26}	0	0	-
X_{10}	0	0	-	X_{27}	0	0	-
X_{11}	0	0	-	X_{28}	0	0	-
X_{12}	0	0	-				
X_{13}	0	0	-	**Random Effects**			
X_{14}	0	0	-		Variance		
X_{15}	0	0	-	Cluster (BS)	0.0215		
X_{16}	-0.0011	0.0114	0.9216				

The results of the selection of variables with bPMM on twenty-eight predictor variables can be seen in Table 4. The bPMM model 2 formed is:

$$\log(\mu_{ij}) = 1.21134 - 0.01621X_{4_{ij}} - 0.00112X_{16_{ij}} + b_i$$

NTT (Nusa Tenggara Timur) is an archipelagic province where much of its territory consists of islands separated by the sea. Consequently, many rural areas differ substantially from urban areas in terms of access to public infrastructure such as healthcare facilities, clean water, electricity, and transportation. These disparities contribute to regional differences in fertility patterns. The diversity of areas selected into the random effect structure reflects this variation, with a calculated variance of 0.0215. When compared to Bali, the selected Census Blocks in NTT appear more homogeneous, suggesting a relatively consistent pattern of contextual influences within rural settings across the province.

Table 4. Distribution of Samples Based on Variables Used

Variable	Bali Model		NTT Model		Variable	Bali Model		NTT Model	
	>2	≤ 2	>2	≤ 2		≤ 2	≤ 2	>2	≤ 2
X_4					X_{17}				
0	51.0	**49.0**	65.0	35.0	0	59.5	40.5	64.0	**36.0**
1	62.9	37.1	62.7	**37.3**	1	58.9	**41.1**	68.2	31.8
X_5					X_{18}				
0	59.8	40.2	64.0	**36.0**	0	76.2	23.8	66.2	34.8
1	51.4	**48.6**	66.9	33.1	1	58.5	**41.5**	63.7	**36.3**
X_6					X_{19}				
0	66.7	33.3	65.6	34.4	0	56.5	**43.5**	64.9	35.1
1	58.9	**41.1**	60.1	**39.9**	1	67.5	32.5	58.4	**41.6**
X_7					X_{20}				
0	66.7	33.3	64.2	**35.8**	0	55.1	44.9	65.7	34.3

Variable	Bali Model		NTT Model		Variable	Bali Model		NTT Model	
	>2	≤ 2	>2	≤ 2		≤ 2	≤ 2	>2	≤ 2
1	54.0	**46.0**	64.3	35.7	1	63.4	**49.2**	62.1	**37.9**
2	60.3	39.7	64.3	35.7	X_{21}				
3	62.3	37.7	66.1	33.9	0	48.7	28.9	64.6	35.4
X_8					1	58.6	**41.4**	67.1	32.9
0	59.0	**41.0**	64.4	**35.6**	2	63.4	36.6	61.0	**39.0**
1	71.4	28.6	76.5	23.5	X_{22}				
X_9					0	65.5	34.5	64.2	**35.8**
0	66.0	34.0	70.7	29.3	1	58.8	**41.2**	64.6	35.4
1	58.4	**41.6**	64.1	**35.9**	X_{23}				
X_{10}					0	54.6	33.8	62.7	**37.3**
0	65.9	34.1	66.2	33.8	1	76.2	23.8	76.1	23.9
1	55.8	**44.2**	64.3	**35.7**	2	60.4	**39.6**	65.4	34.6
X_{11}					X_{24}				
0	73.8	26.2	65.6	34.4	0	52.7	**47.3**	65.0	35.0
1	57.0	**43.0**	64.5	**35.5**	1	60.8	39.2	64.9	35.1
X_{12}					2	57.6	42.4	63.2	**36.8**
0	63.3	36.7	67.9	32.1	X_{25}				
1	58.1	**41.9**	64.2	**35.8**	0	60.0	40.0	63.9	**36.1**
X_{13}					1	58.9	**41.1**	66.3	33.7
0	85.7	14.3	64.5	35.5	X_{26}				
1	58.8	**41.2**	64.5	**35.5**	0	60.6	39.4	64.1	**35.9**
X_{14}					1	51.4	**48.6**	67.7	32.3
0	57.0	**43.0**	64.9	35.1	X_{27}				
1	62.0	38.0	60.2	**39.8**	0	56.8	**43.2**	64.4	**35.6**
X_{15}					1	61.4	38.6	64.7	35.3
0	60.0	40.0	65.2	34.8	X_{28}				
1	59.2	**40.8**	63.6	**36.4**	0	59.5	40.5	64.2	**35.8**
X_{16}					1	56.1	**43.9**	67.2	32.8
0	53.6	**46.4**	64.6	35.4					
1	61.7	38.3	64.1	**35.9**					

Based on Table 4, the distribution of women of reproductive age in Bali and NTT, categorized by whether they have more than two or at most two live-born children, based on selected explanatory variables. For instance, 65% of women in NTT who live in rural areas (X_4) have more than two children, compared to 35% who have two or fewer. This suggests a strong association between rural residence and higher fertility. Similarly, the table indicates that 39.8% of rural women in NTT only completed primary education (X_7), and 96% fall into the poor wealth index category (X_{21}), showing the predominance of low socio-economic status in rural settings.

Despite this, 81.5% of these women report participating in family planning programs, as reflected in the high percentage (X_{22}) among those with more than two children. This apparent contradiction may reflect the limited effectiveness or delayed impact of such programs in high-fertility areas. Notably,

although the program coverage is high, it may not yet translate into a significant reduction in fertility, likely due to underlying cultural or structural factors such as early marriage, limited access to health education, or low contraceptive consistency.

This analysis focuses on NTT, the province with the highest fertility rate in Indonesia. However, comparisons to Bali also reveal important contrasts. For example, in Bali, urban women (X_4) are more prevalent among those with fewer than two children (37.1%), while rural women dominate in the higher fertility group (51%). These patterns suggest that geographical, educational, and economic disparities contribute significantly to fertility outcomes.

4. CONCLUSION

The boosted Poisson mixed model (bPMM) was applied to analyze the number of live births among women of reproductive age in the provinces of Bali and East Nusa Tenggara (NTT). Using twenty-eight explanatory variables as fixed effects and one random effect at the Census Block level, the study identified key predictors for each region. In Bali, residential area, motorcycle ownership, and access to family planning information via radio significantly influenced the number of live births, while in NTT, only residential area emerged as a significant factor. The model achieved optimal fit at step 63 for Bali and step 6 for NTT, based on the minimum AIC values. The results suggest that women residing in rural areas with lower economic status and limited exposure to family planning tend to have higher fertility rates, particularly in Bali. Although the family planning program in rural NTT appears effective based on high participation rates, the province still reports the highest fertility rate in Indonesia. It is important to note that this study focused only on two provinces and relied on cross-sectional data, which may not fully reflect temporal changes or national patterns. Additionally, the assumption of a Poisson distribution without adjusting for potential overdispersion and the challenges of variable selection in high-dimensional settings may affect the precision of the estimates. Future research should consider extending the model to accommodate overdispersion and zero-inflation, incorporating longitudinal or more recent data, and integrating spatial and measurement error adjustments to enhance model robustness and policy relevance.

Author Contributions

Ita Wulandari: Conceptualization, Methodology, Writing-Original Draft, and Data Curation. Khairil Anwar Notodiputro: Draft Preparation and Formal Analysis. Bagus Sartono: Formal Analysis and Visualization. Anwar Fitrianto: Formal Analysis and Writing-Review and Editing. Anang Kurnia: Formal Analysis and Writing-Review. All authors discussed the results and contributed to the final manuscript.

Funding Statement

This research was funded by Statistics Indonesia (BPS) through a doctoral scholarship grant from the national budget (APBN) of Indonesia.

Acknowledgment

The author gratefully acknowledges the guidance and support provided by academic supervisors and lecturers at IPB University throughout the doctoral program. Appreciation is also extended to Statistics Indonesia (BPS) for awarding the doctoral scholarship funded by the national budget (APBN). Their contributions were instrumental in the completion of this research. Any remaining errors are the sole responsibility of the author.

Declarations

We declare that we have no conflicts of interest to report in this study. The research was conducted without any commercial or financial relationships that could be construed as a potential conflict of interest.

REFERENCES

[1] T. Hastie, R. Tibshirani, and J. Friedman, *THE ELEMENTS OF STATISTICAL LEARNING: DATA MINING, INFERENCE, AND PREDICTION*, 2nd ed. New York: Springer, 2009. [Online]. Available: https://link.springer.com/book/10.1007/978-0-387-84858-7?utm_source=chatgpt.com

[2] M. Balzer, E. Bergherr, S. Hutter, and T. Hepp, *GRADIENT BOOSTING FOR DIRICHLET REGRESSION MODELS*, no. 0123456789. Springer Berlin Heidelberg, 2025. doi: https://doi.org/10.1007/s10182-025-00526-5

[3] A. Alsahaf, N. Petkov, V. Shenoy, and G. Azzopardi, "A FRAMEWORK FOR FEATURE SELECTION THROUGH BOOSTING", *Expert Syst. Appl.*, vol. 187, no. Sept. 2021, p. 115895, 2022. doi: https://doi.org/10.1016/j.eswa.2021.115895.

[4] G. Schultz Lindenmeyer and H. da Silva Torrent, "*BOOSTING AND PREDICTABILITY OF MACROECONOMIC VARIABLES: EVIDENCE FROM BRAZIL*", vol. 64, no. 1. Springer US, 2024. doi: https://doi.org/10.1007/s10614-023-10421-3.

[5] P. Bühlmann and T. Hothorn, "BOOSTING ALGORITHMS: REGULARIZATION, PREDICTION AND MODEL FITTING", *Stat. Sci.*, vol. 22, no. 4, pp. 477–505, 2007. doi: https://doi.org/10.1214/07-STS242.

[6] Y. Freund and R. E. Schapire, "EXPERIMENTS WITH A NEW BOOSTING ALGORITHM", *Proc. 13th Int. Conf. Mach. Learn.*, pp. 148–156, 1996, doi: https://doi.org/10.1.1.133.1040.

[7] R. Wang, "ADABOOST FOR FEATURE SELECTION, CLASSIFICATION AND ITS RELATION WITH SVM, A REVIEW", *Phys. Procedia*, vol. 25, pp. 800–807, 2012. doi: https://doi.org/10.1016/j.phpro.2012.03.160.

[8] L. Pebrianti, F. Aulia, H. Nisa, and K. Saputra S, "IMPLEMENTATION OF THE ADABOOST METHOD TO OPTIMIZE THE CLASSIFICATION OF DIABETES DISEASES WITH THE NAÏVE BAYES ALGORITHM", *J. Sist. dan Teknol. Inf.*, vol. 7, no. 2, pp. 122–127, 2022, [Online]. Available: http://jurnal.unmuhjember.ac.id/index.php/JUSTINDO

[9] P. Beja-Battais and C. Borelli, "OVERVIEW OF ADABOOST : RECONCILING ITS VIEWS TO BETTER UNDERSTAND ITS DYNAMICS", *arXiv:2310.18323v1*, pp. 3–31, 2023, [Online]. Available: http://arxiv.org/abs/2310.18323

[10] J. Friedman, T. Hastie, and R. Tibshirani, "ADDITIVE LOGISTIC REGRESSION: A STATISTICAL VIEW OF BOOSTING", *Ann. Stat.*, vol. 28, no. 2, pp. 337–374, 2000. doi: https://doi.org/10.1214/aos/1016120463.

[11] A. S. Suggala, B. Liu, and P. Ravikumar, "GENERALIZED BOOSTING," in *Advances in Neural Information Processing Systems 33*, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, Eds. Vancouver, Canada: NeurIPS, 2020.

[12] L. Knieper, T. Hothorn, E. Bergherr, and C. Griesbach, "GRADIENT BOOSTING FOR GENERALISED ADDITIVE MIXED MODELS", *Stat. Comput.*, vol. 35, no. 4, 2025. doi: https://doi.org/10.1007/s11222-025-10612-y.

[13] G. Tutz and A. Groll, "LIKELIHOOD-BASED BOOSTING IN BINARY AND ORDINAL RANDOM EFFECTS MODELS", *J. Comput. Graph. Stat.*, vol. 22, no. 2, pp. 356–378, Jan. 2013. doi: https://doi.org/10.1080/10618600.2012.694769.

[14] E. Fammaldo and M. Lestari, "GRADIENT BOOSTING TREES UNTUK PEMODELAN DAN PREDIKSI BIAYA KERUGIAN ASURANSI MOBIL", no. 01, pp. 634–642, 2024.

[15] Y. Yang, W. Qian, and H. Zou, "A BOOSTED TWEEDIE COMPOUND POISSON MODEL FOR INSURANCE PREMIUM", *arXiv:1508.06378v2 [stat.ME]*, p. 11, Aug. 2016. doi: https://doi.org/10.1080/07350015.2016.1200981.

[16] Badan Pusat Statistik, "HASIL SENSUS PENDUDUK 2020", in *Statistik Demografi Indonesia*, Badan Pusat Statistik, 2025.

[17] Kementrian Kesehatan RI, Badan Pusat Statistik RI, and USAID, *SURVEI DEMOGRAFI DAN KESEHATAN INDONESIA TAHUN 2017*. 2018. [Online]. Available: https://ia802800.us.archive.org/30/items/LaporanSDKI2017/Laporan SDKI 2017.pdf

[18] G. Willame, J. Trufin, and M. Denuit, "BOOSTED POISSON REGRESSION TREES: A GUIDE TO THE BT PACKAGE IN R", *Ann. Actuar. Sci.*, vol. 18, no. 3, pp. 605–625, 2024. doi: https://doi.org/10.1017/S174849952300026X.

[19] F. Hadiji, A. Molina, S. Natarajan, and K. Kersting, "POISSON DEPENDENCY NETWORKS: GRADIENT BOOSTED MODELS FOR MULTIVARIATE COUNT DATA", *Mach. Learn.*, vol. 100, no. 2–3, pp. 477–507, 2015. doi: https://doi.org/10.1007/s10994-015-5506-z.

[20] G. Gao, H. Wang, and M. V. Wüthrich, "BOOSTING POISSON REGRESSION MODELS WITH TELEMATICS CAR DRIVING DATA", *Mach. Learn.*, vol. 111, no. 1, pp. 243–272, 2022. doi: https://doi.org/10.1007/s10994-021-05957-0.

[21] L. Breiman, "ARCING CLASSIFIERS", *Ann. Stat.*, vol. 26, no. 3, pp. 801–824, Jun. 1998, [Online]. Available: http://www.jstor.org/stable/120055. doi: https://doi.org/10.1214/aos/1024691079

[22] J. H. Friedman, "GREEDY FUNCTION APPROXIMATION: A GRADIENT BOOSTING MACHINE", *Ann. Stat.*, vol. 29, no. 5, pp. 1189–1232, 2001, [Online]. Available: http://www.jstor.org/stable/2699986?origin=JSTOR-pdf. doi: https://doi.org/10.1214/aos/1013203451

[23] P. Bühlmann and B. Yu, "BOOSTING WITH THE L2 LOSS: REGRESSION AND CLASSIFICATION", *J. Am. Stat. Assoc.*, vol. 98, no. 462, pp. 324–339, 2003. doi: https://doi.org/10.1198/016214503000125.