pijmath
Pattimura International Journal of Mathematics (PIJMath)
Volume 04, issue 01, pages 017-028, May 2025
E-ISSN : 2830-6791
RESEARCH ARTICLE ◙ *OPEN ACCESS*
DOI: https://doi.org/10.30598/pijmathvol4iss1pp17-28

Classification of Poverty Status in Maluku Province using SMOTE-Random Forest Algorithm

Ferina L. Damamain[1], Lexy J. Sinay[2*], Sanlly J. Latupeirissa[3], Lusye Bakarbessy[4]

[1,2,3]Statistics Study Program, Faculty of Science and Technology, Pattimura University
[4]Mathematics Study Program, Faculty of Science and Technology, Pattimura University
[2]Integrated Mathematics Laboratory, Faculty of Science and Technology, Pattimura University
Jl. Ir. M. Putuhena, Ambon, Maluku, 97233, Indonesia

Corresponding author's e-mail: [2]lexyjz@gmail.com / lexy.sinay@lecturer.unpatti.ac.id*

Article History

Received: November 14th, 2024
Revised: March 30th, 2025
Accepted: April 24th, 2025
Published: May 1st, 2025

Keywords

Classification;
Maluku Province;
Poverty;
Random Forest;
SMOTE;

ABSTRACT

Poverty is a complex issue. According to BPS publications, in 2023, 9.36% of the Indonesian Population lives below the poverty line. Maluku is one of the provinces with a high poverty rate, reaching 16.23%. This research aims to classify poverty status in Maluku Province using the SMOTE-random forest algorithm. This research uses SUSENAS 2022 data, where the data is not balanced. SMOTE is used to overcome this problem. The model is based on a training data proportion of 75% and testing 25%, with Random Forest parameters consisting of $\boldsymbol{m}$, the number of variables randomly selected at each split, set to 4, and $\boldsymbol{r}$, the number of trees in the forest, set to 100. The critical factor influencing poverty status in Maluku Province is the number of household members. Although based on survey data of 5,972 households, which may limit the generalizability of the findings, this study provides a novel application of SMOTE-Random Forest in poverty classification at the provincial level in Indonesia.

How to cite this article:
Damamain F. L., Sinay L. J., Latupeirissa S. J., Bakarbessy L., "CLASSIFICATION OF POVERTY STATUS IN MALUKU PROVINCE USING SMOTE-RANDOM FOREST ALGORITHM", *Pattimura Int. J. Math. (PIJMATH).*, vol. 04, iss. 01, pp. 017-028, May, 2025.

1. Introduction

Poverty is a complex issue that encompasses not only low income and limited assets but also restricted access to essential services such as education, healthcare, sanitation, clean water, and public infrastructure. Furthermore, poverty is intricately linked to social inequalities, including disparities in educational attainment, gender, religious, and ethnic differences, as well as the prevalence of discrimination and limited access to economic opportunities, among other factors [1]. These various dimensions highlight that poverty is a deeply rooted problem that affects both economic and social aspects of life.

The complexity of this poverty issue is clearly reflected in Indonesia, where poverty remains a significant and persistent challenge despite numerous efforts by both the central and local governments. However, despite these initiatives, poverty continues to persist and requires comprehensive and sustained interventions. According to the 2023 publication by Statistics Indonesia (BPS), approximately 9.36% of the Indonesian population lives below the poverty line [2].

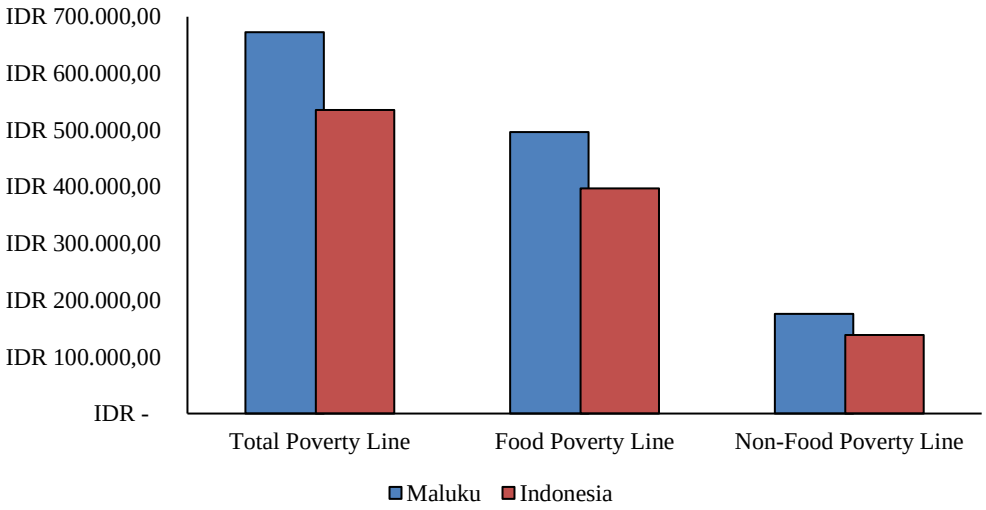

Figure 1. Comparison of Poverty Lines Between Maluku and Indonesia

Maluku Province is one of the regions requiring special attention, as poverty in Maluku remains one of the critical issues that need to be addressed. Based on BPS data from September 2022, Maluku's poverty line is higher than the national average, indicating that the minimum living costs in Maluku are relatively greater compared to Indonesia as a whole [3][4]. A higher poverty line suggests that, to meet basic food and non-food needs, individuals in Maluku require more resources than the average Indonesian citizen. Although the expenditure patterns between food and non-food components are relatively similar at both the national and regional levels, the overall standard for basic living necessities in Maluku remains higher.

The factors contributing to poverty in Maluku Province are highly specific, as they are closely related to Maluku's geographical conditions as a vast archipelagic region. Several factors contribute to the region's poverty, including low educational attainment [5], unequal access to resources and infrastructure such as clean water, sanitation, and electricity, limited access to quality healthcare services [6], and a reliance on agriculture and fisheries sectors, which are vulnerable to climate change and natural disasters [7]. These interrelated factors create a complex web that drives the high poverty rates in Maluku Province.

The Maluku Provincial Government's efforts to address poverty include expanding access to basic services, providing social assistance and economic programs, and improving infrastructure. However, challenges such as limited resources and infrastructure, low community participation, and the ineffectiveness of certain government programs continue to hinder progress. Targeted government programs are crucial in reducing poverty rates. Well-targeted programs enable the government to allocate resources efficiently and provide assistance that directly addresses the needs of the most vulnerable populations. This ensures that the benefits of the programs reach the right recipients, allowing the intended impact to be achieved directly and effectively.

The poverty proportion data produced by BPS uses the Poverty Line as a standard to classify the population's poverty level, representing the minimum basic needs required for a healthy and decent living. To analyze this data, a method or model that can accurately identify the poverty status of each household based on the factors influencing it is needed. The use of an appropriate model greatly aids in reducing the risk of misallocating aid to those who are truly

eligible. A well-designed model has high classification accuracy, ensuring that assistance is directed to the right recipients, thereby enhancing the effectiveness of poverty reduction efforts.

Classification is a process used to identify patterns or functions that describe and differentiate data into distinct classes, enabling the prediction of previously unclassified data [8]. One of the widely recognized classification methods is the random forest, introduced by [9]. Random forest is an extension model of the decision tree methodology, where multiple decision trees are constructed, each trained using a subset of data samples, and a random selection of features is considered at each split of the tree. This ensemble method enhances predictive accuracy by aggregating the results of several decision trees, thus improving model robustness. Random forest serves two key purposes: classification and regression. It is particularly effective in detecting interactions between dependent and independent variables while offering flexibility in handling complex datasets. Its ability to reduce overfitting and handle large datasets makes the random forest method a suitable model for diverse applications, particularly in contexts requiring high classification accuracy and robust predictive performance.

The random forest method offers several advantages, including an accurate feature selection process that enhances the performance of classification models, making it particularly effective for addressing big data problems with more complex parameters [10]. The random forest algorithm can handle data with a large number of attributes, identifying the significant features that influence predictions. Additionally, random forest is easy to implement and applicable to a wide range of data types and problems [11]. [12] compared the random forest method with other decision tree methods for classifying household poverty status in Central Java Province. The study found that random forest outperformed other decision tree methods, achieving an accuracy of 93.95%. Similar findings were reported by [13] and [14], where the random forest method demonstrated excellent performance, yielding very high accuracy rates. These results further affirm the suitability of random forest as a reliable classification tool in various domains.

However, since Random Forest decision are based on classification error, the presence of imbalanced data has been shown to substantially reduce the effectiveness of this classification models [15]. Data imbalance occurs when one class of data has significantly more observations than the other class. The class with the larger number of observations is referred to as the majority class, while the class with fewer observations is called the minority class. If this imbalance is ignored, the classification process may become biased, with the model favoring the majority class and often neglecting the minority class [16]**.** One method that can address the issue of data imbalance is the Synthetic Minority Over-sampling Technique (SMOTE). The principle of this method is to increase the number of observations in the minority class to equalize it with the number of observations in the majority class by generating new synthetic data points based on the k-nearest neighbors' algorithm [17].

Several studies have demonstrated the effectiveness of SMOTE when combined with the Random Forest (RF) algorithm. For instance, Weller et al. [18] reported that the application of SMOTE resulted in higher classification accuracy compared to oversampling or no resampling approaches, with Random Forest emerging as one of the top-performing classifiers. Similarly, Ismail et al. [19] showed that integrating SMOTE with Random Forest improved prediction performance significantly, particularly in highly imbalanced datasets, outperforming traditional undersampling techniques, such as random undersampling. Furthermore, Fayz et al. [20] highlighted that using SMOTE alongside Random Forest enhanced the classification accuracy, sensitivity, and positive predictive accuracy in the diagnosis of cervical cancer, compared to only using Random Forest. These findings collectively support the use of SMOTE as an effective method to address data imbalance issues when using Random Forest classifiers.

This study aims to develop a classification model using the random forest method to accurately identify the poverty status of each household in Maluku Province. In addition, it seeks to analyze key variables influencing household poverty status at the district and city levels. The resulting model is expected to assist the government in effectively classifying poor and non-poor households, thereby supporting more precise targeting of poverty alleviation programs. By improving classification accuracy and providing insights into regional characteristics, the model can help ensure that aid is distributed to the households that need it most, minimize mistargeting, and enhance the overall effectiveness of poverty reduction efforts.

2. Research Methods

2.1. Random Forest Method

Random forest is a highly effective classification method that builds on the CART approach by using bootstrap aggregating (bagging) and random feature selection. This method relies on selecting the largest number of trees, making it a form of ensemble data mining. The error rate in the random forest algorithm depends on the strength and correlation among the decision trees [21]. Specifically, increasing the strength of individual trees, for example by considering more features at each split, can enhance predictive accuracy. However, such an increase may simultaneously raise the

correlation among trees, which reduces the benefit of averaging their predictions. To address this, random forests apply bootstrap sampling and random feature selection at each split node, thereby lowering the correlation between trees and decreasing the variance of the ensemble prediction [22]. Bagging creates several independent decision trees by using multiple random samples from the training data, enhancing model robustness. Meanwhile, random selection of features avoids the dominance of strong predictors and makes the model more robust to noise.

The working principle of the random forest method involves combining multiple decision trees to generate more accurate and stable predictions. As a result, the random forest method is considered simpler compared to other ensemble techniques such as boosting. This simplicity comes from the fact that random forests require fewer parameters to set, such as only the number of trees ($ntree$) and the number of features considered at each split ($mtry$). Furthermore, random forests are less sensitive to noisy data and outliers, allowing them to achieve good predictive performance without a lot of data preparation [23].

There are at least four key parameters used to build the random forest algorithm [24]:

$a_n \in \{1,2,\dots,n\}$ = the number of data points randomly selected from the original dataset to grow each individual tree.

$mtry \in \{1,2,\dots,m\}$ = the number of variables randomly chosen to be considered for splitting at each node during tree construction.

$ntree \in \{1,2,\dots,r\}$ = the total number of trees generated within the random forest.

$nodesize \in \{1,2,\dots,a_n\}$ = the minimum number of data points a node must have to be allowed to split further

Before constructing each decision tree, a random selection of observations a_n is drawn, with or without replacement, from the original dataset. Then, at each node of the tree, a random subset of m variables is selected from the total p available explanatory variables, where $m \leq p$. Typically, the best size of mmm is approximated by the square root of the total number of variables $\left(m = \sqrt{p}\right)$ although it can also be adjusted to twice $\left(m = 2\sqrt{p}\right)$ or half $\left(m = \frac{1}{2}\sqrt{p}\right)$ of that value depending on specific needs. Among the selected m variables, the best split is determined based on a predefined splitting criterion, such as Gini Importance. The node is then partitioned into two child nodes according to the selected split. This process is recursively repeated for each child node: at every split, a new subset of m variables is randomly selected and the optimal split is determined, until a stopping criterion is satisfied, such as all observations in a node belonging to the same class or the node size falling below a minimum threshold [23].

2.1.1. Variable Importance

Most data processing using machine learning is not only focused on finding an accurate model but also on identifying which predictor variables are the most important in the predictive model being used [25].Understanding the significance of each variable helps in interpreting the model, improving its efficiency, and potentially guiding decision-making processes based on the insights derived from the data. By identifying the key predictors, models can be refined to focus on the most impactful variables, ultimately enhancing their practical utility and performance in real-world applications.

Variable importance in the random forest algorithm can be calculated using the Gini Importance. Gini Importance is used to indicate the stability of each explanatory variable, where a higher Gini Importance value signifies a more stable variable [9]. Let p represent the number of predictor variables, with $h = 1,2,\dots,p$ representing each variable. The importance of predictor variable X_h can be calculated using the Gini Importance. The formula for calculating Gini Importance is as follows [26]

$$MDG_h = \frac{1}{r}\sum_{t=1}^{N}[z(h,t)I(h,t)]$$

where r is the number of the trees, N is the total number of nodes across the entire forest, h is the feature index, and t is the node index. The function $z(h,t)$ is an indicator variable that equals 1 if node t splits on feature h, and 0 otherwise. Meanwhile, $I(h,t)$ represents the impurity reduction associated with node t when it is split based on feature h.

2.1.2. Model Evaluation

Model evaluation is the process of measuring the performance of the generated model, with the goal of assessing its effectiveness in classification tasks. Typically, a confusion matrix is used to evaluate the performance of a model.

The confusion matrix provides information about the actual and predicted classes from the model (Table 2).

Table 3. Confusion Matrix

Actual	Prediction	
	Positive	**Negative**
Positive	True Positive (a)	False Negative (b)
Negative	False Positive (c)	True Negative (d)

True Positives (TP) and True Negatives (TN) represent the frequencies of observations that are correctly predicted by the model. True Positives refer to cases where the model accurately predicts positive instances, while True Negatives refer to cases where the model accurately predicts negative instances. In contrast, False Positives (FP) occur when the model incorrectly predicts a positive result for an observation that is actually negative, representing a Type I error. Conversely, False Negatives (FN) occur when the model incorrectly predicts a negative result for an observation that is actually positive

The random forest method can evaluate model performance using specificity, sensitivity, and accuracy. Specificity measures the model's ability to correctly identify negative samples from the total number of actual negative samples. Sensitivity (recall or true positive rate) measures the model's ability to correctly identify positive samples from the total number of positive samples. Accuracy represents the model's overall performance, indicating how well the model correctly classifies both positive and negative samples. Sensitivity and specificity are opposing metrics, meaning there is often a trade-off between them. The following equations describe how to calculate specificity, sensitivity, and accuracy **[27]**:

$$Specificity = \frac{d}{c+d};\ Sensitivity = \frac{a}{a+b};\ Accuracy = \frac{a+d}{a+b+c+d}$$

where:

a = True Positive (TP), meaning cases that are positive and correctly identified as positive.

b = False Negative (FN), meaning cases that are actually positive but incorrectly classified as negative.

c = False Positive (FP), meaning cases that are actually negative but incorrectly classified as positive.

d = True Negative (TN), meaning cases that are negative and correctly identified as negative.

2.2. Synthetic Minority Oversampling Technique (SMOTE)

In addressing classification problems, especially those involving imbalanced datasets, one of the most widely adopted techniques is the Synthetic Minority Over-sampling Technique (SMOTE). SMOTE is a widely used technique for handling data imbalance by synthetically increasing the number of minority class observations. It works by generating new data points through the k-nearest neighbors' approach, allowing the minority class size to match that of the majority class. The number of synthetic instances generated can be adjusted based on the practical considerations of implementation. Although there is no universally fixed threshold for when to apply SMOTE, a recommended majority-to-minority sample ratio is 5:1, which has been shown to improve model performance. By balancing the class distribution, SMOTE enhances the model's ability to accurately classify minority instances, resulting in more reliable and robust predictions **[16]**.

The SMOTE algorithm refines the synthetic sampling process by incorporating a normal distribution-based interpolation mechanism **[17]**. Essentially, SMOTE generates new minority class samples through random linear interpolation between existing samples and their neighboring instances. To enhance the classification performance on imbalanced datasets, the algorithm increases the data imbalance ratio by creating artificial minority samples. The specific procedure involves first calculating the distance of each minority sample x_i $(i = 1,2,\dots,n)$ o other samples within the minority class to identify its k nearest neighbors.

Based on the desired level of over-sampling, m neighbors are randomly selected from these k nearest neighbors. New synthetic samples p_{ij} are then generated according to the following equation:

$$p_{ij} = x_i + rand(0,1) \times (x_{ij} - x_i)$$

where $rand(0,1)$ is a random number between 0 and 1, x_{ij} is k-nearest neighbor observation vector, and x_i is the actual observation vector. This process continues until the fused dataset reaches a specified imbalance ratio.

2.3. Research Procedure

The stages in the research are

1. Pre-analysis:
   a. The data classification used aligns with the predefined definitions.
   b. Conduct data exploration and descriptive statistical analysis to obtain an overview of the variables to be analyzed.
   c. Identify whether the data is balanced.
2. Handling unbalanced data with the SMOTE method
3. Modeling using a random forest algorithm.
4. Evaluate the best model produced by looking at accuracy, sensitivity, and specificity values
5. Determine the important variables per district/city generated from the model.

3. Results And Discussion

3.1. Data Description

The data used in this study consist of household data from Maluku Province, obtained from the 2022 National Socioeconomic Survey (SUSENAS), sourced from the Central Statistics Agency (BPS). The variables used in the study are presented in **Table 4**.

The data used in this study consists of 5,972 household observations, comprising 880 poor households and 5,092 non-poor households. A poor household is defined as one whose average monthly per capita expenditure falls below the poverty line. The poverty line applied in this study is the Poverty Line for Maluku Province in September 2022, which was set at Rp. 672,456 per capita per month **[3]**.

The general overview of household poverty status categories in Maluku Province by district/city is presented in **Table 5**. **Table 6** shows that the number of non-poor households is significantly higher than poor households across Maluku Province. The regions with the highest poverty rate are Buru Selatan, Seram Bagian Barat, and Kepulauan Aru, where the percentage exceeds 20%. Meanwhile, the area with the lowest proportion of poor households is Ambon City, with a percentage of less than 20%.

Table 7. Research Variables

Variable		Description
Y:	Household Poverty Status	Poverty classification of the household into two categories (poor or non-poor).
X_1:	Residential Building Ownership Status	Ownership type of the household's residence (five categories).
X_2:	Floor area of a residential building (m^2)	Size of the residential building floor in square meters (nominal variable).
X_3:	The main building material of the widest house wall	Main material used for the widest part of the house wall (seven categories).
X_4:	Main building material of the largest house floor	Main material used for the largest area of the house floor (nine categories).
X_5:	Bowel room facilities	Type of bathroom facility available in the household (six categories).
X_6:	Main water source for bathing/washing/etc.	Main source of water for bathing, washing, and other domestic activities (eleven categories).
X_7:	The main types of fuel used for cooking	Main type of fuel used for cooking activities (eleven categories).
X_8:	Refrigerator ownership	Refrigerator ownership status (two categories).
X_9:	Air Conditioner Ownership	Air conditioner ownership status (two categories).
X_{10}:	Computer/Laptop Ownership	Computer or laptop ownership status (two categories).
X_{11}:	Gold/Jewelry Ownership	Gold or jewelry ownership status (two categories).
X_{12}:	Motorcycle Ownership	Motorcycle ownership status (two categories).
X_{13}:	Car Ownership	Car ownership status (two categories).
X_{14}:	Number of Household Members	Total number of household members (nominal variable).
X_{15}:	Age of Head of Household	Age of the head of household (nominal/continuous variable).
X_{16}:	Higher Education Level for Heads of Households	Highest educational attainment of the head of household (twenty-four categories).
X_{17}:	Household Status	Household status within the dwelling unit (two categories).

Source: Central Statistics Agency (BPS)

Table 8. Poverty Status

Regency/City	Poor		Not Poor	
	Sum of Household	Percentage	Sum of Household	Percentage
Ambon City	12	2.09	563	97.91
Buru Regency	58	10.74	482	89.26
South Buru Regency	134	26.43	373	73.57
Aru Islands Regency	114	21.55	415	78.45
Tanimbar Islands Regency	91	16.98	445	83.02
Southwest Maluku Regency	45	8.36	493	91.64
Central Maluku Regency	75	11.94	553	88.06
Southeast Maluku Regency	67	12.57	466	87.43
West Seram Regency	128	23.27	422	76.73
Eastern Seram Regency	91	16.91	447	83.09
Tual City	65	13.05	433	86.95
Maluku Province	**880**	**14.74**	**5092**	**85.26**

Source: Data Processing

3.2. Unbalanced Data Handling

The data balancing technique used in this study is SMOTE (Synthetic Minority Over-sampling Technique). By applying SMOTE, the imbalance between poor and non-poor household data is mitigated, ensuring that the classification model has sufficient representation from the minority class (poor households). This helps improve the model's accuracy in predicting the poverty status, especially for the minority group. This technique works by generating synthetic samples from the minority class by adding noise to the data, thereby increasing the number of samples in the minority class. The results of the SMOTE analysis are presented in **Table 9**.

Table 10. SMOTE Results

Information	Before SMOTE		After SMOTE	
	Poor	Not Poor	Poor	Not Poor
Frequency	880	5092	3015	2957
Percentage	14.74	85.26	50.49	49.51

Source: Data Processing

Table 11 shows that the initial data proportion (before SMOTE) was 1:6, with non-poor households representing the majority class. After applying SMOTE, the results indicate that the data proportion has become balanced, providing a more equitable distribution between poor and non-poor households compared to the original data.

3.3. Random Forest Modeling

The initial step in the random forest algorithm is to train the data generated from the SMOTE technique. Various combinations of the parameters m (the number of variables randomly sampled as candidates at each split) and r (the number of trees in the forest) are tested during the modeling process to comprehensively evaluate the model's performance. The values of m used are:

$$m = \sqrt{p} = \sqrt{17} = 4.12 \approx 4;\ m = 2\sqrt{p} = 2\sqrt{17} = 8.25 \approx 8; m = \tfrac{1}{2}\sqrt{p} = \tfrac{1}{2}\sqrt{17} = 2.06 \approx 2;$$

The performance of a random forest model depends on the determination of the number m of explanatory variables (features) to be randomly selected and the number r of trees to be constructed. A value of $r = 50$ has been tested and shown to provide good classification results when using the bagging method **[9]**. Furthermore, recent studies suggest that increasing the number of trees generally improves the stability and predictive performance of random forest models. **[28]** found that the matching performance stabilized after constructing 500 trees, leading them to recommend using 500 trees in their application. Similarly, **[29]** emphasized that while most performance gains occur within the first 100 trees, using a larger number of trees further enhances model robustness, particularly for stability and variable importance estimation. However, **[9]** mentioned that the generalization error of a random forest converges as the number of trees increases, indicating that while adding more trees reduces error, most of the predictive performance is captured relatively early in the ensemble growth. Therefore, the evaluation of models with a smaller number of trees remains important, particularly for cases involving limited computational resources or smaller datasets.

The optimal split between training and testing data for random forest can be influenced by several factors, including dataset size, the number of features, and class imbalance. **[16]** recommended using 66% of the total data for training. Further theoretical justification was provided by **[30],** who explained that statistically, allocating 70–80% of the data for training and 20–30% for testing yields the best balance. This split minimizes the risk of overfitting while

ensuring the validity and precision of model performance estimation, as it maximizes the product of the training and testing proportions and satisfies the conditions for reliable error estimation.

Table 12, **Table 6**, **Table 7**, **Table 8**, and **Table 9** present the performance analysis of the random forest classification model based on the split between training and testing data. The results indicate that the best random forest classification model is the one with a 75:25 split between training and testing data, using parameters $m = 4$ and $r = 100$. This conclusion is drawn from the higher accuracy and sensitivity values achieved by this combination compared to other configurations

Table 13. Random forest model performance for 70:30 data proportions

Parameters		m					
		2		4		8	
r	25	Accuracy	0.816	Accuracy	0.835	Accuracy	0.834
		Sensitivity	0.779	Sensitivity	0.809	Sensitivity	0.807
		Specificity	0.862	Specificity	0.865	Specificity	0.864
	50	Accuracy	0.835	Accuracy	0.847	Accuracy	0.834
		Sensitivity	0.805	Sensitivity	0.818	Sensitivity	0.806
		Specificity	0.871	Specificity	0.880	Specificity	0.866
	100	Accuracy	0.833	Accuracy	0.848**	Accuracy	0.838
		Sensitivity	0.802	Sensitivity	0.818	Sensitivity	0.809
		Specificity	0.868	Specificity	0.882**	Specificity	0.871
	500	Accuracy	0.829	Accuracy	0.843	Accuracy	0.812
		Sensitivity	0.796	Sensitivity	0.813	Sensitivity	0.874*
		Specificity	0.870	Specificity	0.879	Specificity	0.840

* Indicates the highest value achieved for each individual evaluation criterion (accuracy, sensitivity, or specificity)
** Indicates the best parameter combination with multiple evaluation criteria reaching their highest values.
Source: Data Processing

Table 14. Random forest model performance for 75:25 data proportions

Parameters		m					
		2		4		8	
r	25	Accuracy	0.816	Accuracy	0.845	Accuracy	0.823
		Sensitivity	0.791	Sensitivity	0.811	Sensitivity	0.795
		Specificity	0.846	Specificity	0.887	Specificity	0.864
	50	Accuracy	0.832	Accuracy	0.849	Accuracy	0.840
		Sensitivity	0.805	Sensitivity	0.816	Sensitivity	0.811
		Specificity	0.864	Specificity	0.889	Specificity	0.874
	100	Accuracy	0.837	Accuracy	0.858**	Accuracy	0.848
		Sensitivity	0.812	Sensitivity	0.830**	Sensitivity	0.818
		Specificity	0.866	Specificity	0.890	Specificity	0.883
	500	Accuracy	0.831	Accuracy	0.852	Accuracy	0.849
		Sensitivity	0.798	Sensitivity	0.817	Sensitivity	0.818
		Specificity	0.873	Specificity	0.896*	Specificity	0.885

* Indicates the highest value achieved for each individual evaluation criterion (accuracy, sensitivity, or specificity)
** Indicates the best parameter combination with multiple evaluation criteria reaching their highest values.
Source: Data Processing

Table 15. Random forest model performance for 80:20 data proportions

Parameters		m					
		2		4		8	
r	25	Accuracy	0.826	Accuracy	0.835	Accuracy	0.834
		Sensitivity	0.795	Sensitivity	0.803	Sensitivity	0.803
		Specificity	0.863	Specificity	0.874	Specificity	0.871
	50	Accuracy	0.823	Accuracy	0.839	Accuracy	0.835
		Sensitivity	0.798	Sensitivity	0.805	Sensitivity	0.808
		Specificity	0.850	Specificity	0.881	Specificity	0.867
	100	Accuracy	0.825	Accuracy	0.849	Accuracy	0.837
		Sensitivity	0.789	Sensitivity	0.813	Sensitivity	0.806
		Specificity	0.871	Specificity	0.892*	Specificity	0.875
	500	Accuracy	0.829	Accuracy	0.852**	Accuracy	0.843
		Sensitivity	0.792	Sensitivity	0.820**	Sensitivity	0.811
		Specificity	0.875	Specificity	0.890	Specificity	0,880

* Indicates the highest value achieved for each individual evaluation criterion (accuracy, sensitivity, or specificity)
** Indicates the best parameter combination with multiple evaluation criteria reaching their highest values.
Source: Data Processing

Table 16. Random forest model performance for 85:15 data proportions

Parameters		m					
		2		4		8	
r	25	Accuracy	0.816	Accuracy	0.830	Accuracy	0.830
		Sensitivity	0.798	Sensitivity	0.797	Sensitivity	0.800
		Specificity	0.836	Specificity	0.872	Specificity	0.868
	50	Accuracy	0.818	Accuracy	0.852**	Accuracy	0.829
		Sensitivity	0.779	Sensitivity	0.823**	Sensitivity	0.799
		Specificity	0.870	Specificity	0.885	Specificity	0.866
	100	Accuracy	0.828	Accuracy	0.838	Accuracy	0.834
		Sensitivity	0.787	Sensitivity	0.807	Sensitivity	0.803
		Specificity	0.883	Specificity	0.876	Specificity	0.871
	500	Accuracy	0.821	Accuracy	0.846	Accuracy	0.836
		Sensitivity	0.787	Sensitivity	0.811	Sensitivity	0.807
		Specificity	0.865	Specificity	0.889*	Specificity	0.872

* Indicates the highest value achieved for each individual evaluation criterion (accuracy, sensitivity, or specificity)
** Indicates the best parameter combination with multiple evaluation criteria reaching their highest values.
Source: Data Processing

3.4. Determination of the Most Important Variables in Maluku Province

The random forest model provides significant insights by identifying important variables. This information can be instrumental in decision-making and policy formulation aimed at improving household welfare in Maluku Province. By highlighting the key factors that influence poverty, policymakers can design more targeted and effective interventions to address the specific needs of households and allocate resources more efficiently.

The most important variables in the model are determined based on Gini Importance, which measures each variable's contribution to the reduction of impurity in the random forest algorithm. This analysis identifies the variables that are most influential in distinguishing between the response categories, namely poor and non-poor households. Variables are ranked according to their Gini Importance values, with higher values indicating greater importance in predicting the poverty status of households in Maluku Province.

Table 17. The level of importance of variables on the best model random forest best

Variable	Gini Importance Values	Variable	Gini Importance Values
X_{14}	80.232	X_3	16.874
X_6	31.238	X_{11}	16.026
X_4	27.975	X_1	13.942
X_{16}	26.484	X_{10}	12.692
X_8	24.603	X_{13}	11.704
X_{15}	22.926	X_{12}	11.607
X_5	19.722	X_{17}	10.906
X_7	19.271	X_9	7.259
X_2	17.838		

Source: Data Processing

Based on **Table 18**, the variable with the highest level of importance in determining household poverty status in Maluku Province is X_{14}, or household size (the number of household members). This variable shows a significantly higher Gini Importance value compared to other variables, indicating that household size is a key indicator in determining poverty status. In other words, the number of household members has a significant influence on the poverty status of households in Maluku Province. Therefore, policies or programs that focus on managing or reducing household size may have a positive impact on overall poverty alleviation efforts in Maluku Province. Addressing household size through targeted interventions could help improve the economic well-being of households, particularly those at higher risk of poverty.

3.5. Determination of the Most Important Variables of the Regency/City

In the previous section, the most important variables for the overall Maluku Province were presented. In this section, the most important variables are examined on a partial basis for each district/city in Maluku Province. The Gini Importance values are used to rank the importance of explanatory variables in each district/city. These results are summarized in **Table 11**.

In this study, a Random Forest model was constructed at the provincial level using combined data from all districts and cities. The district/city-level analysis was conducted by applying the provincial model to the corresponding

subsets of data for each region. The Gini Importance values were calculated based on the provincial model, allowing comparisons of variable importance across districts and cities.

This approach was adopted to ensure methodological consistency and to maintain model robustness, considering that the number of observations within individual districts or cities was relatively limited. Building separate models for each district/city could introduce instability and reduce comparability due to varying sample sizes. By using a single provincial model, the analysis allows for direct comparison of variable importance across regions under a unified learning structure.

The data analysis results presented in **Table 11** provide valuable insights into the factors that influence poverty status in each district/city in Maluku Province. The variable X_{14} (household size) is identified as the most influential variable in determining poverty status in nearly every district/city in Maluku Province, except in Buru Regency. This indicates that household size plays a crucial role in affecting various social, economic, or demographic aspects across these regions. Additionally, in Buru Regency, household size is the second most important variable, with X_8 (ownership of a refrigerator) being the most important determinant of poverty status. This suggests that the ownership of some assets, such as a refrigerator, has a significant positive impact on household economics in Buru Regency, and thus, strongly influences poverty status.

Table 19. Important variables of each Regency/City in Maluku Province

Regency/City	The Most Important Variables		
	1	2	3
Ambon City	X_{14}	X_1	X_6
Buru Regency	X_8	X_{14}	X_6
South Buru Regency	X_{14}	X_{16}	X_4
Aru Islands Regency	X_{14}	X_6	X_5
Tanimbar Islands Regency	X_{14}	X_6	X_3
Southwest Maluku Regency	X_{14}	X_{16}	X_6
Central Maluku Regency	X_{14}	X_5	X_4
Southeast Maluku Regency	X_{14}	X_6	X_8
West Seram Regency	X_{14}	X_6	X_4
Eastern Seram Regency	X_{14}	X_7	X_5
Tual City	X_{14}	X_6	X_4

Source: Data Processing

The data analysis presented in **Table 20** provides critical information regarding the factors that influence poverty status in each district or city in Maluku Province. The variable X_{14} (household size) plays a highly significant role in determining poverty status in nearly every district/city in Maluku, except in Buru Regency. This suggests that household size is a key factor impacting various social, economic, and demographic aspects across these regions. This finding indicates that larger households may face greater economic vulnerability due to the increased consumption needs that must be met with limited resources. Consistent with this observation, **[31]** conclude that households with more than four members were 2.1 times more likely to experience extreme poverty compared to smaller households, as larger family sizes increase the burden on available resources. Similarly, **[32]** found that households with more than four members were 3.8 times more likely to fall into poverty, emphasizing that the strain on per capita income limits opportunities for savings and investment. Together, these studies reinforce the conclusion that household size is a critical factor that increasing the risk of poverty. Additionally, for Buru Regency, household size ranks as the second most important variable, with X_8 (ownership of a refrigerator) emerging as the primary determinant of poverty status. This finding indicates that the ownership of assets, such as a refrigerator, positively contributes to household economic conditions in Buru, significantly influencing poverty levels in the area.

Table 21 also reveals that, for the second and third rankings, the influencing variables for poverty across districts and cities in Maluku Province are quite diverse. For the second place, the most dominant important variable across the districts/cities is X_6 (main water source). In third place, X_4 (main building material for the largest floor area of the house) is prominent, followed by X_5 (access to sanitation facilities). This indicates that poverty status in Maluku's districts and cities is closely tied to health-related issues, such as the availability of clean water, the quality of flooring materials, and proper sanitation facilities, which have yet to meet health standards in many areas. Additionally, there are other variables that significantly impact certain districts/cities within Maluku Province. These variables include homeownership status, the main building material for the largest walls of the house, the primary fuel used for cooking, and the highest level of education attained by the household head. These findings emphasize the multifaceted nature of poverty, with factors related to housing, education, and access to basic resources playing crucial roles in determining household welfare across the region.

Based on the analysis of **Table 22**, the poverty issue in each district/city in Maluku Province has its own unique characteristics, except in West Seram Regency and Tual City, which share the same most important variables. The

variation in key variables across districts/cities in Maluku Province highlights that poverty is a cross-sectoral issue, requiring collaboration or cooperation between various sectors or fields to address it effectively.

4. Conclusions

Based on the data obtained, the ratio of poor to non-poor households was 1:6, indicating an imbalance in the actual data. To address this issue, the SMOTE (Synthetic Minority Over-sampling Technique) method was applied to balance the number of observations between the two classes. The balanced dataset was then used for random forest modeling, allowing the model to learn from poor and non-poor households more equally. As a result, the model's ability to predict poverty status improved, particularly in identifying poor households. Before applying SMOTE, the model tended to favor the majority class, leading to lower predictive accuracy for the minority class. After balancing the data, the model achieved better precision, recall, and overall classification performance, contributing to the development of a more reliable and robust poverty classification model.

The best random forest model obtained is the result of model selection. The optimal model was based on a data split of 75% for training and 25% for testing, with the number of variables per tree set to $m = 4$ and the number of trees generated is set to $r = 100$. The performance of this model is demonstrated by its accuracy, with an accuracy score of 85.8%, specificity of 83.0%, and sensitivity of 89.0%.

Based on the best random forest model, the overall Gini Importance values were obtained for Maluku Province as a whole, as well as partial Gini Importance values for each district/city within the province. The overall Gini Importance analysis for Maluku Province shows that the most important factor influencing the classification of household poverty status is the number of household members. Meanwhile, the partial Gini Importance analysis for each district/city reveals that the key variables influencing poverty classification vary significantly across regions. However, several variables tend to dominate in most districts/cities, including the main water source, the primary building material for the largest floor area of the house, and access to sanitation facilities.

Despite these promising results, this study has certain limitations, particularly related to the scope of the dataset. The analysis was based on survey data comprising 5,972 households, which, while substantial, does not fully represent the entire population of Maluku Province. Consequently, future research could benefit from using census data or expanding the sample size to improve the representativeness and generalizability of the model. Such enhancements would allow for a more comprehensive understanding of the factors influencing household poverty across different regions. Future research is encouraged to expand the dataset by incorporating census data or increasing the sample size to enhance the representativeness of the findings. In addition, applying cross-validation techniques, exploring alternative data balancing methods, and developing specialized models for each district or city would further improve the performance and generalizability of poverty classification models.

References

[1] I. L. Organization, "World Employment Social Outlook," 2019.
[2] BPS, "Indonesia Poverty Profile in March 2023," 2023.
[3] Statistics Indonesia, Maluku Province, "Poverty Profile in Maluku September 2022," Ambon, 2023.
[4] BPS, "Indonesia Poverty Profile in September 2022," 2023.
[5] BPS, "Maluku Province in Figures 2019," 2020.
[6] M. of Health, "Indonesia Health Profile 2017," Jakarta, 2018.
[7] Statistics Indonesia, Maluku Province, "Maluku Province in Figures 2019," Ambon, 2020.
[8] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*. Waltham, MA: Morgan Kaufmann Publishers, 2012.
[9] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, pp. 5–32, 2001.
[10] R. Supriyadi, W. Gata, N. Maulidah, and A. Fauzi, "Penerapan Algoritma Random Forest Untuk Menentukan Kualitas Anggur Merah," *E-Bisnis : Jurnal Ilmiah Ekonomi dan Bisnis*, vol. 13, no. 2, pp. 67–75, 2020, doi: 10.51903/e-bisnis.v13i2.247.
[11] A. Liaw and M. Wiener, "Classification and regression by randomForest," *R Journal*, vol. 2, no. 3, pp. 18–22, 2002, [Online]. Available: http://www.stat.berkeley.edu/
[12] F. Izzati, "Perbandingan Metode CHAID dan Random Forest," *Skripsi*, 2022.
[13] M. I. Putra, "Sistem Rekomendasi Kelayakan Kredit," *Skripsi*, 2019.
[14] L. Fadilah, *Kalsifikasi Random Forest Pada Data Imbalanced*. 2018.
[15] R. O'Brien and H. Ishwaran, "A Random Forests Quantile Classifier for Class Imbalanced Data," *Pattern Recognition*, vol. 90, pp. 232–249, 2019.
[16] N. V Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling

Technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.
[17] S. Wang, Y. Dai, J. Shen, and J. Xuan, "Research on Expansion and Classification of Imbalanced Data Based on SMOTE Algorithm," *Scientific Reports*, vol. 11, no. 24039, 2021.
[18] D. L. Weller, T. M. T. Love, and M. Wiedmann, "Comparison of Resampling Algorithms to Address Class Imbalance when Developing Machine Learning Models," *Frontiers in Environmental Science*, vol. 9, p. 701288, 2021.
[19] E. Ismail, W. Gad, and M. Hashem, "A Hybrid Stacking-SMOTE Model for Optimizing the Prediction of Autistic Genes," *BMC Bioinformatics*, vol. 24, no. 379, 2023.
[20] S. Fayz, M. A. Rizka, and F. Maghraby, "Cervical Cancer Diagnosis Using Random Forest Classifier with SMOTE and Feature Reduction Techniques," *IEEE Access*, vol. 6, pp. 59475–59485, 2018.
[21] C. Zhang, C. Liu, X. Zhang, and G. Almpanidis, "An up-to-date comparison of state-of-the-art classification algorithms," *Expert Systems with Applications*, vol. 82, pp. 128–150, 2017.
[22] J. Hayton, "Predictive Modeling Based on Random Forests. In Predictive Modeling of Drug Sensitivity," *Ranadip: Academic Press*, 2017.
[23] A. E. Hastie, R. Tibshirani, and J. Friedman, *The Element of Statistical Learning: Data Mining, Inference and Prediction*, 2nd ed. Springer, 2009.
[24] G. Biau and E. Scornet, "A random forest guided tour," *TEST*, vol. 25, no. 2, pp. 197–227, 2016.
[25] A. Cutler, D. R. Cutler, and J. R. Stevens, "Random Forest," in *Ensemble Machine Learning: Methods and Applications*, New York: Springer, 2012, pp. 157–175.
[26] P. Sandri, M., & Zuccolotto, "Variable Selection Using Random Forests. In Data Analysis, Classification and the Forward Search," *Springer*, 2006.
[27] F. Gorunescu, *Data Mining: Concepts, Models and Techniques*. Berlin Heidelberg: Springer-Verlag, 2011.
[28] J. Zhao, P., Su, X., Ge, T., & Fan, "Propensity Score and Proximity Matching Using Random Forest," *Contemporary Clinical Trials*, 2016.
[29] P. Probst, M. Wright, and A.-L. Boulesteix, "Hyperparameters and Tuning Strategies for Random Forest," *arXiv preprint arXiv:1804.03515*, 2018.
[30] A. Gholamy, V. Kreinovich, and O. Kosheleva, "Why 70/30 or 80/20 Relation Between Training and Testing Sets: A Pedagogical Explanation," El Paso, 2018.
[31] L. O. F. and N. Agustina, "Analisis Faktor-Faktor yang Memengaruhi Status Kemiskinan Ekstrem," *Seminar Nasional Official Statistics*, 2023.
[32] Suryana and K. Swarniati, "Eradicating Poverty And Human Capital Development In Indonesia: An Approach with Multilevel Logistic Regression Model," *Welfare: Jurnal Ilmu Kesejahteraan Sosial*, vol. 10, no. 2, pp. 107–121, 2021.