BAREKENG: Journal of Mathematics and Its Applications
June 2025 Volume 19 Issue 2 Page 1143-1156
P-ISSN: 1978-7227 E-ISSN: 2615-3017
https://doi.org/10.30598/barekengvol19iss2pp1143-1156

EVALUATION OF MULTIVARIATE ADAPTIVE REGRESSION SPLINES ON IMBALANCED DATASET FOR POVERTY CLASSIFICATION IN BENGKULU PROVINCE

Idhia Sriliana [1*], Sigit Nugroho[2], Winalia Agwil[3], Esther Damayanti Sihombing[4]

[1,2,3,4]*Department of Statistics, Faculty of Mathematics and Natural Sciences, University of Bengkulu*
Jln. WR. Supratman, Kandang Limun, Bengkulu, 38371, Indonesia

*Corresponding author's e-mail: * **idhiasriliana@unib.ac.id***

ABSTRACT

Article History:

Received: 22nd September 2024
Revised: 3rd February 2025
Accepted: 4th March 2025
Published: 1st April 2025

Keywords:

Classification;
Class Imbalance;
MARS;
Poverty.

Classification is a statistical method that aims to predict the class of an object whose class label is unknown. The Multivariate Adaptive Regression Splines (MARS) classification method is a classification model that involves several basis functions with influential predictor variables. The MARS classification model is generally effective in classifying imbalanced data, including poverty data classification. The response variable used is the poverty status of households classified into poor and non-poor households, and the predictor variables consist of several poverty indicators. The problem that often arises in classification methods is a class imbalance in the response variable. Due to the poverty status included in the class imbalance data, the Bootstrap Aggregating (Bagging) and Synthetic Minority Over-sampling Technique (SMOTE) approaches will be used to improve classification accuracy on the MARS model. Bagging works by replicating data to strengthen the stability of classification accuracy, while SMOTE works by synthesizing data from minority data classes. The evaluation results showed that the classification model of poverty in Bengkulu Province using the SMOTE-MARS method provides the best classification accuracy compared to the MARS (25.81%) and Bagging-MARS (32.26%) methods based on the sensitivity value obtained, which is 85.36%.

How to cite this article:

I. Sriliana, S. Nugroho, W. Agwil and E. D. Sihombing., "EVALUATION OF MULTIVARIATE ADAPTIVE REGRESSION SPLINES ON IMBALANCED DATASET FOR POVERTY CLASSIFICATION IN BENGKULU PROVINCE," *BAREKENG: J. Math. & App.*, vol. 19, iss. 2, pp. 1143-1156, June, 2025.

Journal homepage: *https://ojs3.unpatti.ac.id/index.php/barekeng/*
Journal e-mail: *barekeng.math@yahoo.com; barekeng.journal@mail.unpatti.ac.id*

Research Article · Open Access

1. INTRODUCTION

Classification is a statistical method used to group objects into a class. According to [1], classification is the grouping of objects into a predefined class or category. Classification aims to predict the class of an object whose class label is unknown [2]. One of the statistical methods that can be used for classification analysis is the Multivariate Adaptive Regression Splines (MARS) method. The MARS method is a refinement of the truncated spline nonparametric method with recursive partitioning regression [3]. The MARS model overcomes the weaknesses of the two methods to produce a continuous model at knots and be able to identify the presence of linear and additive functions. MARS can also produce more accurate response variable estimates with high-dimensional data problems with the number of observations and predictor variables of $3 \leq n \leq 20$ [4].

The problem that often arises in classification methods is the occurrence of imbalanced classes on the response variable. A high imbalance in the proportion of observations between categories on the response variable leads to the formation of the majority and the minority classes. Unbalanced data between the majority class and the minority class cause misclassification, especially in the minority class, since the classification model is most likely to predict most of the data that fall within the majority class [5]. Methods that can be used to address classification problems when there is an imbalanced dataset are Bootstrap Aggregating (Bagging) and Synthetic Minority Over-sampling Technique (SMOTE). The Bagging method works by replicating data to strengthen the stability of classification accuracy. Meanwhile, the SMOTE method works by synthesizing data from minority classes.

Several studies have been conducted with the MARS method, such as [6] with a focus on classifying village status in East Nusa Tenggara Province using MARS with an imbalanced proportion of data groups which resulted in an accuracy value of 99.40%, sensitivity of 99.84%, and specificity of 92.8% on test data. The results showed that MARS can classify well even in unbalanced data conditions. Furthermore, [7] used MARS and Bagging-MARS to determine the population poverty model in Central Java Province. In addition, [8] conducted research on lecturer performance on private campuses using the Bagging-MARS algorithm. The research shows that Bagging-MARS produces a higher accuracy value than the single-stage MARS method, which is 88.52%. Meanwhile, [9] used MARS to estimate soil pollution. As well as the study by [10] using a combination of Random Forest and MARS Binary Response for Classification of HIV/AIDS Patients in Surabaya with a classification accuracy of 91%

One of the imbalanced datasets occurs in household poverty status. In Indonesia, poverty is a global issue that is of concern to the government because it is a complex and multidimensional population problem. Indonesia has a poverty percentage of 9.36% of the total population, which is 25.90 million people below the poverty line [11]. This figure makes Indonesia the poorest country with a ranking of 73 in the world. Bengkulu is one of the provinces in Indonesia that has a high poverty rate. BPS Socio-Economic Data in March 2023 shows that Bengkulu is ranked as the second poorest province in Sumatra, with a percentage of poor people of 14.04%. This number illustrates that poverty in Bengkulu Province is relatively high compared to the national poverty rate, even though the economy of Bengkulu Province in the second quarter of 2023 increased by 3.03% [12].

Various government programs and policies in the health, social, economic, and other fields have been attempted to reduce the number of poor households. Identifying the characteristics of households that are classified as poor households in Bengkulu Province is an important study to conduct since it can serve as a reference for government programs to be implemented on target. Therefore, this study aims to evaluate the classification model of household poverty status in Bengkulu Province using MARS, Bagging-MARS, and SMOTE-MARS methods. The classification model accuracy is evaluated by comparing MARS and the other two methods based on accuracy rate, sensitivity, and specificity.

2. RESEARCH METHODS

2.1 Multivariate Adaptive Regression Splines (MARS)

Multivariate Adaptive Regression Splines (MARS) is a nonparametric method that can solve regression and classification problems. MARS combines the truncated spline nonparametric method with recursive partitioning regression. MARS can be used in high-dimensional data cases where the number of observations and predictor variables is quite large, specifically $3 \leq n \leq 20$.

The general MARS model is written as follows [3]:

$$f(x_i) = \beta_0 + \sum_{m=1}^{M} \beta_m \prod_{k=1}^{Km} [s_{km}(x_{v(k,m)} - t_{km})]_+ \quad (1)$$

where,

β_0 : main of basis function
β_m : coefficient of basis function $-m$
M : maximum of basis functions
K_m : the maximum degree of interaction
s_{km} : its value ± 1, depending on whether the data is to the right or left of the knot point
$x_{v(k,m)}$: predictor variable $-v$.
t_{km} : knot point of predictor variable $-v$.

The MARS model based on the nonparametric regression function is expressed in the following equation:

$$y_i = \beta_0 + \sum_{m=1}^{M} \beta_m \prod_{k=1}^{Km} [s_{km}(x_{v(k,m)} - t_{km})]_+ + \varepsilon_i \quad (2)$$

Then, **Equation (2)** can be elaborated as follows:

$$y_i = \beta_0 + \sum_{m=1}^{M} \beta_m \left[s_{km}\left(x_{v(k,m)} - t_{km}\right)\right]_+ + \sum_{m=1}^{M} \beta_m \left[s_{km}\left(x_{v(k,m)} - t_{km}\right)\right]_+ \left[s_{km}\left(x_{v(k,m)} - t_{km}\right)\right]_+$$
$$+ \sum_{m=1}^{M} \beta_m \left[s_{km}\left(x_{v(k,m)} - t_{km}\right)\right]_+ \left[s_{km}\left(x_{v(k,m)} - t_{km}\right)\right]_+ \left[s_{km}\left(x_{v(k,m)} - t_{km}\right)\right]_+ + \cdots + \varepsilon_i \quad (3)$$

Equation (3) is the sum of basis functions in which the first summation component involves one variable. The second summation includes the first sum of the basis function for interactions involving two variables. The third sum involves three variables, and so on.

Furthermore, **Equation (2)** can be written in matrix form as follows:

$$\boldsymbol{y} = \boldsymbol{B\beta} + \boldsymbol{\varepsilon} \quad (4)$$

where $\boldsymbol{y} = (y_1, y_2, \ldots, y_N)^T, \boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_M)^T, \boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_N)^T$

$$\boldsymbol{B} = \begin{bmatrix} 1 & \prod_{k=1}^{K_1} [s_{km}(x_{v(k,m)} - t_{km})] & \cdots & \prod_{k=1}^{K_M} [s_{km}(x_{v(k,m)} - t_{km})] \\ 1 & \prod_{k=1}^{K_1} [s_{km}(x_{v(k,m)} - t_{km})] & \cdots & \prod_{k=1}^{K_M} [s_{km}(x_{v(k,m)} - t_{km})] \\ \vdots & \vdots & \vdots & \vdots \\ 1 & \prod_{k=1}^{K_1} [s_{km}(x_{v(k,m)} - t_{km})] & \cdots & \prod_{k=1}^{K_M} [s_{km}(x_{v(k,m)} - t_{km})] \end{bmatrix}$$

The estimation parameters model in **Equation (4)** is carried out using the least squares method such that the following estimator is obtained.

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{B}^T \boldsymbol{B})^{-1} \boldsymbol{B}^T \boldsymbol{Y} \quad (5)$$

The best MARS model selection is conducted using two approaches: forward selection and backward elimination. The forward selection method aims to obtain a function with the maximum number of basis functions [13]. In terms of restricting the model, a maximum number of functions is used. Despite the restriction, the forward selection results in a model with a very large number of basis functions. Therefore, it is necessary to remove some of the basis functions to achieve a simplified model. According to [13], backward elimination aims to obtain a simplified model (parsimony principle) by removing basis functions that have a small contribution to the response at the forward selection stage by minimizing the Generalized Cross Validation (GCV) function.

According to [3], the GCV formula is defined as:

$$GCV = \frac{\frac{1}{N}\sum_{i=1}^{N}\left(y_i - \hat{f}_M(x_i)\right)^2}{\left[1 - \frac{C(M)}{N}\right]^2} \quad (6)$$

where $C(M) = trace[\boldsymbol{B}(\boldsymbol{B}^T\boldsymbol{B})^{-1}\boldsymbol{B}^T] + 1$, $\boldsymbol{B}$ is a matrix of M nonconstant functions with $M \times N$ dimension, and N is the number of observations.

2.2 MARS Classification

Classification is the process of grouping objects into a class or category that has been determined [1]. Classification is the process of finding a model that can describe and differentiate class data. The goal of classification is to predict the class of objects where the class label is not known [2]. [14] stated that if the response variable in regression consists of two categories (binary), it is referred to as binary response regression. One method used for analysis is binary logistic regression. Due to this point, classification in MARS is based on binary logistic regression analysis, allowing for the use of probability models expressed in the following equation:

$$\pi(x) = \frac{e^{\hat{y}}}{1 + e^{\hat{y}}} \quad (7)$$

where $\hat{y} = \beta_0 + \sum_{m=1}^{M}\beta_m \prod_{k=1}^{Km}[s_{km}(x_{v(k,m)} - t_{km})]_+$.

According to [15], the classification of binary response variables (1 and 0) can be carried out using a cutoff point of 0.5. If the estimated probability is greater than 0.5, the prediction will fall into group 1, but if the estimated probability is less than or equal to 0.5, the prediction will fall into group 0. A good classification method can produce few classification errors or a small probability of misclassification [1]. The evaluation of classification performance can be computed through cross-tabulation called confusion matrix [2]. The structure of the confusion matrix is presented in Table 1.

Table 1. Confusion Matrix of Two-Class Classification

Prediction	Actual	
	Class 1	Class 2
Class 1	n_{11}	n_{12}
Class 2	n_{21}	n_{22}

n_{11}: total observations from class 1 that are correctly classified as class 1

n_{12}: total observations from class 1 that are incorrectly classified as class 2

n_{21}: total observations from class 2 that are incorrectly classified as class 1

n_{22}: total observations from class 2 that are correctly classified as class 2

The Apparent Error Rate (APER) is a procedure for evaluating the classification error made by a classification function [1]. The Total Accuracy Rate (TAR) is used to calculate the classification accuracy of the observations. The TAR value represents the proportion of correctly classified observations. The formula to calculate APER and TAR is as follows:

$$APER\ (\%) = \frac{Number\ of\ incorrect\ predictions}{Total\ number\ of\ predictions} = \frac{n_{12} + n_{21}}{n_{11} + n_{12} + n_{21} + n_{22}} \times 100\%$$

$$TAR\ (\%) = \frac{Number\ of\ correct\ predictions}{Total\ number\ of\ predictions} = \frac{n_{11} + n_{22}}{n_{11} + n_{12} + n_{21} + n_{22}} \times 100\%$$

In the classification table, sensitivity is considered, which represents the accuracy of observations in the positive group, and specificity, which represents the accuracy of observations in the negative group. The ability to measure sensitivity and specificity well indicates that the classification method is good at predicting observations in each group. The formulas to calculate sensitivity and specificity are as follows [16]:

$$sensitivity\ (\%) = \frac{n_{11}}{n_{11} + n_{12}} \times 100\%$$

$$specificity\ (\%) = \frac{n_{22}}{n_{21} + n_{22}} \times 100\%$$

2.3 Bootstrap Aggregating (Bagging)

Bootstrap Aggregating (Bagging) is an ensemble learning method introduced by Breiman in 1994 to improve the stability and accuracy of certain algorithms, such as regression or classification. As its name suggests, Bagging consists of two stages: bootstrap and aggregating. The classification process in Bagging starts with bootstrap, a resampling technique where random samples are taken with replacement from the training dataset. This is repeated k times to obtain k models. Then, the aggregation stage is conducted, where the results of the k models are combined [17].

The Bagging algorithm is as follows:

1. A data set L consisting of $\{(y_i, x_i),\ i = 1,2,\dots,n\}$ is sampled with replacement, resulting in a new data set $L^* = \{\{(y_i^*, x_i^*),\ i = 1,2,\dots,n\}$.
2. Perform a classification or regression algorithm on the new data set L^*.
3. Repeat steps 1 and 2 for w bootstrap replications.
4. The predicted response variable is the average of the predictions produced for continuous response variables or the mode (majority vote) for categorical response variables.

2.4 Synthetic Minority Over-sampling Technique (SMOTE)

The Synthetic Minority Over-sampling Technique (SMOTE) was introduced by [18] to address the class imbalance problem in data. SMOTE is a resampling method that increases the minority class samples using the K-Nearest Neighbors (KNN) approach. Synthetic data is generated based on the nearest neighbors of a given data point in the minority class. The procedure for generating synthetic data in SMOTE is described by the following equation:

$$x_{new} = x_i + (\hat{x}_i - x_i) \times rand[0,1] \tag{8}$$

where x_i is the i-th predictor variable value, $rand[0,1]$ is a random value between 0 and 1, and $\hat{x}_i$ is one of the K-Nearest Neighbors (KNN). KNN is used to generate synthetic data for the minority class. The Euclidean distance method is used to calculate the proximity between data points, expressed as follows:

$$d(\boldsymbol{x}, \boldsymbol{y}) = \sqrt{(\boldsymbol{x} - \boldsymbol{y})^t(\boldsymbol{x} - \boldsymbol{y})} = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \tag{9}$$

in which $\boldsymbol{x}$ is the predictor variable vector x and $\boldsymbol{y}$ is the response variable vector y.

The algorithm for generating synthetic data using the SMOTE method is as follows:

1. Determining the value of k, the nearest neighbor.
2. Set the oversampling percentage.
3. Calculating the distance between observations in the minority class.
4. Calculating the synthetic data value using **Equation (8)**.
5. Obtaining new synthetic data consisting of the combination of each minority class and SMOTE-generated results.

2.5 Data and Data Sources

The data used in this study consists of secondary data from the National Socio-Economic Survey (Susenas) conducted in March 2022, obtained from Statistics Indonesia, Bengkulu Province. The objects of this study are households in Bengkulu Province. The data consists of 13 predictor variables and a binary response variable (Y), representing the household's status as either poor or non-poor. The predictor variables used in this study are region of residence (X_1), house ownership status (X_2), roofing material (X_3), main wall material (X_4), main floor material (X_5), defecation facility usage (X_6), lighting source (X_7), main cooking fuel source (X_8), land ownership status (X_9), floor area (X_{10}), number of household members (X_{11}), number of families (X_{12}), and daily calorie consumption (X_{13}).

2.6 Data Analysis Procedure

The data analysis procedure in this study is carried out in the following stages:

1. Data exploration.
2. Identification of the correlation between the response variable and the predictor variables.
3. Establishment and estimation of the MARS classification model with the following steps:
   a. Determining the number of Basis Functions (BF), Maximum Interaction (MI), and Minimum Observation (MO).
   b. Establishing the MARS classification model by combining BF, MI, and MO.
   c. Determining the best MARS model according to the minimum GCV value based on the forward selection and backward elimination.
   d. Forming a confusion matrix table.
   e. Calculating the accuracy of the MARS classification model.
4. Establishment of the Bagging-MARS classification model through the following steps:
   a Performing bootstrap replications on the data.
   b Modelling MARS of each bootstrap replication with the combination of BF, MI, and MO from the best MARS model identified in the previous stage.
   c Selecting the best Bagging-MARS classification model based on the minimum GCV value.
   d Forming a confusion matrix table.
   e Calculating the accuracy of the Bagging-MARS classification model.
5. Establishment of the SMOTE-MARS classification model through the following steps:
   a. Generating synthetic data using the SMOTE algorithm.
   b. Obtaining the SMOTE-MARS classification model similar to step 3.
   c. Forming a confusion matrix table.
   d. Calculating the accuracy of the SMOTE-MARS classification model.
6. Evaluation of the classification model accuracy.

3. RESULTS AND DISCUSSION

3.1 Data Exploration

The National Socio-Economic Survey (Susenas), Statistics Indonesia in March 2022 showed that there were 616 poor households and 5115 non-poor households out of 5731 households surveyed in Bengkulu

Province. A household is poor if its monthly per capita consumption expenditure is below the poverty line. The poverty line in Bengkulu Province in March 2022 was IDR 590,754. Based on Figure 1, it can be seen that there is a considerable difference between poor households and non-poor households. The percentage of poor households is only 10.75%, while the percentage of non-poor households reaches 89.25%. Therefore, there is a class imbalance in this dataset where the poor household group is the minority class while the non-poor household group is the majority class.

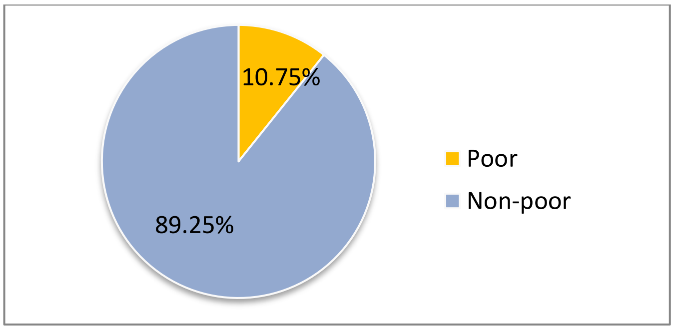

Figure 1. Percentage of Response Variable Categories

Before performing MARS classification modeling, the correlation between the response variable and each predictor variable was identified. The relationship between the categorical predictor variable and the categorical response variable is tested using a chi-square test. Meanwhile, the point biserial correlation test is used to identify the relationship between the numeric predictor variable and the categorical response variable. The testing results are presented in Table 2 below:

Table 2. Correlation Test between Response Variable and Each Predictor Variable

No	Variable	*Chi-square* (χ^2)	r_{pb}	$p-value$	Decision
Categorical Predictor Variable					
1	X_1	8.724	-	0.0031	reject H_0
2	X_2	19.347	-	0.0002	reject H_0
3	X_3	4.894	-	0.2984	Fail to reject H_0
4	X_4	48.618	-	2.66×10^{-9}	reject H_0
5	X_5	194.880	-	$< 2.2 \times 10^{-16}$	reject H_0
6	X_6	107.730	-	$< 2.2 \times 10^{-16}$	reject H_0
7	X_7	76.201	-	$< 2.2 \times 10^{-16}$	reject H_0
8	X_8	71.927	-	4.07×10^{-14}	reject H_0
9	X_9	13.307	-	0.0002	reject H_0
Numeric Predictor Variable					
10	X_{10}	-	0.114	$< 2.2 \times 10^{-16}$	reject H_0
11	X_{11}	-	0.204	$< 2.2 \times 10^{-16}$	reject H_0
12	X_{12}	-	0.092	3.87×10^{-12}	reject H_0
13	X_{13}	-	0.315	$< 2.2 \times 10^{-16}$	reject H_0

The Chi-square (χ^2) test result in Table 2 concludes that the categorical variables $X_1, X_2, X_4, X_5, X_6, X_7, X_8$, and X_9 correlate with the response variable. Meanwhile, the result of the point biserial correlation test for numerical variables shows that there is a correlation between predictor variables $X_{10}, X_{11}, X_{12}, X_{13}$ and the response variable. As a result, 12 predictor variables have a significant correlation with the response variable. These predictor variables will be involved in modeling.

3.2 MARS Classification Model

MARS model is constructed based on a combination of Basis Function (BF), Maximum Interaction (MI), and Minimum Observation (MO). The number of significant predictors is 12 variables such that the BF values used are 24, 36, and 48. The following step is to determine the MARS model by trial and error based on the minimum GCV value. GCV values for all models formed based on each combination of BF, MI, and MO are presented in Table 3.

Table 3. GCV Value of BF, MI, and MO Combination for MARS Model

No.	BF	MI	MO	GCV	No.	BF	MI	MO	GCV
1	24	1	0	0.0754147	19	36	2	2	0.0706785
2	24	1	1	0.0754148	20	36	2	3	0.0706781
3	24	1	2	0.0754147	21	36	3	0	0.0703243
4	24	1	3	0.0754147	22	36	3	1	0.0703271
5	24	2	0	0.0714211	23	36	3	2	0.0703272
6	24	2	1	0.0714341	24	36	3	3	0.0703272
7	24	2	2	0.0714348	25	48	1	0	0.0754147
8	24	2	3	0.0714349	26	48	1	1	0.0754148
9	24	3	0	0.0711379	27	48	1	2	0.0754147
10	24	3	1	0.0712759	28	48	1	3	0.0754147
11	24	3	2	0.0712764	29	48	2	0	0.0701827
12	24	3	3	0.0712765	30	48	2	1	0.0702344
13	36	1	0	0.0754147	31	48	2	2	0,0702342
14	36	1	1	0.0754148	32	48	2	3	0.0701208
15	36	1	2	0.0754147	33	48	3	0	0.0697111
16	36	1	3	0.0754147	34	48	3	1	0.0694715
17	36	2	0	0.0706916	35	48	3	2	0.0694711
18	36	2	1	0.0706785	**36**	**48**	**3**	**3**	**0.0694710**

The best MARS classification model is obtained in the combination of BF = 48, MI = 3, and MO = 3 with a minimum GCV value of 0.0694710. Furthermore, the selection of the optimum basis function at the backward elimination obtained the minimum GCV value in the subset model containing 26 basis functions. Hence, the MARS classification model is obtained as follows:

$$\begin{aligned}\hat{f}(x) = 0.01559 &+ 0.00101BF3 - 0.00154BF6 - 0.00008BF7 + 0.00001BF8 + 0.43111BF9\\ &+ 0.10295BF10 - 0.38639BF11 - 0.06394BF12 - 0.00029BF13 - 0.00062BF14\\ &- 0.00001BF15 + 0.00033\,BF16 - 0.00024BF17 + 0.00043BF18 - 0.25398BF19\\ &- 0.00843BF21 + 0.00008BF22 + 0.00057BF25 + 0.30454BF27 - 0.18019BF29\\ &- 0.00009BF31 + 0.00035BF32 + 0.00262BF33 + 0.000002BF35 + 0.00192BF36\end{aligned}$$

where,

BF3 : $\text{Max}(0, 1915.24 - X_{13})$
BF6 : $\text{Max}(0, X_{11} - 8)\text{Max}(0, 1915.24 - X_{13})$
BF7 : $\text{Max}(0, 8 - X_{11})\text{Max}(0, 1915.24 - X_{13})$
BF8 : $(X_1 = 1)\text{Max}(0, 96 - X_{10})\text{Max}(0, 1915.24 - X_{13})$
BF9 : $X_6 = 4$
BF10 : $X_8 = 5$
BF11 : $(X_6 = 4)\text{Max}(0, X_{11} - 5)$
BF12 : $(X_6 = 4)\text{Max}(0, 5 - X_{11})$
BF13 : $(X_4 = 0)\text{Max}(0, 1915.24 - X_{13})$
BF14 : $(X_2 = 3)\text{Max}(0, 1915.24 - X_{13})$
BF15 : $(X_4 = 5)\text{Max}(0, 96 - X_{10})\text{Max}(0, 1915.24 - X_{13})$
BF16 : $\text{Max}(0, X_{12} - 2)\text{Max}(0, 1915.24 - X_{13})$
BF17 : $\text{Max}(0, 2 - X_{12})\text{Max}(0, X_{13} - 1915.24)$
BF18 : $(X_4 = 0)(X_7 = 1)\text{Max}(0, 1915.24 - X_{13})$
BF19 : $(X_1 = 1)(X_6 = 4)$
BF21 : $\text{Max}(0, 5 - X_{11})$
BF22 : $(X_6 = 3)\text{Max}(0, X_{10} - 96)\text{Max}(0, 1915.24 - X_{13})$
BF25 : $\text{Max}(0, X_{11} - 6)\text{Max}(0, 2 - X_{12})\text{Max}(0, 1915.24 - X_{13})$
BF27 : $(X_1 = 1)(X_6 = 4)\text{Max}(0, X_{11} - 5)$
BF29 : $(X_8 = 2)\text{Max}(0, X_{11} - 5)$

BF31 : $\text{Max}(0, 48 - X_{10})\text{Max}(0, X_{12} - 2)\text{Max}(0, 1915.24 - X_{13})$
BF32 : $\text{Max}(0, X_{10} - 80)\text{Max}(0, X_{11} - 5)$
BF33 : $\text{Max}(0, 80 - X_{10})\text{Max}(0, X_{11} - 5)$
BF35 : $\text{Max}(0, 77 - X_{10})\text{Max}(0, 2781.34 - X_{13})$
BF36 : $(X_2 = 3)(X_4 = 5)\text{Max}(0, 1915.24 - X_{13})$

Measurement of classification accuracy in the MARS model is carried out through the calculation of the confusion matrix value. In the following the confusion matrix based on the MARS classification model is given in Table 4.

Table 4. Confusion Matrix of MARS Classification Model

Prediction	Actual	
	Non-poor	**Poor**
Non-poor	253	23
Poor	3	8
Percentage Correct (%)	98.83	25.81

3.3 Bagging-MARS Classification Model

The Bagging-MARS classification model is performed with significant pairs of the response and the predictor variables obtained from the best MARS model. In this method, the bootstrap technique is performed on classification modeling with a combination of BF, MI, and MO likely to the MARS model in the previous stage, namely single-stage MARS. The best Bagging-MARS model was obtained at the 60th replication of 100 replications. This model provides a minimum GCV value of 0.06063. The following is the Bagging-MARS classification model obtained

$$\begin{aligned}\hat{f}(x) = 0.00165 &+ 0.51333BF2 + 0.00113\text{BF3} - 0.17474BF4 - 0.00042BF5 - 0.00054BF6 \\ &+ 0.11263BF7 + 0.00178BF8 - 0.22986BF9 - 0.10585BF10 - 0.00069BF11 \\ &+ 0.00002BF12 - 0.00197BF13 - 0.00037BF14 - 0.00103BF15 + 0.00133BF16 \\ &+ 0.00005BF17 + 0.00002BF18 + 0.00143BF19 + 0.00024BF20 + 0.00002BF21 \\ &+ 0.00025BF22 + 0.00002BF23 - 0.00001BF24 - 0.00002BF25 + 0.00094BF26\end{aligned}$$

in which,

BF2 : $X_8 = 5$
BF3 : $\text{Max}(0, 1680.29 - X_{13})$
BF4 : $(X_7 = 0)(X_8 = 5)$
BF5 : $(X_1 = 1)\text{Max}(0, 1680.29 - X_{13})$
BF6 : $(X_4 = 0)\text{Max}(0, 1680.29 - X_{13})$
BF7 : $(X_4 = 1)\text{Max}(0, X_{11} - 4)$
BF8 : $(X_4 = 1)\text{Max}(0, 1680.29 - X_{13})$
BF9 : $(X_8 = 5)\text{Max}(0, X_{11} - 5)$
BF10 : $(X_8 = 5)\text{Max}(0, 5 - X_{11})$
BF11 : $(X_8 = 5)\text{Max}(0, 1629.68 - X_{13})$
BF12 : $\text{Max}(96 - X_{10})\text{Max}(0, 2722.89 - X_{13})$
BF13 : $\text{Max}(0, 4 - X_{11})\text{Max}(0, 1285.76 - X_{13})$
BF14 : $\text{Max}(0, 7 - X_{11})\text{Max}(0, 1680.29 - X_{13})$
BF15 : $(X_1 = 1)(X_2 = 3)\text{Max}(0, 1680.29 - X_{13})$
BF16 : $(X_4 = 0)(X_7 = 1)\text{Max}(0, 1680.29 - X_{13})$
BF17 : $(X_1 = 1)\text{Max}(0, X_{10} - 77)\ \text{Max}(0, 1680.29 - X_{13})$
BF18 : $(X_1 = 1)\text{Max}(0, 77 - X_{10})\ \text{Max}(0, 1680.29 - X_{13})$
BF19 : $(X_1 = 1)\text{Max}(0, X_{11} - 6)\ \text{Max}(0, 1680.29 - X_{13})$
BF20 : $(X_1 = 1)\text{Max}(0, 6 - X_{11})\ \text{Max}(0, 1680.29 - X_{13})$
BF21 : $(X_6 = 4)\text{Max}(96 - X_{10})\text{Max}(0, 2722{,}89 - X_{13})$
BF22 : $(X_7 = 0)\text{Max}(0, 7 - X_{11})\text{Max}(0, 1680.29 - X_{13})$
BF23 : $\text{Max}(96 - X_{10})\text{Max}(0, X_{11} - 4)\text{Max}(0, 2722.89 - X_{13})$
BF24 : $\text{Max}(96 - X_{10})\text{Max}(0, X_{12} - 2)\text{Max}(0, 2722.89 - X_{13})$

BF25 : $\text{Max}(96 - X_{10})\text{Max}(0, 2 - X_{12})\text{Max}(0, 2722.89 - X_{13})$
BF26 : $\text{Max}(0, 7 - X_{11})\text{Max}(0, X_{12} - 2)\text{Max}(0, 1680.29 - X_{13})$

Afterward, the confusion matrix table formed based on the Bagging-MARS classification model is given in Table 5 below:

Table 5. Confusion Matrix of Bagging-MARS Classification Model

Prediction	Actual	
	Non-poor	Poor
Non-poor	253	21
Poor	3	10
Percentage Correct (%)	98.83	32.26

3.4 SMOTE-MARS Classification Model

The data used in the study is an imbalanced dataset because there are majority and minority class categories in the response variable, household poverty status. Therefore, over-sampling was performed on the minority class data using the SMOTE method. The over-sampling percentage used is 800%, in which the number of data replications generated for the minority class is eight times the number of original samples in the class. The aim is to increase the sample size in the minority class such that the minority class becomes more balanced with the majority class. The new data generated yielded nearly identical class proportions, with the majority class numbering 5115 and the over-sampled minority class being 4928. After balancing the data, the classification model is performed using the MARS method. The stages carried out are similar to the stages of MARS modeling. According to the analysis results in Table 6, the best SMOTE-MARS classification model is obtained, the model with a combination of BF = 48, MI = 3, and MO = 1, which gives a minimum GCV value of 0.130805. A maximum interaction of 3 means that in the SMOTE-MARS model, there is interaction between predictor variables with a maximum interaction of 3. A minimum observation of 3 means that in the SMOTE-MARS model, the distance between knot points is three observations.

Table 6. GCV Value of BF, MI, and MO Combination for SMOTE-MARS Model

No.	BF	MI	MO	GCV	No.	BF	MI	MO	GCV
1	24	1	0	0.138608	19	36	2	2	0.132756
2	24	1	1	0.138603	20	36	2	3	0.132756
3	24	1	2	0.138603	21	36	3	0	0.132755
4	24	1	3	0.138603	22	36	3	1	0.132755
5	24	2	0	0.135910	23	36	3	2	0.132756
6	24	2	1	0.135910	24	36	3	3	0.132756
7	24	2	2	0.135910	25	48	1	0	0.138105
8	24	2	3	0.135910	26	48	1	1	0.138100
9	24	3	0	0.135910	27	48	1	2	0.138100
10	24	3	1	0.135910	28	48	1	3	0.138100
11	24	3	2	0.135910	29	48	2	0	0.130963
12	24	3	3	0.135910	30	48	2	1	0.130953
13	36	1	0	0.138105	31	48	2	2	0.130954
14	36	1	1	0.138100	32	48	2	3	0.130962
15	36	1	2	0.138100	33	48	3	0	0.130813
16	36	1	3	0.138100	**34**	**48**	**3**	**1**	**0.130805**
17	36	2	0	0.132755	35	48	3	2	0.130806
18	36	2	1	0.132755	36	48	3	3	0.130812

Parameter estimation of the SMOTE-MARS model was performed on the selected MARS model through forward selection and backward elimination methods. The basis functions obtained in the forward selection amounted to 35, with or without interaction. Furthermore, backward elimination produces the minimum GCV value in the subset model containing 32 basis functions. Consequently, the best SMOTE-MARS classification model obtained based on backward elimination is as follows:

$$\hat{f}(x) = 0.18608 - 0.00021BF2 + 0.00049BF3 - 0.19781BF4 + 0.00001BF5 + 0.00001BF6 - 0.20044BF7 - 0.09224BF8 + 0.30604BF9 + 0.51466BF10 + 0.00010BF11 + 0.00015BF12 - 0.00019BF14 - 0.00037BF15 + 0.00015BF16 + 0.25984BF17 + 0.00004BF18 - 0.00043BF19 + 0.00129BF20 - 0.02645BF21 - 0.00025BF22 - 0.00041BF24 - 0.23068BF25 - 0.19510BF26 + 0.00034BF27 - 0.12909BF28 - 0.11497BF29 + 0.000001BF30 - 0.00012BF31 - 0.00018BF32 + 0.00025BF34 - 0.00022BF35$$

in which,

BF2 : $\text{Max}(0, X_{13} - 2682.88)$
BF3 : $\text{Max}(0, 2682.88 - X_{13})$
BF4 : $X_5 = 1$
BF5 : $\text{Max}(0, X_{10} - 120)\text{Max}(0, 2682.88 - X_{13})$
BF6 : $\text{Max}(0, 120 - X_{10})\text{Max}(0, 2682.88 - X_{13})$
BF7 : $\text{Max}(0, X_{11} - 7.55)$
BF8 : $\text{Max}(0, 7.55 - X_{11})$
BF9 : $X_8 = 5$
BF10 : $X_6 = 4$
BF11 : $(X_1 = 1)\text{Max}(0, 2682.88 - X_{13})$
BF12 : $(X_5 = 1)\text{Max}(0, X_{13} - 1976.48)$
BF14 : $(X_8 = 5)\text{Max}(0, X_{13} - 2074.3)$
BF15 : $(X_8 = 5)\text{Max}(0, 2074.3 - X_{13})$
BF16 : $(X_4 = 1)\ \text{Max}(0, 2682.88 - X_{13})$
BF17 : $X_5 = 3$
BF18 : $\text{Max}(0, 7.55 - X_{11})\text{Max}(0, X_{13} - 1156.46)$
BF19 : $\text{Max}(0, 7.55 - X_{11})\text{Max}(0, 1156.46 - X_{13})$
BF20 : $(X_6 = 4)\ \text{Max}(0, X_{10} - 22.21)$
BF21 : $(X_6 = 4)\ \text{Max}(0, 22.21 - X_{10})$
BF22 : $(X_8 = 2)\text{Max}(0, 2682.88 - X_{13})$
BF24 : $(X_6 = 4)\text{Max}(0, 2065.03 - X_{13})$
BF25 : $(X_1 = 1)(X_6 = 4)$
BF26 : $(X_4 = 2)(X_5 = 3)$
BF27 : $(X_4 = 2)(X_5 = 1)\text{Max}(0, 1976.48 - X_{13})$
BF28 : $(X_6 = 4)\text{Max}(0, X_{11} - 4.74)$
BF29 : $(X_6 = 4)\text{Max}(0, 4.74 - X_{11})$
BF30 : $\text{Max}(0, X_{10} - 253.56)\text{Max}(0, 2682.88 - X_{13})$
BF31 : $\text{Max}(0, X_{12} - 1.53)\text{Max}(0, 2682.88 - X_{13})$
BF32 : $\text{Max}(0,1.53 - X_{12})\text{Max}(0,2682.88 - X_{13})$
BF34 : $(X_6 = 4)\text{Max}(0, 4.54 - X_{11})\text{Max}(0, 2065.03 - X_{13})$
BF35 : $(X_1 = 1)(X_2 = 3)\text{Max}(0, 2682.88 - X_{13})$

Furthermore, Table 7 presents the confusion matrix values formed by the SMOTE-MARS classification model.

Table 7. Confusion Matrix of SMOTE-MARS Classification Model

Prediction	Actual	
	Non-poor	**Poor**
Non-poor	203	36
Poor	53	210
Percentage Correct (%)	79.29	85.36

3.5 Evaluation of Classification Model Accuracy

Evaluation of the accuracy of the household poverty status classification model in Bengkulu Province based on the APER, TAR, Specificity, and Sensitivity values obtained from the confusion matrix table of the three methods is presented in **Table 8** below:

Table 8. Evaluation of Classification Model Accuracy

Rate	Method		
	MARS	Bagging-MARS	SMOTE-MARS
APER (%)	9.06	8.36	17.73
TAR (%)	90.94	91.64	82.87
Specificity (%)	98.83	98.83	79.29
Sensitivity (%)	25.81	32.26	85.36

In the case of class imbalance, the sensitivity value often takes precedence over the accuracy value (TAR) due to the accuracy value of an imbalanced class distribution can be biased. Unbalanced data between the majority class and the minority class can lead to misclassification because of relying on the majority class in the classification. The sensitivity value can measure the model's capability to correctly detect positive cases or minority classes from all actual positive case data. As shown in **Table 8**, the SMOTE-MARS method has a higher sensitivity value than the MARS and Bagging-MARS methods. The sensitivity value obtained in the SMOTE-MARS method is 85.36%. The SMOTE worked by generating new synthesized data to balance the minority class and the majority class such that in the classification with the MARS method, high classification accuracy is obtained in the minority class. Thus, it can be concluded that the SMOTE-MARS method is more accurate in classifying household status in Bengkulu Province, which has a class imbalance problem.

The modeling results show that there are ten significant predictor variables in the SMOTE-MARS classification model, which are region of residence (X_1), house ownership status (X_2), main wall material (X_4), main floor material(X_5), defecation facilities usage (X_6), main cooking fuel source (X_8), floor area (X_{10}), number of household members (X_{11}), number of families (X_{12}), and daily calorie consumption (X_{13}). This indicates that these predictor variables have a relationship with the response variable of household poverty status in Bengkulu Province. Meanwhile, the other three predictor variables, main roofing material, lighting source, and land ownership status, do not affect the poverty status of households in Bengkulu Province.

The classification model obtained can be interpreted on each basis function formed using the odds model and the odds ratio value. The probability of poor households $\pi(x)$ and the odds ratio value for each basis function in the SMOTE-MARS classification model can be seen in **Table 9**. For illustration, Basis function 3 (BF3) is $\text{Max}(0, 2682.88 - X_{13})$ with a coefficient of 0.00049 and $\pi(x) = 0.54651$. This means that assuming the other predictor variables are constant, households with a daily calorie consumption of less than 2682.88 Kcal have a 0.54651 chance of becoming a poor household. Meanwhile, households with a daily calorie consumption of more than or equal to 2682.88 Kcal have zero basis function coefficient. The odds ratio value of 1.00049 can be interpreted that households with a daily calorie consumption of less than 2682.88 Kcal are 1.00049 times more likely to become poor households than households with a daily calorie consumption of more than or equal to 2682.88 Kcal.

Table 9. Probability Value and Odds Ratio of the SMOTE-MARS Classification Model

No		Basis Function	Probability	Odds Ratio
1	BF1	–	0.54639	1.20452
2	BF2	$\text{Max}(0, X_{13} - 2682.88)$	0.54633	0.99979
3	BF3	$\text{Max}(0, 2682.88 - X_{13})$	0.54651	1.00049
4	BF4	$X_5 = 1$	0.49707	0.82053
5	BF5	$\text{Max}(0, X_{10} - 120)\text{Max}(0, 2682.88 - X_{13})$	0.54639	1.00001
6	BF6	$\text{Max}(0, 120 - X_{10})\text{Max}(0, 2682.88 - X_{13})$	0.54639	1.00001
7	BF7	$\text{Max}(0, X_{11} - 7.55)$	0.49641	0.81837
8	BF8	$\text{Max}(0, 7.55 - X_{11})$	0.52344	0.91189
9	BF9	$X_8 = 5$	0.62061	1.35804
10	BF10	$X_6 = 4$	0.66835	1.67307
11	BF11	$(X_1 = 1)\text{Max}(0, 2682.88 - X_{13})$	0.54641	1.00010
12	BF12	$(X_5 = 1)\text{Max}(0, X_{13} - 1976.48)$	0.54642	1.00015

No		Basis Function	Probability	Odds Ratio
13	BF14	$(X_8 = 5)\text{Max}(0, X_{13} - 2074.3)$	0.54634	0.99981
14	BF15	$(X_8 = 5)\text{Max}(0, 2074.3 - X_{13})$	0.54629	0.99963
15	BF16	$(X_4 = 1)\ \text{Max}(0, 2682.88 - X_{13})$	0.54642	1.00015
16	BF17	$X_5 = 3$	0.60967	1.29672
17	BF18	$\text{Max}(0, 7.55 - X_{11})\text{Max}(0, X_{13} - 1156.46)$	0.54639	1.00004
18	BF19	$\text{Max}(0, 7.55 - X_{11})\text{Max}(0, 1156.46 - X_{13})$	0.54628	0.99957
19	BF20	$(X_6 = 4)\ \text{Max}(0, X_{10} - 22.21)$	0.54671	1.00129
20	BF21	$(X_6 = 4)\ \text{Max}(0, 22.21 - X_{10})$	0.53982	0.97390
21	BF22	$(X_8 = 2)\text{Max}(0, 2682.88 - X_{13})$	0.54632	0.99975
22	BF24	$(X_6 = 4)\text{Max}(0, 2065.03 - X_{13})$	0.54628	0.99959
23	BF25	$(X_1 = 1)(X_6 = 4)$	0.48885	0.79399
24	BF26	$(X_4 = 2)(X_5 = 3)$	0.49774	0.82275
25	BF27	$(X_4 = 2)(X_5 = 1)\text{Max}(0, 1976.48 - X_{13})$	0.54647	1.00034
26	BF28	$(X_6 = 4)\text{Max}(0, X_{11} - 4.74)$	0.51424	0.87889
27	BF29	$(X_6 = 4)\text{Max}(0, 4.74 - X_{11})$	0.51777	0.89139
28	BF30	$\text{Max}(0, X_{10} - 253.56)\text{Max}(0, 2682.88 - X_{13})$	0.54638	1.00000
29	BF31	$\text{Max}(0, X_{12} - 1.53)\text{Max}(0, 2682.88 - X_{13})$	0.54636	0.99988
30	BF32	$\text{Max}(0, 1.53 - X_{12})\text{Max}(0, 2682.88 - X_{13})$	0.54634	0.99982
31	BF34	$(X_6 = 4)\text{Max}(0,4.54 - X_{11})\text{Max}(0, 2065.03 - X_{13})$	0.54645	1.00025
32	BF35	$(X_1 = 1)(X_2 = 3)\text{Max}(0, 2682.88 - X_{13})$	0.54633	0.99978

4. CONCLUSIONS

By three methods, MARS, Bagging-MARS, and SMOTEl-MARS, for the accuracy of household poverty status classification in Bengkulu Province, the best MARS model for the accuracy of household poverty status classification in Bengkulu Province is obtained from the SMOTE-MARS model with the combination of BF = 48, MI = 3, and MO = 1 which produces the minimum GCV value of 0.130805. The SMOTE-MARS classification model provides a higher sensitivity value than MARS and Bagging-MARS, which is 85.36%. This implies that SMOTE-MARS has the best accuracy in predicting observations, especially in the minority class. Moreover, the results showed that there are ten significant predictor variables in the SMOTE-MARS classification model, such as region of residence, house ownership status, main wall material, main floor material, defecation facilities usage, main cooking fuel source, floor area, number of households members, number of families, and daily calorie consumption. This indicates that these predictor variables have a significant influence on the response variable of household poverty status in Bengkulu Province.

ACKNOWLEDGMENT

This work is supported by LPPM Bengkulu University, PNBP Funding Fundamental Research Scheme [Grant No. 2942/UN30.15/PT/2024].

REFERENCES

[1] D. Barry and W. Hardle, *APPLIED NONPARAMETRIC REGRESSION.*, 1st ed. Cambridge University Press., 1994. doi: 10.2307/2982873.

[2] J. Han, J. Pei, and H. Tong, *DATA MINING CONCEPTS AND TECHNIQUES*, 4th ed. Morgan Kaufmann, 2023.

[3] W. Zhang, A. T. C. Goh, and Y. Zhang, “MULTIVARIATE ADAPTIVE REGRESSION SPLINES APPLICATION FOR MULTIVARIATE GEOTECHNICAL PROBLEMS WITH BIG DATA,” *Geotech. Geol. Eng.*, vol. 34, no. 1, pp. 193–204, 2016, doi: 10.1007/s10706-015-9938-9.

[4] D. Çanga, “USE OF MARS DATA MINING ALGORITHM BASED ON TRAINING AND TEST SETS IN

DETERMINING CARCASS WEIGHT OF CATTLE IN DIFFERENT BREEDS," *Tarim Bilim. Derg.*, vol. 28, no. 2, pp. 259–268, 2022, doi: 10.15832/ankutbd.818397.

[5] J. M. Johnson and T. M. Khoshgoftaar, "SURVEY ON DEEP LEARNING WITH CLASS IMBALANCE," *J. Big Data*, vol. 6, no. 1, pp. 1–54, 2019, doi: 10.1186/s40537-019-0192-5.

[6] Tamonob, Onisimus, Sumertajaya, I. Made, Rahman, and L. O. Abdul, "ANALISIS MULTIVARIATE ADAPTIVE REGRESSION SPLINES (MARS) UNTUK MENGKLASIFIKASIKAN STATUS DESA DI PROVINSI NUSA TENGGARA TIMUR," Institute Pertanian Bogor, 2020.

[7] R. D. L. N. Karisma, J. Juhari, and R. A Rosa, "POVERTY IN CENTRAL JAVA USING MULTIVARIATE ADAPTIVE REGRESSION SPLINES AND BOOTSTRAP AGGREGATING MULTIVARIATE ADAPTIVE REGRESSION SPLINES," *CAUCHY J. Mat. Murni dan Apl.*, vol. 6, no. 4, pp. 238–245, 2021, doi: 10.18860/ca.v6i4.10871.

[8] M. Hasyim *et al.*, "BOOTSTRAP AGGREGATING MULTIVARIATE ADAPTIVE REGRESSION SPLINES (BAGGING MARS) TO ANALYSE THE LECTURER RESEARCH PERFORMANCE IN PRIVATE UNIVERSITY," *J. Phys. Conf. Ser.*, vol. 1114, no. 1, 2018, doi: 10.1088/1742-6596/1114/1/012117.

[9] B. K. Kilinc, S. Malkoc, A. S. Koparal, and B. Yazici, "USING MULTIVARIATE ADAPTIVE REGRESSION SPLINES TO ESTIMATE POLLUTION IN SOIL," *Int. J. Adv. Appl. Sci.*, vol. 4, no. 2, pp. 10–16, 2017, doi: 10.21833/ijaas.2017.02.002.

[10] Nidhomuddin and B. W. Otok, "RANDOM FOREST DAN MULTIVARIATE ADAPTIVE REGRESSION SPLINE (MARS) BINARY RESPONSE UNTUK KLASIFIKASI PENDERITA HIV/AIDS DI SURABAYA," *Stat. Fak. Mat. dan Ilmu Pengetah. Alam Inst. Teknol. Sepuluh Novemb.*, vol. 1, no. 3, pp. 50–57, 2015.

[11] B. P. Statistik, "PROFIL KEMISKINAN DI INDONESIA MARET 2023," 2023. [Online]. Available: https://www.bps.go.id/pressrelease/2018/07/16/1483/persentase-penduduk-miskin-maret-2018-turun-menjadi-9-82-persen.html

[12] Badan Pusat Statistik, "PROFIL KEMISKINAN PROVINSI BENGKULU MARET 2023," 2023.

[13] M. A. Sahraei, H. Duman, M. Y. Çodur, and E. Eyduran, "PREDICTION OF TRANSPORTATION ENERGY DEMAND: MULTIVARIATE ADAPTIVE REGRESSION SPLINES," *Energy*, vol. 224, pp. 1–9, 2021, doi: 10.1016/j.energy.2021.120090.

[14] D. R. Cox and E. J. Snell, *ANALYSIS OF BINARY DATA*, 2nd ed. CRC PRess, 1989.

[15] B. C. L. Huang, Y. Xiang, and Z. H. Huang, "USE LOGISTIC REGRESSION TO PREDICT USER' BEHAVIORS," *Appl. Mech. Mater.*, vol. 651–653, pp. 1695–1698, 2014, doi: 10.4028/www.scientific.net/AMM.651-653.1695.

[16] A. Agresti, *AN INTRODUCTION TO CATEGORICAL DATA ANALYSIS*, 2nd ed., vol. 28, no. 11. Florida: A John Wiley & Sons, Inc, 2009. doi: 10.1002/sim.3564.

[17] W. Agwil, D. Agustina, H. Fransiska, and N. Hidayati, "KLASIFIKASI KARAKTERISTIK KEMISKINAN DI PROVINSI BENGKULU TAHUN 2020 MENGGUNAKAN METODE POHON KLASIFIKASI GABUNGAN," *J. Apl. Stat. Komputasi Stat.*, vol. 14, no. 2, pp. 23–32, 2022.

[18] D. Elreedy and A. F. Atiya, "A COMPREHENSIVE ANALYSIS OF SYNTHETIC MINORITY OVERSAMPLING TECHNIQUE (SMOTE) FOR HANDLING CLASS IMBALANCE," *Inf. Sci. (Ny).*, vol. 505, pp. 32–64, 2019, doi: 10.1016/j.ins.2019.07.070.