STATISTICS, OPTIMIZATION AND INFORMATION COMPUTING
Stat., Optim. Inf. Comput., Vol. **16**, October 2026, pp 3808–3832.
Published online in International Academic Press (www.IAPress.org)

A Categorical Approach to Classifying Poor Households in East Java: Utilizing Support Vector Machine (SVM) for SDG 1

Vita Ratnasari [1,*], Kadek Adi Surya Negara [1], Andrea Tri Rian Dani [2, 3], Muhammad Fikry Al Farizi [1]

[1] *Department of Statistics, Faculty of Science and Data Analytics, Institut Teknologi Sepuluh Nopember, Surabaya 60111, Indonesia*
[2] *Doctoral Study Program of Mathematics and Natural Sciences, Faculty of Science and Technology, Universitas Airlangga, Surabaya 60115, Indonesia*
[3] *Statistics Study Program, Department of Mathematics, Faculty of Mathematics and Natural Sciences, Universitas Mulawarman, Samarinda 75119, Indonesia*

Abstract Persistent inaccuracies in household poverty classification impede Indonesia's progress toward Sustainable Development Goal 1 (No Poverty). In March 2023, East Java Province reported the highest absolute number of poor residents nationwide, underscoring the need for more reliable, data-driven targeting mechanisms. This study introduces a novel categorical modeling framework that integrates binary logistic regression, based variable selection with Support Vector Machine (SVM) classification, enhanced by Synthetic Minority Oversampling Technique (SMOTE) to address severe class imbalance. From the 2023 National Socioeconomic Survey (SUSENAS), eleven socioeconomic indicators were identified as statistically significant predictors of poverty status: household size, urban–rural residence, head-of-household employment, literacy, wall material, floor area, sanitation facility, primary drinking-water source, land ownership, motorcycle ownership, and ICT asset ownership. We then constructed linear and Radial Basis Function (RBF) SVM classifiers, each trained on an 80% SMOTE-augmented dataset. The linear-kernel SVM, $(C = 0.1)$ emerged as the best performer, achieving 78.73% overall accuracy, 77.07% sensitivity (poor-household recall), 78.82% specificity (non-poor recall), a geometric mean of 77.94%, and an AUC of 77.94%. Compared to RBF-kernel alternatives, the linear model delivered superior balance between minority-class and majority-class discrimination without incurring additional computational complexity. This research offers policymakers a robust, scalable tool for rapid poverty monitoring and targeted intervention design by demonstrating that a rigorously tuned, categorical SVM framework can substantially improve subnational poverty classification. Future work should explore alternative cross-validation schemes and nonlinear kernels to optimize predictive performance further.

Keywords Household of Poverty Classification, Poor Households, Support Vector Machine, Socioeconomic Indicators, SDGs 1

AMS 2010 subject classifications 62H30, 62J12, 62P25

DOI: 10.19139/soic-2310-5070-3755

1. Introduction

Indonesia, as a developing country with the largest population in Southeast Asia, faces serious challenges in reducing poverty [1]. According to data from the Central Statistics Agency (*Badan Pusat Statistik* (BPS)), Indonesia's national poverty rate in March 2023 reached 9.36%, significantly higher than the target set in the 2020–2024 National Medium-Term Development Plan (*Rencana Pembangunan Jangka Menengah Nasional* (RPJMN)), which is 6.5%–7.5%. Despite various efforts, many people still have not felt the full benefits of

*Correspondence to: Vita Ratnasari (Email: vitaratnasari.its@gmail.com). Department of Statistics, Faculty of Science and Data Analytics, Institut Teknologi Sepuluh Nopember. Teknik Mesin Street No. 175, Keputih, Sukolilo, Surabaya 60111, Indonesia.

ISSN 2310-5070 (online) ISSN 2311-004X (print)

poverty alleviation programs [2–4]. Furthermore, inaccuracies in classifying household poverty status constitute a significant obstacle to the success of this policy [5, 6]. One increasingly emerging approach is using machine learning technology to more accurately classify household poverty status, focusing on more holistic and measurable socioeconomic factors.

Poverty is one of the main challenges in the Sustainable Development Goals (SDGs), especially the first goal, which aims to end poverty in all its forms and dimensions worldwide by 2030 [7–9]. In the context of the SDGs, poverty is measured by income and access to basic services such as education, health, housing, and opportunities to participate in the economy [10]. SDG 1 also includes broader indicators such as access to productive resources and resilience to economic risks. Therefore, to achieve the SDG 1 target, Indonesia needs to improve the government's classification method of poor households [11, 12]. More precisely, data-driven methods will support more efficient and targeted aid distribution [13, 14].

Poverty is when an individual or household cannot meet the basic needs for a decent life, such as food, clothing, health care, and access to other basic services [15]. BPS defines the poverty line based on the minimum expenditure required to meet these basic needs [16]. However, this poverty line only considers economic factors, while non-economic factors, such as education level, employment, and access to resources, also greatly influence a household's poverty status [17–19]. Thus, a poverty classification based solely on the poverty line may not be sufficient to capture the true complexity of poverty. Therefore, a more comprehensive approach is needed to determine household poverty status, which includes a broader range of socio-economic factors [20].

To address these challenges, machine learning technology, specifically the Support Vector Machine (SVM), offers the opportunity to improve the accuracy of classifying household poverty status [21]. SVM works by finding a hyperplane that maximizes the margin between two distinct classes, in this case, poor and non-poor households. The main advantage of SVM lies in its ability to handle high-dimensional and complex data [22–24]. Using kernels, SVM can transform non-linear data into a higher-order feature space, enabling more accurate separation between classes that are difficult to distinguish in the original space [25, 26]. In the context of household poverty classification, SVM can effectively utilize various socioeconomic variables with non-linear relationships. Compared with conventional classification methods such as logistic regression or Naive Bayes, SVM has advantages in terms of accuracy and generalization capabilities [27]. Although frequently used in classification problems, logistic regression tends to be limited to linear relationships between variables and outputs, making it less effective in handling data with more complex relationships. Meanwhile, Naive Bayes assumes independence between features, which often does not correspond to real-world data with dependencies between variables. With its ability to handle linearly non-separable data and consider optimal separation margins, SVM provides more accurate and robust results in classifying poor and non-poor households.

Factors determining poverty can be divided into three main categories: regional, community, and individual and household characteristics [6]. These characteristics are interrelated and influence the socioeconomic conditions of households [20]. For example, education and access to decent work can affect household income and economic resilience. In addition, social characteristics such as marital status and family size can also impact a household's ability to meet basic needs. Therefore, it is essential to consider these factors in poverty classification. BPS has established 14 criteria for identifying poor households eligible for social assistance, but in practice, many households are not accurately identified due to the limitations of these criteria.

Poverty reduction remains a significant challenge in East Java, the province with the largest number of poor people in Indonesia. In 2023, the number of poor people in East Java reached 4.19 million. Although various social assistance programs have been launched, their distribution has not been optimal due to inaccuracies in classifying poor households. Therefore, optimizing the classification of low-income families is crucial to ensure more targeted social assistance.

This research employs Support Vector Machine (SVM) to explore the key factors influencing poverty status in East Java, while simultaneously comparing the performance of two SVM approaches: linear classification using a linear kernel and non-linear classification through a Radial Basis Function (RBF) kernel. Given the imbalanced nature of the dataset, wherein non-poor households significantly outnumber low-income families, the study will implement the Synthetic Minority Oversampling Technique (SMOTE) to mitigate class imbalance. By

incorporating SMOTE into the SVM classification process, the study seeks to improve the model's accuracy in classifying poor and non-poor households, thereby enhancing predictive outcomes.

This approach represents a novel contribution to the field. Additionally, the research aims to identify the most effective SVM method for distinguishing between poor and non-poor households in East Java and to uncover the primary socioeconomic factors that influence poverty status. The results of this study are expected to provide valuable insights for policymakers, enabling the development of more targeted and effective poverty reduction strategies. Unlike the official poverty measurement approach used by Statistics Indonesia (BPS), which determines poverty status by comparing household per capita expenditure with the official poverty line, this study proposes a predictive classification model based on household socioeconomic characteristics. Therefore, the proposed model is intended to complement, rather than replace, the existing official methodology by providing an alternative tool for identifying households at risk of poverty when detailed expenditure data are unavailable or for supporting the initial targeting of social assistance programs. In doing so, this research will contribute to advancing the scientific understanding of data classification techniques and support the formulation of evidence-based poverty alleviation policies in Indonesia, aligning with the United Nations Sustainable Development Goal 1, which aims to eradicate poverty globally by 2030.

2. Materials and Methods

2.1. Literature Review

2.1.1 Data Scale Standardization

Data preprocessing plays a vital role in enhancing the accuracy of machine learning models. Min-Max normalization is commonly used to scale continuous variables by transforming their values into a range between 0 and 1 [28]. This ensures that all variables are on the same scale, preventing any one variable from disproportionately affecting the model's performance. The Min-Max normalization is expressed in Equation (1).

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{1}$$

One-Hot Encoding is used to convert categorical variables into binary format, where each category is represented by a separate binary column. This allows machine learning models to process categorical data effectively, without implying any ordinal relationships between categories.

2.1.2 Binary Logistic Regression

Binary logistic regression is a regression analysis used to describe the relationship between a set of predictor variables either continuous or categorical and a binary response variable [29]. The binary logistic regression equation is derived from the interpretation of the probability function $\pi(\boldsymbol{X}_i) = \mathbb{E}(Y_i \mid \boldsymbol{X}_i)$, where $\pi(\boldsymbol{X}_i)$ is expressed in Equation (2) [30].

$$\pi(\boldsymbol{X}_i) = \frac{\exp\left(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_p X_{ip}\right)}{1 + \exp\left(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_p X_{ip}\right)} \tag{2}$$

In binary logistic regression, hypothesis testing is conducted to examine the significance of the parameter coefficients β in the model [31, 32]. The testing is carried out in two stages: simultaneous (overall) testing and partial (individual) testing [30].

1. **Simultaneous Test**
   The simultaneous test evaluates the overall significance of the coefficient $\boldsymbol{\beta}$ in the model. This statistical test assesses whether the estimated coefficient significantly explains the dependent variable. The test is based on formulating null and alternative hypotheses, where the null hypothesis generally posits that the coefficient equals zero (indicating no effect). In contrast, the alternative hypothesis suggests that the coefficient differs from zero (indicating a significant effect) [33]. This approach is crucial in determining the relevance of the

predictor variables in the model.

$\mathrm{H}_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0 \ (\boldsymbol{\beta} = \mathbf{0})$
(The predictor variables do not have a significant effect on the response variable);
H_1 : at least one $\beta_j \neq 0, \ j = 1, \ldots, p \ (\exists \ j \in \{1, \ldots, p\}$ such that $\beta_j \neq 0)$
(At least one predictor variable has a significant effect on the response variable).

The test statistic for the simultaneous test is presented in Equation (3).

$$G = -2 \ln \left(\frac{\left[\frac{n_1}{n}\right]^{n_1} \left[\frac{n_0}{n}\right]^{n_0}}{\prod_{i=1}^{n} (\pi_i)^{(y_i)} (1 - \pi_i)^{(1-y_i)}} \right) \tag{3}$$

The decision for the simultaneous test is determined using a significance level of α, where $\alpha = 0.05$ and the degrees of freedom are p. The null hypothesis H_0 is rejected when the value of $G > \chi^2_{(\alpha;\, p)}$ or when the $p - value < \alpha$.

2. **Partial Test**
The partial test is conducted to identify which specific variables significantly affect the response variable. Unlike the simultaneous test, which evaluates the overall significance of the coefficients, the partial test assesses the significance of each coefficient β individually for each predictor variable. In this test, the null hypothesis typically states that the coefficient for a given predictor is equal to zero (indicating no effect). In contrast, the alternative hypothesis posits that the coefficient differs from zero (indicating a significant effect). This method helps in isolating the contribution of each variable to the model and determining which factors play a significant role in explaining the variation in the response variable [34, 35].

$\mathrm{H}_0 : \beta_j = 0, \ j = 1, \ldots, p$
(The j^{th} predictor variable does not have a significant effect on the response variable);
$\mathrm{H}_1 : \beta_j \neq 0, \ j = 1, \ldots, p$
(The j^{th} predictor variable has a significant effect on the response variable).

The test statistic for the partial test is presented in Equation (4).

$$W^2 = \frac{\beta_j^2}{\left[SE\left(\beta_j\right)\right]^2} \tag{4}$$

2.1.3 K–Fold Cross Validation

K–Fold Cross Validation is a widely used method for evaluating machine learning models. The dataset is divided into k equal-sized folds, and the model is trained and tested k times, with each fold serving as the test set once while the remaining folds are used for training. This technique helps ensure that each data point is used for both training and testing, which provides a more reliable estimate of the model's generalization ability [36, 37]. By averaging the performance metrics from all iterations, K–Fold Cross Validation reduces bias and ensures that the model is not overly influenced by a particular train-test split. This is especially important for datasets with limited samples or variability [38]. For Support Vector Machines (SVM), K–Fold Cross Validation is particularly valuable as it helps optimize hyperparameters such as the regularization parameter C and kernel functions while minimizing the risk of overfitting. Overall, K–Fold Cross Validation provides a more comprehensive evaluation, making it an essential tool for assessing the performance and generalization ability of models, especially in classification tasks like those using SVM [22].

2.1.4 Synthetic Minority Over-Sampling Technique (SMOTE)

SMOTE (Synthetic Minority Over-sampling Technique) is a widely used method to address class imbalance in machine learning datasets [39]. This oversampling technique aims to balance the dataset by generating synthetic instances for the minority class, thereby equalizing its quantity with that of the majority class [40]. Unlike simple random replication, SMOTE creates new data points by interpolating between existing instances of the minority class. Specifically, it uses the k-nearest neighbors (k-NN) algorithm to identify similar instances and generates synthetic samples by taking the difference between feature vectors of neighboring instances, scaling this difference by a random value, and then adding it to the feature vector of the minority class sample. The Euclidean distance can be expressed in Equation (5).

$$x_{knn} = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \ldots + (x_n - y_n)^2} \tag{5}$$

In general, the process of generating synthetic data can be expressed in Equation (6).

$$x_{syn} = x_i + (x_{knn} - x_i) \times \delta \tag{6}$$

Here, x_{syn} represents the synthetic data generated through replication, x_i denotes the data to be replicated, x_{knn} refers to the data point that is closest to x_i based on distance, and δ is a random number between 0 and 1.

2.1.5 Support Vector Machine (SVM)

SVM is a supervised learning algorithm used to solve classification problems by finding a hyperplane that can separate two sets of data belonging to different classes [22, 41]. SVM is known for its high level of accuracy and performs well in both linear and nonlinear cases, making it a commonly used method for classification tasks, especially when the target variable is binary [42].

1. **Support Vector Machine in Linear Classification**
   Linear SVM classification uses a decision boundary that determines the classification outcome, allowing the linear model or hyperplane to be formed optimally [24]. Linear SVM separates paired classes by grouping each pair of data points that belong to different categories [43]. An illustration of linear SVM with separated data is shown in Figure 1.

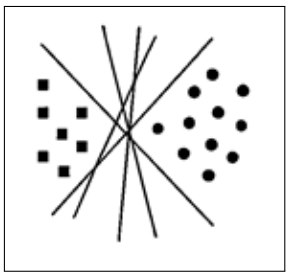

Figure 1. Classification of Linearly Separable Data

Each classified data point that is separated by the decision boundary, also known as the separating hyperplane, is calculated using Equation (7).

$$f(x) = x^\top w + b = 0 \tag{7}$$

If the value of $f(x) > 0$, the data point is classified into class 1, whereas if $f(x) < 0$, it is classified into class 0. The linear SVM applies a classification rule that is expressed in Equation (8).

$$f(x) = \text{sign}\left(x^\top \hat{w} + \hat{b}\right) \tag{8}$$

The values of $\hat{w}$ and $\hat{b}$ are obtained from Equation (9) and Equation (10), respectively.

$$\hat{w} = \sum_{i=1}^{n} a_i y_i x_i \tag{9}$$

$$\hat{b} = \frac{1}{n}\left(\sum_{i=1}^{n} y_i - x_i^{\top} w\right) \tag{10}$$

In Equation (9) the value of a_i is controlled by the Cost Parameter (C). The Cost Parameter is a turning parameter used in the modelling of linear SVM. The value of a_i falls within the range $0 \leq a_i \leq C$.

2. **Support Vector Machine in Nonlinear Classification**
Most real-world problems are nonlinear in nature, making it difficult to optimally separate training data using a linear approach [22, 27]. SVM was therefore developed to address this issue by utilizing kernel functions, allowing it to effectively separate nonlinear data [42, 44]. An illustration of nonlinear SVM mapping such as mapping into three dimensions is shown in Figure 2.

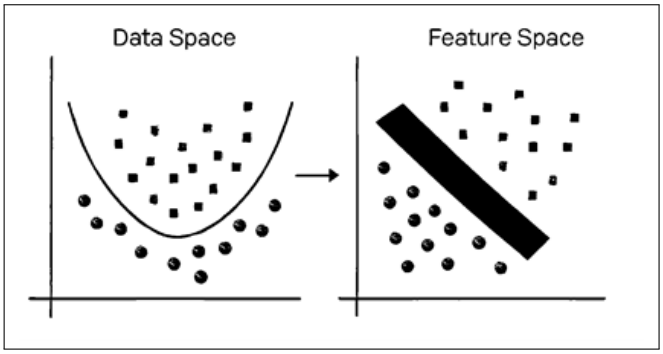

Figure 2. Mapping to Three-dimensional Space in Nonlinear SVM

The classification rule in nonlinear SVM can be expressed in Equation (11).

$$f(x) = \text{sign}\left(\sum_{i=1}^{n}\sum_{j=1}^{n} a_i y_i \,\mathrm{K}\left(x_i,\, x_j\right) + b\right) \tag{11}$$

The value of b is obtained from Equation (12).

$$\hat{b} = \frac{1}{n}\left(\sum_{i=1}^{n} y_i - \sum_{i=1}^{n}\sum_{j=1}^{n} a_i y_i \,\mathrm{K}\left(x_i,\, x_j\right)\right) \tag{12}$$

In Equation (11) and (12), the value of a_i like that in linear SVM falls within the range $0 \leq a_i \leq C$, where the Cost Parameter (C) serves to control the value of a_i. SVM provides several types of kernels, including the Polynomial Kernel, Radial Basis Function (RBF) Kernel, Sigmoid Kernel, and others. In this study, the RBF kernel is used, as expressed in Equation (13).

$$\mathrm{K}\left(x_i,\, x_j\right) = \exp\left(-\frac{\|x_i - x_j\|}{2\sigma^2}\right) \tag{13}$$

2.1.6 Classification Accuracy

The performance of a classification model can be measured and evaluated using the confusion matrix method [18]. A confusion matrix compares the classification results of the model, specifically the predicted values and the actual values. These values are obtained through the calculation of the confusion matrix, as shown in Table 1.

Table 1. Confusion Matrix

Actual Value	**Predicted Value**	
	1 (Positive Class)	0 (Negative Class)
1 (Positive Class)	True Positive (TP)	False Negative (FN)
0 (Negative Class)	False Positive (FP)	True Negative (TN)

The confusion matrix provides the basis for calculating five classification accuracy metrics, which are expressed in Equations (14) through (18).

$$\text{Accuracy} = \frac{TP + TN}{TP + FN + FP + TN} \tag{14}$$

$$\text{Sensitivity} = \frac{TP}{TP + FN} \tag{15}$$

$$\text{Specificity} = \frac{TN}{FP + TN} \tag{16}$$

$$G - Mean = \sqrt{\text{Sensitivity} \times \text{Specificity}} \tag{17}$$

$$\text{AUC} = \frac{1}{2}\left(\text{Sensitivity} + \text{Specificity}\right) \tag{18}$$

Sensitivity, also known as Recall or the True Positive Rate (TPR), measures the model's ability to correctly identify positive instances by comparing the true positives to the actual positive instances. Specificity, or the True Negative Rate (TNR), reflects the model's ability to correctly identify negative instances by comparing the true negatives to the total negative ones.

The $G - Mean$ is the geometric mean of sensitivity and specificity, offering a balanced performance measure, especially in imbalanced datasets. It combines both the ability to correctly identify positive and negative cases. Finally, the Area Under the Curve (AUC) is the average of sensitivity and specificity, providing a single measure of the model's ability to distinguish between the positive and negative classes, with an AUC of 1 indicating perfect classification and an AUC of 0.5 suggesting no better performance than random guessing [45].

2.2. Research Methodology

2.2.1 Data Source

In this study, the data were obtained from the National Socio-Economic Survey (*Survei Sosial Ekonomi Nasional* (*Susenas*)) conducted by Statistics Indonesia (*Badan Pusat Statistik* (BPS)) of East Java Province in March 2023. The predictor variables were sourced from the *Susenas*, while the response variable was obtained from the *Susenas* Consumption and Expenditure Module. The unit of analysis in this study consists of 32,544 households spread across 38 regencies/ cities in East Java.

2.2.2 Research Variable

The response variable (Y) and predictor variables (X) used in this study are described in Table 2.

Table 2. Research Variables and Operational Definitions

Variables	Name	Scale	Description
Y	Household of Poverty Status	Nominal	0: Non-Poor Household 1: Poor Household
X_1	Gender of Household Head	Nominal	0: Male 1: Female
X_2	Age of Household Head	Ratio	Age in years
X_3	Highest Education Certificate of Household Head	Nominal	0: Primary School or Lower 1: Junior High School 2: Senior High School 3: Higher Education
X_4	Number of Household Members	Ratio	Number of persons
X_5	Area of Residence	Nominal	0: Urban 1: Rural
X_6	Employment Status of Household Head	Nominal	1: Laborer/ Employee 2: Casual Worker 3: Family Worker/ Unpaid 4: Unemployed
X_7	Functional Disability Status of Household Head	Nominal	0: No Disability 1: Has Disability
X_8	Literacy Status of Household Head	Nominal	0: Illiterate 1: Literate
X_9	Type of Roof Material	Nominal	0: Concrete/ Tiles 1: Zinc/ Asbestos 2: Bamboo/ Wood/ Thatch/ Leaves/ Rumbia/ Others
X_{10}	Type of Wall Material	Nominal	0: Bricks and Others 1: Bamboo and Others
X_{11}	Main Lighting Source	Nominal	0: No Electricity 1: PLN/Non-PLN Electricity
X_{12}	Floor Area	Ratio	Area in square meters (m^2)
X_{13}	Type of Toilet	Nominal	0: Private 1: Shared 2: None/ Open Defecation
X_{14}	Main Source of Drinking Water	Nominal	0: Bottled/ Refilled Water 1: Piped/ Spring/ Rainwater/ etc.
X_{15}	Main Source of Water for Bathing/ Washing	Nominal	0: Bottled/ Refilled Water 1: Well/ Pump/ Piped Water
X_{16}	Land Ownership	Nominal	0: Does Not Own 1: Owns

Variables	Name	Scale	Description
X_{17}	Ownership of Motorbike	Nominal	0: Does Not Own 1: Owns
X_{18}	Ownership of Computer/ Laptop/ Tablet	Nominal	0: Does Not Own 1: Owns

2.2.3 Analysis Step

The steps taken to achieve the research objectives are as follows:

1. Classify households into poor and non-poor categories using the poverty line specific to each regency/ city.
2. Conduct descriptive statistical analysis to summarize the characteristics of the data.
3. Normalize continuous variables using the Min-Max normalization method and apply One-Hot Encoding for categorical variables.
4. Perform variable selection using significance testing based on binary logistic regression.
5. Split the data into training and testing sets using the K-Fold Cross-Validation method.
6. Apply oversampling to the training data using the SMOTE method, following its standard procedures.
7. Perform classification using the Support Vector Machine (SVM) method, with the following steps:
   (a) Build SVM models using the training data with both linear and Radial Basis Function (RBF) kernels.
   (b) Make predictions on the test data using both SVM models.
   (c) Select the best model based on a comparison of the performance of the two models.
8. Analyze classification accuracy and interpret the results of the best-performing model.
9. Conclude and provide recommendations based on the research findings.

3. Results and Discussion

3.1. Characteristics of Research Variables

Household poverty status is divided into two categories: poor households and non-poor households. The comparison between these two categories is presented in Figure 3.

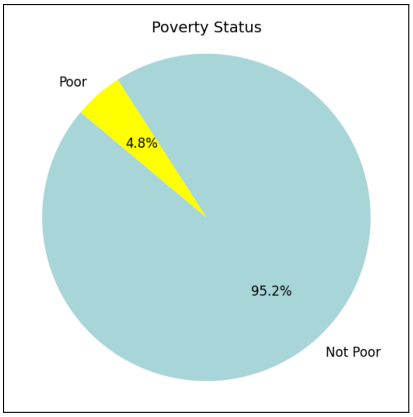

Figure 3. Comparison of Household Poverty Status in East Java

Based on Figure 3, data from East Java reveal that 4.8% of households, or $1,570$ families, are classified as poor, while the remaining 95.2%, or $30,974$ households, are categorized as not poor. This significant disparity in the distribution of households highlights the imbalance between the two categories, with the "not poor" group vastly outnumbering the "poor" group. To address this class imbalance in the predictive modeling process, techniques such as SMOTE (Synthetic Minority Over-sampling Technique) are essential to prevent the model from being biased towards the majority class, ensuring more accurate and balanced predictions.

Based on Table 2, a total of 18 predictor variables were used, representing three dimensions. The variables head of household's gender (X_1), age of the head of household (X_2), and number of household members (X_4) represent the demographic dimension. The variables main employment status of the head of household (X_6), ownership of land/plots (X_{16}), ownership of motorbike (X_{17}), and ownership of computer/ laptop/ tablet (X_{18}) represent the economic dimension. Meanwhile, the variables highest education certificate of household head (X_3), residence area status (X_5), functional disability status of household head (X_7), literacy status of household head (X_8), type of roof material (X_9), type of wall material (X_{10}), floor area of the house (X_{12}), main lighting source (X_{11}), type of toilet (X_{13}), main source of drinking water (X_{14}), and main source of water for bathing/ washing/ etc (X_{15}) represent the social dimension.

Before conducting the analysis, multicollinearity among the predictor variables was assessed using the Variance Inflation Factor (VIF). The results showed that all 18 predictor variables had VIF values below 5, indicating that there was no evidence of multicollinearity among the predictor variables used in this study.

3.2. Variable Selection Using Binary Logistic Regression Significance Testing

The significance testing of binary logistic regression is conducted on the parameter β_i for each variable. Before testing each variable, scale standardization is performed to ensure the data has the same scale and prevent any disparities. The first test conducted is the simultaneous test.

Table 3. Simultaneous Test Results of Binary Logistic Regression

G	$P-value$	$\chi^2_{(0.05,\ 17)}$	**Decision**
2237.271	0.000	27.587	Reject H_0

Table 3 shows that H_0 (Null Hypothesis) is rejected, indicating that there is at least one predictor variable that significantly affects the response variable. After performing the simultaneous test and obtaining a significant result, the next step is to conduct a partial test.

Table 4. Partial Test Results After Elimination

Variable	W^2	$\chi^2_{(0.05,\ 1)}$	$P-value$	**Decision**
X_4	1253.913		0.000	Reject H_0
X_5	25.450		0.000	Reject H_0
$X_{6(0)}$	27.828		0.000	Reject H_0
$X_{6(1)}$	46.858		0.000	Reject H_0
$X_{6(2)}$	26.853		0.000	Reject H_0
$X_{6(3)}$	4.006		0.045	Reject H_0
X_8	33.498		0.000	Reject H_0
$X_{10(0)}$	85.601		0.000	Reject H_0
$X_{10(1)}$	7.818	3.841	0.005	Reject H_0
X_{12}	126.337		0.000	Reject H_0
$X_{13(0)}$	7.308		0.007	Reject H_0
$X_{13(1)}$	2.214		0.000	Reject H_0

Variable	W^2	$\chi^2_{(0.05,\ 1)}$	$P-value$	Decision
$X_{14(0)}$	50.158		0.000	Reject H_0
$X_{14(1)}$	7.893		0.005	Reject H_0
X_{16}	4.982		0.026	Reject H_0
X_{17}	239.513		0.000	Reject H_0
X_{18}	89.696		0.000	Reject H_0

Table 4 shows the results of the partial test after the elimination process, revealing several predictor variables that are significant. There are 11 predictor variables that significantly affect the response variable. These variables are: number of household members (X_4), residential area status (X_5), employment status of the head of household (X_6), literacy status of the head of household (X_8), main wall material of the house (X_{10}), floor area of the house (X_{12}), toilet usage status (X_{13}), main source of drinking water (X_{14}), ownership of land/plots (X_{16}), ownership of motorcycles (X_{17}), and ownership of computers/ laptops/ tablets (X_{18}). These variables are considered significant as indicated by the test statistic value $W > \chi^2_{(0.05,\ 1)}$ and $p-value$ less than 0.05. These significant variables will be further analyzed using SVM modeling to classify household poverty status in East Java Province.

The results of the significance test show that household poverty in East Java Province in 2023 is influenced by demographic conditions, housing quality, access to basic facilities, and asset ownership. The number of household members (X_4) is one of the important predictors because the number of household members will affect the household poverty status. The employment status of the head of the household (X_6) and the literacy status of the head of the household (X_8) also play an important role in determining poverty conditions. Households with heads who have stable jobs and good literacy skills generally have a greater chance of obtaining stable income and better access to information, education, and job opportunities. Conversely, limitations in these aspects can reduce the household's ability to improve their economic well-being. Characteristics of the residence, such as the main wall material (X_{10}), floor area (X_{12}), status of toilet facility usage (X_{13}), and main source of drinking water (X_{14}), are indicators that reflect the quality of life and the level of household welfare. This shows that the characteristics of the residence can influence the classification of whether a household falls into the poor or non-poor category. In addition, asset ownership such as land (X_{16}), motorcycles (X_{17}), and computers, laptops, or tablets (X_{18}) also serve as important indicators in distinguishing between poor and non-poor households. The ownership of these assets indicates the economic condition of the household. The economic condition will certainly influence the classification of poor and non-poor households.

3.3. Classification Using Support Vector Machine (SVM)

The classification of household poverty status in East Java Province is carried out using the Support Vector Machine (SVM) method with two models: the linear kernel model and the Radial Basis Function (RBF) kernel model. The linear kernel requires a Cost (C) parameter, while the RBF kernel requires both Cost (C) and Gamma (γ) parameters. The percentage of oversampling using the SMOTE method was tested at increases of 40%, 60%, and 80% with the set k value being $k = 5$. In addition, combinations of the parameters C and α are also explored. The dataset is divided using 10-fold Cross-validation, where each experiment produces 10 values for $\mathrm{Accuracy}$, $\mathrm{Sensitivity}$, $\mathrm{Specificity}$, $G-Mean$, and AUC, which are then averaged to obtain a stable classification result. The optimal parameters for C and γ are those that can classify poor and non-poor classes in a balanced manner.

3.3.1 SVM Classification with Linear Kernel

The parameter C in the SVM model with a linear kernel is tuned on the training data and tested on the testing data. The values of C used are 0.01, 0.1, 1, and 10, resulting in 12 combinations when combined with different SMOTE oversampling variations. The results of the modeling with combinations of SMOTE and parameter C are presented in Table 5.

Table 5. Linear Kernel Classification on Testing Data

C	Accuracy	Sensitivity	Specificity	$G-Mean$	AUC
Oversampling 40%					
0.01	92.35%	25.67%	95.73%	49.52%	60.70%
0.1	89.87%	38.92%	92.45%	59.90%	65.68%
1	89.18%	41.34%	91.60%	61.45%	66.47%
10	89.02%	41.53%	91.43%	61.53%	66.48%
Oversampling 60%					
0.01	85.53%	53.19%	87.17%	68.07%	70.18%
0.1	83.91%	57.58%	85.25%	70.02%	71.41%
1	83.52%	58.28%	84.8%	70.26%	71.54%
10	83.51%	58.79%	84.76%	70.56%	71.78%
Oversampling 80%					
0.01	78.99%	66.62%	79.61%	72.81%	73.12%
0.1	79.00%	68.41%	79.53%	73.73%	73.97%
1	**79.01%**	**67.96%**	**79.57%**	**73.50%**	**73.77%**
10	78.97%	67.90%	79.53%	73.45%	73.72%

Table 5 shows that an 80% oversampling yields higher average classification performance in terms of Sensitivity and $G-Mean$ compared to other oversampling levels. With 80% oversampling, the best-performing Cost parameter (C) for classifying "poor" and "non-poor" households is $C=0.1$. This model is considered the best because it achieves the highest average $G-Mean$ of 73.73% and a reasonably high average sensitivity of 68.41%. Regarding the addition of 80% oversampling applied to the minority class training data, it is visualized in Figure 4.

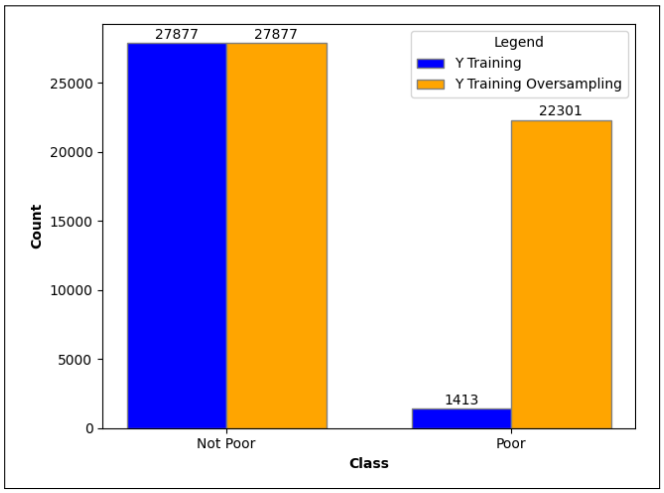

Figure 4. Number of Oversampling of Linear Kernel Training Data

Figure 4 illustrates the increase in data for the minority class, where the number of poor households rises from 1,413 to 22,301. This increase was achieved by applying 80% oversampling using data from the majority class. The 80% increase in the minority class, which represents poor households, enhances the classification accuracy of the model in identifying poverty status. However, for non-poor households, there is a decrease in the average specificity value. Despite this, the higher average $G-Mean$ value, compared to other oversampling techniques, suggests that the 80% increase in minority data leads to a more balanced classification accuracy, further supported

by a high average AUC value. Subsequently, the performance of this best model will be examined across each fold of the cross-validation, as presented in Table 6.

Table 6. Classification Accuracy of Linear Kernel ($C = 0.1$) with 80% SMOTE Oversampling

Fold	Accuracy	Sensitivity	Specificity	$G - Mean$	AUC
1	79.48%	64.33%	80.25%	71.85%	72.29%
2	79.94%	68.15%	80.54%	74.09%	74.34%
3	78.37%	69.43%	78.83%	73.98%	74.13%
4	79.66%	70.12%	80.12%	75.26%	75.64%
5	79.59%	69.43%	80.11%	74.58%	74.77%
6	79.95%	68.54%	80.53%	74.24%	74.44%
7	79.72%	63.69%	80.53%	71.65%	72.11%
8	78.73%	77.07%	78.82%	77.94%	77.94%
9	78.67%	67.67%	79.76%	73.71%	73.59%
10	77.41%	64.97%	78.04%	71.21%	71.51%
Average	79.00%	68.41%	79.53%	73.73%	73.97%

Table 6 presents the classification results using parameter $C = 0.1$ across folds 1 to 10, showing relatively consistent classification performance. However, a notable difference appears in the Sensitivity and $G - Mean$ values, particularly in fold 8, which records the highest scores in both metrics. Overall, the model using a linear kernel with $C = 0.1$ and 80% oversampling performs well in classifying "poor" and "non-poor" households. The model achieves a high average $G - Mean$ of 73.73%, supported by strong values in Accuracy, Sensitivity, Specificity, and AUC.

3.3.2 Classification Using Support Vector Machine with RBF Kernel

The tuning of parameters C and γ in the SVM model with an RBF kernel is performed on the training data and evaluated on the testing data. The values of C used are the same as those in the linear kernel model, while the values for γ are set at 0.3, 0.5, 0.7, and 0.9. In addition, the SMOTE oversampling technique is applied, resulting in a total of 48 experimental combinations. The results of the modeling using combinations of SMOTE, C, and γ parameters are presented in Table 7.

Table 7. Radial Basis Function Kernel Classification on Testing Data

C	γ	Accuracy	Sensitivity	Specificity	$G - Mean$	AUC
Oversampling 40%						
0.01	0.3	94.42%	6.75%	98.86%	25.60%	52.81%
0.1		92.00%	24.08%	95.87%	47.85%	60.19%
1		90.81%	30.32%	93.87%	53.00%	62.17%
10		90.25%	31.91%	93.76%	54.41%	62.56%
0.01	0.5	94.62%	5.80%	99.14%	24.74%	52.47%
0.1		92.37%	20.83%	95.74%	44.53%	59.28%
1		90.26%	33.32%	93.44%	55.87%	62.45%
10		89.78%	33.25%	92.66%	55.41%	62.95%
0.01	0.7	94.07%	4.98%	99.14%	22.58%	51.36%
0.1		92.04%	17.86%	96.12%	41.72%	57.96%
1		90.02%	29.11%	93.48%	52.05%	60.98%
10		89.67%	32.87%	92.55%	55.06%	62.71%

C	γ	Accuracy	Sensitivity	Specificity	$G-Mean$	AUC
0.01	0.9	92.99%	17.58%	96.37%	40.85%	56.52%
0.1		92.09%	17.36%	95.78%	40.58%	56.69%
1		90.29%	31.19%	92.48%	53.72%	61.10%
10		89.53%	32.99%	92.39%	55.13%	62.69%
Oversampling 60%						
0.01	0.3	86.22%	35.10%	88.81%	55.78%	61.95%
0.1		85.42%	39.68%	87.72%	58.94%	63.70%
1		85.10%	49.24%	86.92%	65.35%	68.08%
10		83.90%	50.38%	85.61%	65.31%	68.00%
0.01	0.5	86.57%	32.23%	89.32%	53.57%	60.77%
0.1		86.31%	40.15%	88.11%	59.46%	64.13%
1		85.22%	46.69%	87.32%	63.75%	66.93%
10		83.42%	49.55%	85.86%	65.37%	67.44%
0.01	0.7	88.94%	25.52%	91.56%	48.36%	58.84%
0.1		85.94%	36.54%	89.64%	57.26%	63.16%
1		84.93%	45.41%	86.94%	62.31%	65.98%
10		84.03%	47.11%	85.72%	63.74%	66.80%
0.01	0.9	90.54%	42.19%	94.16%	62.12%	66.57%
0.1		85.82%	39.88%	88.52%	59.34%	64.20%
1		84.14%	47.52%	85.99%	63.85%	66.76%
10		86.22%	35.10%	88.81%	55.78%	61.95%
Oversampling 80%						
0.01	**0.3**	76.95%	58.92%	77.86%	67.70%	68.39%
0.1		**79.03%**	**62.87%**	**79.85%**	**70.82%**	**71.30%**
1		78.43%	61.12%	79.09%	69.52%	70.86%
10		77.59%	61.52%	77.26%	68.45%	69.54%
0.01	0.5	75.43%	58.89%	76.13%	66.95%	67.57%
0.1		78.48%	59.90%	80.38%	69.10%	69.53%
1		78.44%	57.33%	80.92%	68.09%	69.15%
10		78.51%	58.99%	79.62%	68.59%	68.76%
0.01	0.7	74.10%	58.60%	74.89%	66.22%	66.74%
0.1		77.64%	58.92%	78.02%	67.80%	68.30%
1		79.64%	55.35%	80.97%	66.84%	68.11%
10		78.62%	55.86%	79.85%	66.70%	67.92%
0.01	0.9	73.37%	58.47%	74.12%	65.81%	66.30%
0.1		77.94%	56.21%	78.92%	66.82%	67.97%
1		79.49%	55.22%	80.72%	66.69%	67.97%
10		76.95%	58.92%	77.86%	67.70%	68.39%

Table 7 shows that using 80% oversampling results in higher average Sensitivity and $G-Mean$ values compared to other oversampling levels. These results are consistent with those obtained using the linear kernel. The best combination of parameter C and parameter γ for classifying "poor" and "non-poor" households is $C=0.1$ and $\gamma=0.3$. This model is considered the best because it yields a high average $G-Mean$ of 70.82% and a good average Sensitivity for classifying the "poor" class, which is 62.87%. Further analysis of the best model will be presented based on the modeling results for each fold, as shown in Table 8.

Table 8. Classification Accuracy of Linear Kernel ($C = 0.1$) with 80% SMOTE Oversampling

Fold	Accuracy	Sensitivity	Specificity	$G-Mean$	AUC
1	78.89%	58.60%	79.92%	68.44%	69.26%
2	79.54%	57.96%	80.63%	68.68%	69.30%
3	78.28%	64.33%	78.99%	71.28%	71.66%
4	79.20%	66.88%	79.92%	73.11%	73.40%
5	80.10%	64.97%	81.50%	72.77%	72.63%
6	78.93%	64.31%	79.61%	71.41%	71.31%
7	79.10%	61.15%	79.85%	70.17%	70.09%
8	79.10%	68.15%	77.95%	72.91%	73.09%
9	80.12%	68.15%	79.72%	73.67%	73.09%
10	77.57%	56.69%	78.72%	66.76%	67.66%
Average	78.89%	58.60%	79.92%	68.44%	69.26%

Table 8 shows that the classification results using the parameter combination of $C = 0.1$ and $\gamma = 0.3$ across folds 1 to 10 exhibit relatively consistent classification accuracy. A notable variation is observed in the Sensitivity values, particularly in fold 8. Overall, the modeling using the RBF kernel with the parameter combination of $C = 0.1$ and $\gamma = 0.3$ and 80% oversampling performs quite well in classifying both "poor" and "non-poor" households. This is evident from the relatively high average $G - Mean$ of 70.82%, supported by other classification performance metrics.

3.4. Comparison of SVM with Linear and RBF Kernel

The next step after obtaining the best classification accuracy from each kernel is to compare the performance of the kernels to determine which one yields the best results in classifying "poor" and "non-poor" households. The comparison is primarily based on the $G - Mean$ value, followed by other classification performance metrics. The comparison of the best classification performance between the linear kernel SVM and the RBF kernel SVM is presented in Table 9.

Table 9. Comparison of Classification Accuracy

Kernel	Accuracy	Sensitivity	Specificity	$G-Mean$	AUC
Linear	79.00%	**68.41%**	79.53%	**73.73%**	**73.97%**
RBF	**79.03%**	62.87%	**79.85%**	70.82%	71.36%

Table 9 compares the performance of two kernel types in SVM, namely the Linear kernel and the RBF kernel, in classifying poverty status in East Java Province. Based on the results shown, the SVM with the Linear kernel performs better than the RBF kernel. This can be observed from the higher $G - Mean$ value in the Linear kernel SVM (73.73%) compared to the RBF kernel SVM (70.82%). A higher $G - Mean$ indicates a better balance between Sensitivity and Specificity, which is crucial in imbalanced datasets like in this study. Although the Accuracy of the RBF kernel is slightly higher (79.03% compared to 79.00% for the Linear kernel), this difference is minimal and does not outweigh the Linear kernel's better performance in detecting the poverty class.

In terms of Sensitivity, the Linear kernel has a higher value (68.41%) compared to the RBF kernel (62.87%), indicating that the Linear kernel SVM is more effective in identifying poor individuals, an aspect that is especially important in poverty classification. The Specificity of the RBF kernel is slightly higher (79.85%) compared to the Linear kernel (79.53%), but this difference is not significant in the context of this application, where it is more critical to accurately detect poor individuals. The AUC for the Linear kernel is also higher (73.97%) compared to the RBF kernel (71.36%), indicating that the Linear kernel has a better ability to differentiate between poverty

and non-poverty classes. The use of SMOTE has proven effective in addressing the issue of imbalanced data by improving the classification accuracy of the minority class (poverty), which also contributes to the improved performance of both SVM models.

Overall, while the RBF kernel shows a slight advantage in terms of Accuracy and Specificity, the SVM with the Linear kernel excels in Sensitivity, $G-Mean$, and AUC, making it the optimal choice for classifying poverty status in East Java Province. The successful application of SMOTE in handling the data imbalance further strengthens the performance of this model, particularly in detecting the minority class.

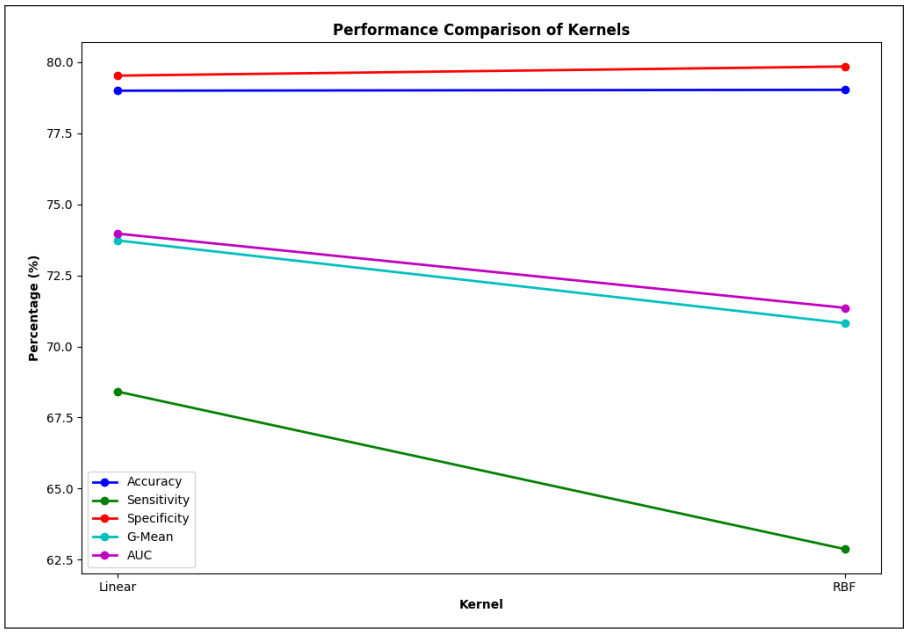

Figure 5. Performance Comparison of Kernel

The performance comparison of the SVM kernels, as shown in the Figure 5, highlights the superior performance of the linear kernel over the RBF kernel. The Accuracy, Sensitivity, Specificity, $G-Mean$, and AUC values for the linear kernel remain consistently higher than those for the RBF kernel, as demonstrated by the graphical representation. The linear kernel shows a slight increase in Accuracy and $G-Mean$ compared to the RBF kernel, indicating its robustness in classifying poverty status in East Java. The use of SMOTE has effectively mitigated the data imbalance, especially for the minority class, which is reflected in the improved $G-Mean$ and AUC scores. Overall, the linear kernel SVM model provides a more balanced and accurate classification of household poverty status.

After determining that the best SVM method is the linear kernel SVM, the next step is to present the hyperplane function generated from the linear kernel SVM. The model formed using the linear kernel SVM with 11 predictor variables is as follows.

$$f(x_i) = \text{sign}\left(\sum_{j=1}^{n} w_j x_{ij}\right) \tag{19}$$

where:

$$\begin{aligned}\sum_{j=1}^{n} w_j x_{ij} = {} & w_0 + w_4 x_{i4} + w_5 x_{i5} + w_{6(0)} x_{i6(0)} + w_{6(1)} x_{i6(1)} + w_{6(2)} x_{i6(2)} + w_{6(3)} x_{i6(3)} \\ & + w_{6(4)} x_{i6(4)} + w_8 x_{i8} + w_{10(0)} x_{i10(0)} + w_{10(1)} x_{i10(1)} + w_{10(2)} x_{i10(2)} + w_{12} x_{i12} \\ & + w_{13(0)} x_{i13(0)} + w_{13(1)} x_{i13(1)} + w_{13(2)} x_{i13(2)} + w_{14(0)} x_{i14(0)} + w_{14(1)} x_{i14(1)} \\ & + w_{14(2)} x_{i14(2)} + w_{16} x_{i16} + w_{17} x_{i17} + w_{18} x_{i18}\end{aligned} \tag{20}$$

The weight values (w) for each variable are obtained from the calculation in Equation (11), which is influenced by the regularization parameter $C = 0.1$.

The function presented here is derived from the best-performing fold of the linear kernel SVM. The reason for presenting the function from the best fold is that K-fold Cross-validation generates a separate function for each fold, and fold 8 represents how well the model with parameter $C = 0.1$ classifies "poor" and "non-poor" households. The average result of the 10-fold cross-validation is the main result of the research, while Fold 8 is only used as an illustration of a representative model to explain the characteristics of the generated model. The following is the hyperplane function from fold 8 of the linear kernel SVM with parameter $C = 0.1$.

$$f(x_i) = \text{sign}\left(\sum_{j=1}^{n} w_j x_{ij}\right) \tag{21}$$

where:

$$\begin{aligned}\sum_{j=1}^{n} w_j x_{ij} = {} & 5.749 + 7.690x_{i4} + 0.109x_{i5} - 1.309x_{i6(0)} - 1.623x_{i6(1)} \\ & - 1.561x_{i6(2)} - 1.662x_{i6(3)} - 0.923x_{i6(4)} - 0.530x_{i8} - 1.914x_{i10(0)} \\ & - 1.140x_{i10(1)} - 1.546x_{i10(2)} - 5.905x_{i12} - 1.496x_{i13(0)} - 1.497x_{i13(1)} \\ & - 1.207x_{i13(2)} - 1.902x_{i14(0)} - 0.992x_{i14(1)} - 1.230x_{i14(2)} - 0.264x_{i16} \\ & - 1.099x_{i17} - 1.744x_{i18}\end{aligned} \tag{22}$$

The hyperplane function is constructed based on 11 variables that have an influence in classifying household poverty status. If $f(x_i) < 0$, the household is categorized as non-poor, whereas if $f(x_i) > 0$, it is categorized as poor.

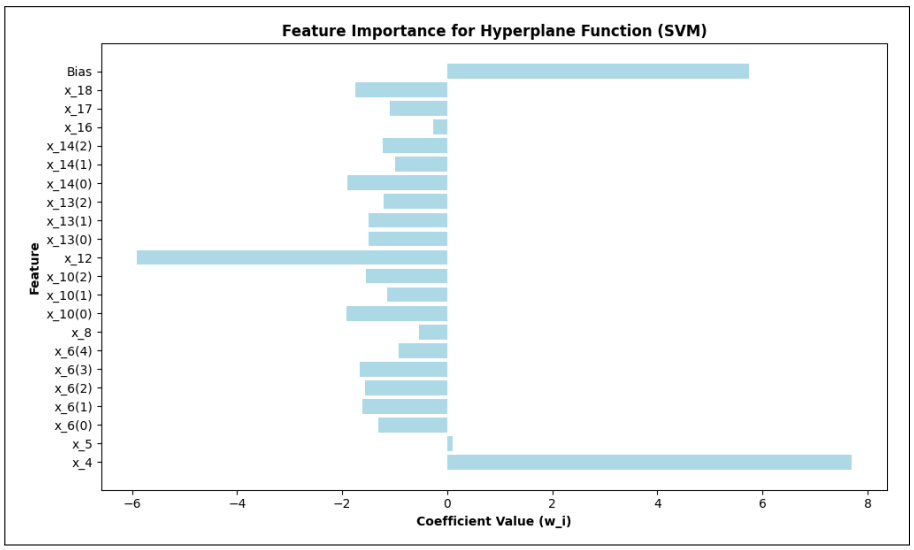

Figure 6. Performance Comparison of Kernel

The feature importance plot for the hyperplane function in the SVM from Figure 6, reveals the relative contributions of each predictor variable in determining household poverty status classification. The horizontal bars represent the magnitude of the coefficients (w_j) for each feature, where the length of the bar indicates the strength of each variable's influence on the decision boundary. Larger absolute coefficient values suggest greater importance of the corresponding feature in influencing the classification outcome.

The linear model used for classifying household poverty status involves a series of coefficients associated with different predictor variables. Among these, x_{i4} stands out with a significant positive coefficient of 7.690,

meaning that as the value of x_{i4} increases, the output of the hyperplane function $w_j x_{ij}$ also increases, pushing the classification towards the poor category. Similarly, x_{i5}, though having a smaller positive coefficient of 0.109, still exerts a positive influence on the classification. On the other hand, the other variables all have negative coefficients. These negative coefficients indicate that as the values of these features increase, the result of $w_j x_{ij}$ decreases, making the classification more likely to fall into the "non-poor" category.

If we look at each variable coefficient in the SVM hyperplane function, the variables with the greatest influence are number of household members (X_4), floor area (X_{12}), and main source of drinking water (X_{14}). These findings indicate that the demographic characteristics of households and the conditions of their living environments are the most distinguishing factors between poor and non-poor households in East Java Province. The Number of Household Members variable has a significant contribution because the more household members there are, the greater the consumption needs that must be met. If the increase in the number of household members is not matched by an increase in income, the economic burden on the household will become heavier and the risk of poverty will increase. The Number of Household Members variable has a significant contribution because the more household members there are, the greater the consumption needs that must be met. The Floor Area variable also becomes one of the dominant predictors. The floor area of a house is an indicator that reflects the quality of the dwelling and the level of household welfare. In addition, the Main Source of Drinking Water becomes a variable with a significant influence because access to a proper drinking water source is one of the basic indicators of community welfare. Overall, these three variables indicate that the classification of poverty is not only influenced by economic aspects but also by demographic characteristics and access to basic services.

The bias term w_0, which is 5.749, is another important component of the model. It acts as an offset, adjusting the decision boundary to help the model correctly classify the data based on the sum of all the feature contributions. The bias term ensures that the classification rule $f(x_i) > 0$ for "poor" households and $f(x_i) < 0$ for "non-poor" households is properly positioned, thus improving the accuracy of the model's predictions. Together, these variables and their coefficients construct a robust model that classifies households into poverty and non-poverty categories, with certain features having a more significant impact on the classification outcome than others. Next, the confusion matrix from fold 8 of the linear kernel SVM with parameter $C = 0.1$ will be presented. The confusion matrix for fold 8 is shown in Table 10.

Table 10. Confusion Matrix Results

Actual Value	**Predicted Value**	
	1 (Positive Class)	0 (Negative Class)
1 (Positive Class)	121	36
0 (Negative Class)	656	2441

After obtaining the confusion matrix, the classification performance metrics, namely Accuracy, Sensitivity, Specificity, $G - Mean$, and AUC can be manually calculated using Equations (14) through (18). The model accuracy value is as follows.

$$\begin{aligned} \text{Accuracy} &= \frac{TP + TN}{TP + FN + FP + TN} \\ &= \frac{121 + 2441}{121 + 36 + 656 + 2441} \\ &= 0.7873 \end{aligned}$$

The correct prediction of "poor" households is reflected by the Sensitivity value. The Sensitivity of the model is as follows.

$$\text{Sensitivity} = \frac{TP}{TP + FN} = \frac{121}{121 + 36} = 0.7707$$

The correct prediction of "non-poor" households is reflected by the Specificity value. The Specificity of the model is as follows.

$$\text{Specificity} = \frac{TN}{FP + TN} = \frac{2441}{656 + 2441} = 0.7882$$

The model ability to balance the classification of the positive and negative classes is indicated by the $G - Mean$ value. The $G - Mean$ of the model is as follows.

$$\begin{aligned} G - Mean &= \sqrt{\text{Sensitivity} \times \text{Specificity}} \\ &= \sqrt{0.7707 \times 0.7881} \\ &= 0.7794 \end{aligned}$$

Finally, the best model ability to successfully separate the positive and negative classes is indicated by the AUC value. The AUC of the model is as follows.

$$\begin{aligned} \text{AUC} &= \frac{1}{2}\left(\text{Sensitivity} + \text{Specificity}\right) \\ &= \frac{1}{2}\left(0.7707 + 0.7881\right) \\ &= 0.7794 \end{aligned}$$

The five manually calculated classification performance metrics are then summarized as percentages in Table 11.

Table 11. Classification Performance of Linear Kernel SVM

Accuracy	**Sensitivity**	**Specificity**	$\boldsymbol{G - Mean}$	**AUC**
78.73%	77.07%	78.82%	77.94%	77.94%

Table 11 shows that the linear kernel SVM model can correctly classify "poor" households with an Accuracy of 77.07% and "non-poor" households with an Accuracy of 78.82%. It also achieves a balanced classification, with a G-Mean of 77.94%. These results suggest that the SVM model with a linear kernel and parameter $C = 0.1$, combined with 80% oversampling, is effective in overcoming the bias that typically favors classifying only "non-poor" households. This performance is better than that of previous studies, demonstrating that the SVM model, when combined with the SMOTE method, can be a viable alternative for classifying household poverty status. While the overall Accuracy of 78.73% indicates a relatively strong performance, the misclassification of "poor" households still requires further improvement.

The Sensitivity of 77.07% indicates that the model still faces challenges in accurately identifying "poor" households. This lower Sensitivity reflects the difficulty in classifying the minority class in a highly imbalanced dataset. Conversely, the Specificity of 78.82% indicates that the model successfully identified "non-poor" households, consistent with the larger proportion of these households. This suggests that the model has adequately addressed most of the imbalance, although errors in predicting poor households remain.

The $G - Mean$ value of 77.94% indicates that the model successfully balances the two classes, although data imbalance remains a challenge. This relatively high $G - Mean$ value suggests that the model is not excessively biased toward a single class and is capable of distinguishing between "poor" and "non-poor" households to a satisfactory extent. Nevertheless, further improvement is still possible, particularly in enhancing the model's sensitivity to low-income families. Similarly, the AUC value also reached 77.94%, indicating that the model possesses a moderate ability to discriminate between the two classes, although the result still reflects a slight overlap in the predicted classifications.

The classification results on fold 8 are a representative example in the application of the SVM model for classification and the formation of the confusion matrix in one iteration of 10-fold cross-validation, applied to real cases, particularly the status of poor households. The main results of the model are still based on the results from all 10 folds, as presented in Table 6. The average results indicate that the SVM model with SMOTE has

consistent performance across various data splits, so the reported classification ability reflects the overall model performance and is not dependent on a specific fold.

The results of this research are not only based on the improvement of the accuracy of classifying poor households but also as a new potential in utilizing it as a decision support system in poverty program evaluation. The developed model is capable of identifying household characteristics related to poverty status based on socio-economic predictor variables that have not been previously applied. Thus, the model's prediction results can be used as supporting information in the initial identification (pre-screening) process of households that are potentially poor before the administrative verification or field verification process by the relevant agencies.

Although the proposed model shows good classification performance, this study still has several limitations that need to be considered when interpreting the research results. First, this study only uses micro data from the March 2023 SUSENAS in East Java Province. Therefore, the developed model represents the characteristics of household poverty during the observation period and has not yet evaluated the consistency of the model's performance against changes in socio-economic conditions over time. Second, the research results are influenced by several methodological decisions, such as the choice of oversampling level, the type of SVM kernel, and the hyperparameter values used. Although these parameters were selected based on experimental evaluation results and yielded the best performance in this study, different parameter configurations have the potential to produce different performance outcomes as well. Third, the results of this study were specifically developed using the characteristics of households in East Java Province in 2023, so the model's generalizability to other provinces or regions with different socio-economic conditions has not yet been determined. Fourth, the implementation of the model in practice requires the availability of household data with a variable structure comparable to SUSENAS data. In addition, the model needs to be periodically updated (retraining) using the latest survey data to remain capable of accommodating changes in poverty characteristics over time. On the other hand, the process of model training and parameter optimization requires relatively greater computational resources as the amount of data and model complexity increase, so this aspect needs to be considered if the model is to be widely implemented as a decision support system.

Based on these limitations, future research is expected to address various existing limitations through the use of more diverse data, optimization of model parameters, and validation in different regions and periods, thereby producing a more robust model with better generalization capabilities.

4. Conclusion

Based on the analysis and discussion, modeling using the linear-kernel SVM $(C = 0.1)$ augmented by 80% SMOTE oversampling emerged as the most effective strategy for binary classification of household poverty status in East Java. This approach achieved a balanced trade-off between $\mathrm{Sensitivity}$ (77.07%) and $\mathrm{Specificity}$ (78.82%), yielding an overall $\mathrm{Accuracy}$ of 78.73% and a $G - Mean$ of 77.94%. These results demonstrate that, by converting socioeconomic indicators into categorical predictors and applying a carefully regularized SVM, one can reliably distinguish between "poor" and "non-poor" households. Such a model not only supports timely, cost-effective monitoring of SDG 1 (No Poverty) at sub-national scales but also provides policymakers with a robust tool for targeting interventions in regions with heterogeneous poverty profiles. Overall, this model is proposed in a process that can assist the government in distributing aid, which includes the complete stages of household data collection, prediction using the model, prioritization, verification and validation, and determination of aid recipients.

Acknowledgement

This work was supported in part by the Institut Teknologi Sepuluh Nopember, under project scheme of the Publication Writing and IPR Incentive Program (PPHKI) 2025.

The Declaration of Conflict of Interest/ Common Interest

The authors declare no conflicts of interest.

REFERENCES

[1] A. Gai, R. Ernan, A. Fauzi, B. Barus, and D. Putra. "Poverty Reduction Through Adaptive Social Protection and Spatial Poverty Model in Labuan Bajo, Indonesia's National Strategic Tourism Areas". en. In: *Sustainability* vol. 17, no. 2 (Jan. 2025), pp. 555. ISSN: 2071-1050. DOI: 10.3390/su17020555. URL: https://www.mdpi.com/2071-1050/17/2/555 (visited on 02/21/2026).

[2] A. Ciucu, V. Vargas, C. Păuna, and A.-I. Jigani. "Poverty, Education, and Decent Work Rates in Central and Eastern EU Countries". en. In: *Standards* vol. 5, no. 2 (June 2025), pp. 16. ISSN: 2305-6703. DOI: 10.3390/standards5020016. URL: https://www.mdpi.com/2305-6703/5/2/16 (visited on 02/21/2026).

[3] A. M. Balisacan, E. M. Pernia, and A. Asra. "Revisiting growth and poverty reduction in Indonesia: what do subnational data show?" In: *Bulletin of Indonesian Economic Studies* vol. 39, no. 3 (Dec. 2003). _eprint: https://doi.org/10.1080/0007491032000142782, pp. 329–351. ISSN: 0007-4918. DOI: 10.1080/0007491032000142782. URL: https://doi.org/10.1080/0007491032000142782 (visited on 02/21/2026).

[4] L. Sugiharti, R. Purwono, M. A. Esquivias, and H. Rohmawati. "The Nexus between Crime Rates, Poverty, and Income Inequality: A Case Study of Indonesia". en. In: *Economies* vol. 11, no. 2 (Feb. 2023), pp. 62. ISSN: 2227-7099. DOI: 10.3390/economies11020062. URL: https://www.mdpi.com/2227-7099/11/2/62 (visited on 02/21/2026).

[5] A. Alsharkawi, M. Al-Fetyani, M. Dawas, H. Saadeh, and M. Alyaman. "Poverty Classification Using Machine Learning: The Case of Jordan". en. In: *Sustainability* vol. 13, no. 3 (Jan. 2021), pp. 1412. ISSN: 2071-1050. DOI: 10.3390/su13031412. URL: https://www.mdpi.com/2071-1050/13/3/1412 (visited on 02/21/2026).

[6] S. M. Ali, A. Appolloni, F. Cavallaro, I. D'Adamo, A. Di Vaio, F. Ferella, M. Gastaldi, M. Ikram, N. M. Kumar, M. A. Martin, A.-S. Nizami, I. Ozturk, M. P. Riccardi, P. Rosa, E. Santibanez Gonzalez, C. Sassanelli, D. Settembre-Blundo, R. K. Singh, M. Smol, G. A. Tsalidis, I. Voukkali, N. Yang, and A. A. Zorpas. "Development Goals towards Sustainability". en. In: *Sustainability* vol. 15, no. 12 (Jan. 2023), pp. 9443. ISSN: 2071-1050. DOI: 10.3390/su15129443. URL: https://www.mdpi.com/2071-1050/15/12/9443 (visited on 02/21/2026).

[7] M. d. C. Pérez-Peña, M. Jiménez-García, J. Ruiz-Chico, and A. R. Peña-Sánchez. "Analysis of Research on the SDGs: The Relationship between Climate Change, Poverty and Inequality". en. In: *Applied Sciences* vol. 11, no. 19 (Jan. 2021), pp. 8947. ISSN: 2076-3417. DOI: 10.3390/app11198947. URL: https://www.mdpi.com/2076-3417/11/19/8947 (visited on 02/21/2026).

[8] H. Thamrin. "Educational aspects in efforts to realize SDGs in Indonesia". In: *Journal of Advances in Education and Philosophy* vol. 4, no. 11 (2020), pp. 473–477. URL: https://saudijournals.com/media/articles/JAEP_411_473-477.pdf (visited on 02/21/2026).

[9] M. F. Cordova and A. Celone. "SDGs and Innovation in the Business Context Literature Review". en. In: *Sustainability* vol. 11, no. 24 (Jan. 2019), pp. 7043. ISSN: 2071-1050. DOI: 10.3390/su11247043. URL: https://www.mdpi.com/2071-1050/11/24/7043 (visited on 02/21/2026).

[10] Q.-Q. Liu, M. Yu, and X.-L. Wang. "Poverty reduction within the framework of SDGs and Post-2015 Development Agenda". In: *Advances in Climate Change Research* vol. 6, no. 1 (Mar. 2015), pp. 67–73. ISSN: 1674-9278. DOI: 10.1016/j.accre.2015.09.004. URL: https://www.sciencedirect.com/science/article/pii/S1674927815000489 (visited on 02/21/2026).

[11] R. Purwono, W. W. Wardana, T. Haryanto, and M. Khoerul Mubin. "Poverty dynamics in Indonesia: empirical evidence from three main approaches". In: *World Development Perspectives* vol. 23 (Sept. 2021), pp. 100346. ISSN: 2452-2929. DOI: 10.1016/j.wdp.2021.100346. URL: https://www.sciencedirect.com/science/article/pii/S245229292100062X (visited on 02/21/2026).

[12] L. M. Fonseca, J. P. Domingues, and A. M. Dima. "Mapping the Sustainable Development Goals Relationships". en. In: *Sustainability* vol. 12, no. 8 (Jan. 2020), pp. 3359. ISSN: 2071-1050. DOI: 10.3390/su12083359. URL: https://www.mdpi.com/2071-1050/12/8/3359 (visited on 02/21/2026).

[13] X. Zheng, W. Zhang, H. Deng, and H. Zhang. "County-Level Poverty Evaluation Using Machine Learning, Nighttime Light, and Geospatial Data". en. In: *Remote Sensing* vol. 16, no. 6 (Jan. 2024), pp. 962. ISSN: 2072-4292. DOI: 10.3390/rs16060962. URL: https://www.mdpi.com/2072-4292/16/6/962 (visited on 02/21/2026).

[14] V. Ratnasari, S. H. Utama, and A. T. Rian Dani. "Toward Sustainable Development Goals (SDGs) with Statistical Modeling: Recursive Bivariate Binary Probit." In: *IAENG International Journal of Applied Mathematics* vol. 54, no. 8 (2024). URL: https://www.iaeng.org/IJAM/issues_v54/issue_8/IJAM_54_8_04.pdf (visited on 02/21/2026).

[15] Y. Wang, S. Jia, W. Qi, and C. Huang. "Examining Poverty Reduction of Poverty-Stricken Farmer Households under Different Development Goals: A Multiobjective Spatio-Temporal Evolution Analysis Method". en. In: *International Journal of Environmental Research and Public Health* vol. 19, no. 19 (Jan. 2022), pp. 12686. ISSN: 1660-4601. DOI: 10.3390/ijerph191912686. URL: https://www.mdpi.com/1660-4601/19/19/12686 (visited on 02/21/2026).

[16] N. P. A. M. Mariati, I. N. Budiantara, and V. Ratnasari. "Modeling Poverty Percentages in the Papua Islands using Fourier Series in Nonparametric Regression Multivariable". en. In: *Journal of Physics: Conference Series* vol. 1397, no. 1 (Dec. 2019), pp. 012071. ISSN: 1742-6596. DOI: 10.1088/1742-6596/1397/1/012071. URL: https://doi.org/10.1088/1742-6596/1397/1/012071 (visited on 02/21/2026).

[17] J. Symonds. "The Poverty Trap: Or, Why Poverty is Not About the Individual". en. In: *International Journal of Historical Archaeology* vol. 15, no. 4 (Dec. 2011), pp. 563–571. ISSN: 1573-7748. DOI: 10.1007/s10761-011-0156-8. URL: https://doi.org/10.1007/s10761-011-0156-8 (visited on 02/21/2026).

[18] G. E. Halkos and P.-S. C. Aslanidis. "Addressing Multidimensional Energy Poverty Implications on Achieving Sustainable Development". en. In: *Energies* vol. 16, no. 9 (Jan. 2023), pp. 3805. ISSN: 1996-1073. DOI: 10.3390/en16093805. URL: https://www.mdpi.com/1996-1073/16/9/3805 (visited on 02/21/2026).

[19] A. Usmanova, A. Aziz, D. Rakhmonov, and W. Osamy. "Utilities of Artificial Intelligence in Poverty Prediction: A Review". en. In: *Sustainability* vol. 14, no. 21 (Jan. 2022), pp. 14238. ISSN: 2071-1050. DOI: 10.3390/su142114238. URL: https://www.mdpi.com/2071-1050/14/21/14238 (visited on 02/21/2026).

[20] U. Grzybowska, A. Wojewódzka-Wiewiórska, G. Vaznonienė, and H. Dudek. "Households Vulnerable to Energy Poverty in the Visegrad Group Countries: An Analysis of Socio-Economic Factors Using a Machine Learning Approach". en. In: *Energies* vol. 17, no. 24 (Jan. 2024), pp. 6310. ISSN: 1996-1073. DOI: 10.3390/en17246310. URL: https://www.mdpi.com/1996-1073/17/24/6310 (visited on 02/21/2026).

[21] A. Razaque, M. Ben Haj Frej, M. Almi'ani, M. Alotaibi, and B. Alotaibi. "Improved Support Vector Machine Enabled Radial Basis Function and Linear Variants for Remote Sensing Image Classification". en. In: *Sensors* vol. 21, no. 13 (Jan. 2021), pp. 4431. ISSN: 1424-8220. DOI: 10.3390/s21134431. URL: https://www.mdpi.com/1424-8220/21/13/4431 (visited on 02/21/2026).

[22] F. Wang, Z. Zhen, B. Wang, and Z. Mi. "Comparative Study on KNN and SVM Based Weather Classification Models for Day Ahead Short Term Solar PV Power Forecasting". en. In: *Applied Sciences* vol. 8, no. 1 (Jan. 2018), pp. 28. ISSN: 2076-3417. DOI: 10.3390/app8010028. URL: https://www.mdpi.com/2076-3417/8/1/28 (visited on 02/21/2026).

[23] G. Latif, G. Ben Brahim, D. N. F. A. Iskandar, A. Bashar, and J. Alghazo. "Glioma Tumors' Classification Using Deep-Neural-Network-Based Features with SVM Classifier". en. In: *Diagnostics* vol. 12, no. 4 (Apr. 2022), pp. 1018. ISSN: 2075-4418. DOI: 10.3390/diagnostics12041018. URL: https://www.mdpi.com/2075-4418/12/4/1018 (visited on 02/21/2026).

[24] R. Guido, S. Ferrisi, D. Lofaro, and D. Conforti. "An Overview on the Advancements of Support Vector Machine Models in Healthcare Applications: A Review". en. In: *Information* vol. 15, no. 4 (Apr. 2024), pp. 235. ISSN: 2078-2489. DOI: 10.3390/info15040235. URL: https://www.mdpi.com/2078-2489/15/4/235 (visited on 02/21/2026).

[25] S. E. Jozdani, B. A. Johnson, and D. Chen. "Comparing Deep Neural Networks, Ensemble Classifiers, and Support Vector Machine Algorithms for Object-Based Urban Land Use/Land Cover Classification". en. In: *Remote Sensing* vol. 11, no. 14 (Jan. 2019), pp. 1713. ISSN: 2072-4292. DOI: 10.3390/rs11141713. URL: https://www.mdpi.com/2072-4292/11/14/1713 (visited on 02/21/2026).

[26] P. Thanh Noi and M. Kappas. "Comparison of Random Forest, k-Nearest Neighbor, and Support Vector Machine Classifiers for Land Cover Classification Using Sentinel-2 Imagery". en. In: *Sensors* vol. 18, no. 1 (Jan. 2018), pp. 18. ISSN: 1424-8220. DOI: 10.3390/s18010018. URL: https://www.mdpi.com/1424-8220/18/1/18 (visited on 02/21/2026).

[27] M. A. Nanda, K. B. Seminar, D. Nandika, and A. Maddu. "A Comparison Study of Kernel Functions in the Support Vector Machine and Its Application for Termite Detection". en. In: *Information* vol. 9, no. 1 (Jan. 2018), pp. 5. ISSN: 2078-2489. DOI: 10.3390/info9010005. URL: https://www.mdpi.com/2078-2489/9/1/5 (visited on 02/21/2026).

[28] H. Anysz, A. Zbiciak, and N. Ibadov. "The Influence of Input Data Standardization Method on Prediction Accuracy of Artificial Neural Networks". In: *Procedia Engineering*. XXV Polish – Russian – Slovak Seminar "Theoretical Foundation of Civil Engineering" vol. 153 (Jan. 2016), pp. 66–70. ISSN: 1877-7058. DOI: 10.1016/j.proeng.2016.08.081. URL: https://www.sciencedirect.com/science/article/pii/S1877705816322238 (visited on 02/21/2026).

[29] V. Ratnasari, P. -, I. C. Aviantholib, and A. T. R. Dani. "Parameter estimation and hypothesis testing the second order of bivariate binary logistic regression (S-BBLR) model with Berndt Hall-Hall-Hausman (BHHH) iterations". In: *Commun. Math. Biol. Neurosci.* vol. 2022, no. 0 (Nov. 2022), pp. Article ID 35. ISSN: 2052-2541. DOI: 10.28919/cmbn/7258. URL: https://scik.org/index.php/cmbn/article/view/7258 (visited on 02/21/2026).

[30] L. Benhlima and M. E. H. Tirari. "Improving representativeness in big data analysis through weighted machine learning methods: A case study on the logistic regression model". en. In: *Statistics, Optimization & Information Computing* vol. 13, no. 2 (2025), pp. 780–794. ISSN: 2310-5070. DOI: 10.19139/soic-2310-5070-2015. URL: https://iapress.org/index.php/soic/article/view/2015 (visited on 07/19/2026).

[31] S. Cessie and J. C. Houwelingen. "Logistic Regression for Correlated Binary Data". In: *Journal of the Royal Statistical Society Series C: Applied Statistics* vol. 43, no. 1 (Mar. 1994), pp. 95–108. ISSN: 0035-9254. DOI: 10.2307/2986114. URL: https://doi.org/10.2307/2986114 (visited on 02/21/2026).

[32] M. Fathurahman, Purhadi, Sutikno, and V. Ratnasari. "Hypothesis testing of Geographically weighted bivariate logistic regression". en. In: *Journal of Physics: Conference Series* vol. 1417, no. 1 (Dec. 2019), pp. 012008. ISSN: 1742-6596. DOI: 10.1088/1742-6596/1417/1/012008. URL: https://doi.org/10.1088/1742-6596/1417/1/012008 (visited on 02/21/2026).

[33] P. Royston and D. G. Altman. "Regression Using Fractional Polynomials of Continuous Covariates: Parsimonious Parametric Modelling". en. In: *Journal of the Royal Statistical Society: Series C (Applied Statistics)* vol. 43, no. 3 (1994). _eprint: https://rss.onlinelibrary.wiley.com/doi/pdf/10.2307/2986270, pp. 429–453. ISSN: 1467-9876. DOI: 10.2307/2986270. URL: https://onlinelibrary.wiley.com/doi/abs/10.2307/2986270 (visited on 02/21/2026).

[34] P. Purhadi and M. Fathurahman. "A Logit Model for Bivariate Binary Responses". en. In: *Symmetry* vol. 13, no. 2 (Feb. 2021), pp. 326. ISSN: 2073-8994. DOI: 10.3390/sym13020326. URL: https://www.mdpi.com/2073-8994/13/2/326 (visited on 02/21/2026).

[35] T. M. Kaombe, J. C. Banda, G. A. Hamuza, and A. S. Muula. "Bivariate logistic regression model diagnostics applied to analysis of outlier cancer patients with comorbid diabetes and hypertension in Malawi". en. In: *Scientific Reports* vol. 13, no. 1 (May 2023), pp. 8340. ISSN: 2045-2322. DOI: 10.1038/s41598-023-35475-z. URL: https://www.nature.com/articles/s41598-023-35475-z (visited on 02/21/2026).

[36] K. Phinzi, D. Abriha, and S. Szabó. "Classification Efficacy Using K-Fold Cross-Validation and Bootstrapping Resampling Techniques on the Example of Mapping Complex Gully Systems". en. In: *Remote Sensing* vol. 13, no. 15 (Jan. 2021), pp. 2980. ISSN: 2072-4292. DOI: 10.3390/rs13152980. URL: https://www.mdpi.com/2072-4292/13/15/2980 (visited on 02/21/2026).

[37] S. Akakpo, P. Dambra, R. Paz, T. Smyth, F. Torre, and C. Yu. "Optimization of the K-Nearest Neighbor Algorithm to Predict Bank Churn". en. In: *Statistics, Optimization & Information Computing* vol. 12, no. 5 (July 2024), pp. 1397–1408. ISSN: 2310-5070. DOI: 10.19139/soic-2310-5070-2098. URL: https://iapress.org/index.php/soic/article/view/2098 (visited on 07/19/2026).

[38] Z. Lyu, Y. Yu, B. Samali, M. Rashidi, M. Mohammadi, T. N. Nguyen, and A. Nguyen. "Back-Propagation Neural Network Optimized by K-Fold Cross-Validation for Prediction of Torsional Strength of Reinforced Concrete Beam". en. In: *Materials* vol. 15, no. 4 (Jan. 2022), pp. 1477. ISSN: 1996-1944. DOI: 10.3390/ma15041477. URL: https://www.mdpi.com/1996-1944/15/4/1477 (visited on 02/21/2026).

[39] T. Wongvorachan, S. He, and O. Bulut. "A Comparison of Undersampling, Oversampling, and SMOTE Methods for Dealing with Imbalanced Classification in Educational Data Mining". en. In: *Information* vol. 14, no. 1 (Jan. 2023), pp. 54. ISSN: 2078-2489. DOI: 10.3390/info14010054. URL: https://www.mdpi.com/2078-2489/14/1/54 (visited on 02/21/2026).

[40] J. H. Joloudari, A. Marefat, M. A. Nematollahi, S. S. Oyelere, and S. Hussain. "Effective Class-Imbalance Learning Based on SMOTE and Convolutional Neural Networks". en. In: *Applied Sciences* vol. 13, no. 6 (Jan. 2023), pp. 4006. ISSN: 2076-3417. DOI: 10.3390/app13064006. URL: https://www.mdpi.com/2076-3417/13/6/4006 (visited on 02/21/2026).

[41] B. W. Heumann. "An Object-Based Classification of Mangroves Using a Hybrid Decision Tree—Support Vector Machine Approach". en. In: *Remote Sensing* vol. 3, no. 11 (Nov. 2011), pp. 2440–2460. ISSN: 2072-4292. DOI: 10.3390/rs3112440. URL: https://www.mdpi.com/2072-4292/3/11/2440 (visited on 02/21/2026).

[42] E. D. Madyatmadja, C. P. M. Sianipar, C. Wijaya, and D. J. M. Sembiring. "Classifying Crowdsourced Citizen Complaints through Data Mining: Accuracy Testing of k-Nearest Neighbors, Random Forest, Support Vector Machine, and AdaBoost". en. In: *Informatics* vol. 10, no. 4 (Dec. 2023), pp. 84. ISSN: 2227-9709. DOI: 10.3390/informatics10040084. URL: https://www.mdpi.com/2227-9709/10/4/84 (visited on 02/21/2026).

[43] O. J. Egwom, M. Hassan, J. J. Tanimu, M. Hamada, and O. M. Ogar. "An LDA–SVM Machine Learning Model for Breast Cancer Classification". en. In: *BioMedInformatics* vol. 2, no. 3 (Sept. 2022), pp. 345–358. ISSN: 2673-7426. DOI: 10.3390/biomedinformatics2030022. URL: https://www.mdpi.com/2673-7426/2/3/22 (visited on 02/21/2026).

[44] S. Hu, Y. Ge, M. Liu, Z. Ren, and X. Zhang. "Village-level poverty identification using machine learning, high-resolution images, and geospatial data". In: *International Journal of Applied Earth Observation and Geoinformation* vol. 107 (Mar. 2022), pp. 102694. ISSN: 1569-8432. DOI: 10.1016/j.jag.2022.102694. URL: https://www.sciencedirect.com/science/article/pii/S0303243422000204 (visited on 02/21/2026).

[45] M. Yusoff, T. Haryanto, H. Suhartanto, W. A. Mustafa, J. M. Zain, and K. Kusmardi. "Accuracy Analysis of Deep Learning Methods in Breast Cancer Classification: A Structured Review". en. In: *Diagnostics* vol. 13, no. 4 (Jan. 2023), pp. 683. ISSN: 2075-4418. DOI: 10.3390/diagnostics13040683. URL: https://www.mdpi.com/2075-4418/13/4/683 (visited on 02/21/2026).