Media Statistika

p-ISSN 1979 – 3693 e-ISSN 2477 – 0647
MEDIA STATISTIKA 16(2) 2023: 112-123
http://ejournal.undip.ac.id/index.php/media_statistika

# COMPARISON OF LOGISTIC MODEL TREE AND RANDOM FOREST ON CLASSIFICATION FOR POVERTY IN INDONESIA

**Sukarna[1,2], Khairil Anwar Notodiputro[2], Bagus Sartono[2]**
[1] Departement of Mathematics, Universitas Negeri Makassar
[2] Department of Statistics, IPB University

**e-mail**: khairil@apps.ipb.ac.id

**DOI: 10.14710/medstat.16.2.112-123**

**Article Info:**
Received: 2 June 2023
Accepted: 17 December 2023
Available Online: 22 December 2023

**Keywords:**
*Decision Tree; Logistic Model Tree; Random Forest; Poverty; Model Classification*
.

**Abstract:** Classification methods are commonly employed to ensure homogeneous data within each group, facilitating the prediction of specific categories. The most frequently used classification models are Logistic Model Tree (LMT) and Random Forest (RF). This study aims to assess the accuracy rate in predicting the poverty status of regencies or towns across Indonesia, utilizing eight independent variables. The entire dataset was obtained from the official Central Bureau of Statistics website. The study investigates the accuracy of various iterations and combinations of training data. The results indicate that RF outperforms LMT in terms of accuracy, achieving a 100% improvement in iterations k=10 and k=500 and a 75% improvement in iteration k=100. Consequently, the RF proves to be more effective than the LMT for analyzing Indonesian poverty data, especially when incorporating all eight independent variables.

## 1. INTRODUCTION

Classification procedures may be accomplished using parametric and nonparametric approaches (Waluyo et al., 2014). One well-known parametric technique is logistic regression analysis. Logistic regression (Hosmer et al., 2013) combines a categorical response variable with independent variables that might be interval or categorical. Classification and Regression Trees (CART) (Breiman et al., 1984) are a popular nonparametric classification approach. CART (Breiman et al., 1984) may have a categorical response variable, resulting in classification trees, or a continuous response variable, leading to regression trees. The CART classification technique consists of four major parts in the analysis (Waluyo et al., 2014): (1) the tree-building process, (2) pausing the tree-building process, (3) trimming the tree, and (4) computing the tree's accuracy and interpretation.

The fundamental advantage of decision trees (Sari, 2021; Sartono & Dharmawan, 2023; Tan et al., 2019) is their capacity to simplify complex decision-making procedures. This advantage presents decision trees as a method that facilitates understanding in classification modeling. Several decision tree-based approaches are extensively used in machine learning for classification and regression. There are some popular tree-building methods or algorithms (Breiman et al., 1984; Chen et al., 2017; Prasetya & Abdulrakhman, 2022; Priyam et al., 2013; Sartono & Dharmawan, 2023; Tan et al., 2019; Waluyo et al., 2014), such as (1) CART, (2) Iterative Dichotomiser 3 (ID3), (3) an extension of ID3 (C4.5),

(4) Chi-squared Automatic Interaction Detector (CHAID), (5) Multivariate Adaptive Regression Splines (MARS), (6) Random Forest (RF), (7) Logistic Model Trees (LMT), and (8) Gradient Boosted Trees (GBT). This study focuses on two primary decision tree methods, namely LMT and RF.

LMT and RF use ensemble learning approaches, combining many models to improve prediction performance. They can handle overfitting scenarios successfully, analyze vast and complicated data, and are unaffected by multicollinearity among independent variables. However, these two approaches vary because LMT is based on logistic regression, whereas RF is based on decision trees. LMT generates a single tree that uses feature relationships, whereas RF constructs separate trees. LMT is less complicated than RF and is thought to be more stable.

Many researchers have studied LMT and RF, namely (1) Chen et al. (2017) in forecasting landslide susceptibility claim that RF outperforms LMT and CART. (2) In developing Digital Marketing, Gao & Ding (2022) advocate RF as particularly successful. (3) Afrianto & Wasesa (2020) discovered that RF is superior in prediction models for peer-to-peer accommodation. (4) Mukodimah & Fauzi (2021) claim that Random Forest is the best classifier for recognizing Iris species. (5) Mohammadi et al. (2022) established that LMT gives the most optimal performance in nanofluids solutions based on viscosity value. (6) Nhu et al. (2020a) in detecting landslides and vulnerability in Cameron Highlands, Malaysia, recommending LMT as the best model. (7) Sari (2021) determined that RF outperforms other methods when modeling data with significant multicollinearity, a large number of observations, and a high percentage of missing data; however, LMT with missing data elimination performed better for data with strong multicollinearity. (8) Nhu et al. (2020b) found that the LMT technique is the most accurate when used in shallow landslide susceptibility mapping. Four of the eight studies concluded that LMT is better, while the other four supported RF. As a result, the performance of these approaches is still dependent on the particular data applied. As a result, this study aims to evaluate both methods in the context of poverty in Indonesia, considering eight independent factors.

Indonesia has the most islands in the world, with around 17,500 islands. It has a highly populated population of roughly 273.52 million people as of January 31, 2023, with a land size of 1.91 million square kilometers and a coastline length of 81,000 kilometers (Finaka, 2018). This enormous population provides problems, and benefits, and may sometimes be a burden. Figure 1 depicts the population distribution in Indonesia by province.

Poverty is a challenge for a country with a considerable population. The number of impoverished people in Indonesia increased (in thousands) from 2017 to 2022 (Figure 2). The BPS-Statistics Indonesia published a report on the number of people living in poverty in Indonesia in 2022 (Figure 2), which totaled 26.16 million. In 2021, the poverty rate decreased from 5.3% to 27.54 million. However, the number of people living in poverty has increased steadily from 25.14 million in 2019 to 27.54 million in 2021, representing an 8.7% rise over the two years (2019-2021). In 2022, the number of impoverished individuals reduced by roughly 5.3% compared to 2021. As a result, the estimated poverty rate in Indonesia for 2022 is about 9.56%.

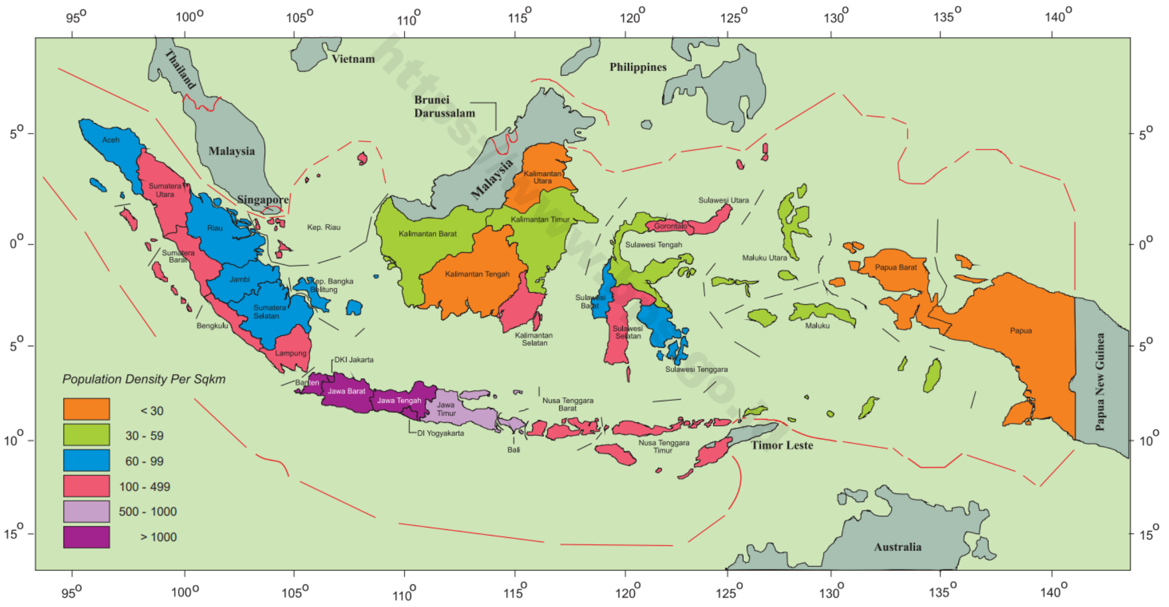


**Figure 1**. Geographical Conditions and Population Density in Indonesia 2021 (BPS, 2022b)

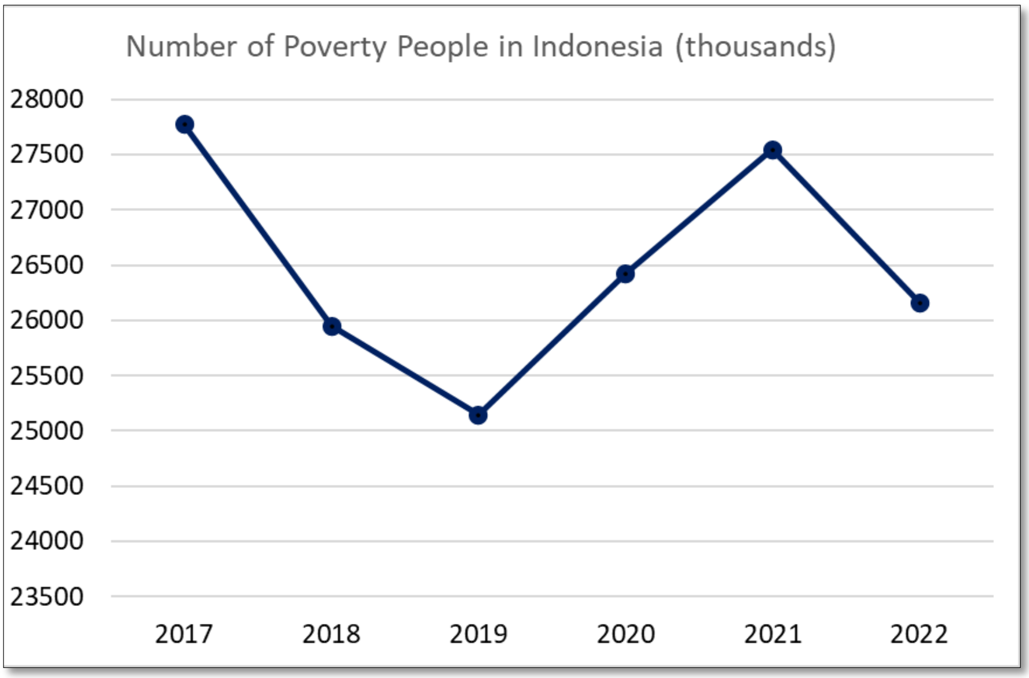


**Figure 2.** The Number of Impoverished People in Indonesia 2017 – 2022 (https://www.bps.go.id/en)

This study aims to provide a straightforward answer by comparing two prominent models for defining the elements that cause poverty in Indonesia. LMT and RF are two common approaches under consideration. Several factors have been investigated and are thought to be connected to the number of people living in poverty, including (1) the Gender Development Index (GDI), (2) Women's Income Contribution (Percentage), (3) Adjusted Per Capita Expenditure (Thousand Rupiah/Person/Year), (4) Male Life Expectancy, (5) Female Life Expectancy, (6) Human Development Index (HDI), (7) Male Expected Years of Schooling, (8) Female Expected Years of Schooling.

## 2. LITERATURE REVIEW

### 2.1. Decision Tree

Decision Trees (DTs), a nonparametric classification tool, are similar to Regression intends (Breiman, 2021; Sari, 2021) in that they learn basic decision rules derived from data attributes to form a model or prediction. A tree resembles a piecewise function in terms of constant approximation. There are several positive aspects of applying decision trees, involving the ability to handle both numerical and categorical data, requiring little data preparation and often data normalization, requiring dummy variables to be created and blank values to be removed.

### 2.2. Random Forest (RF)

The Random Forest classification approach mixes numerous independent decision trees (Prasetya & Abdulrakhman, 2022). A voting mechanism determines the categorization with the most votes from the individual trees. Random forests are an extension of Breiman's ensemble approach (Breiman, 2021) that aims to improve classification accuracy. Breiman (Breiman, 2021) presented Random Forest as a practical ensemble learning approach. It may be used for classification, Regression, and unsupervised learning (Liaw & Wiener, 2002) and has been widely employed with outstanding results in various disciplines (Chen et al., 2014).

### 2.3. Logistic Model Tree (LMT)

The standard form of binary logistic regression is

$$P(Y=1|x)=\frac{\exp\left(\beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_pX_p\right)}{1+\exp\left(\beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_pX_p\right)} \quad (1)$$

where Y is the response variable, and X represents the independent variables (Sartono & Dharmawan, 2023). The categorization of observation into a given class is decided by comparing the value of $P(Y=1|x)$ with a threshold value (commonly referred to as the threshold value) (Sartono & Dharmawan, 2023).

The LMT classification model combines decision tree learning techniques and Logistic Regression (LR) (W. Chen et al., 2017). In the logistic variation, information gain is utilized for splitting, the LogitBoost process is used to generate an LR model at each node in the tree, and the tree is trimmed using the CART algorithm (Breiman et al., 1984). To avoid training data overfitting, the LMT uses cross-validation to determine the number of LogitBoost iterations. The LogitBoost technique use additive LR of least-squares fits for each class $M_i$ (Doetsch et al., 2009):

$$L_M(x)=\beta_0+\sum_{i=1}^{n}\beta_ix_i \quad (2)$$

where $\beta_i$ denotes the coefficient of the i-th component of vector $\boldsymbol{x}$ and $n$ represent the number of samples.

In the LMT model, the linear LR approach is utilized to determine the posterior probability of leaf nodes (Doetsch et al., 2009; Tien Bui et al., 2016):

$$P(M|x) = \frac{\exp(L_M(x))}{\sum_{M=1}^{D} \exp(L_M(x))} \quad (3)$$

where $D$ is the number of classes.

### 2.4. Model Assessment

Classification is often evaluated using a test data set of a particular size not utilized in the classification's training. The classification model may be assessed or evaluated using a variety of metrics, including accuracy, error rate, recall/sensitivity/true positive rate, specificity/true negative rate, and precision (Prasetya & Abdulrakhman, 2022; Sartono & Dharmawan, 2023).

**Table 1.** Confusion Matrix

| | Prediction: Yes | Prediction: No | Total |
|---|---|---|---|
| Actual: Yes | True Yes (TY) | False No (FN) | Actual Yes (AY) |
| Actual: No | False Yes (FY) | True No (TN) | Actual No (AN) |
| Total | Prediction Yes (PY) | Prediction No (PN) | Total (T) |

Some of the model assessment measurements (Sartono & Dharmawan, 2023) are:

1. Accuracy: Part of an instance that is correctly classified

$$Accuracy = \frac{TY + TN}{T}$$

2. Error rate: part of an instance that is misclassified

$$Error\ rate = \frac{FY + FN}{T}$$

3. Sensitivity/Recall: the percentage of positive instances that are correctly classified.

$$Sensitifity = Recall = \frac{TP}{AY}$$

4. Specificity: the percentage of negative instances that are correctly classified.

$$Specificity = \frac{TN}{AN}$$

5. Precision: the proportion of the relevant outcomes

$$Precision = \frac{TY}{PY}$$

## 3. MATERIAL AND METHOD

### 3.1. Material

All the data used in this study were collected online from BPS's official website, www.bps.go.id. The dependent data is the status of the rise in the number of individuals living in poverty in each district/city from 2019 to 2022. Several factors that impact the number of individuals living in poverty have been investigated, namely: (1) the GDI (as $X_1$), (2) Women's Income Contribution (as $X_2$), (3) Adjusted Per Capita Expenditure (as $X_3$), (4) Male Life Expectancy (as $X_4$), (5) Female Life Expectancy (as $X_5$), (6) the HDI (as $X_6$),

(7) Male Expected Years of Schooling (as $X_7$), (8) Female Expected Years of Schooling (as $X_8$).

The $X_5$ is frequently used as an indicator to assess a region's development in the field of health (BPS, 2022a, 2022b). The $X_6$ is an indicator that depicts people's possibilities to access products as part of their right to income, health, education, and levels of expenditure and consumption to reach a higher quality of life (BPS, 2022a, 2022b). According to the $X_1$, women are involved in paid labor and contribute to domains such as economics, politics, and decision-making processes (Sekjend Kemenkes RI, 2012).

One of the primary assumptions of regression analysis is the presence of multicollinearity, which ensures that there is no correlation among the independent variables. Figure 3 illustrates the bivariate correlation among the independent variables.

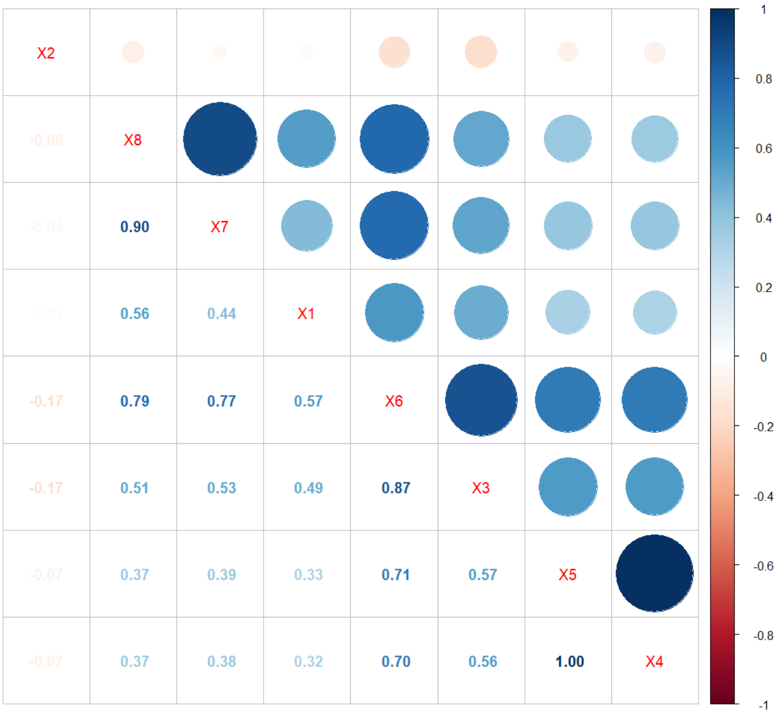


**Figure 3.** The Correlation between Variables

Fortunately, one of lucky the benefits for model trees or random forest is ridge from multicollinearity assumptions.

Figure 4 shows boxplot of all variables affecting poverty status. Several independent variables show significant correlations, indicating the existence of multicollinearity. The variables with the highest correlation are $X_4$ and $X_5$ ($r$ = 1.00), $X_7$ and $X_8$ ($r$ = 0.90), $X_3$ and $X_6$ ($r$ = 0.87), $X_6$ and $X_8$ ($r$ = 0.79), $X_6$ and $X_7$ ($r$ = 0.77), $X_5$ and $X_6$ ($r$ = 0.71), $X_4$ and $X_6$ ($r$ = 0.70). However, the benefit of LMT and RF is their capacity to overlook multicollinearity. As a result, the analysis will proceed while ignoring this multicollinearity.

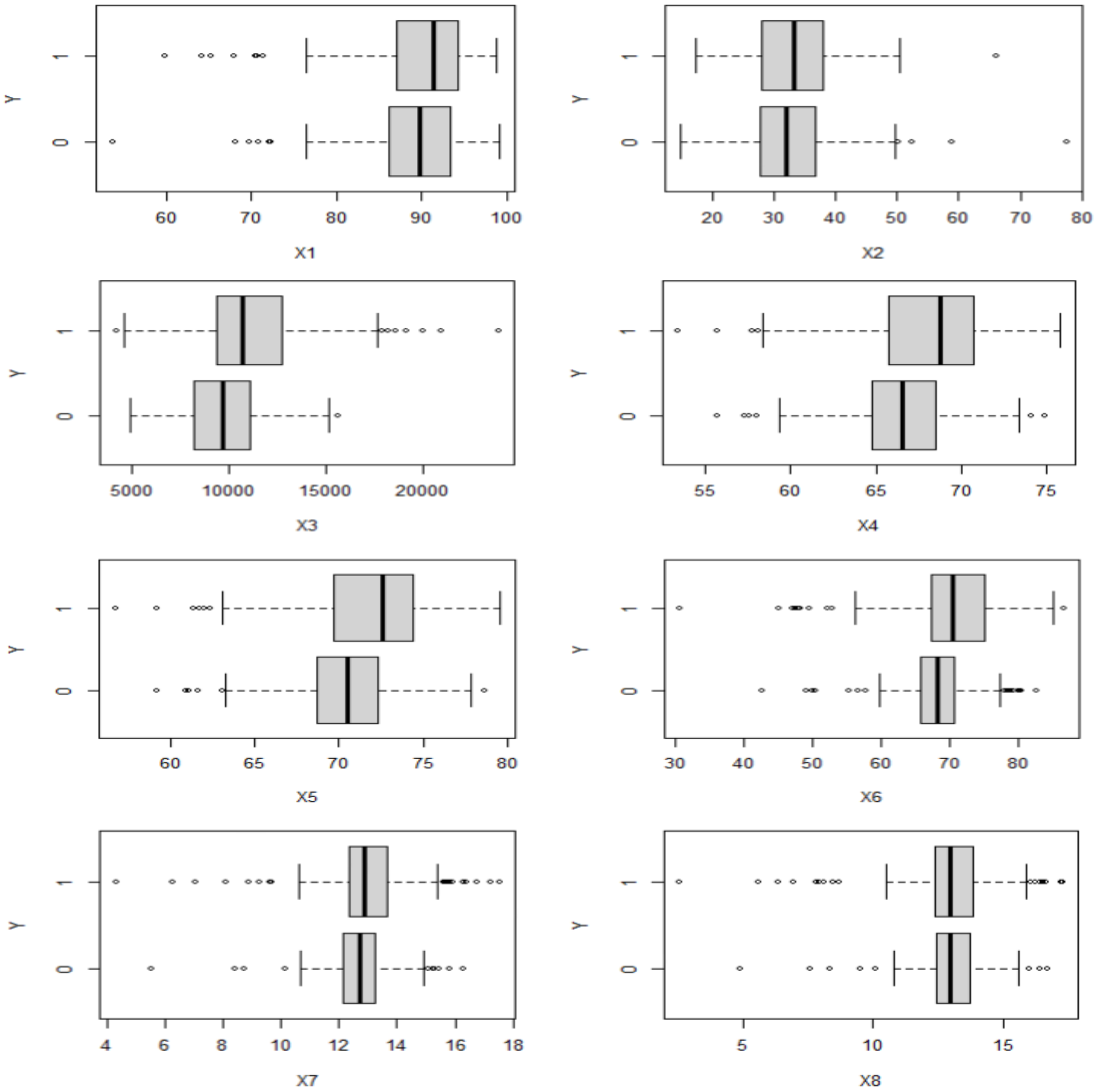


**Figure 4**. A Boxplot of All Variables Affecting Poverty Status

### 3.2. Method

LMT and RF analysis steps are as follows:

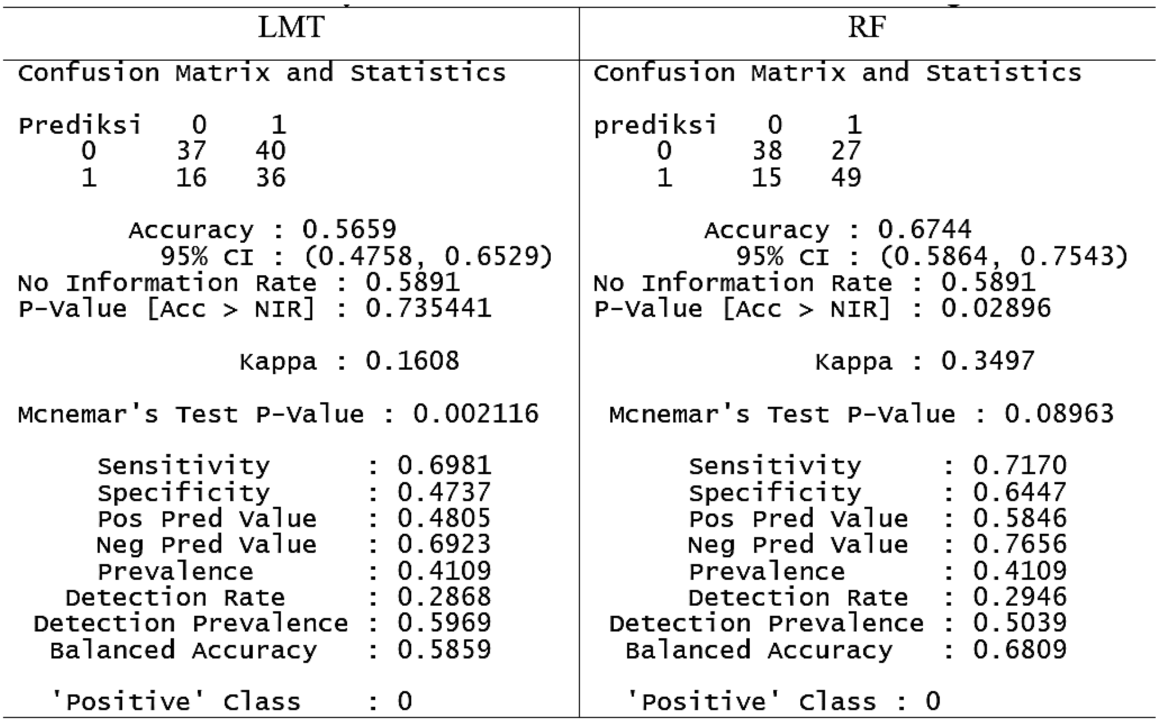
a. Data review. Investigate the data by discovering missing data for each regency or city, filtering data to ensure data appropriateness (such as outliers) and inspecting the data format in appropriate forms.
b. Partitioning of data. Divide the data into two parts: training and testing data. Training data is used to develop the LMT and RF models, and it is split into 70%, 80%, or 90%. The remainder of the data is testing data for model assessment.
c. Model construction. Use the default settings to build the LMT and RF classification models.
d. Determining evaluation metrics. The first step is to compute the prediction rate based on the assessment model value. The accuracy values for the LMT and RF models are determined in the second phase.
e. Selecting the best model. Determine the optimum model based on the average accuracy values acquired from different iterations and changes in training data volume.
f. Interpreting the outcomes. Comparing the accuracy model of each portion of the combinations between each iteration and various quantities of training data.

## 4. RESULTS AND DISCUSSION

### 4.1. Results

### 4.1.1.Sample Result

Figure 5 compares LMT and RF on programs run result just once. Figure 5 highlights that the LMT findings were able to estimate the feasibility of predicting 37 respondents who were consistent in their answers "NO" and 36 for "YES". The RF model, on the other hand, is compatible with 38 "NO" and 49 "YES". These findings imply the RF model predicts categorization more consistently than the LMT model. In keeping with this, the LMT accuracy number of 0.5659 is mathematically less than the RF accuracy value of 0.6744. As a result of these outcomes, RF surpasses LMT in predicting poverty status in Indonesia. Is this true after 10, 50, 100, or 500 repetitions? The true answer can be found in the following results.



LMT

```
Confusion Matrix and Statistics

Prediksi   0    1
    0     37   40
    1     16   36

       Accuracy : 0.5659
         95% CI : (0.4758, 0.6529)
No Information Rate : 0.5891
P-Value [Acc > NIR] : 0.735441

              Kappa : 0.1608

Mcnemar's Test P-Value : 0.002116

          Sensitivity      : 0.6981
          Specificity      : 0.4737
          Pos Pred Value   : 0.4805
          Neg Pred Value   : 0.6923
          Prevalence       : 0.4109
        Detection Rate     : 0.2868
  Detection Prevalence : 0.5969
    Balanced Accuracy      : 0.5859

    'Positive' Class    : 0
```

RF

```
Confusion Matrix and Statistics

prediksi   0    1
    0     38   27
    1     15   49

       Accuracy : 0.6744
         95% CI : (0.5864, 0.7543)
No Information Rate : 0.5891
P-Value [Acc > NIR] : 0.02896

              Kappa : 0.3497

 Mcnemar's Test P-Value : 0.08963

        Sensitivity     : 0.7170
        Specificity     : 0.6447
        Pos Pred Value  : 0.5846
        Neg Pred Value  : 0.7656
        Prevalence      : 0.4109
        Detection Rate  : 0.2946
 Detection Prevalence : 0.5039
  Balanced Accuracy     : 0.6809

  'Positive' Class : 0
```

**Figure 5**. The Analysis Results of The LMT and RF Models for A Single Run

### 4.1.2.Assessment with Controlling Training Data

This first investigation illustrates the importance of iteration and the determination of the data training presentation in evaluating the model. Errors in establishing these two points can lead to erroneous findings.

This article includes a preliminary example of using "set.seed(123)" in the R syntax used in this initial study. This syntax, "set.seed," restricts the randomization range. The program uses three separate data training presentations, namely 70%, 80%, and 90%, with an extra expression of 82% included as a unique instance.

**Table 1**. The Accuracy Results of the LMT and RF Models in The "set.seed(123)" Space

| Training data (Percentages) | 70% | 80% | 90% | 82% |
|---|---|---|---|---|
| LMT Accuracy value | 0.6580645 | 0.6407767 | 0.6923077 | 0.655914 |
| RF Accuracy value | 0.6516129 | 0.6699029 | 0.6538462 | 0.655914 |
| Conclusion | LMT | RF | LMT | LMT = RF |

Table 1 compares the accuracy values obtained using LMT and RF and finishes with some observations, such as (1) When 80% of the training data is chosen, RF (accuracy = 0.6699029) surpasses LMT (accuracy = 0.6407767), (2) When 90% of the training data is chosen, on the contrary, LMT (accuracy = 0.6923077) outperforms RF (accuracy = 0.6538462), (3) Similarly to the previous resume, when 70% of the training data is chosen, LMT (accuracy = 0.6580645) outperforms RF (accuracy = 0.6516129), dan (4) Both LMT and RF attain the same accuracy of 0.655914 while training the data with an 82% percentage.

In summary, it can be inferred that there are no definitive conclusions regarding the differences between LMT and RF models, as their performance heavily relies on the amount of training data provided. To mitigate this dependency, it is advisable to employ improved iteration and randomization strategies when selecting the best model.

### 4.1.3.Result

Table 2 shows how the average value of the accuracy of the LMT and RF models varies with the number of iterations and amount of training data.

**Table 2.** The Average Accuracy Value At Various Iterations And The Quantity Of Training

| Iteration | 10 | | 50 | | 100 | | 500 | |
|---|---|---|---|---|---|---|---|---|
| Prob. data | LMT | **RF** | **LMT** | RF | LMT | **RF** | LMT | RF |
| 0.7 | 0.6226 | **0.6265** | 0.5983 | **0.6244** | 0.6159 | **0.6261** | 0.6137 | **0.6223** |
| 0.8 | 0.6263 | **0.6379** | **0.6237** | 0.6204 | **0.6327** | 0.6262 | 0.6245 | **0.6303** |
| 0.9 | 0.6250 | **0.6365** | **0.6392** | 0.6362 | 0.6277 | **0.6281** | 0.6296 | **0.6312** |

The average accuracy values for LMT on 70% training data are 0.6226 in the first type of iterations (k=10), whereas RF obtains an accuracy of 0.6265. This result means that the average accuracy of RF is 0.0039 (=0.6265-0.6226), more than that of LMT. In the 80% training data, the average accuracy of LMT is 0.6263, whereas RF achieves an accuracy of 0.6379, resulting in a 0.0116 difference. As a result, RF is roughly 1.16% more accurate than LMT. A similar pattern can be seen with 90% of training data, where RF beats LMT with about 1.15% accuracy. As a result, throughout ten iterations, the RF model outperforms the control model by around 1%.

The comparison of LMT and RF provides various outcomes for total iterations at $k$ =50 and $k$ =100. In the case of 50 iterations, the LMT model outperforms the RF model in 80% and 90% of the training data, respectively, but RF outperforms LMT only in 70% of the cases. For 100 iterations, however, RF outperforms LMT on 70% and 90% of training data, respectively, whereas the LMT model is only better on 80%.

The findings are consistent at k=500 iterations, as they were at $k$ =10. The average accuracy numbers for RF are consistently more significant than those for LMT. Throughout 500 iterations, RF surpasses LMT in all three permutations of training data percentages.

### 4.2. Discussion

The findings of this study align with previous studies conducted by multiple experts (at least four experts) who investigated the performance of LMT or RF models. Four expert-led studies (Afrianto & Wasesa, 2020; Chen et al., 2017; Gao & Ding, 2022; Mukodimah & Fauzi, 2021) found that RF outperforms LMT in predicting case categorization. However, there have been instances where LMT has been found to perform better (Mohammadi et al., 2022; Nhu et al., 2020a; Nhu et al., 2020b; Sari, 2021). These inconsistencies suggest that the superiority of LMT or RF models depends on specific circumstances or cases.

The findings of this study add considerably to the current amount of knowledge regarding the influence of specific data applications on the performance of LMT or RF models. Four experts found RF superior in the circumstances above, applying it to particular conditions such as landslide susceptibility (Chen et al., 2017), digital marketing (Gao & Ding, 2022), peer-to-peer accommodation (Afrianto & Wasesa, 2020), and Iris species (Mukodimah & Fauzi, 2021). In the current study, RF performance outperforms LMT in the context of Indonesian poverty.

The variables contributing to discrepancies in model evaluation findings are a fascinating field for future research. This study suggests three primary potential causes: the type of cases, the number of iterations, and the probability of training data.

## 5. CONCLUSION

The RF model is the best developed for the poverty situation, based on the average accuracy values for 10 and 500 iterations. However, for 50 iterations, LMT outperforms RF, whereas RF outperforms LMT for 100 iterations. As a result, RF outperforms LMT in three iteration circumstances ($k$=10, 100, and 500), but LMT surpasses it just in iteration $k$ =50. As a result, in this scenario, RF outperforms LMT. When considering eight independent variables, such as (1) the GDI (as $X_1$), (2) Women's Income Contribution (as $X_2$), (3) Adjusted Per Capita Expenditure (as $X_3$), (4) Male Life Expectancy (as $X_4$), (5) Female Life Expectancy (as $X_5$), (6) the HDI (as $X_6$), (7) Male Expected Years of Schooling (as $X_7$), (8) Female Expected Years of Schooling (as $X_8$) in the instance of poverty in Indonesia, the recommended model to use is Random Forest instead of Logistic Model Tree.

Future research should focus on the elements that influence the performance of LMT and RF models. This research will serve as the foundation for developing models for real applications.

## REFERENCES

Afrianto, M. A. & Wasesa, M. (2020). Booking Prediction Models for Peer-to-peer Accommodation Listings using Logistics Regression, Decision Tree, K- Nearest Neighbor, and Random Forest Classifiers. *Journal of Information Systems Engineering and Business Intelligence*, *6*(2), 123–132. http://e-journal.unair.ac.id/index.php/JISEBI

BPS. (2022a). Kabupaten Gowa Dalam Angka 2022. In *Kabupaten Gowa dalam Angka*. BPS Kabupaten Gowa.

BPS. (2022b). Statistik Indonesia (Statistical Yearbook of Indonesia) 2022. In *Statistik Indonesia* (p. 790). BPS-Statistics Indonesia.

Breiman, L. (2021). Random Forest. *Machine Learning*, *45*, 5–32. https://doi.org/10.1023/A:1010933404324

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). Classification and Regression Trees. *Biometrics*, *40*(3).

Chen, W., Xie, X., Wang, J., Pradhan, B., Hong, H., Bui, D. T., Duan, Z., & Ma, J. (2017). A Comparative Study of Logistic Model Tree, Random Forest, and Classification and Regression Tree Models for Spatial Prediction of Landslide Susceptibility. *Catena*, *151*, 147–160. https://doi.org/10.1016/j.catena.2016.11.032

Chen, Y., Naud, C. M., Rangwala, I., Landry, C. C., & Miller, J. R. (2014). Comparison of The Sensitivity of Surface Downward Longwave Radiation to Changes in Water Vapor at Two High Elevation Sites. *Environmental Research Letters*, *9*(11). https://doi.org/10.1088/1748-9326/9/11/114015

Doetsch, P., Buck, C., Golik, P., Hoppe, N., Kramp, M., Laudenberg, J., Oberdorfer, C., Steingrube, P., Forster, J., & Mauser, A. (2009). Logistic Model Trees with AUC Split Criterion for the KDD Cup 2009 Small Challenge. *Journal on Machine Learning Research: Workshop and Conference Proceedings*, *7*, 77–88. http://jmlr.csail.mit.edu/proceedings/

Finaka, A. W. (2018). *Indonesia Kaya Potensi Kelautan dan Perikanan*. Indonesiabaik.Id. https://indonesiabaik.id/infografis

Gao, W., & Ding, Z. (2022). Construction of Digital Marketing Recommendation Model Based on Random Forest Algorithm. *Security and Communication Networks*, *2022*. https://doi.org/10.1155/2022/1871060

Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied Logistic Regression*. John Wiley & Sons, Inc. https://doi.org/10.1002/9781118548387

Liaw, A., & Wiener, M. (2002). Classification and Regression by Random Forest. *R News*, *2*(3), 18–22.

Mohammadi, M., Khorrami, M. K., & Ghasemzadeh, H. (2022). Classification of Nanofluids Solutions Based on Viscosity Values: A Comparative Study of Random Forest, Logistic Model Tree, Bayesian Network, and Support Nector Machine Models. *Infrared Physics & Technology*, *125*, 104273. https://doi.org/10.1016/j.infrared.2022.104273

Mukodimah, S., & Fauzi, C. (2021). Comparison of Tree Implementation, Regression Logistics, and Random Forest To Detect Iris Types. *Jurnal TAM (Technology Acceptance Model)*, *12*(2), 149–157. https://doi.org/10.56327/jurnaltam.v12i2.1074

Nhu, V. H., Mohammadi, A., Shahabi, H., Ahmad, B. Bin, Al-Ansari, N., Shirzadi, A., Geertsema, M., Kress, V. R., Karimzadeh, S., Kamran, K. V., Chen, W., & Nguyen, H. (2020a). Landslide Detection and Susceptibility Modeling on Cameron Highlands (Malaysia): A Comparison between Random Forest, Logistic Regression and Logistic Model Tree Algorithms. *Forests*, *11*(8). https://doi.org/10.3390/F11080830

Nhu, V., Shirzadi, A., Shahabi, H., Singh, S. K., Al-Ansari, N., Clague, J. J., Jaafari, A., Chen, W., Miraki, S., Dou, J., Luu, C., Gorski, K., Pham, B. T., Nguyen, H. D., & Ahmad, B. Bin. (2020b). Shallow Landslide Susceptibility Mapping: A Comparison between Logistic Model Tree, Logistic Regression, Naïve Bayes Tree, Artificial Neural Network, and Support Vector Machine Algorithms. *International Journal of Environmental Research and Public Health*, *17*(2749), 1–30. https://doi.org/10.3390/ijerph17082749

Prasetya, J., & Abdulrakhman. (2022). Comparison of Smote Random Forest and Smote k-Nearest Neighbors Classification Analysis on Imbalanced Data. *Media Statistika*, *15*(2), 198–208. https://doi.org/10.14710/medstat.15.2.198-208

Priyam, A., Abhijeet, Gupta, R., Rathee, A., & Srivastava, S. (2013). Comparative Analysis of Decision Tree Classification Algorithms. *International Journal of Current*

*Engineering and Technology*, *3*(2), 334–337. http://inpressco.com/category/ijce

Sari, P. (2021). *Perbandingan Performa Metode Pohon Model Logistik dan Random Forest pada Pengklasifikasian Data* [IPB]. https://doi.org/10.29244/xplore.v12i1.858

Sartono, B., & Dharmawan, H. (2023). *Pemodelan Prediksi Berbasis POHON Klasifikasi*. PT Penerbit IPB Press.

Sekjend Kemenkes RI. (2012). *Profil Kesehatan Indonesia 2012*.

Tan, P.-N., Steinbach, M., Karpatne, A., & Kumar, V. (2019). *Introduction to Data Mining* (Second). Pearson Education, Inc.

Tien Bui, D., Pham, B. T., Nguyen, Q. P., & Hoang, N. D. (2016). Spatial Prediction of Rainfall-Induced Shallow Landslides using Hybrid Integration Approach of Least-Squares Support Vector Machines and Differential Evolution Optimization: a Case Study in Central Vietnam. *International Journal of Digital Earth*, *9*(11), 1077–1097. https://doi.org/10.1080/17538947.2016.1169561

Waluyo, A., Mukid, M. A., & Wuryandari, T. (2014). Perbandingan Klasifikasi Nasabah Kredit Menggunakan Regresi Logistik Biner dan CART (Classification and Regression Trees). *Media Statistika*, *7*(2), 95–104.