CAUCHY – Jurnal Matematika Murni dan Aplikasi
Volume 10 (2) (2025), Pages 507-518
p-ISSN: 2086-0382; e-ISSN: 2477-3344

Identification and Modelling Tuberculosis Incidence Risk Factors in West Java with Negative Binomial Mixed Model Random Forest

Restu Arisanti[1*], Resa Septiani Pontoh[1], Sri Winarni[1], Nisa Akbarilah Putri[2], and Stefany Maurin[2]

[1] *Department of Statistics, Universitas Padjadjaran*
[2] *Bachelor Program of Statistics' Department, Universitas Padjadjaran*

Abstract

Tuberculosis (TB) is a major health problem in many parts of the world, including in West Java Province, Indonesia. Accurate assessment of TB risk factors can improve overall TB control efforts. This study introduces modelling by integrating Negative Binomial Mixed Models (NBMM) and Random Forest (RF) called the Negative Binomial Mixed Model Random Forest (NBMMRF) model. This model is used to identify and assess risk factors associated with the incidence of tuberculosis. Firstly, using NBMM to add fixed effects and random effects in the model and compensate for overdispersion. Afterwards, we included a Random Forest component in the model, which helped us detect relevant predictive features and change model weights accordingly. The resulting Negative Binomial Mixed Model Random Forest has a high accuracy value of up to 0.915. After obtaining the appropriate model, forecasting of tuberculosis cases in 2024 was carried out and the forecasting result was 1904 cases, which increased from the number of cases in the previous year. The results of this study show the importance of various related parties to continue to be vigilant, pay attention to various related risk factors, and continue to make various efforts to study, prevent, and control Tuberculosis disease effectively.

Keywords: NBMMRF; Negative Binomial Mixed Model; Random Forest; Tuberculosis

1 Introduction

Infectious diseases arise as a result of the interaction of various factors from the agent, host and environment. Disease transmission can be transmitted in many ways including through viruses and bacteria [1]. Tuberculosis is one example of a disease transmitted through the Mycobacterium tuberculosis bacteria. These bacteria can spread through the air, for example if a tuberculosis patient sneezes or coughs, the Mycobacterium bacteria can be spread and infected to other individuals [2]. There is a 5% to 10% risk of not recovering for tuberculosis patients. A person has a higher risk of developing tuberculosis if they have HIV disease, malnutrition, or diabetes [1], [2].

Tuberculosis is the 13th leading cause of death in the world, according to the WHO Indonesia is responsible for 8.4% of tuberculosis cases worldwide after China and India and is the third largest country with the highest number of tuberculosis cases [3]. West Java is the province with the most tuberculosis cases in Indonesia in 2021, with a total of 85,681 cases reported [4]. In West Java, there were nearly

[*] **Corresponding author**. E-mail: r.arisanti@unpad.ac.id

Submitted: November 07, 2024 Reviewed: May 04, 2025 Accepted: June 03, 2025
DOI: https://doi.org/10.18860/cauchy.v10i2.29750

60,000 TB cases in 2019, with a prevalence rate of 280 cases per 100,000 population, according to data from the West Java Provincial Health Office.

Modelling of tuberculosis cases in West Java province was carried out through the integration of the generalised linear mixed model (GLMM) model with extreme neural networks [1]. Following that, in this research use negative binomial mixed model and Random Forest method were combination of the generalised linear mixed model (GLMM) and Random Forest approaches. Values such as R-Square are used to measure the accuracy of the prediction results from the Random Forest method. In this study, modelling of tuberculosis case data in West Java is done with a negative binomial mixed model. Furthermore, Random Forest will be applied for forecasting tuberculosis cases in the following year based on the risk factors identified in the study. Thus, the information obtained from this study can be used as a basis for the government and the community to improve the quality of tuberculosis prevention and control. This is especially true in West Java Province.

This study presents a novel approach by developing a hybrid model named Negative Binomial Mixed Model Random Forest (NBMMRF), which integrates the Negative Binomial Mixed Model (NBMM) with Random Forest techniques. The novelty of this research lies in combining a statistical model capable of handling overdispersed count data with random effects (NBMM) and a machine learning algorithm (Random Forest) that enhances predictive performance and identifies important predictors. This integration has not been previously applied in the context of tuberculosis incidence modeling in Indonesia. The main contribution of this study is the formulation and implementation of the NBMMRF framework for longitudinal and region-based epidemiological data, supported by empirical evidence showing a high predictive accuracy with an R^2 value of 0.915. The model provides a practical and data-driven basis for forecasting TB cases and informing targeted public health interventions in West Java.

The remainder of this paper is structured as follows. Section 2 describes the data sources, modeling framework, and methodology including the Negative Binomial Mixed Model and Random Forest integration. Section 3 presents the results and discussion, including model selection, evaluation, and forecasting outcomes. Section 4 concludes the study by summarizing the findings and outlining potential implications for public health policy and future research.

2 Methods

This section describes the methodological framework adopted in this study. It begins by outlining the modeling approach used for longitudinal tuberculosis data, followed by the specification of the Negative Binomial Mixed Model, parameter estimation techniques, and the integration of Random Forest for prediction and forecasting. Evaluation metrics used to assess model performance are also presented.

2.1 Modelling Longitudinal Data

Longitudinal effects of several risk variables on TB progression were modelled in this study using GLMM. Both fixed effects and random effects, as well as correlations between repeated observations over time, can be accounted for through GLMM [1]. In addition, GLMM is also able to handle non normal distribution of response variables and missing data [5]. Then, investigate the data distribution using Cullen and Frey [6].

2.2 Generalized Linear Mixed Models (GLMM)

A method for handling non normal distributions of response variables, generalized linear mixed models (GLMM) combine generalized linear models with linear mixed models to produce precise variance estimates for complex data. Fixed effects and random effects on linear predictors are included in GLMM, an extension of Generalized Linear Models (GLM) [1]. Regression models that offer various distributions and linking functions are GLMMs. The purpose of the linking function is to adjust the value of the dependent variable to fit the scale of the linear predictor. The relationship with the predictor variables will then be linearized as a result. The general form of the GLMM is as follows (in matrix notation):

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}\mathbf{b} + \varepsilon \tag{1}$$

with:

- $\mathbf{y}$: vector of response variables ($N \times 1$),
- $\mathbf{X}$: matrix obtained according to the fixed effects ($N \times p$),
- β: column vector for fixed effect parameters ($p \times 1$),
- $\mathbf{Z}$: design random effects matrix ($N \times 1$),
- $\mathbf{b}$: random effects vector,
- ε: column vector of residuals ($N \times 1$) [7].

2.3 Negative Binomial Mixed Model

The selection of a suitable distribution and link function for the supplied data is the first stage in the modeling procedure. For count data, the natural distributions are Poisson or, in the case of overdispersion, the negative binomial distribution [8]. The negative binomial distribution was used in this research to account for overdispersion. We presumptively observe the negative binomial distribution for the counting response y_i:

$$y_i \sim \mathrm{NB}(y_i \mid \mu_i, \theta) = \frac{\Gamma(y_i + \theta)}{\Gamma(\theta) y_i!} \left(\frac{\theta}{\mu_i + \theta}\right)^{\theta} \left(\frac{\mu_i}{\mu_i + \theta}\right)^{y_i} \tag{2}$$

where:

- μ_i: mean,
- θ: parameter shape,
- Γ: function gamma [9].

The predictor variable X is as many as units, the random variable Z is as many as random factors, and the total sequence T, as read by the logarithm of the link function, are all related to the mean parameter (μ) in the negative binomial mixed model.

$$\log(\mu) = \log(T) + \mathbf{X}\beta + \mathbf{Z}\mathbf{b} \tag{3}$$

where:

- $\log(\mu)$: the offset, which accounts for differences in the total number of sequence reads across samples,
- β: the host factor factors' fixed effects vector X,
- $\mathbf{b}$: the vector of random effects for Z.

The correlation between the samples and the various sources of variation are modeled using random effects, helping to prevent biased inference on the impact of the predictor variable X. Typically, it is believed that the random effects vector will be assumed to be the multivariate normal distribution [10].

2.4 Parameter Estimation for GLMM

GLMM can be used to study longitudinal data because it can model both within-subject correlation and between-subject variation. Correlation between measurements occurs because in longitudinal studies, the same subjects are measured repeatedly over time. Such correlations must be considered in statistical analyses [1]. GLMM analysis will be conducted using `R` software and with the help of `glmmADMB` package.

Finding parameters that optimize the total likelihood of a data set is the basic objective of MLE. This is accomplished in the setting of the exponential family by maximizing the log-likelihood function $\ell(\theta; y, \phi)$, over the canonical parameter θ depending on the observation y and the scale parameter ϕ [11], [12].

The parameter vector determines the link function based on the distribution of dependent variables; the link function then determines the mean, i.e. μ, where the inverse link function. The canonical parameter is a function of the mean ($\theta(\mu)$). So, the following is a general formulation of the exponential family's log-likelihood.

$$\ell(\beta; y, \phi) = \frac{y\{\theta[g^{-1}(X\beta)]\} - u\left(\theta[g^{-1}(X\beta)]\right)}{a(\phi)} + c(y, \phi) \tag{4}$$

2.5 Forecasting

Forecasting plays a significant role in decision making for all companies that are concerned with the future. A conventional time series forecasting model's standard operating method is to find the pattern that most closely matches the previous data. In other words, a functional form is chosen that best captures the relationship between the input (past observations) and output (prediction) of the system [13]. Large and complicated data sets, like demographic, clinical, and laboratory data sets, can be analyzed using machine learning techniques. Machine learning algorithms can be trained to forecast a variety of outcomes, including the likelihood that a patient will contract TB, the efficacy of their treatment, and other outcomes. The Feedforward Neural Network approach is used in this study to make this prediction. These algorithms can also spot data patterns and linkages that conventional statistical methods might miss out on.

2.6 Random Forest

Random forest is a supervised learning that is an extension of the decision tree method where each tree in this method depends on the random vector values that are sampled freely and evenly [14]. Random Forest was introduced by Breiman in 2001 as an ensemble method that combines bagging and random feature selection to construct multiple diverse decision trees. Each tree is built using bootstrap samples from the original dataset, and at each split, only a random subset of features is considered, with final predictions obtained by aggregating the outputs of all trees [15]. In a more formal setting, for a p dimensional random vector $X = (X_1, \ldots, X_P)^T$ representing real valued input variables or predictor variables and a random variable Y representing real valued responses, an unknown joint distribution $P_{XY} = (X, Y)$ is assumed. The objective is to find a prediction function $f(X)$ to predict Y. The prediction function is the determinant of the loss function $L(Y, f(X))$ and is defined to minimize the loss value [16].

$$\mathbb{E}_{XY}\left(L(Y, f(X))\right) \tag{5}$$

where the subscript denotes the expectation with respect to the joint distribution of X and Y.

In a classification situation, if the set of possible values of Y is denoted by γ, minimising $\mathbb{E}_{XY}\left(L(Y, f(X))\right)$ for zero one loss gives

$$f(x) = \arg\max P(Y = y \mid X = x) \tag{6}$$

The ensemble constructs f in terms of a collection of so called 'base learners' $h_1(x), \ldots, h_J(x)$ and these base learners are combined to produce an 'ensemble predictor' $f(x)$. In regression, the base learners are averaged

$$f(x) = \frac{1}{J}\sum_{j=1}^{J} I(y = h_j(x)) \tag{7}$$

while in classification, $f(x)$ is the most frequently predicted class ("voting")

$$f(x) = \arg\max \frac{1}{J}\sum_{j=1}^{J} I(y = h_j(x)) \tag{8}$$

The infinite forest estimator is obtained by taking the limit as $M \to \infty$ and equals

$$h(x, L) = \mathbb{E}_{\Theta}\left[h(x, \Theta, L)\right] \tag{9}$$

In Random Forests, the jth base learner is a tree denoted by $h_j(X,\Theta_j)$, where Θ_j is a set of random variables and Θ_j is independent for $j = 1,\dots,J$. Although the definition of Random Forest is very general, random forests are implemented in specific ways. To understand the Random Forest algorithm, it is important to have a basic knowledge of the type of tree used as a base learner.

The lag of the significant predictor variables and the lag of the number of tuberculosis cases were included in Random Forest modelling. Forecasting the number of future tuberculosis cases was predicted using Random Forest. Random Forest analysis will be conducted using R software and with the help of `randomForest` package.

2.7 Model Evaluation

We assess the effectiveness of each forecasting model using R-squared (R^2) is utilized to confirm the accuracy of prediction curve fitting. The result is that each metric is run $m-p-q+1$ times [1].

$$R^2 = 1 - \frac{\sum_{i=1}^{m}(Y_i-\hat{Y}_i)^2}{\sum_{i=1}^{m}(Y_i-\bar{Y})^2} \tag{10}$$

where:

- m : number of observations,
- Y_i : the actual values,
- $\hat{Y}_i$: the predicted values,
- $\bar{Y}$: the average of actual value.

3 Results and Discussion

3.1 Data

The data used in this study is secondary data obtained from the website of the Indonesian Public Health Agency. The unit of observation for this data is 27 cities/districts in West Java from 2016 to 2021 with 6 variable factors affecting tuberculosis cases in West Java [1]. The following are the variable factors affecting tuberculosis cases in West Java presented in the Table 1.

Table 1: Variables and Data Sources

Variables	Source
Dependent Variables	
Number of Tuberculosis Case (y)	West Java Provincial Health Office
Independent Variables	
Population by Age Group (X_1)	Statistics Indonesia and Open Data Jabar
0-14 (X_{11})	
15-44 (X_{12})	
45-64 (X_{13})	
65+ (X_{14})	
Infant BCG Immunization Coverage (X_2)	West Java Provincial Health Office and Open Data Jabar
Population Density (X_3)	Statistics Indonesia and West Java Provincial Health Office
Healthcare Facility (X_4)	West Java Provincial Health Office
The Number of Public Hospitals (X_{41})	
The Number of Health Center (X_{42})	
The Implementation of Community-based Sanitation (X_5)	West Java Provincial Health Office

[1] https://app-diskes.jabarprov.go.id/drive/s/TyZTzEqnm5TfrM4

3.2 Data exploration

Exploring the Distribution of Dependent Variables with the Cullen and Frey Method Exploration of the distribution of the dependent variable was carried out with the help of the fitdistrplus[x] package and the R programme was used to explore the distribution of the variable [6].

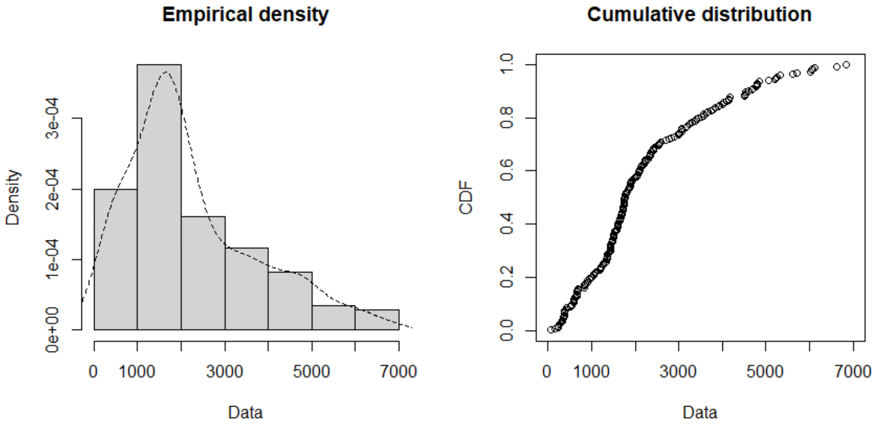

Figure 1: Histogram and CDF of the Number of Tuberculosis Cases

Figure 1 shows the histogram and Cumulative Distribution Function (CDF) of the dependent variable in this study (number of Tuberculosis cases).

In addition to exploration by looking at the histogram and CDF of the number of tuberculosis cases, a review of the Cullen-Frey plot and comparison of Akaike Information Criteria (AIC) values between distributions can be used to identify the distribution that best fits the data.

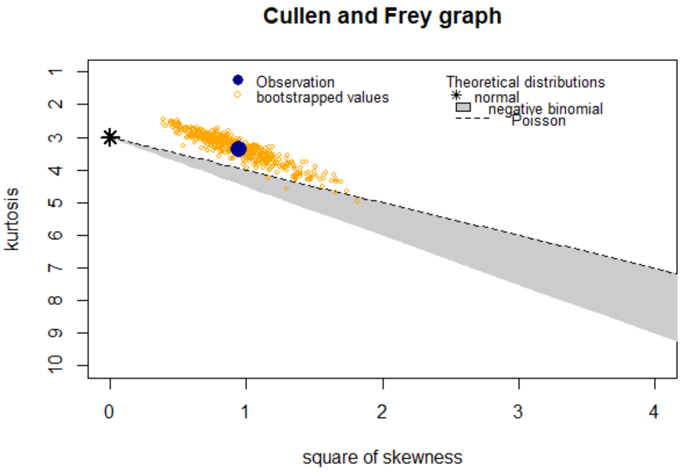

Figure 2: Cullen and Frey Graph of the Number of Tuberculosis Cases used in the Study

In Figure 2, the Cullen and Frey plot is shown, which shows that the variable number of Tuberculosis cases in this study with a bootstrap of 500 best fits the negative binomial distribution.

Table 2: Comparison of AIC Values of Various Distributions to Fit the Distribution of Number of Tuberculosis Cases

Distribution	AIC
Normal	3582.424
Poisson	200729.5
Negative Binomial	3523.444

In Table 2, the AIC values of various distributions are listed. Based on the AIC values in Table 2, it can be concluded that the negative binomial distribution with the smallest AIC value is the distribution that best fits the data on the number of Tuberculosis cases in West Java Province from 2016 to 2023. Both analyses (Cullen & Frey plot and AIC value) show that the negative binomial distribution assumption is the best to be used for this variable in further analysis procedures.

3.3 Correlation between Variables

Each of the predictor variables and the response variable have correlations that can give an indication of significance. Factors were included in the model to give an indication of significance. The closeness of the relationship between predictor variables or what is referred to as multicollinearity in this study can be seen from the value of.

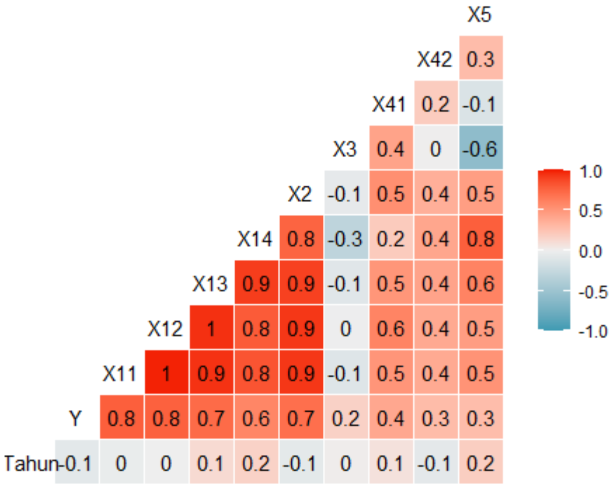

Figure 3: Correlation between the variables of this study

In Figure 3 above, it can be seen that all predictor factors are correlated with the response variable. In addition, it is also found that. The various categories in the population variable by age group are correlated with each other.

3.4 Negative Binomial Mixed Model

Various tuberculosis factors that affect the number of tuberculosis cases in various districts/cities in West Java in 2016-2023 can be seen through GLMM modelling. Based on the results of previous analyses related to the distribution of the dependent variable, it is assumed that the dependent variable has a negative binomial distribution and the application of this assumption is used in the construction of the GLMM model in the research conducted.

Table 3: Comparison of AIC Values for Poisson GLMM and Negative Binomial GLMM Models

GLMM Distribution Assumptions	GLMM Models	AIC	Chi Square	df	p-value
Poisson	$g(\mu) = \beta_0 + \beta_1 D_1 X_1 + \beta_2 D_2 X_1 + \beta_3 D_3 X_1 + \beta_4 X_2 + \beta_5 X_3 + \beta_6 D_1 X_4 + \beta_7 X_5 + bZ$	7482.4	1834943999	151	0
Negative Binomial	$g(\mu) = \beta_0 + \beta_1 D_1 X_1 + \beta_2 D_2 X_1 + \beta_3 D_3 X_1 + \beta_4 X_2 + \beta_5 X_3 + \beta_6 D_1 X_4 + \beta_7 X_5 + bZ$	2642	127.2684986	151	0.920

By including all variables, Table 3 compares the AIC values of the Poisson and Negative Binomial GLMM models and the overdispersion test results for both models. By comparing the two tables, it can be seen that the negative binomial assumption is more suitable for use in the data on the number of tuberculosis cases in this study because it tends to have a smaller AIC value and is also better able to overcome the overdispersion problem found in the Poisson GLMM model when applied to the case of this study. This is in accordance with the findings of the distribution exploration stage conducted previously, as well as the findings of [17], [18]. which state that although discrete data, such as the number of disease cases, can usually be assumed to have a Poisson distribution, overdispersion (having a variance greater than the mean) can occur in some cases. One of these cases can be overcome by using the negative binomial distribution assumption. Therefore, from now on, GLMM with negative binomial assumption will be used.

A model can be said to be good if it has the smallest AIC and all predictor variables are significant. Therefore, insignificant predictor variables will not be included in the model. After obtaining significant variables, a model will be formed from these significant variables and it will be seen which model has the smallest AIC value. The model that has the smallest AIC value is the best negative binomial GLMM model so that this model will be used for different predictor variables on the number of tuberculosis cases in various districts and cities in West Java. The estimation results of the negative binomial GLMM model with all significant components are shown in Table 4.

Table 4: Best Negative Binomial GLMM Model Selection Included Random Effects of Districts/Cities for Each Model

Model	Covariates Included	Significant Covariates ($\alpha = 0.05$)	AIC
1	X_{12}, X_{13}, X_3	X_{12}, X_{13}, X_3	3372.4
2	X_2, X_3, X_5	X_2, X_3, X_5	3393.6
3	X_{11}, X_3	X_{11}, X_3	3356.4
4	X_{12}, X_{13}	X_{12}, X_{13}	3375.9
5	X_{11}, X_3, X_5	X_{11}, X_3, X_5	3358.3
6	X_{13}, X_{14}, X_3	X_{13}, X_{14}, X_3	3372

Based on Table 4 above, model 3 has the smallest AIC. While model 5 has an AIC that is not much different from the AIC of model 3, with a difference of 2.1 and is still included in the small AIC. Therefore, in the next analysis, model 3 is selected as the best model to use in random forest analysis. The selected negative binomial GLMM model did not show any symptoms of overdispersion, as seen in Table 5. In addition, the predictor variables included in the best model did not show any symptoms of multicollinearity as indicated by VIF values that did not exceed 10 in Table 6.

Table 5: Overdispersion Test on Selected Negative Binomial GLMM Models 3

Chi Square	Ratio (Chi Square/df)	df	p-value	Description
153.0464813	0.7614253	201	0.9950568	Insignificant, no symptoms of overdispersion

Table 6: VIF of various Predictor Variables in Model 3

Variables	VIF
Population by Age Group 0-14 (X_{11})	1.000697
Population Density (X_3)	1.000697

Table 7: Regression Coefficient and Significance for the Negative Binomial GLMM Model

Variables	Coefficient Estimate	Standard Error	z-score	p-value	Description
Intercept	6.58×10^0	1.26×10^{-1}	5202	$< 2 \times 10^{-16}$	Significant ($\alpha < 0.001$)
Population by Age Group (0-14)	1.88×10^{-6}	2.17×10^{-7}	8.68	$< 2 \times 10^{-16}$	Significant ($\alpha < 0.05$)
Population Density (X_3)	4.05×10^{-5}	1.28×10^{-5}	3.17	0.0057	Significant ($\alpha < 0.05$)
Variance			0.07398		
Standard Deviation			0.272		

The selected negative binomial GLMM model did not show any symptoms of overdispersion, as seen in Table 5. In addition, the predictor variables included in the best model did not show any symptoms of multicollinearity as indicated by VIF values that did not exceed 10 in Table 6. The following is the best general linear mixed-model (GLMM) model to model multiple factors on the number of TB cases in West Java:

$$\hat{Y} = g(\mu) = \beta_0 + \beta_1 X_{\text{Age 0-14}} + \beta_3 X_{\text{Population Density}} + bZ_{\text{Districts/Cities}} \tag{11}$$

The natural logarithm function corresponding to the negative binomial assumption is the link function for the mean of the response variable with $g(.)$ the link function for the mean of the response variable is the natural logarithm function that conforms to the negative binomial assumption as described in Equation 3 in the introduction above and b is the random effect intercept coefficient for each district/city in West Java. The regression coefficient estimation results for these variables can be seen in Table 7.

In this study, it was found that categories of population variables based on age group for the age range of 0-14 years and population density variable that significantly influenced the number of Tuberculosis cases in each district/city in West Java Province. The population for the age range of 0-14 years has a coefficient of 0.00000188. This indicates that the variable has a positive influence on the number of Tuberculosis cases, meaning that every additional 1 person of age range of 0-14 years in a district/city will increase the log of the expected number of Tuberculosis cases in the district/city by 0.00000188 units (if other variables are constant).

This study also found that, with a coefficient of 0.0000405, population density in each district or city has a positive and significant influence on the number of tuberculosis cases. Thus, assuming other variables remain constant, an increase in population density in a district or city by 1 person/km2 can lead to an increase of 0.0000405 in the predicted log number of tuberculosis cases. This result is in line with literature studies that found that population density can contribute to an increased risk of tuberculosis infection [19].

Then, using the Random Forest method, the two significant variables will be included in the next step of the analysis.

3.5 Negative Binomial Mixed Model Random Forest

In the previous section, we have obtained significant predictor variables that can be used to produce forecasting of the number of tuberculosis cases. The input data underwent a normalization step initially before being fed into the model. The dataset is divided into two parts, with 87.5% of the data used for training the model as training data and 12.5% for testing the model as testing data. The training data and testing data are used to test the model by making predictions using the pre-trained model, restoring the standardization process, and comparing the predicted results with the actual data. Table 8 shows the model assessment results for the testing data.

Table 8: model assessment for testing data

No Model	ntree	R Square	No Model	ntree	R Square
1	350	0.907	7	600	0.97
2	100	0.915	8	700	0.91
3	200	0.905	9	800	0.909
4	300	0.913	10	900	0.908
5	400	0.908	11	1000	0.908
6	500	0.913	12	1050	0.909

Based on Table 8, it can be seen that the R Square value is in the range of 0.9–0.92 and the R-Square value has started to approach the number 1. For determining the best model, it is seen from the model that has the highest R Square and is closest to number 1. Because model 2 of the testing data has the highest R Square compared to other models, which is 0.915, the model chosen for forecasting is model 2 with ntree of 100.

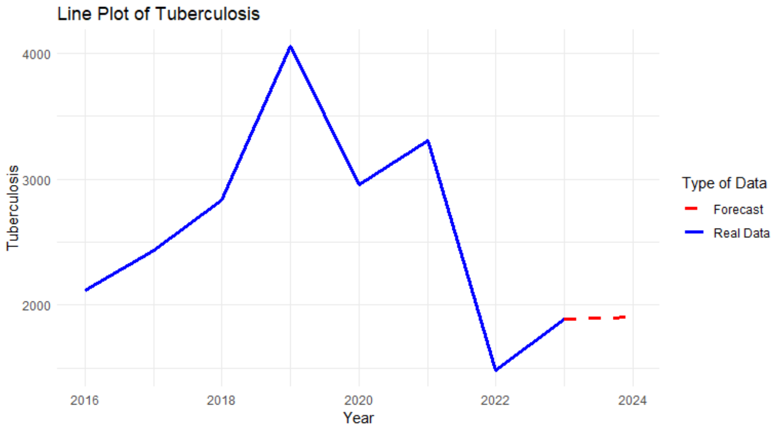

Figure 4: Line plot of Average Tuberculosis Cases in West Java Province, Indonesia

Figure 4 shows the forecasting results indicated by the red dashed line. Forecasting is done for a period of one year, i.e. projecting the number of TB cases for the year 2024. As illustrated in Figure 4, the forecasting results show a slight increase in the number of TB cases compared to the previous year. However, the projected increase is still relatively low, with an average number of TB cases in 2023 of 1889 cases and a projected number of TB cases in 2024 of 1904 cases.

4 Conclusion

In this study, longitudinal data on tuberculosis cases and other risk variables were examined over the 2016–2023 period in 27 districts and cities in West Java Province. This study demonstrates that overdispersion in Poisson GLMM can be overcome by GLMM under the assumption of a negative binomial distribution. According to this study, Population by Age Group 0-14, and Population Density all significantly affect the number of tuberculosis cases in each district and city.

Based on the analysis performed in this study, it can be concluded that the GLMM model may be used to explain the number of tuberculosis cases in the West Java Province in 2016–2023 under the assumption of a negative binomial distribution, as shown in the Equation 12:

$$\hat{Y} = g(\mu) = \beta_0 + \beta_1 X_{(\text{Age 0-14})} + \beta_3 X_{(\text{Population Density})} + bZ_{(\text{Districts/Cities})} \tag{12}$$

Then in the random forest model assessment for testing data, the model evaluation results were obtained with $R^2 = 0.915$. This model was then used to forecast the number of Tuberculosis cases in 2024 with the forecasting results show a slight increase in the number of TB cases compared to the previous year. The results of this study show the importance of various related parties to continue to be vigilant, pay attention to various related risk factors, and continue to make various efforts to study, prevent, and control Tuberculosis disease effectively.

CRediT Authorship Contribution Statement

Restu Arisanti: Conceptualization, Methodology, Supervision, Project Administration, Funding Acquisition, Writing–Original Draft. **Resa Septiani Pontoh:** Supervision, Project Administration, Funding Acquisition. **Sri Winarni:** Supervision, Project Administration, Funding Acquisition. **Nisa Akbarilah Putri:** Software, Data Curation, Formal Analysis, Writing–Original Draft, Visualization. **Stefany Maurin:** Data Curation, Writing–Review, & Editing.

Declaration of Generative AI and AI-assisted technologies

No generative AI or AI-assisted technologies were used during the preparation of this manuscript.

Declaration of Competing Interest

The authors declare no competing interests.

Funding and Acknowledgments

This research received no external funding.

Data Availability

The dataset analyzed during the current study is publicly available in the open data Jabar and West Java Provincial Health Office: Health Profile of West Java Province.

References

[1] R. Arisanti, R. S. Pontoh, S. Winarni, Y. Nurhasanah, A. P. Pertiwi, and S. D. N. Aini, "Integrating generalized linear mixed models with extreme neural network: Enhancing pulmonary tuberculosis risk modeling in west java, indonesia," *Commun. Math. Biol. Neurosci.*, vol. 2024, Article–ID, 2024. DOI: 10.28919/cmbn/8748.

[2] World Health Organization, *Global tuberculosis report 2023*, en. Genève, Switzerland: World Health Organization, Nov. 2023. Available online.

[3] S. Sulistyawati and A. W. Ramadhan, "Risk Factors for Tuberculosis in an Urban Setting in Indonesia: A Case-control Study in Umbulharjo I, Yogyakarta," *Journal of UOEH*, vol. 43, no. 2, pp. 165–171, Jun. 2021. DOI: 10.7888/juoeh.43.165.

[4] Dinas Kesehatan Provinsi Jawa Barat, *Profil kesehatan provinsi jawa barat tahun 2022*, Dinas Kesehatan Provinsi Jawa Barat, Accessed: june 3, 2025, 2023. Available online.

[5] N. Tang, M. Yuan, Z. Chen, *et al.*, "Machine Learning Prediction Model of Tuberculosis Incidence Based on Meteorological Factors and Air Pollutants," *International Journal of Environmental Research and Public Health*, vol. 20, no. 5, p. 3910, Feb. 2023. DOI: 10.3390/ijerph20053910.

[6] A. C. C. dan H. C. Frey, *Probabilistic Techniques in Exposure Assessment: A Handbook for Dealing with Variability and Uncertainty in Models and Inputs*. Springer, New York, 1999. Available online.

[7] J. Bruin, *Newtest: Command to compute new test @Online*, Feb. 2011. Available online.

[8] C. E. McCulloch, "Maximum Likelihood Algorithms for Generalized Linear Mixed Models," *Journal of the American Statistical Association*, vol. 92, no. 437, p. 162, Mar. 1997. DOI: 10.2307/2291460.

[9] J. M. Hilbe, *Negative Binomial Regression*. Cambridge University Press, 2011. DOI: 10.1017/CBO9780511973420.

[10] X. Zhang, H. Mallick, Z. Tang, *et al.*, "Negative binomial mixed models for analyzing microbiome count data," *BMC Bioinformatics*, vol. 18, no. 1, pp. 1–10, 2017. DOI: 10.1186/s12859-016-1441-7.

[11] S. D. Kachman, "An introduction to generalized linear mixed models," *Statistics*, vol. 24, pp. 59–73, 2008. DOI: 10.4148/2475-7772.1352.

[12] E. P. Liski, "Generalized linear mixed models: Modern concepts, methods and applications," *International Statistical Review*, no. 3, pp. 482–483, Dec. 2013. DOI: 10.1111/insr.12042_24.

[13] B. Charbuty and A. Abdulazeez, “Classification Based on Decision Tree Algorithm for Machine Learning,” *Journal of Applied Science and Technology Trends*, vol. 2, no. 01, pp. 20–28, 2021. DOI: 10.38094/jastt20165.

[14] Lakshmi Prasanna and S. Mehrotra, “Comparative Analysis of Machine Learning Algorithms on Mental Health Dataset,” *Lecture Notes in Networks and Systems*, vol. 719 LNNS, no. 2, pp. 599–606, 2023. DOI: 10.1007/978-981-99-3758-5_54.

[15] R. H. P. Y. Damayanti, S. Astutik, and A. B. Astuti, “Geographically Weighted Random Forest Model for Addressing Spatial Heterogeneity of Monthly Rainfall with Small Sample Size,” *CAUCHY: Jurnal Matematika Murni dan Aplikasi*, vol. 10, no. 1, pp. 442–456, May 2025. DOI: 10.18860/cauchy.v10i1.32161.

[16] A. Cutler, D. R. Cutler, and J. R. Stevens, “Ensemble Machine Learning,” *Ensemble Machine Learning*, no. January, 2012. DOI: 10.1007/978-1-4419-9326-7.

[17] T. W. Utami, “Analisis Regresi Binomial Negatif Untuk Mengatasi Overdispersion Regresi Poisson Pada Kasus Demam Berdarah Dengue,” *Jurnal Statistika*, vol. 1, no. 2, pp. 59–65, 2013. Available online.

[18] A. A. Yirga, S. F. Melesse, H. G. Mwambi, and D. G. Ayele, “Negative binomial mixed models for analyzing longitudinal CD4 count data,” *Scientific Reports*, vol. 10, no. 1, p. 16 742, Oct. 2020. DOI: 10.1038/s41598-020-73883-7.

[19] P. R. Donald, B. J. Marais, and C. E. Barry, “Age and the epidemiology and pathogenesis of tuberculosis,” *The Lancet*, vol. 375, no. 9729, pp. 1852–1854, May 2010. DOI: 10.1016/S0140-6736(10)60580-6.