CAUCHY – Jurnal Matematika Murni dan Aplikasi
Volume 10 (2) (2025), Pages 644-659
p-ISSN: 2086-0382; e-ISSN: 2477-3344

Random Forest Classification of Infant Mortality Rate in Indonesia: A Gini-Based Analysis

Ria Dhea L.N. Karisma[*] Usman Pagalay, and Muhammad Khudzaifah

Department of Mathematics, Faculty of Science and Technology, Universitas Islam Negeri Maulana Malik Ibrahim Malang, Indonesia

Abstract

One of the indicators used to measure the success of development programs in Indonesia is the Infant Mortality Rate (IMR). IMR is a sensitive indicator and represents maternal and child health problems in a country. Random forest is an ensemble machine learning method that combines multiple decision trees using bootstrap aggregation. It aims to improve the prediction accuracy and robustness of the model. In addition, it can be applied to both case classification and regression because it can handle high-dimensional and complex cases and non-linear relationships. In this study, Random Forest is used to solve the classification of IMR cases in Indonesia, making them easy to interpret and related to policy relevance. The aim of this study is to predict infant mortality factors using the Gini Index to determine which variables need to be improved. The Gini Index is used to identify key factors, enabling targeted policy interventions. It highlights the most influential variables, helping policymakers focus on areas that require improvement for more effective outcomes. The evaluation model in this study uses out-of-bag estimation and k-fold validation. The model achieves an overall accuracy of 99.97%, with a sensitivity of 99.87% and specificity of 100%, indicating excellent performance. The most important variables in this study are breastfeeding, type of birth (single and twin), and birth weight of the baby. The parent node in IMR is breastfeeding, where live IMRs that are breastfed have a greater chance of survival than dead IMRs that are not breastfed.

Keywords: Accuracy; Gini Index; Infant Mortality Rate; Random Forest; Sensitivity; Specificity

1 Introduction

Several indicators, including the infant mortality rate, under-five mortality rate, maternal mortality rate, and nutritional levels, can measure the success of development programs in Indonesia. According to the World Health Organization (WHO) [1], IMR is a sensitive indicator. IMR represents health problems in the country; apart from that, it is related to infant mortality and reflects maternal health. In developing countries, such as Indonesia, IMR is the focus of policy and program planning so that it can provide basic services to reduce IMR and diseases in children and babies. Indonesia's IMR in each province is quite high compared to several ASEAN countries. Indonesia is ranked 8th in countries with high IMR compared to Malaysia and Vietnam. In 2015, Indonesia's IMR was 126 per 100,000 deaths, meaning that among 100,000

[*] **Corresponding author**. E-mail: riadhea@uin-malang.ac.id

Submitted: October 18, 2024 Reviewed: June 26, 2025 Accepted: July 2, 2025
DOI: https://doi.org/10.18860/cauchy.v10i2.29508

live births, 126 babies died before they were one year old [2]. The target for achieving sustainable development that the world has set regarding IMR is 12 infant deaths among 100,000 live births [3]. Meanwhile, from 1991 to 2017, the death rate was still relatively high and far from the target set until 2030.

The high IMR in Indonesia is influenced by individual factors such as family income and household factors. In the case of the population group with low income, the number of IMR incidents was 40, while for those with high income, it was 20. Thus, the higher the socio-economic class of a family, the lower the infant mortality rate [4]. Based on the theory, factors that can influence infant mortality in Indonesia can be measured by maternal education, maternal employment, maternal age at birth, family welfare, residence, birth order, birth interval, type of birth, place of birth, breastfeeding, frequency of ANC (Antenatal Care) or pregnancy checks, baby's gender, and baby's birth weight.

Random Forest or RF is a machine learning technique that is currently popular in the world of science. This method is known for its non-parametric capabilities for large-scale datasets [5]. Random forests can produce models in the form of levels of importance or measures of importance. Breiman introduced this method, developed to improve the use of bagging or bootstrap aggregation methods. Random Forest is a classification method that consists of a classification tree structure in which there is an additional layer in the form of resampling in the bagging process [6].

The study contributes to the growing literature on IMR by offering a classification approach. Some of the existing studies apply regression to estimate continuous IMR values. However, this research categorizes IMR into discrete classes. It makes the results easier to interpret and more relevant for policy design. The model includes a combination of social, economic, and demographic variables, including maternal income, employment, education, and regional disparities, to reflect real conditions better. To address the class imbalance in the dataset, Random Walk Oversampling (ROSE) was applied, helping to improve model accuracy and sensitivity. Additionally, the use of Gini Importance allows for identifying key factors that can serve as a basis for targeted health interventions.

Previous research on the prediction of preterm infant mortality using the Random Forest approach showed that it is particularly effective in identifying critical risk factors associated with preterm infant mortality due to its robustness in handling complex and non-linear interactions in the data [7]. The sensitivity of the model is 88%. [8] used Random Forest and several other machine-learning methods to predict infant mortality in Brazil. The results showed that Random Forest can effectively identify socio-economic and health-related factors that affect infant mortality. The variables are maternal education and access to health services.

Previous research in Indonesia still needs to pay attention to comprehensive analysis, especially on social and economic factors, such as family income, maternal income, maternal education and employment status in influencing infant mortality rates. In this study, these factors are integrated using the Random Forest model. This method is expected to be able to handle various predictor variables and identify their relative importance. Some studies provide a generalized view of IMR across Indonesia without focusing on the unique regional disparities among provinces. Moreover, this research addresses the regional variations by including province-specific data in the Random Forest model. Research on Indonesia's IMR still needs to be improved compared to other ASEAN countries such as Malaysia and Vietnam. Therefore, this study attempts to analyze more broadly the IMR in Indonesia which is still high even though the government has attempted to create development programs to reduce the IMR in Indonesia.

The study uses machine learning techniques, a powerful non-parametric method, to predict factors affecting infant mortality rates. In addition, this study uses Gini Importance to determine which variables are the most influential (important variables). Furthermore, it provides a more detailed analysis compared to previous classical methods. The Gini Index is used to identify the most influential variables, which in turn facilitates effective policy planning. Primarily, the intervention can be directed especially at groups socially and economically.

The study aims to determine the prediction of factors that influence the Infant Mortality Rate using Random Forest and accuracy. Gini Importance determines the most important variables in the Random Forest model for infant mortality cases in Indonesia. The conclusions of this research are evaluations of related parties to reduce the infant mortality rate based on the model produced using the Random Forest model.

The paper is organized as follows. Section 2 explains the methods, which is the Random Forest classification model. Then, it describes the including data preprocessing, handling imbalance class using ROSE, model construction, and evaluation metrics. Section 3 presents and discusses the results. Finally, the conclusion is presented in Section 4, along with the key findings and suggestions for health policy.

2 Methods

To systematically address the objectives of this research, a series of methods and techniques were applied. This section provides a comprehensive explanation of the machine learning algorithms, statistical approaches, and evaluation metrics used in this study. Each methodological step is designed to ensure the accuracy, robustness, and interpretability of the rainfall forecasting model. The discussion begins with the Random Forest algorithm, which serves as the core classification technique in this study.

2.1 Random Forest

[6] developed the Random Forest concept from the bagging or bootstrap aggregating process when selecting a classification process. This method is an ensemble classification method whose decision is based on voting from the most popular classes to classify [9]. Moreover, it provides high accuracy compared to individual classification. The classification process is done by selecting predicted values based on basic classification techniques. Tree classification has a high accuracy value compared to random classification. However, combining the tree classification with certain randomly generated characteristics will increase accuracy [6].

The advantage of RF compared to other classification methods is that it can be used for large-scale data with high accuracy. In addition, it provides estimates based on variables in the classification. It can be used with other methods for data with an imbalanced population with stable errors. It also produces unbiased estimates by generating errors in the process of building classification trees [10]. Figure 1 is an illustration of RF [11].

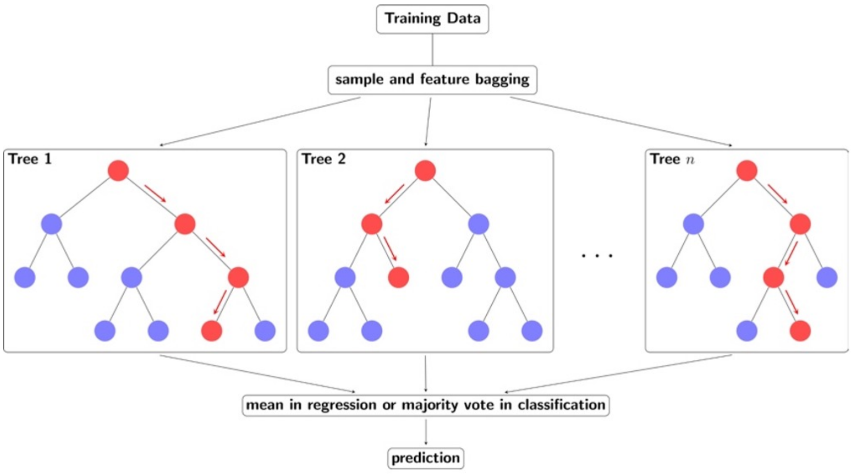

Training Data
sample and feature bagging
Tree 1
Tree 2
Tree n
...
mean in regression or majority vote in classification
prediction

Figure 1: Random Forest Illustration

The tree classification procedure is a general kth tree, generally a random vector θ_k, which is independent of the past random vectors $\theta_1, \dots, \theta_{k-1}$ but has the same distribution. A resulting tree comes from training data or a training set, and θ_k is the result of a classifier, namely $h(x, \theta_k)$. Then, x is an input vector. Hence, θ is the number of training sets. In random split choice selection, θ is the number of random integers between 1 and k. Eq. 1 shows the definition of Random Forest [6].

$$\{h(x, \theta_k),\ k = 1, \dots\} \tag{1}$$

Random Forest is a classification method that contains a collection of trees built from Eq. 1, where θ_k is an independently and identically distributed random vector. Each tree assigns a unit vote to the most popular class in the input x. Given the ensemble classification $h_1(x), h_2(x), \dots, h_K(x)$, the training data or training set is taken randomly from the random vector distribution Y, and X is defined as the margin of the following function in Eq. 2.

$$\mathrm{mg}(X, Y) = P_\Theta(h(X, \Theta) = Y) - \max_{j \neq Y} P_\Theta(h(X, \Theta) = j) \tag{2}$$

where:

- $\mathrm{mg}(X, Y)$: the margin function.
- $P_\Theta(h(X, \Theta) = Y)$: probability that X is grouped into class Y using model $h(X, \Theta)$.
- $\max_{j \neq Y} P_\Theta(h(X, \Theta) = j)$: maximum probability that X is grouped into a class other than Y using $h(X, \Theta)$.

In Random Forest, the margin is measured by the extent to which the average voting for class X, while Y is the correct class, exceeds the average voting for the other classes. If the margin is greater, the confidence level is also greater, leading to more accurate predictions.

Random Forest has generalization errors. It is the expected distance between testing data and training data, which is used to measure whether the algorithm is overfitting the training data [12]. $\hat{PE}^*$ in Eq. 3 is the generalization error in a Random Forest.

$$\hat{PE}^* = P_{(X,Y)}(\mathrm{mg}(X, Y) < 0) \tag{3}$$

where $P_{(X,Y)}$ denotes the joint probability over the input-output space. In Random Forest, $h_k(X) = h(X, \Theta_k)$.

As the number of trees increases over $\Theta_1, \dots, \Theta_K$, $\hat{PE}^*$ will converge, as shown in Eq. 4.

$$P_{(X,Y)}\left(P_\Theta(h(X, \Theta) = Y) - \max_{j \neq Y} P_\Theta(h(X, \Theta) = j) < 0\right) \tag{4}$$

$P_{(X,Y)}$ is the joint probability from X and Y. $P_\Theta(h(X, \Theta) = Y)$ is the probability of X being grouped into class Y using the model $h(X, \Theta)$. The maximum probability that X is grouped into a class other than Y using $h(X, \Theta)$ is $\max_{j \neq Y} P_\Theta(h(X, \Theta) = j)$. Then, $P_{(X,Y)}(\mathrm{mg}(X, Y) < 0)$ represents the probability that X and Y co-occur with the margin function $\mathrm{mg}(X, Y) < 0$.

Random Forest continues to be refined and applied in a wide range of fields, and its relevance is expected to remain strong as machine learning research increasingly intersects with advances in quantum computing and hybrid intelligence frameworks [13].

2.2 Strength and Correlation

Random Forest has an upper limit on error that can be reduced by generalizing the error on two parameters: the measure of accuracy in individual classification and the correlation between them. Optimizing Random Forest models involves balancing the accuracy of individual trees and minimizing the correlation between them. In order, the result robust and accurate models. Eq. 5 developed by [14]

Definition 1 (Margin function of Random Forest)**.** *The margin function of a Random Forest is defined as*

$$mg(X,Y) = P_\Theta(h(X,\Theta) = Y) - \max_{j \neq Y} P_\Theta(h(X,\Theta) = j) \tag{5}$$

It is modified to examine at average accuracy on a single tree. Whenever the strength is large then the prediction accuracy value has a high value. However, increasing the accuracy value by minimizing the correlation value ($\bar{\rho}$) while maintaining the strength value (s) is necessary. The strength of the classifier set $\{h(X,\Theta)\}$ is

$$s = E_{(X,Y)}\, mr(X,Y) \tag{6}$$

Then, $\hat{j}(X,Y)$ is a a prediction of the target variable Y based on the input variable X. Assume $s \geq 0$ in Chebychev's inequality

$$PE^* \leq \frac{\text{var}(mg)}{s^2} \tag{7}$$

In Eq. 7 $\hat{j}(X,Y) = \arg\max_{j \neq Y} P_\Theta(h(X,\Theta) = j)$, then the raw margin function as

$$rmg(\Theta,X,Y) = I(h(X,\Theta) = Y) - I(h(X,\Theta) = \hat{j}(X,Y)) \tag{8}$$

Eq. 8 describe as

$$mg(X,Y) = P_\Theta(h(X,\Theta) = Y) - P_\Theta(h(X,\Theta) = \hat{j}(X,Y)) \tag{9}$$

$$= E_\Theta[I(h(X,\Theta) = Y) - I(h(X,\Theta) = \hat{j}(X,Y))] \tag{10}$$

Then, $mg(X,Y)$ in Eq. 10 is expectation of $rmg(\Theta,X,Y)$ for Θ. For any density function of f, then

$$[E_\Theta f(\Theta)]^2 = E_{(\Theta,\Theta')} f(\Theta) f(\Theta') \tag{11}$$

where in Eq. 11 Θ, Θ' is independent that has an identical distribution. The following equation is

$$[mg(X,Y)]^2 = E_{(\Theta,\Theta')} rmg(\Theta,X,Y) rmg(\Theta',X,Y) \tag{12}$$

For $E_{(\Theta,\Theta')} rmg(\Theta,X,Y)$ is double expectation of the function $rmg(\Theta,X,Y)$ to Θ and Θ'. Using Eq. 10 it get variance from margin function

$$\begin{aligned} \text{var}(mr) &= \mathbb{E}_{\Theta,\Theta'}\left[mr(X,Y)^2\right] \\ &= \mathbb{E}_{\Theta,\Theta'}\left[\text{rmg}(\Theta,X,Y) \cdot \text{rmg}(\Theta',X,Y)\right] \\ &= \mathbb{E}_{\Theta,\Theta'}\left[\rho(\Theta,\Theta') \cdot \text{sd}(\Theta) \cdot \text{sd}(\Theta')\right] \\ &= \bar{\rho} \cdot (\mathbb{E}_\Theta[\text{sd}(\Theta)])^2 \end{aligned} \tag{13}$$

Eq. 13 shows that the generalization error in the Random Forest model is controlled by two main components: the strength of each decision tree and the correlation between trees. Mathematically, the variance of the margin function states that to minimize the generalization error, it is necessary to maximize the accuracy of each tree. Consequently, at the same time, it also minimizes the average correlation between trees $\bar{\rho}$. Random Forest achieves optimal performance when each decision tree in the ensemble is both individually accurate and contributes to prediction diversity.

2.3 Out of Bag Estimate (OOB)

Error estimation in Random Forest models can be completed without cross-validation or validation set deletion. A model built from a bootstrapped tree without pruning is performed and fitted to a subset of bootstrapped observations. Furthermore, on average, each tree is formed using about two-thirds of the observations (called the data in the bag) [15]. Meanwhile, the angle of the remaining observations that is not used to load the tree is called out-of-bag (OOB).

Researchers can predict the response to the i-th observation by using each tree observed with OOB. About $B/3$ will produce a prediction for the i-th observation. However, if it is used for case classification, the majority vote can be used. Where in a single OOB prediction for the i-th observation can be obtained by calculating the classification error (for a classification case). For regression case, the mean squared error (MSE) of the OOB prediction can be used.

The resulting OOB error serves as a valid estimate of the generalization or test error in the bagged model. This occurs because the response for each observation is derived from predictions across trees, rather than being individually computed per observation. Compared to using cross-validation, the OOB approach to test error estimation is particularly useful when performing analysis on large-scale data.

2.4 Random Forest Model Algorithm

Generally, the RF model is a bootstrap aggregating (bagging) process in the sorter selection process. Bootstrap aggregating is an ensemble method that improves classification by combining random classification methods on training data. The aim is to reduce variance and prevent overfitting. The following is a random forest model algorithm.

1. Taking sample data from actual data with bootstrap resampling with replacement. Here is the bootstrap resampling process:
   (a) Construct an empirical distribution $\hat{F}_n$ from samples that assigns probability $1/n$ to each X_i where $i = 1, 2, \ldots, n$
   (b) Take "n" random bootstrap samples at random with returns from the empirical distribution $\hat{F}_n$ called the first bootstrap $X^{(*1)}$
   (c) Select and calculate the statistic $\hat{\theta}$ from the bootstrap sample $X^{(*1)}$ called $\hat{\theta}_1^*$
   (d) Repeat steps a to c until B times; then it will get $\hat{\theta}_1^*, \hat{\theta}_2^*, \ldots, \hat{\theta}_B^*$
   (e) Form a probability distribution of $\hat{\theta}_B^*$ with a given probability as $1/B$ at each $\hat{\theta}_1^*, \hat{\theta}_2^*, \ldots, \hat{\theta}_B^*$ which is a bootstrap estimator for the sampling distribution of $\hat{\theta}$ called as $\hat{F}^*$
   (f) The bootstrap estimate is approximated by

$$\hat{\theta}^* = \frac{1}{B} \sum_{b=1}^{B} \hat{\theta}_b^*$$

2. Conducting the formation of a classification tree from the results of bootstrap resampling for each result with the best predictor variable as the best classifier determinant taken randomly without pruning. The following is the process of compiling a classification tree [6]:
   (a) Determining the root node by calculating the entropy and gain information values:

$$\text{Entropy}(S) = -\sum_{i=1}^{c} p_i \log_2 p_i$$

Where:

- S: Sample data in that used in the training data
- p_i: ratio of the number of samples from the subset and the value of the i-th attribute
- c: Number of classes

$$\text{Gain}(S, A) = \text{Entropy}(S) - \sum_{v \in \text{values}(A)} \frac{|S_v|}{|S|} \text{Entropy}(S_v)$$

Where:

- S: class from data
- A: feature
- Entropy(S): Entropy from class S
- values(A): Number of attributes in A
- S_v: Attribute Representation of A

(b) The largest information gain value will be the root node

(c) Performing process a to b to obtain branches

3. Make a prediction of the classification model based on the results of the classification tree formed
4. Carry out the process one to three until the desired number of classification trees is obtained with n repetitions
5. Make a final classification prediction by combining the results of the classification trees that have been obtained based on the majority vote.

2.5 Importance Variable

Variable importance is a metric that quantifies the influence of each predictor variable within a classification model. In Random Forests, the Gini index is commonly used to assess the importance of each variable [6]. This measure helps to identify and interpret the variables most critical for accurate classification.

The importance of variable X_m is defined as:

$$\text{Imp}(X_m) = \frac{1}{N_T} \sum_T \sum_{\substack{t \in T \\ v(s_t)=X_m}} p(t)\, \Delta i(s_t, t) \tag{14}$$

where:

- X_m is a predictor variable,
- N_T is the number of trees in the Random Forest,
- T indexes each tree in the forest,
- t is a node in tree T,
- $v(s_t)$ is the variable used for splitting at node t,
- $p(t)$ is the proportion of samples that reach node t,
- $\Delta i(s_t, t)$ is the decrease in impurity at split s_t at node t.

[6] applied the Strong Law of Large Numbers to show that, as the number of trees increases, the following holds:

Theorem 1. *Let C be a zero-probability set in the sequence space $\Theta_1, \Theta_2, \ldots$ such that, for all x, it is outside C. Then,*

$$\frac{1}{N} \sum_{n=1}^{N} I(h(\Theta_n, x) = j) \to P_\Theta(h(\Theta, x) = j)$$

as $N \to \infty$.

Proof. For a fixed training dataset and Θ, consider the set of all x such that $h(\Theta, x) = j$. This set can be represented as a finite union of hyper-rectangles, denoted $S_1, \ldots, S_K$. Define $\varphi(\Theta_n) = k$ if $\{x : h(\Theta, x) = j\} = S_k$, and let N_k be the number of times $\varphi(\Theta_n) = k$ in the first N trials. Then,

$$\frac{1}{N} \sum_{n=1}^{N} I(h(\Theta_n, x) = j) = \frac{1}{N} \sum_{k=1}^{K} N_k I(x \in S_k)$$

By the Law of Large Numbers,

$$\frac{N_k}{N} = \frac{1}{N}\sum_{n=1}^{N} I(\varphi(\Theta_n) = k)$$

converges to $P_\Theta(\varphi(\Theta) = k)$ as N grows. For those k corresponding to sets with zero probability (i.e., values in C), convergence does not occur, but these are negligible. Therefore,

$$\frac{1}{N}\sum_{n=1}^{N} I(h(\Theta_n, x) = j) \to \sum_{k=1}^{K} P_\Theta(\varphi(\Theta) = k) I(x \in S_k)$$

The right side simplifies to $P_\Theta(h(\Theta, x) = j)$. This result demonstrates that as the number of trees increases, Random Forests are unlikely to overfit and will yield a small error value. □

2.6 K-Fold Cross Validation

K-Fold cross-validation is one of the methods used to validate machine learning methods primarily to evaluate model performance more accurately [15]. The technique divides data into several subsets or folds that can be used alternately to test the model, especially on training and testing data. Then, the model will be trained using fold $k-1$ of the data. Therefore, tested using testing data, which is the remaining part that is not used during training. This process is repeated up to k times, where each part will take turns becoming training data. One of the advantages of using k-fold cross-validation is that it reduces variance by using all data alternately for training and testing. Thus, the variance of the model firm estimate becomes lower. The selection of k can consider several values. If the k value is small (2 or 5), it can provide a biased estimate but has low variance. Large k values (10 or more) can provide a more accurate estimate of model performance but with higher variance and longer computation time [16]. Figure 2 is an illustration of k-fold cross-validation [17].

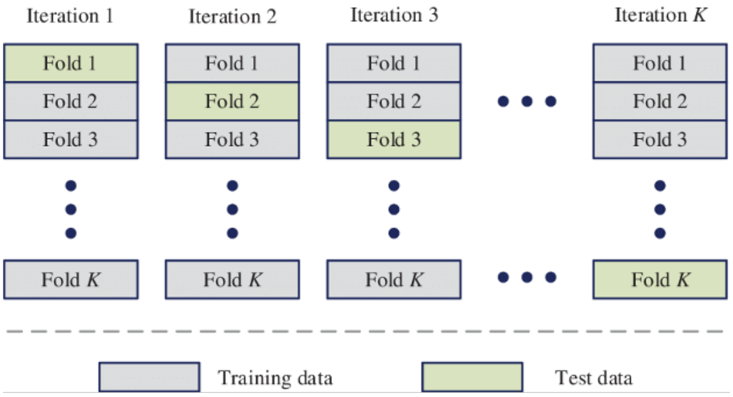


Figure 2: K-Fold Cross Validation

2.7 Classification Accuracy Level

The evaluation model in grouping analyzed data is by using apparent error. This type of error is directly observable from the results, allowing for straightforward identification and correction without the need for complex evaluation procedures [10]. One of these methods can use a confusion matrix table. Confusion matrix is a table used to evaluate the performance of a classification model by comparing model predictions and actual labels. In the confusion matrix, there are four basic elements, namely True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). True Positive is a positive case that is predicted correctly by the model. True Negative is the number of negative cases predicted correctly by the model. False

Positive is the number of negative cases predicted as positive or type I errors. False Negative is the number of positive cases predicted as negative or type II errors. Table 1 below is a confusion matrix table for classification problems [18].

Table 1: Confusion Matrix

Actual	**Prediction**	
	Positive	**Negative**
Positive	True Positive (TP)	False Negative (FN)
Negative	False Positive (FP)	True Negative (TN)

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} \tag{15}$$

$$\text{Sensitivity} = \frac{TP}{TP + FN} \tag{16}$$

$$\text{Specificity} = \frac{TN}{FP + TN} \tag{17}$$

Accuracy, sensitivity, and specificity of the model are obtained by entering into the Eq. 15, Eq. 16, and Eq. 17 True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). Accuracy indicates the accuracy or precision of the classification results, while sensitivity is used to measure the ability of the model to detect positive cases correctly. A high sensitivity value can indicate that the model can detect almost all important positive cases, where failure to detect positive conditions can have serious consequences. Specificity can be used to measure the ability of the model to recognize negative cases correctly, where it can indicate the proportion of all negative cases that the model correctly detects. Assuming a high specificity value, this indicates that the model performs effectively in identifying negative cases. Moreover, misclassifying negative cases as positive can cause problems, then false positive results must be avoided.

2.8 Random Over Sampling Examples (ROSE)

Class imbalance in classification problems can lead to model bias toward the majority class, thereby impairing the model's ability to identify instances of the minority class accurately [19]. The ROSE (Random Over Sampling Examples) method balances classes by generating synthetic data using a smoothed bootstrap approach based on kernel density estimation (KDE). The synthetic data are generated around the existing minority class observations. The kernel in the Eq. 18 describes the process of generating these synthetic examples.

$$x^* \sim \sum_{i=1}^{n_1} \frac{1}{n_1} K_h(x - x_i) \tag{18}$$

where, n_1 is number of observations in minority class. Then, K_h is kernel function with parameter bandwidth h. In this study, use Gaussian kernel [20].

$$K_h(x - x_i) = \frac{1}{(2\pi h^2)^{p/2}} \exp\left(-\frac{\|x - x_i\|^2}{2h^2}\right) \tag{19}$$

Eq. 19 ensures that each synthetic point x^* is generated from the local distribution around the minority sample. Thus, the approach preserves the original data distribution while improving class balance, enhancing overall classification performance.

2.9 Data Sources

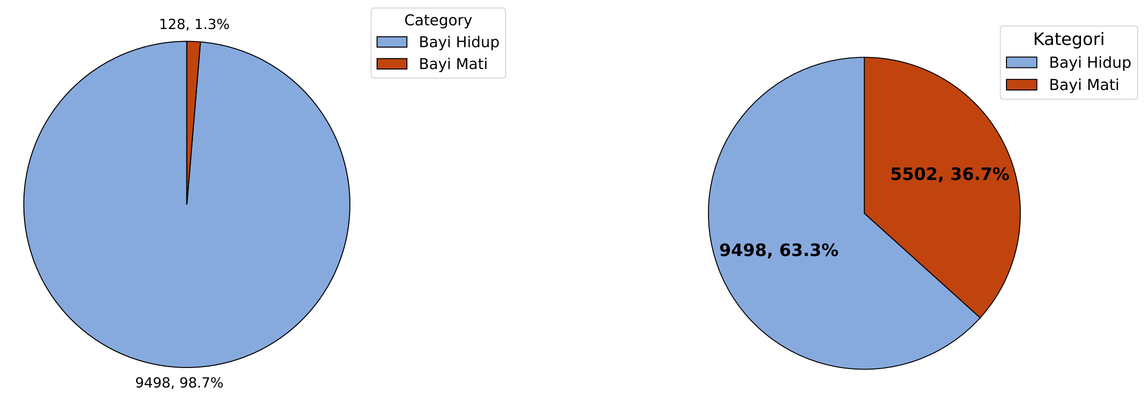
The data was used from the 2017 Indonesian Demographic Survey (SDKI 2017) by the Central Statistics Agency (BPS), with a sample of 9,626 babies born from 2012 to 2017 and born to women of childbearing age [4]. The definition of women of childbearing age based on the implementation of the SDKI is women aged 15-49 years. The variables used in this study consist of two dependent variables (Y) and independent variables. The dependent variable is categorical data, the Infant Mortality Rate (IMR). Meanwhile, the independent variables consist of 13 variables (Birth Weight, Mother's Age at Delivery, Mother's Education, Mother's Occupation, Birth Interval, Frequency of Antenatal Care, Place of Delivery, Infant Gender, Breastfeeding, Birth Type, Birth Order, Residence, Family Wealth Index). The IMR has two categories, which are live infants and infant mortality.

The following are systematic research steps to obtain accurate IMR classification results. Each stage is designed to ensure the validity and robustness of the model used. The main research steps include data collection, pre-processing, handling class imbalance, model construction using Random Forest, model evaluation, and interpretation of the most influential variables. These steps are explained as follows.

1. Data Collection
   The study uses secondary data from Indonesia's 2017 Infant Mortality Rate (IMR) dataset, which includes 9,626 observations with various socio-economic and demographic variables such as breastfeeding, type of birth, and birth weight.
2. Pre-processing
   Data exploration: Descriptive statistics and pie charts are generated to visualize the distribution of infant mortality cases before and after oversampling.
3. Handling Class Imbalance using ROSE
   (a) The original dataset is highly imbalanced, with only 1.3% infant deaths.
   (b) To address this, the ROSE (Random Over Sampling Examples) method is applied, which synthetically generates minority class samples using a smoothed bootstrap approach based on Kernel Density Estimation (KDE).
   (c) This step increases the total data to 15,000 balanced cases.
4. Model Development with Random Forest
   (a) The Random Forest algorithm is applied for classification, where the number of trees (ntree) is set to 500
   (b) The optimal number of predictors at each split (mtry) is tested with values 3 and 4, and $mtry = 4$ is selected due to its lower error rate (0.02%).
5. Model Evaluation
   (a) Out-of-Bag (OOB) error estimation is used to assess model generalization
   (b) K-Fold Cross-Validation ($k = 5$) is conducted to evaluate performance
   (c) Confusion matrix analysis shows accuracy, sensitivity, specificity
   (d) Variable Importance Analysis using the Gini Index
6. Interpretation

3 Results and Discussion

Descriptive Statistics Mortality Infant Mortality Rate (IMR) in 2017 has 9,626 data. It shows a characteristic used in this research. The data has been resampled using oversampling. Since it has an imbalance, especially in IMR which the number of live babies is greater than the number of dead babies. Imbalanced data provides errors in the analysis. However, it takes effort to have sensitivity values. On the other hand, the model used in this study, which applies resampling techniques such as oversampling, has limited interpretability. It excludes data information but increases minority data based on natural similarities. The previous data was 9.926 cases; after oversampling, it became 15,000 cases.



(a) Descriptive Statistics IMR Before ROSE

(b) Descriptive Statistics IMR After ROSE

Figure 3: Perbandingan statistik deskriptif IMR sebelum dan sesudah ROSE.

Figure 3a shows that 128 babies out of 9,626 births, or 1.3% of cases, died. The rest, 9,498 babies out of 9,626 births, or 98.7% of cases, were alive. It shows that there were 14 deaths out of 1,000 births. Besides that, Figure 3b shows descriptive statistics IMR after oversampling using ROSE.

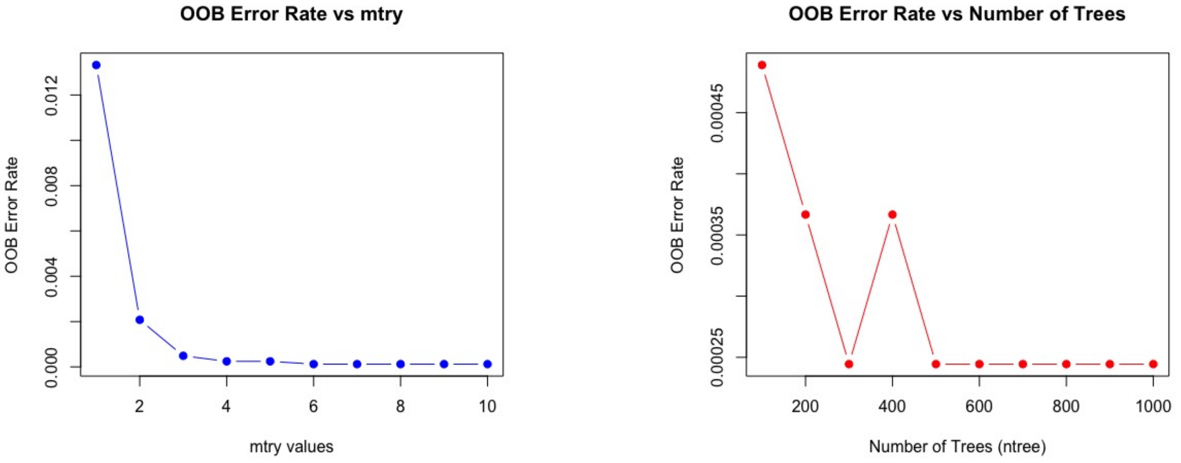
The data used has been resampled because there is a data imbalance where the case of living babies more than dead babies, causing data imbalance. Continuing the analysis without addressing the imbalance may result in biased outcomes, difficulty in generating sensitivity values, and limited model interpretability. The study uses resampling for handling oversampling. It retains the original data while enhancing the minority data by adding similar instances. The previous data was 9,926 cases; after oversampling, it became 15,000 cases. The oversampling used is Random Walk Oversampling (ROSE).

In the Random Forest method, there is no pruning process, which is a pruning process to obtain an optimal tree. However, this process can affect the classification tree's accuracy level [10]. Before conducting the Random Forest analysis, select the best separator between the predictor variables, randomly with M or $\sqrt{p}$, where M or p is the number of predictors [6]. The predictor variables used in this study are 14, then the M value is 3 or 4 with 500 trees or a k value of 500 trees. Values of M and k have been determined, then choose one that produces the smallest error. Table 2 is a table of classification errors using 500 trees and M values of 3 and 4.

Table 2: Classification error rate for different values of M and k

M	k	Error Rate (%)
3	500	0.12
4	500	0.02

Table 2 shows that the smallest classification error value is $M = 4$. This value indicates that using $M = 4$ can minimize the out-of-bag (OOB) estimate from the error rate to 0.02% compared to using $M = 3$. If depicted with an OOB plot, then with $M = 4$, the error value has converged. Figure 4a presents a plot of the OOB vs. error rate.



(a) OOB error for different *mtry* values **(b)** OOB error rate with varying number of trees

Figure 4: Comparison of OOB error under different settings.

Figure 4a shows that using $M = 4$, the error rate becomes convergent compared to using $M = 3$. Then, the optimum value for the number of tree replications is examined. Figure 4b shows that using 500 trees yields the most optimal OOB value.

Since four features are selected, it means that at each node of the tree in the Random Forest, the algorithm randomly considers four features from all the existing features to determine the best split. The larger the value of M, the more features are considered at each branch, which can increase the possibility of finding a better split. However, it may also increase model complexity and the risk of overfitting.

Meanwhile, if the number of trees is large, the model predictions tend to be more stable and accurate because the results are averaged across all trees. Nevertheless, beyond a certain point, increasing the number of trees does not significantly improve accuracy but only increases computing time. Based on the analysis, this study uses 500 trees with $M = 4$.

Following the determination of the number of variables and trees, the next step is to perform classification using the Random Forest method. Figure 5 presents the resulting Random Forest classification tree.

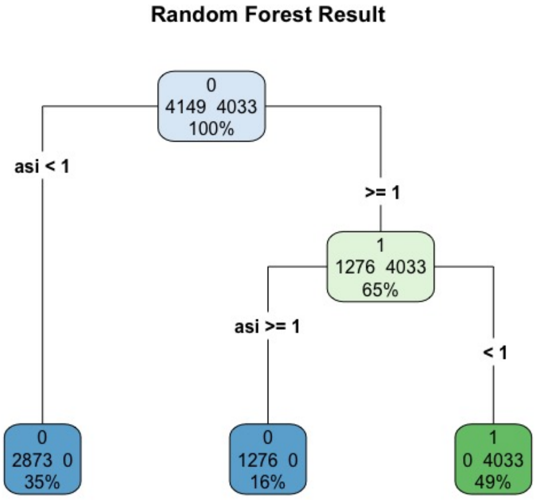


Figure 5: Random Forest Result

Figure 5 shows the results of the Random Forest, which has five nodes. Node 2 (left branch of Node 1) has 2,873 observations with predicted class 0. Predicted class 0 means that all observations in this node are predicted as class 0. In addition, it has an expected loss of 0, where there is no error because all observations in this node are class 0. A total of 2,873 observations in this node fall into class 0. The probability value is 1.0 for class 0 and 0.0 for class 1, meaning that all observations are predicted as class 0.

Node 3, or the right branch of Node 1, has 5,309 observations. The predicted class is 1, indicating that most observations in this node are class 1. The expected loss is 0.2403466, meaning the prediction error when predicting all observations in this node as class 1 is 0.2403466. Class counts include 1,276 observations in class 0 and 4,033 observations in class 1. The probabilities are 0.240 for class 0 and 0.760 for class 1, indicating that 76% of observations in this node are class 1.

Node 6, or the left branch of Node 3, has 1,276 observations. The predicted class is 0, meaning all observations in this node are predicted as class 0. The expected loss is 0, indicating no error since all observations are class 0. The class count is 1,276 observations in class 0. The probability is 1.0 for class 0, meaning all observations are predicted as class 0.

Node 7, or the right branch of Node 3, has 4,033 observations. The predicted class is 1, so all observations in this node are predicted as class 1. The expected loss is 0, or no error, because all observations are class 1. Class counts show that 4,033 observations fall into class 1. The probability is 1.0 for class 1, meaning all observations are predicted as class 1.

Figure 6 shows important variables in the case of Infant Mortality Rate based on the Gini index. The Gini index characterizes the ability to distinguish classes with the largest number of observations or the most important variables in a node. *Class* is a category of important variables in the response, where the response variables used in this study are the number of live and dead babies.

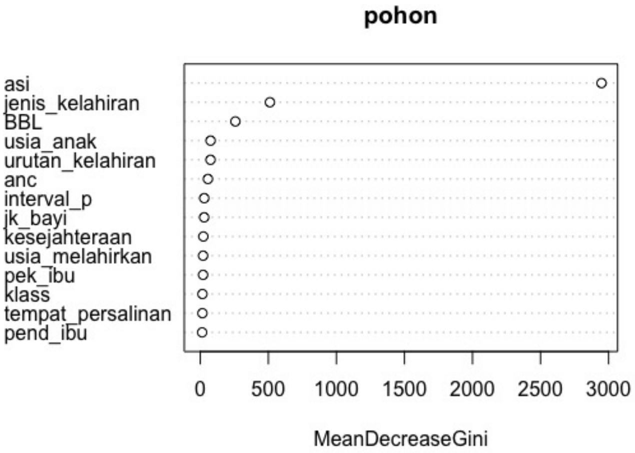


Figure 6: Importance Variable of IMR

Figure 6 highlights that breastfeeding is the most significant factor affecting IMR (Infant Mortality Rate), with an average Gini index close to 3,000. The influence of breastfeeding is further reinforced by other variables, each with Gini indices around 600, reflecting their moderate yet distinct importance in affecting IMR. However, their impact is less substantial compared to breastfeeding.

Additionally, birth weight is the next most influential variable, with a mean decrease in the Gini index of around 250, indicating its moderate contribution to improving the model's predictive accuracy. However, it plays a lesser role compared to breastfeeding. As shown in Figure 6, the parent node in the IMR decision tree is breastfeeding, indicating that breastfed infants have a higher likelihood of survival than those not breastfed, where mortality is higher.

Random Forest model evaluation can be performed using k-fold validation to assess the model's goodness of fit. In this study, the evaluation is conducted using 5-fold cross-validation with `mtry = 4`.

Table 3: Random Forest classification accuracy with $mtry = 4$ and $ntree = 500$

mtry	ntree	Accuracy (%)
4	500	0.9997557

Table 3 shows that the best level of accuracy tested using 5-fold cross-validation is 0.9553649 using as many as four variables.

Table 4: Confusion matrix and APER values for the testing data

	Prediction Class		Total	APER (%)	1-APER (%)
Actual Class	Positive	Negative			
Data Testing Positive	True Positive: 698	False Positive: 0	698	6.925208×10^{-4}	99.93
Data Testing Negative	False Negative: 1	True Negative: 745	746		
Total	699	745	1,444		

Table 4 shows the True Positive (TP) value of 745 cases, where the actual class is 1 and the model also predicts 1. This indicates that the model successfully recognizes 745 examples of class 1 correctly. The True Negative (TN) value is 698 cases, where the actual class is 0 and the model also predicts 0, meaning the model correctly identifies 698 examples of class 0. The False Positive (FP) value is 0, indicating that there are no cases where the actual class is 0 but the model incorrectly predicts 1. This means the model made no mistakes in predicting class 1 when the actual class was 0. Furthermore, the False Negative (FN) value is only 1 case, where the actual class is 1 but the model predicts 0. This shows that the model fails to recognize only one example of class 1 and incorrectly classifies it as class 0.

Table 5: Accuracy, Sensitivity, and Specificity of the model

Accuracy (%)	Sensitivity (%)	Specificity (%)
0.999308	1	0.998569

Table 5 shows an accuracy value of 99.93%, meaning that this model accurately predicts both class 0 and class 1. The sensitivity value of 99.87% indicates that the model is highly effective in identifying class 1 instances, with only a single case misclassified. The specificity value of 100% means that this model is perfect in identifying examples of class 0 because there are no false positive prediction errors. These values indicate that overall, this model has very good performance with almost perfect predictions for class 0 and class 1, with only one false negative error occurring.

4 Conclusion

Conclusions based on the problems in the study are the predictive factors that affect the IMR based on the Random Forest method, namely breastfeeding, where the average Gini Index is almost 3,000. Then, supported by other variables, which are around 600. The next variable is Birth Weight around 250. Mean Increase Gini is a measure of the contribution of each variable to the homogeneity of nodes and leaves in the resulting random forest. The higher the average value of the decrease in accuracy or the average reduction in the Gini score, the higher the importance of the variable in the model. The parent node in IMR is breastfeeding, where live IMRs that are given breastfeeding have a greater chance of survival than dead IMRs that are not given breastfeeding.

The accuracy of IMR using Random Forest is 99.98%, which means that the classification tree formed in the observation is appropriate. Meanwhile, validation to determine the classification of a feasible tree or the accuracy of the classification error is 6.925208×10^{-4}.

Random Forest was selected for its strengths in handling class imbalance, modeling non-linear relationships, and identifying variable importance through the Gini Index. Although comparing the performance of alternative models such as logistic regression, support vector machines, or decision trees could offer additional perspectives, such experiments were not conducted in this study. These comparisons are recommended for future research to validate and enrich the current findings.

The results showed that the model used had high accuracy and could identify important variables such as breastfeeding and birth weight as the main factors influencing infant life expectation. The findings have implications for policymakers, especially in the health sector in Indonesia. Programs that need to be encouraged are encouraging parents to provide exclusive breastfeeding and improving maternal and infant health services need to be prioritized, especially in areas with high infant mortality rates. In addition, using the ROSE oversampling technique has proven effective in dealing with data imbalance. This approach is hoped to be applied in other studies with similar challenges. Future research can consider the use of longitudinal data so that it can see the impact of policies over time.

CRediT Authorship Contribution Statement

Ria D.L.N.Karisma: Conceptualization, Methodology, Writing–Original Draft, Validation, Visualization. **Usman Pagalay:** Data Curation, Formal Analysis, Writing–Review & Editing, Supervision, Project Administration, Funding Acquisition. **Muhammad Khudzaifah:** Software, Validation, Visualization.

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 funding from the Bantuan Penelitian Berbasis Standar Biaya Masukan Program (SBM), LP2M UIN Maulana Malik Ibrahim Malang, 2024. We acknowledge the Department of Mathematics, Faculty Science and Technology, UIN Maulana Malik Ibrahim Malang, for the data analysis laboratory facilities support provided during this research.

References

[1] World Health Organization. “Unicef-who-wb joint child malnutrition estimates group released new data for 2021.” Accessed: June 26, 2025. (2021), Available online.

[2] UNICEF. “Levels & trends in child mortality: Report 2019.” Accessed: June 26, 2025. (2019), Available online.

[3] United Nations. "Ensure healthy lives and promote well-being for all at all ages (sdg goal 3)." Retrieved January 2024. (2020), Available online.

[4] BKKBN and BPS and Ministry of Health and USAID, "Survei demografi dan kesehatan indonesia 2017," BKKBN, BPS, Ministry of Health, and USAID, Jakarta, Indonesia, 2017, Accessed: June 26, 2025. Available online.

[5] T. Bylander, "Estimating generalization error on two-class datasets using out-of-bag estimates," *Machine Learning*, vol. 48, no. 1-3, pp. 287–297, 2002. DOI: 10.1023/A:1013964023376.

[6] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001. DOI: 10.1023/A:1010933404324.

[7] J. C. Lee, "Predicting mortality risk for preterm infants using random forest," *Scientific Reports*, vol. 11, no. 1, p. 7308, 2021. DOI: 10.1038/s41598-021-86748-4.

[8] L. M. Frota, M. Hasegawa, and P. Jacinto, "Infant mortality in brazil: A survival analysis using machine learning models," *ResearchGate*, pp. 1–46, 2024. DOI: 10.13140/RG.2.2.32819.64805.

[9] T. G. Dietterich, "Ensemble methods in machine learning," *Lecture Notes in Computer Science*, vol. 1857, pp. 1–15, 2000. DOI: 10.1007/3-540-45014-9_1.

[10] R. D. Karisma, "Random forest of modified risk factor on ischemic and hemorrhagic (case study: Medicum clinic, tallinn, estonia)," in *Proceedings of the International Conference on Science and Science Education*, Accessed: June 26, 2025, 2015, pp. 26–41. Available online.

[11] Janosh. "Illustrating the random forest algorithm in tikz." Retrieved January 2024 from https://tex.stackexchange.com/. (Aug. 13, 2019), Available online.

[12] S. W. He, "Predictive modeling of groundwater nitrate pollution and evaluating its main impact factors using random forest," *Chemosphere*, vol. 290, p. 133 388, Mar. 2022. DOI: 10.1016/j.chemosphere.2021.133388.

[13] M. I. Irawan and M. Jamhuri, "State of the art of machine learning: An overview of the past, current, and the future research trends in the era of quantum computing," in *AIP Conference Proceedings*, AIP Publishing, vol. 2641, 2022.

[14] Y. Amit and D. Geman, "Shape quantization and recognition with randomized trees," *Neural Computation*, vol. 9, 1997, Communicated by Shimon Ullman.

[15] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning*. Springer, 2013. Available online.

[16] V. K. Verma, *Analysis Effect of K Values Used in K Fold Cross Validation for Enhancing Performance of Machine Learning Model with Decision Tree*. Switzerland AG: Springer, Cham, 2024.

[17] Q. L. Ren, "Tectonic discrimination of olivine in basalt using data mining techniques based on major elements: A comparative study from multiple perspectives," *Big Earth Data*, vol. 3, no. 1, pp. 8–25, 2019. DOI: 10.1080/20964471.2019.1572452.

[18] G. M. Foody, "Challenges in the real world use of classification accuracy metrics: From recall and precision to the matthews correlation coefficient," *PLOS ONE*, vol. 18, no. 10, 2023. DOI: 10.1371/journal.pone.0291908.

[19] N. Lunardon, G. Menardi, and N. Tore, "Rose: A package for binary imbalanced learning," *The R Journal*, vol. 6, pp. 82–92, 2014.

[20] J. Zhang and L. Chen, "Clustering-based undersampling with random over sampling examples and support vector machine for imbalanced classification of breast cancer diagnosis," *Computer Assisted Surgery*, vol. 24, no. 52, pp. 62–72, 2019. DOI: 10.1080/24699322.2019.1649074.