JURNAL MATEMATIKA, STATISTIKA DAN KOMPUTASI e-ISSN: 2614-8811
Published by Departement of Mathematics, Hasanuddin University, Indonesia
https://journal.unhas.ac.id/index.php/jmsk/index
Vol. 22, No. 1, September 2025, pp. 90-101
DOI: 10.20956/j.v22i1.44488

Comparison of Basic Statistics and Machine Learning Classification Algorithms in Kalimantan Poverty Prediction with Handling Missing Data

Khusnia Nurul Khikmah[1,*], A'yunin Sofro[2]
[1]*Mathematics Study Program, Faculty of Mathematics and Natural Sciences, Universitas Palangka Raya, Palangka Raya, Central Kalimantan, 74874 Indonesia*
[2]*Actuarial Sciences Study Program, Faculty of Mathematics and Natural Sciences, Universitas Negeri Surabaya, Surabaya, East Java, 60231 Indonesia*
***Email**: khusnia.nurulkhikmah@mipa.upr.ac.id[1,*] and ayuninsofro@unesa.ac.id[2]*
**Corresponding Author*

Abstract

Poverty is a crucial development challenge in Indonesia, including in regencies/cities in Kalimantan that require more attention. In reality, poverty is influenced by various factors. Therefore, this research proposes an analysis comparing the accuracy of basic and statistical machine learning models in predicting poverty rates and finding factors that affect poverty rates. The advance of this research is the performance comparison combined with the handling of missing data. The three models proposed in this study are binary logistic regression with backward stepwise selection, random forest, and extremely randomized trees (extra trees). The data used in this study is secondary data taken from the Indonesian Statistics (BPS) of five provinces in Kalimantan, where the pre-processing is done by handling missing data with a k-nearest neighbor (KNN). The results of the poverty prediction analysis show that the binary logistic regression model is the most accurate compared to random forest and extra trees, with a balanced accuracy of 75%. In addition, based on the best model with the highest accuracy, this study also found significant predictor variables that affect the poverty rate of regencies/cities in Kalimantan: population density, average years of schooling, and per capita expenditure on food.

Keywords: *Binary Logistic Regression. Extra Trees, Kalimantan, Poverty, Random Forest.*

1. INTRODUCTION

Multidimensional and complex issues that pose challenges to development in Indonesia, including in the Kalimantan, are related to poverty. These demographic and socio-economic differences cause disparities in development between regions due to limited access to infrastructure and uneven income distribution [18]. These factors cause commodity price volatility to be high and drive poverty [10], [17]. Generally, poverty in Kalimantan is influenced by regional development disparities, especially between urban and rural areas. This contributes to income inequality and

This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License

poverty. According to previous studies, differences in development capacity, natural resource ownership, infrastructure, and geography can exacerbate this inequality [25]. Therefore, understanding the factors that significantly influence poverty status and developing accurate predictive models are crucial and fundamental for formulating effective and targeted policies. This study focuses on poverty modeling in regencies/cities in Kalimantan, which are unique in their demographic and socio-economic characteristics.

Fundamental statistical analysis for modeling poverty issues includes parametric statistical methods such as binary logistic regression. Generally, this model provides clear interpretations of the influence of each predictor factor on the probability of poverty levels based on their odds ratios [14]. Previous research conducted by [11] on population growth in Java demonstrated that binary logistic regression has good predictive capabilities for the data. Another study on [4] also showed that the logistic model could explain the variables causing the status of educated unemployment in the DKI Jakarta province in 2021. However, the assumption of log-odds linearity and the limited handling of non-linear relationships often pose challenges, especially in complex real-world data with interactions between variables. The ability to explain the relationship between predictor variables and response variables or poverty levels in regencies/cities in Kalimantan led to the selection of binary logistic regression in this study.

Furthermore, statistical analysis approaches are no longer limited to basic statistics; advancements in statistics, particularly those incorporating machine learning approaches, have developed rapidly. This study provides a new perspective on poverty prediction aimed at producing more accurate and robust predictions. This study selected two ensemble algorithms based on trees: random forests and extremely randomized trees (extra trees) [12]. Both methods proposed offer novel comparisons of methods related to poverty issues in regencies/cities in Kalimantan. Both machine learning methods offer solutions to the weaknesses of single trees. In random forest, this is achieved by building multiple trees from bootstrap samples and randomly selecting feature subsets at each node split, resulting in a robust model against overfitting [22].

Meanwhile, extra trees are proposed in this study with further consideration that in randomizing tree construction, the entire data set is used, and node splits are selected randomly [16]. As a result, the computations run faster, and random randomization enables this approach to reduce bias. Previous research in [12] comparing several machine learning methods showed that the random forest method is the best method for addressing food insecurity in West Java and outperforms the gradient boosting method. Another study in [1] showed that the extra trees method is also the best classification method with accuracy about 80.23% for addressing breast cancer issues.

Therefore, the comparative study of binary logistic regression, random forest, and extra trees methods, which is the core of this research, aims to provide insights into the results of poverty predictions for regencies/cities in Kalimantan by considering the accuracy of predictions and the ability of the three methods to handle data complexity. A comparison of evaluation metrics such as balanced accuracy, sensitivity, specificity, and accuracy is used in this study to identify the most effective method for modeling and predicting poverty in regencies/cities in Kalimantan while also identifying the demographic and socio-economic factors that contribute significantly. The results of the analysis are expected to provide evidence-based recommendations for policymakers in their efforts to alleviate poverty in regencies/cities in Kalimantan.

2. METHODS

2.1 Binary Logistic Regression

Binary logistic regression is one of statistical analysis technique that aims to model the relationship between a response variable with two categories (binary) with one and more predictor variables. This model is specifically used to analyze response variables that are categorical and binary. Suppose the logistic regression model has an intercept parameter β_0 and a slope parameter,

where $f(x)$ can be estimated using the maximum likelihood method for x as a quantitative and qualitative variable. In that case, mathematically, parameter estimation using the cumulative logistic function with a transformation of the logit $\pi(x)$ symbolized by $f(x)$ can be defined as follows [4], [11].

$$f(x) = ln\ ln\left[\frac{\pi(x)}{1-\pi(x)}\right] = ln\ ln\left[\frac{e^{\beta_0+\sum_{i=1}^{j}\beta_i x_i}}{\left(1-\frac{e^{\beta_0+\sum_{i=1}^{j}\beta_i x_i}}{1+e^{\beta_0+\sum_{i=1}^{j}\beta_i x_i}}\right)}\right] = \beta_0 + \sum_{i=1}^{j}\beta_i x_i \qquad 2.1$$

The interpretation of the model obtained in this binary logistic regression is through the odds ratio association measure [24], [26]. This measure is intended to compare the probability of an event happening with that of it not happening, where the odds ratio value is mathematically written as follows.

$$odds\ ratio = \frac{odds_1}{odds_2} = \frac{\left(\frac{\pi(1)}{1-\pi(1)}\right)}{\left(\frac{\pi(2)}{1-\pi(2)}\right)} \qquad 2.2$$

The binary logistic regression calculated demonstration was then subjected to hypothesis testing to show the model fit, simultaneously and partially. The model fit test selected in this study was the Hosmer-Lemeshow test, which aims to assess how the binary logistic regression model fits the observed data [8]. Where H_0 is that there's no noteworthy contrast between the watched recurrence of occasions and the anticipated recurrence of occasions, or the show fits the information, while the alternative hypothesis H_1 is that there's a noteworthy distinction between the watched recurrence of occasions and the anticipated recurrence of occasions, or the demonstrate does not fit the information. The model will reject H_0 if the $p - value \leq \alpha = 5\%$.

The simultaneous test (LRT) used in this study aims to simultaneously determine the role of all predictor variables on the response variable. The hypothesis in this simultaneous test is H_0: $\beta_0 = \beta_1 = \cdots = \beta_p$, or the predictor variables have no effect on the response variable. The alternative hypothesis is H_1: at least one $\beta_j \neq 0$, where $j = 0, 1, \ldots, p$, or at least one predictor variable affects the response variable. Mathematically, the test statistic can be written as follows [7].

$$G = -2\ ln\ ln\left[\frac{likelihood\ tanpa\ variabel\ prediktor}{likelihood\ dengan\ variabel\ prediktor}\right] \qquad 2.3$$

The partial test used in this study is the Wald test, which aims to determine which predictor variables influence the response variable. The initial hypothesis is H_0: $\beta_j = 0$ and the alternative hypothesis is H_1: $\beta_j \neq 0$ where $j = 0,1, \ldots, p$. The decision to reject H_0 is made if the $p - value < \alpha = 5\%$. Mathematically, the Wald test statistic is as follows [5].

$$W^2 = \left[\frac{\hat{\beta}}{SE_{\hat{\beta}}}\right]^2 \qquad 2.4$$

2.2 Model Evaluation

Model evaluation that used of this study is Akaike's information criterion (AIC) to measures model goodness in statistics to provide information on selecting the best model. In general, this measure aims to balance model goodness-of-fit with model complexity [2]. Model selection is based on AIC values, which are based on the smallest AIC value. For models with k estimated parameters, $\ln(L)$ is the maximum log-likelihood of the fitted model. Mathematically, the AIC equation is as follows [3], [19].

$$AIC = 2k - 2\ ln\ \ln(L) \qquad 2.5$$

2.3 Random Forest (RF)

Random forest is one of the ensemble algorithms in machine learning used for classification and regression. This algorithm is built using the concept of decision trees and bootstrap aggregating (bagging) techniques, which aim to produce more accurate classification models [21]. This algorithm is quite efficient, as it reduces variance and is robust against overfitting and outliers compared to a single decision tree. In general, the bagging technique in random forest selects n samples randomly from the training data, which then generates n new samples that are returned randomly to train the decision tree. Therefore, hyperparameter values are key in random forests. These hyperparameters are $mtry$ (the number of features considered at each node) and $ntree$ (the number of trees). The $mtry$ value is the square root of the number of predictor variables for classification. In general, the random forest algorithm works in the following stages [12].

1. Determine the number (N) of trees to be built in the forest.
2. For each tree where $n = \frac{1}{N}$, then:
   a. Create a bootstrap sample from the training data to be used as a subset of the training data for the tree.
   b. Build a decision tree on the bootstrap sample so that at each node of the tree:
   - Select (m) features randomly from the total (M) features where $(m < M)$.
   - Find the best-split value using the selected (m) features.
   Build the tree to its full depth.
3. Repeat step two for k trees.
4. The random forest prediction result is calculated using the majority vote of the classification results from N trees.

2.4 Extremely Randomized Trees (Extra Trees)

Extremely randomized trees or extra trees are one of the ensemble algorithms in machine learning, which uses a more extreme level of randomization in tree construction. This results in faster computational performance and competitive accuracy. In extra trees, each tree built is a tree from the entire training data (unlike random forests, which are samples from bootstrap). Therefore, extra trees utilize a strong randomization process in node separation to achieve tree diversity [4], [16], [20]. The removal of the bootstrapping process in this algorithm aims to reduce bias in modeling. As a result, the classification produced will be robust against overfitting. Randomization in extra trees is performed when selecting predictor variables and determining cut points to separate nodes—the best cut-point value results from evaluating the Gini coefficient and entropy of each variable value. As a result, model deviation can be minimized. The extremely randomized trees algorithm has the following analysis stages [12].

1. Determine the number of trees N_{tree} to be built.
2. Select the best separation:
   a. Randomly select m independent variables.
   b. Randomly select k cut points.

c. Determine the best separation criteria.
d. Repeat steps a to c until the stopping criteria are met so that the prediction results from one tree is obtained.

3. Repeat step 1 until M trees are formed.
4. Combine the estimation results of each classification tree using majority voting.

2.5 Matrix Evaluation

The accuracy of the classification can be measured by looking at the confusion matrix value, where balanced accuracy scores are used to measure model performance on unbalanced data. Mathematically, the balanced accuracy (BA) value has the following equation [11].

$$BA = \frac{sensitivity + specificity}{2} \quad 2.6$$

Where $sensitivity = \frac{TP}{TP+FN}$ and $specificity = \frac{TN}{TN+FP}$. These values are based on the following confusion matrix table.

Table 2.1. Confusion matrix

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

2.6 Handling Missing Values: K-Nearest Neighbor (KNN)

K-nearest neighbor is one of the factual strategies utilized to handle lost information. For the most part, lost information values are anticipated based on the values of the k-nearest perceptions [9], [15]. The calculation for ascribing lost information utilizing the k-nearest neighbor strategy takes after.

1. Set the esteem of k (the k-nearest observations to the lost information or data).
2. Calculate the Euclidean or the separate between perceptions containing lost information and those without lost information. For x_{ai}, $i = 1,2,\dots,n$ is the i-th variable containing lost information in each perception, while x_{bi}, $i = 1,2,3,\dots,n$ is the i-th variable without missing data. The Euclidean distance is numerically calculated utilizing the taking after condition.

$$d = \sqrt{\sum_{i}^{n} (x_{ai} - x_{bi})^2} \quad 2.7$$

3. Sort the distances based on the smallest esteem and determine the value of k-nearest perceptions.
4. Insert the lost information, with the correlation between the kth neighbor and the lost information being r_k and $e = 10^{-6}$. Then, mathematically, the value of the missing data (ω_k) is as follows.

$$\omega_k = \left(\frac{r_k^2}{(1 - r_k^2 + e)} \right)^2 \quad 2.8$$

2.7 Data and Analysis Stages

This research uses secondary data from the Indonesian Statistics (Badan Pusat Statistik) website of five provinces in Kalimantan. The data was from 56 regencies/cities in Kalimantan in 2024. In general, the analysis used seven predictor variables and one response variable.

Table 2.1. Data description used in the analysis

Symbol	Data	Explanation
Y	Poverty	0: Not poor (%poverty$<$ 10.8%) 1: Poor (%poverty$\geq$ 10.8%)
X_1	Total population	In thousand
X_2	Population growth rate	%
X_3	Percentage of population	%
X_4	Population density	/km^2
X_5	Years of schooling	In Year
X_6	Per capita expenditure on food	%
X_7	Working population	In thousand

This research begins by exploring the data with descriptive analysis, which aims to describe the characteristics of the research data. Furthermore, missing data was preprocessed. Next, inferential analysis will be carried out using three models, in which the data used has been divided into training data and test data, and missing data has been handled. Training data is data used to train the model, while test data is used to evaluate the model. The data splitting ratio uses the hold out, which has been widely used, namely dividing the data into two parts: training data 80% of the total data, and test data 20%. Before modeling, a correlation test was conducted to determine the relationship between each independent variable. The next step is modeling using binary logistic regression models, random forests, and extremely randomized trees with the value of hyperparameter show at Table 2.2. After modeling, parameter tests were carried out, including model fit, simultaneous, and partial tests. The best model selection is the model with the highest balanced accuracy value.

Table 2.2. Hyperparameter of Machine Learning Approach

Model	Hyperparameter	Value
Random Forest	mtry	2
	Node size	1
	Cross-validation	5
Extremely Randomized Trees	n-estimator size	100
	Max-depth	5
	Cross-validation	5

Following, inferential investigation will be carried out utilizing three models, in which the information utilized has been separated into preparing information and test information, and lost information has been dealt with. Preparing information is information utilized to prepare the demonstrate, whereas test information is utilized to assess the show. The information part proportion employments the Pareto Rule, which has been broadly utilized, to be specific isolating the information into two parts: preparing information, as much as 80% of the overall information, and test information, as much as 20%. Some time recently modeling, a relationship test was conducted to decide the relationship between each autonomous variable. The following step is modeling utilizing double calculated relapse models, arbitrary woodlands, and amazingly randomized trees. After modeling, parameter tests were carried out, counting demonstrate fit, synchronous, and halfway tests. The leading show choice is the show with the most elevated adjusted precision esteem.

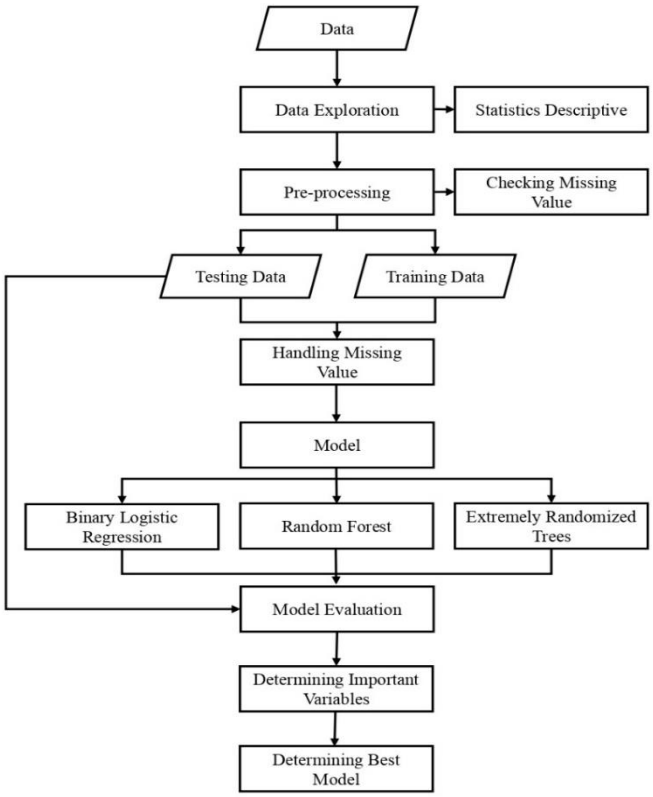

Figure 2.1. Flowchart of research analysis

3. RESULTS AND DISCUSSION

The data used in this study are from 56 regencies/cities in Kalimantan, where the descriptive statistics of the response and predictor variables to be used for analysis are summarized in Tables 3.1 and 3.2.

Table 3.1. Descriptive statistics of response variables

Category	Class Size
0: Not poor (% poverty< 10.8%)	44
1: Poor (% poverty ≥ 10.8%)	12

Table 3.2. Descriptive statistics of predictor variables

	X_1	X_2	X_3	X_4	X_5	X_6	X_7
Min	28.8	0.390	0.810	2.0	6.540	41.43	14640
First quantile	171.9	1.177	4.665	16.0	7.935	50.12	78164
Median	269.7	1.360	6.360	37.0	8.390	53.31	118709
Mean	320.6	1.711	8.928	382.4	8.660	52.51	148769
Third quantile	427.1	1.590	11.150	128.5	9.239	55.72	214701
Max	865.3	18.490	34.250	6819.0	11.705	61.98	403138
Missing Value	-	-	-	-	-	-	13

This information was at that point partitioned for modeling and approval testing. For modeling, 80% of the information or training data was utilized, and 20% of the information or testing data was

utilized for approval testing. The partitioned information was at that point taken care of for lost values utilizing the k-nearest neighbor strategy, where the closest neighbor esteem utilized was $k = 5$. The information taken care of will be proceeded for the investigation prepare by conducting a relationship test. The Pearson relationship strategy (r) was chosen to decide the quality of the indicator variables' direct relationship. The comes about of the Pearson relationship test are appeared in Figure 3.1.

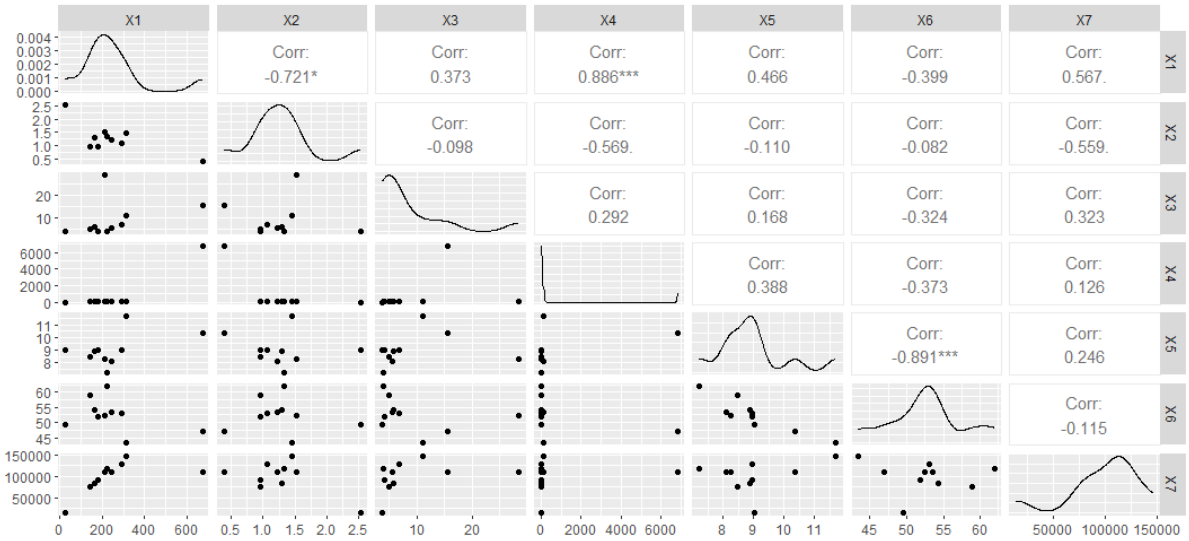

Figure 3.1. Correlation plot of predictor variable

Figure 3.1 shows the robust relationship between two predictor variables with $|r| \in (0.7,1]$, such as the correlation between variables X_1 with X_7 and X_5 with X_6. Then, the relationship is strong for $|r| \in (0.5,0.7]$, such as the correlation between variables X_1 with X_3 and X_4 with X_6. Variables with values of $|r| \in (0.3,0.5]$, such as the correlation between variables X_1 with X_4, X_1 with X_6, X_3 with X_5, X_3 with X_6, and X_4 with X_5, are moderate. The last, variables with $|r| \in (0,0.3]$ are weak.

The training data obtained from the data division is used for modeling with binary logistic regression. The binary logistic regression conducted in this study uses the backward stepwise method to compare the use of predictor variables in the model that are significant in determining the best model based on the smallest AIC value. The results of comparing the binary logistic regression model with the backward stepwise model are presented in Table 3.3.

Table 3.3. Binary logistic regression model with backward stepwise

	Model	AIC Value
1	$Y = X_1 + X_2 + X_3 + X_4 + X_5 + X_6 + X_7$	40.87
2	$Y = X_1 + X_2 + X_4 + X_5 + X_6 + X_7$	39.27
3	$Y = X_1 + X_4 + X_5 + X_6 + X_7$	37.66
4	$Y = X_4 + X_5 + X_6 + X_7$	35.81
5	$\boldsymbol{Y = X_4 + X_5 + X_6}$	**35.65**

The best binary logistic regression model based on backward stepwise is model 5, which includes the predictor variables of population density, years of schooling, and per capita expenditure on food. The parameter estimates from this best model are presented in Table 3.4.

Table 3.4. Parameter estimation of the best binary logistic regression model using backward stepwise

	Parameter Estimator Coefficient	Wald	p-value	Odds Ratio
Intercept	-32.8062	-2.08	0.037*	5.657×10^{-15}
X_4	-0.0332	-1.73	0.083	0.967
X_5	1.6253	1.96	0.050*	5.079
X_6	0.3595	1.83	0.067	1.433

Note: *) significant at $\alpha = 5\%$.

Based on Table 3.4, the mathematical form of the binary logistic regression model with the best model selection using backward stepwise is as follows.

$$Y = -32.8062 - 0.0332X_4 + 1.6253X_5 + 0.3595X_6$$

The odds ratio values in Table 3.4 above can be understood as a measure of the association of the probability of an event occurring. Thus, the odds ratio interpretation for the population density variable (X_4) is that the population density of regencies/cities in Kalimantan has a 0.967 times higher risk of influencing poverty status. The variable of schooling (X_5), which is 5.079, means that the years of schooling in regencies/cities in Kalimantan have a 5.079 times higher risk of influencing poverty status. Furthermore, the variable of per capita expenditure on food (X_6), which is 1.433, means that per capita expenditure in regencies/cities in Kalimantan has a 1.433 times higher risk of influencing poverty status.

The results of the best model parameter estimation were then tested for model goodness-of-fit using the Hosmer Lemeshow test, which produced a p-value of 0.9965, failing to reject H_0 or providing no significant statistical evidence to suggest that the best model obtained was not suitable for the data or that the model obtained was well-suited or consistent with the actual results observed. The next test is a simultaneous test using the G test. The G test results show that the p-value generated is 1.323×10^{-4}, thus rejecting H_0, or there is sufficient statistical evidence to state that the best model obtained has predictor variables that simultaneously influence the response variable. The second further test is a partial test with Wald, see Table 3.4, where based on the p-value generated, only the length of schooling (X_5) rejects H_0, which means that the length of schooling significantly affects the response variable.

The second analysis was conducted using a machine learning approach, where the models used were random forests and extra trees. These two models do not have standard sizes for determining their hyperparameters. Therefore, the hyperparameter values used in this research analysis were determined through hyperparameter tuning. The results of hyperparameter tuning for the random forest model yielded the best mtry value of 2 with a minimum node size of 1. Meanwhile, the best n-estimator size obtained for the extra trees model was 100, with a max-depth size of 5. This hyperparameter tuning was also combined with cross-validation where $k = 5$. Next, model training was performed using training data, and model validation was performed using test data.

The results of the binary logistic regression model testing meet the necessary assumptions, so that the model can be applied to the test data for validation. Evaluation of the training data and test data was performed with balanced accuracy. This accuracy measure was chosen because the response variable categories or poverty data for regencies/cities in Kalimantan are imbalanced, as shown in Table 3.1, with 44 regencies/cities classified as non-poor and 12 classified as poor. Therefore, a concise comparison of the predictive performance of the binary logistic regression model with the backwards stepwise, random forest, and extra trees methods is presented in Table 3.5.

Table 3.5. Comparison of prediction results

Evaluation Matrix	Model		
	Binary Logistic Regression	Random Forest	Extra Trees
Sensitifity	50.00%	62.50%	25.00%
Specificity	100.00%	50.00%	83.33%
Accuracy	90.00%	60.00%	60.00%
Balanced Accuracy	75.00%	56.25%	54.165%

The comparison comes about in Table 3.5 over appear that the binary logistic regression calculated relapse show with in backward stepwise is the best-performing show gotten by this for foreseeing destitution levels in regencies/cities in Kalimantan. This choice is based on the adjusted precision assessment network, where the parallel calculated relapse demonstrate with in backward stepwise has the most elevated esteem compared to the other two models. The advantage of the binary logistic regression calculated relapse demonstrate with in backward stepwise can be credited to selecting critical indicator factors for the reaction variable utilizing in ackward stepwise. The general advantage of basic statistical models is that the analysis process with binary logistic regression can be explained systematically and clearly, unlike machine learning, which tends to be a black box.

The results of this study also show that other factors influencing poverty can be identified. Population density, length of schooling, and per capita expenditure on food are consistently identified as the most important factors in predicting poverty levels in regencies/cities in Kalimantan. These significant variables align with the theory of poverty carried out by previous studies on [6], which states that poverty is closely related to emphasizing the importance of education. Other studies on [23] show that length of schooling is also a significant factor in poverty status. The population density variable that is significant to poverty status in this study is also in line with the results of research on [13], which shows that population density affects poverty.

The comparison of models for classification performance proposed by this study aims to provide new insights into comparing classification performance with basic statistical approaches and machine learning statistics. So, further research is expected to explore classification methods combined with multiclass data and compare missing data handling methods to determine the performance of the proposed classification method. In addition, showing the reliability of the approach based on considerations of the number of observations to the response categories owned by the data is also important. Revealing significant behavior and interactions between the proposed response variables and their features. Further research is expected to be able to overcome the weaknesses found in this study. In addition to selecting the features used in future research, a model approach is expected to be used that only produces the best accuracy.

4. CONCLUSION

This study uses three models to compare basic statistical approaches and machine learning statistical predictions. The data used for analysis was handled using the k-nearest neighbor ($k = 5$) to address missing data. The poverty level prediction analysis results for regencies/cities in Kalimantan indicate that the binary logistic regression model with backward stepwise selection is the most accurate model, with an accuracy of 90% and balanced accuracy of 75% compared to random forest with an accuracy of 60% and balanced accuracy of 56.25% and extra trees with an accuracy of 60% and balanced accuracy of 54.165%. Along with obtaining the best model with the highest accuracy, significant predictor variables influencing the poverty rate of regencies/cities in Kalimantan were also identified: population density, years of schooling, and per capita expenditure on food. Recommendations for addressing this issue are not straightforward and quite complex, requiring collaboration among various stakeholders to improve community well-being. Further

research could be conducted by considering novelty in statistics, such as addressing data imbalance in the response variable. Additionally, for the novelty of the problem, selecting predictor variables that may influence poverty levels.

CONFLICT OF INTEREST

All of authors declare that there is no conflict of interest

REFERENCES

[1] Alfian, G., Syafrudin, M., Fahrurrozi, I., Fitriyani, N. L., Atmaji, F. T. D., Widodo, T., Bahiyah, N., Benes, F., & Rhee, J., 2022. Predicting breast cancer from risk factors using SVM and extra-trees-based feature selection method. *Computers*, *11*(9), 136.

[2] Cavanaugh, J. E., & Neath, A. A., 2019. The Akaike information criterion: Background, derivation, properties, application, interpretation, and refinements. *Wiley Interdisciplinary Reviews: Computational Statistics*, *11*(3), e1460.

[3] De, H., & Acquah, G., 2010. Comparison of Akaike information criterion (AIC) and Bayesian information criterion (BIC) in selection of an asymmetric price relationship. *Journal of Development and Agricultural Economics*, *2*, 1–6.

[4] Fatimah, F., Fitrianto, A., Indahwati, I., Erfiani, E., & Khikmah, K. N., 2023. Synthetic Minority Oversampling Technique Pada Model Logit dan Probit Status Pengangguran Terdidik. *Jambura Journal of Mathematics*, *5*(1), 166–178.

[5] Geurts, P., Ernst, D., & Wehenkel, L., 2006. Extremely randomized trees. *Machine Learning*, *63*, 3–42.

[6] Harris, J. K. (2021). Primer on binary logistic regression. *Family Medicine and Community Health*, *9*(Suppl 1), e001290.

[7] Hassan, S. T., Batool, B., Zhu, B., & Khan, I., 2022. Environmental complexity of globalization, education, and income inequalities: New insights of energy poverty. *Journal of Cleaner Production*, *340*, 130735.

[8] Hilbe, J. M., 2016. *Practical guide to logistic regression*. CRC Press, Taylor & Francis Group Boca Raton, USA.

[9] Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X., 2013. *Applied logistic regression* (Vol. 398). John Wiley & Sons.

[10] Jena, M., & Dehuri, S., 2022. An integrated novel framework for coping missing values imputation and classification. *IEEE Access*, *10*, 69373–69387.

[11] Kajiita, R. M., & Kang'ethe, S. M., 2024. Socio-Economic Dynamics Inhibiting Inclusive Urban Economic Development: Implications for Sustainable Urban Development in South African Cities. *Sustainability*, *16*(7), 2803.

[12] Khikmah, K. N., Indahwati, I., Fitrianto, A., Erfiani, E., & Amelia, R., 2022. Backwards stepwise binary logistic regression for determination population growth rate factor in Java Island. *Jambura Journal of Mathematics*, *4*(2), 177–187.

[13] Khikmah, K. N., Sartono, B., Susetyo, B., & Dito, G. A., 2024. Performance Comparative Study of Machine Learning Classification Algorithms for Food Insecurity Experience by Households in West Java. *Jurnal Online Informatika*, *9*(1), 128–137.

[14] Kisiała, W., & Rącka, I., 2021. Spatial and statistical analysis of urban poverty for sustainable city development. *Sustainability*, *13*(2), 858.

[15] Kumar, S., & Gota, V., 2023. Logistic regression in cancer research: A narrative review of the concept, analysis, and interpretation. *Cancer Research, Statistics, and Treatment*, *6*(4), 573–578.

[16] Lalande, F., & Doya, K., 2022. Numerical data imputation: Choose kNN over deep learning. *International Conference on Similarity Search and Applications*, 3–10.

[17] Mathew, T. E., 2022. An optimized extremely randomized tree model for breast cancer classification. *Journal of Theoretical and Applied Information Technology*, *100*(16), 5234–5246.
[18] Okpala, E. F., Manning, L., & Baines, R. N., 2023. Socio-economic drivers of poverty and food insecurity: Nigeria a case study. *Food Reviews International*, *39*(6), 3444–3454.
[19] Poblete-Cazenave, M., & Pachauri, S., 2021. A model of energy poverty and access: Estimating household electricity demand and appliance ownership. *Energy Economics*, *98*, 105266.
[20] Portet, S., 2020. A primer on model selection using the Akaike Information Criterion. *Infectious Disease Modelling*, *5*, 111–128.
[21] Saeed, U., Jan, S. U., Lee, Y.-D., & Koo, I., 2021. Fault diagnosis based on extremely randomized trees in wireless sensor networks. *Reliability Engineering & System Safety*, *205*, 107284.
[22] Suleiman, T. A., Anyimadu, D. T., Permana, A. D., Ngim, H. A. A., & Scotto di Freca, A., 2024. Two-step hierarchical binary classification of cancerous skin lesions using transfer learning and the random forest algorithm. *Visual Computing for Industry, Biomedicine, and Art*, *7*(1), 15.
[23] Thomas, N. S., & Kaliraj, S., 2024. An Improved and Optimized Random Forest Based Approach to Predict the Software Faults. *SN Computer Science*, *5*(5), 530.
[24] Uralovich, K. S., Toshmamatovich, T. U., Kubayevich, K. F., Sapaev, I. B., Saylaubaevna, S. S., Beknazarova, Z. F., & Khurramov, A., 2023. A primary factor in sustainable development and environmental sustainability is environmental education. *Caspian Journal of Environmental Sciences*, *21*(4), 965–975.
[25] Wahyuningsih, D., Yunaningsih, A., Priadana, M. S., Wijaya, A., Darma, D. C., & Amalia, S., 2020. The dynamics of economic growth and development inequality in Borneo Island, Indonesia. *Journal of Applied Economic Sciences*, *1*(67), 135–143.
[26] Zaidi, A., & Al Luhayb, A. S. M., 2023. Two statistical approaches to justify the use of the logistic function in binary logistic regression. *Mathematical Problems in Engineering*, *2023*(1), 5525675.