Journal of Applied Informatics and Computing (JAIC)
Vol.9, No.4, August 2025, pp. 1451~1462
e-ISSN: 2548-6861

Implementation of The Logistic Regression Algorithm to Analyze Poverty Factors in Aceh Province

Mursyidah[1]*, Rozzi Kesuma Dinata[2], Zara Yunizar[3]
[123] Department of Informatics, Faculty of Engineering, Universitas Malikussaleh, Aceh, Indonesia
mursyidah.180170009@mhs.unimal.ac.id[1], rozzi@unimal.ac.id [2], zarayunizar@unimal.ac.id [3]

Article Info

Article history:

Received 2025-06-02
Revised 2025-06-27
Accepted 2025-07-06

Keyword:

Aceh,
Logistic regression,
Inference and Prediction,
Multicollinearity,
Poverty.

ABSTRACT

Aceh Province continues to face a high poverty rate despite its abundant natural resources. This study aims to analyze the factors influencing poverty status in Aceh Province by applying a binary logistic regression algorithm. The research specifically focuses on an inferential analytical approach to reveal significant relationships among socioeconomic variables. Secondary data were obtained from the Aceh Provincial Statistics Agency (Badan Pusat Statistik/BPS) for the period 2019–2023. Inferential analysis was conducted using the entire dataset through the *statsmodels* library to identify variables that are statistically significant to poverty status. In addition, a classification approach was implemented using *scikit-learn*, with a data split between training data (2019–2022) and testing data (2023), yielding an accuracy of 0.70, precision of 0.81, recall of 0.70, F1-score of 0.66, and AUC of 0.69. These findings provide empirical evidence that improving access to education and equitable infrastructure development in densely populated areas can serve as effective policy focuses in efforts to alleviate poverty in Aceh Province.

This is an open access article under the CC–BY-SA license.

I. INTRODUCTION

Poverty is a multidimensional issue that affects the quality of life and poses a major challenge to economic development in Indonesia. One region that continues to face serious poverty problems is Aceh Province. According to data from the Central Statistics Agency (Badan Pusat Statistik/BPS), the poverty rate in Aceh reached 14.64% in 2022, making it one of the provinces with the highest poverty rates in Sumatra. This condition indicates that poverty in Aceh is structural in nature and requires a more in-depth and targeted analytical approach.

Various socio-economic factors such as the Open Unemployment Rate (OUR), Gross Regional Domestic Product (GRDP), expenditure per capita, and average years of schooling have been examined in relation to poverty by several previous studies [1], [2]. In addition, variables such as total population have also been identified as factors that may influence poverty levels [3]. However, most of these studies have predominantly employed multiple linear regression or descriptive approaches, and have yet to adopt a binary classification of poverty status.

The study by Aini et al. [4] applied binary logistic regression in the context of West Papua; however, it did not address the evaluation of classification model performance. Azis et al. [3] utilized ordinal logistic regression to analyze national poverty levels, but did not explicitly combine predictive and inferential approaches. Meanwhile, the study by Safrina and Hasanah [1], which focused on Aceh, did not implement a training-testing data split or classify poverty status.

International studies affirm that logistic regression is an appropriate statistical approach for modeling categorical dependent variables, such as poverty status. Doherty and Wells [5] demonstrated that logistic regression can reveal social disparities more accurately than linear regression in social epidemiology research. Ratnasari et al. [6] developed the Bivariate Polynomial Binary Logit Regression (BPBLR) model to analyze poverty depth and emphasized

the importance of regression parameter interpretation in data-driven policymaking.

In this context, poverty status in this study is classified in a binary manner (poor = 1, not poor = 0) based on a specific threshold of the percentage of poor population in each district/city. Therefore, binary logistic regression was chosen, as it can directly model the probability of poverty while identifying statistically significant variables. In addition, this approach allows for the evaluation of classification model performance using metrics such as accuracy, precision, recall, F1-score, and AUC as complementary information.

The aim of this study is to analyze the factors that influence poverty status in Aceh Province through a binary logistic regression approach, by combining inferential analysis and model performance evaluation. The findings are expected to serve as a foundation for more targeted and data-driven poverty reduction policies. Specifically, this study seeks to answer two main research questions:
a. What are the factors that influence poverty status in Aceh Province?
b. How can the implementation of the logistic regression algorithm be used to analyze and evaluate the influence of these factors?

II. LITERATUR REVIEW

Poverty is influenced by various socio-economic indicators. This review discusses the key factors contributing to poverty and the relevant analytical approaches used to analyze it.

A. Binary Logistic Regresion

Binary logistic regression is a statistical method used to model the relationship between a categorical (binary) dependent variable and a set of independent variables. This model estimates the probability of an event occurring (Y = 1) using the following equation:

$$p = \frac{1}{1+e^{-(\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_n X_n)}} \quad \text{.....(1)}$$

Information:

- p = the probability of the desired event (e.g., the probability of Y=1).
- e = the base of the natural logarithm.
- β1,β2,…,βn = regression coefficients for each independent variable X1,X2,…,Xn
- β0 = the intercept (constant).

This model has been widely used in various studies to analyze the likelihood of an event based on its input characteristics. Doherty and Wells [5] employed logistic regression in a study on social inequality and concluded that this method is more appropriate than linear regression when dealing with categorical dependent variables.

B. Factors Influencing Poverty

Socioeconomic factors have long been recognized as contributors to regional poverty levels. A study by Safrina and Hasanah [1] indicated that increases in per capita expenditure and average years of schooling significantly reduce poverty levels in Aceh Province. Another study by Salong et al. [2] revealed that education, health, and per capita income play crucial roles in shaping poverty conditions in Maluku Province. Furthermore, Azis et al. [3] emphasized that population size, the Human Development Index (HDI), and economic equity have significant impacts on poverty at the national level.

Several previous studies have identified key determinants of poverty, including education, population size, and household expenditure levels. According to research [7], education has a significant negative impact on poverty in Indonesia. Similar findings were presented by [8], which stated that an increase in average years of schooling leads to a reduction in poverty levels. A study by Ratnasari et al. [6] further supported these findings by asserting that education is one of the primary factors in poverty classification models.

Meanwhile, population size is often associated with resource availability and overall welfare. Regions with high population density tend to face greater social and economic pressures, which may worsen poverty conditions if not supported by equitable development [9]. Azis et al. [3] also highlighted that population growth without improved access to public services and employment opportunities can exacerbate social inequality and deepen poverty.

These findings indicate that socioeconomic variables such as education, expenditure, and population size are highly relevant for analysis in the context of poverty in Aceh. Therefore, the selection of variables in this study is based on empirical evidence from previous research that demonstrates a strong relationship between socioeconomic factors and poverty status.

C. Model Evaluation

Although binary logistic regression is generally used in an inferential context to examine the influence of independent variables on a categorical dependent variable, this study also includes classification performance evaluation as supplementary information. The purpose of this evaluation is to assess how well the model distinguishes poverty status in the test data. The evaluation metrics used include precision, recall, F1-score, and area under the curve (AUC), which are commonly applied in classification analysis to measure accuracy, sensitivity, and the model's ability to identify the poor and non-poor categories [10]. The basic structure of the confusion matrix is presented in Table 1.

TABLE I
CONFUSION MATRIX.

	Predicted Positive	**Predicted Negative**
Actual Positive	True Positive (TP)	False Negative (FN)
Actual Negative	False Positive (FP)	True Negative (TN)

From this confusion matrix, several evaluation metrics can be calculated as follows, based on the study by [11]:

1. Accuracy: Indicates the proportion of correct classifications relative to the total number of observations.
   Akurasi $= \frac{TP+TN}{TP+TN+FP+FN}$.....(2)
2. Precision: Indicates how accurately the model predicts the positive class.
   Precision $= \frac{TP}{TP+FP}$ .....(3)
3. Recall (Sensitivity): Measures the model's ability to correctly identify the actual positive cases.
   Recall $= \frac{TP}{TP+FN}$ .....(4)
4. F1-Score: The harmonic mean of precision and recall, useful in cases of class imbalance.
   1-Score $= 2$ x $\frac{Precision \text{ x } Recall}{Precision+Recall}$ .....(5)
5. AUC (Area Under the Curve): Measures the model's ability to distinguish between the positive and negative classes. An AUC value close to 1 indicates excellent classification performance.

Ratnasari et al. [6] applied this evaluation in the development of the BPBLR (Bivariate Polynomial Binary Logit Regression) model to analyze poverty depth and successfully demonstrated high model performance.

D. Multicollinearity

Multicollinearity occurs when two or more independent variables in a model are highly correlated. This can lead to instability in the estimation of regression coefficients. To detect multicollinearity, the Variance Inflation Factor (VIF) is used, calculated using the following formula:

$VIF_j = \frac{1}{1-R_j^2}$ .....(6)

Where R^2_j is the coefficient of determination from the regression of the independent variable against all other independent variables. A VIF value greater than 10 indicates high multicollinearity, which may potentially reduce the reliability of regression coefficient estimates [12].

E. Previous Studies

The binary logistic regression approach has been widely applied in poverty analysis in Indonesia, as demonstrated by study [13], which found a significant effect of average years of schooling on poverty depth in East Java, and study [14], which identified the Human Development Index (HDI) and Gini Ratio as key national-level predictors through the backward elimination method. Study [15] also confirmed the significant influence of education-related variables (MYS and EYS), with the model achieving an accuracy of 85.3%. Ratnasari et al. [6] developed the BPBLR model and emphasized the importance of selecting relevant variables for accurate poverty classification. In the international context, Doherty and Wells [5] demonstrated the effectiveness of logistic regression in identifying socio-economic risks in epidemiological studies. Meanwhile, Salong et al. [2] and Azis et al. [3] highlighted the significance of education, expenditure, and population density as key factors in explaining poverty.

Considering the findings from various studies, it can be concluded that binary logistic regression is not only relevant for statistical inference but can also be utilized as a predictive tool in data-driven decision-making. This forms the basis of the present study to integrate both inferential and classification approaches in evaluating poverty status in Aceh Province more comprehensively.

III. RESEARCH METHODOLOGY

A. Data and Variables

This study uses secondary data from the Aceh Provincial Statistics Agency (Badan Pusat Statistik/BPS) for the period 2019 to 2023. The dataset covers 23 districts/cities, with a total of 115 samples.

The model was built based on one dependent variable and five independent variables. The variable representing the number of poor people was excluded from the model due to its direct correlation with the target variable (poverty status), which could potentially lead to data leakage. The variables are detailed as follows:

1. Dependent variable (Y)

Poverty status is categorized in binary form (1 = poor, 0 = not poor). To convert numerical data into binary labels, a threshold was applied based on the average percentage of the poor population in Aceh Province for each respective year. The cut-off determination was carried out as follows:

- If the percentage of the poor population in a district/city is higher than the provincial average in that year, it is classified as poor (1).
- If the percentage is equal to or lower than the provincial average, it is classified as not poor (0).

Example:
In 2022, the average percentage of the poor population in Aceh Province was 14.64%. Therefore:

- District A with a rate of 15.10% → categorized as poor (1)
- District B with a rate of 13.80% → categorized as not poor (0)

The cut-off was determined based on the average poverty percentage in Aceh Province for each year to better reflect the regional context and accommodate annual variations. This approach also aims to maintain class balance, ensuring that the model does not favor the majority class and can perform classification more equitably.

2. Independent variables (X)

X1: Open Unemployment Rate (%)
X2: Gross Regional Domestic Product (GRDP) (Thousands of Rupiah)
X3: Expenditure Per Capita(Thousands of Rupiah)
X4: Total Population (People)
X5: Number of Poor Population (People)
X6: Average Length of Schooling (Years)

The selection of variables was based on empirical literature and development theory, which suggest that economic, educational, and demographic factors influence poverty status. Supporting references can be found in [1]–[3].

B. Data Collection and Preprocessing Techniques

The data were obtained through a documentation method from the official website of the Aceh Central Statistics Agency (bps.go.id), in the form of socio-economic indicator tables for each district/city. All preprocessing steps were carried out to ensure data quality before applying logistic regression modeling, both for inferential purposes (using statsmodels) and predictive purposes (using scikit-learn). The preprocessing stages included:

1. Handling of missing data

The dataset was examined to identify missing values. If any were found, the missing entries were replaced with the median value of the corresponding column. This approach was chosen to maintain the stability of the data distribution and avoid distortion caused by extreme imputations.

2. Outlier Detection and Cleaning

Outliers in the numerical variables were detected using the Interquartile Range (IQR) method, with the lower bound defined as (Q1 – 1.5×IQR) and the upper bound as (Q3 + 1.5×IQR). Data points exceeding these bounds were not removed but were handled using a capping method, in which extreme values were replaced with the nearest boundary values. This approach preserves data integrity while preventing model distortion caused by outliers.

The boxplot visualization prior to outlier handling revealed the presence of extreme values in several numerical variables, which could affect the stability of model estimation. This condition is illustrated in Figure 1.

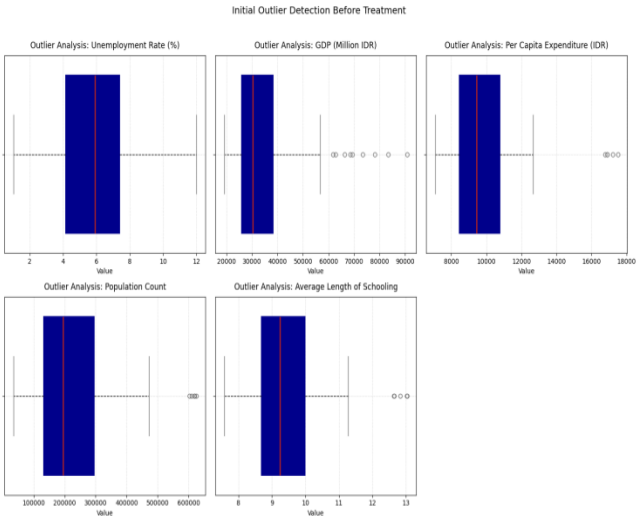

Fig 1. Outlier detection before handling.

Figure 1 presents the outlier detection results for the five independent variables using boxplots. Data points falling outside the interquartile range (IQR) indicate extreme values that may affect model stability. The most prominent outliers were found in the variables GRDP, expenditure per capita, and population size, while the variable average years of schooling showed outliers only on the upper side, and the unemployment rate did not exhibit any significant outliers.

After handling outliers using the Interquartile Range (IQR) method, the data distribution became more controlled and symmetrical, as shown in Figure 2.

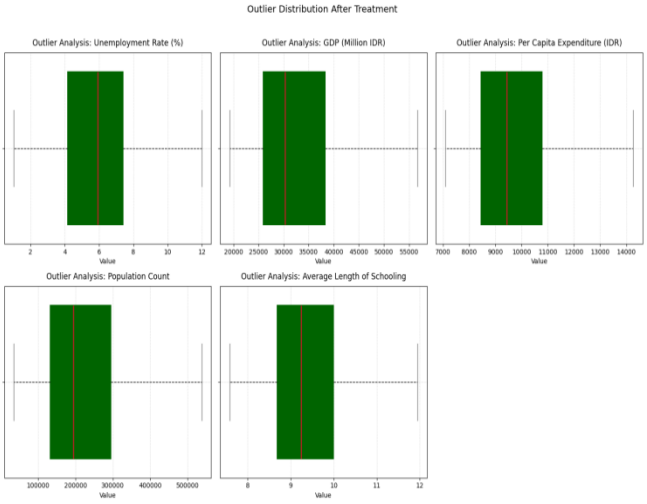

Fig 2. Outlier detection after handling.

This figure 2 illustrates the distribution of independent variables after addressing extreme values using the interquartile range (IQR) method. Compared to the previous visualization (Figure 1. Outlier Detection Before

Handling), the data distributions across all variables now appear more symmetrical and no longer exhibit extreme outliers. This treatment aims to reduce distortion in the estimation of regression coefficients and to minimize the influence of extreme observations on model performance.

2. Training and Test Data Splitting

The data split was applied exclusively to the predictive approach (scikit-learn), using a time-based split based on the year column. Data from 2019 to 2022 were used as the training set (80%), while data from 2023 served as the test set (20%) to evaluate the model's performance on previously unseen data (out-of-sample evaluation).

This approach aims to prevent data leakage and reflect a realistic prediction scenario. Meanwhile, the inferential approach (using *statsmodels*) employs the entire dataset from 2019 to 2023 to obtain stable parameter estimates that are representative of the overall population.

3. Preprocessing Specific to the Predictive Model (scikit-learn)

For the prediction-based approach using *scikit-learn*, two additional preprocessing steps were performed as follows:

a. Standardization: All independent variables were normalized using *StandardScaler*, resulting in a mean of zero and a standard deviation of one. This step is necessary because logistic regression in *scikit-learn* is sensitive to the scale of the variables.
b. Handling Class Imbalance: If an imbalance between the 'poor' and 'non-poor' classes was identified, oversampling of the minority class was performed on the training data using resampling with replacement. This step ensures that the model is not biased toward the majority class.

The inferential approach using *statsmodels* does not involve standardization or oversampling, as its primary objective is to estimate coefficients that can be interpreted on the original scale of the data and to preserve the distributional properties of the predictor variables.

C. Analysis Method

This study employs binary logistic regression to analyze the influence of socioeconomic variables on poverty status at the district/city level in Aceh Province. The primary approach used is inferential, implemented through the *statsmodels* library in Python. Parameter estimation is conducted using the Maximum Likelihood Estimation (MLE) method, which enables statistical significance testing of each independent variable through p-values. Furthermore, coefficient interpretation is carried out by calculating the odds ratios and confidence intervals, aiming to understand the direction and strength of each variable's influence on poverty status.

As a complement, a predictive approach was also implemented using *scikit-learn* to evaluate the classification performance of the model. The data were split by year, with the training set covering 2019–2022 and the test set comprising 2023. For the predictive model, features were standardized using *StandardScaler*, and class imbalance was addressed through oversampling. Model performance was evaluated using accuracy, precision, recall, F1-score, and area under the curve (AUC). Nevertheless, the primary focus remains on inferential analysis to support evidence-based understanding and policy formulation. Figure 3 presents the methodological framework applied in this study.

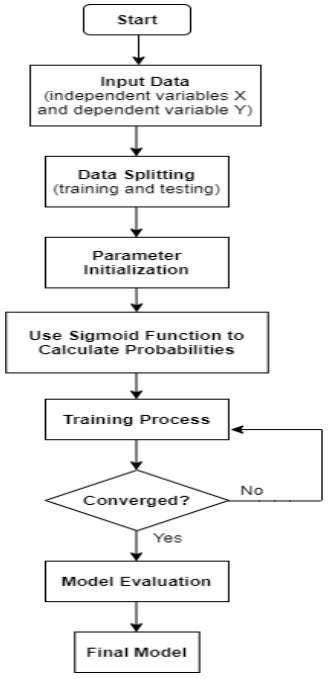

Fig 3. Method Flowchart.

Figure 3 illustrates the stages involved in developing the binary logistic regression model used in this study. The explanation of each stage is as follows:

1. Start
   This represents the starting point of the binary logistic regression modeling process.
2. Input Data (Independent Variables X and Dependent Variable Y)
   In this stage, data are input into the modeling process. The dataset consists of independent variables (X), representing the socioeconomic factors being analyzed, and a dependent variable (Y), which is binary in nature (0 = non-poor, 1 = poor).
3. Data Splitting (Training and Testing)
   The dataset is divided into training and testing sets. This split is intended to allow the model to be trained on one portion of the data and tested on unseen data to evaluate its performance.
4. Parameter Initialization

Initial values are assigned to the regression weights/coefficients. These initial values are typically zero or small random numbers before being updated through the optimization process.

5. Use Sigmoid Function to Calculate Probabilities
   The sigmoid function is applied to transform the linear combination of input variables into probabilities ranging from 0 to 1, representing the likelihood of an observation falling into a particular category (poor/non-poor).
6. Training Process
   The model is trained using the training data to minimize the loss function, typically through optimization methods such as gradient descent.
7. Converged?
   The training process is conducted iteratively until the model parameters reach stable values or the changes between iterations become negligible. The model is considered converged at this point.
8. Model Evaluation
   The trained model is then evaluated using the testing data. Evaluation is conducted using metrics such as accuracy, precision, recall, F1-score, and AUC.
9. Final Model
   After evaluation, the model is finalized and ready to be used for prediction and interpretation within the context of policy-making.

IV. RESULT AND DISCUSSION

A. Data Description

An initial descriptive analysis was conducted to illustrate the socioeconomic characteristics of districts/cities in Aceh Province during the 2019–2023 period. Table 2 presents a summary of statistics for the five main independent variables used in the logistic regression modeling, including the mean, minimum, maximum, and standard deviation values, as shown in Table 2.

TABLE II
DESCRIPTIVE STATISTICS OF VARIABLES.

Variabel	Mean	Min	Max	Standard Deviation
Open Unemployment Rate (TPT) (%)	5.86	1.03	11.99	2.23
Gross Regional Domestic Product (GRDP) (Rp)	34.777	19.229	90.764	13.777
Expenditure per Capita (Rp)	9.831	7.085	17.521	2.057
Total Population (people)	233.678	34.874	624.899	142.692
Average Length of Schooling (years)	9.50	7.58	13.04	1.16

As part of the initial exploration prior to logistic regression modeling, Mutual Information (MI) was measured between each independent variable and the target variable, poverty status. Although MI is not utilized in the inferential approach of logistic regression with statsmodels, this information remains useful for identifying the potential relevance of features before the main analysis. The Mutual Information scores are presented in Table 3.

TABLE III
MUTUAL INFORMATION SCORE.

Variabel	Skor Mutual Information
Open Unemployment Rate (TPT)	0,264850
Gross Regional Domestic Product (GRDP)	0,073293
Expenditure per Capita	0,163274
Total Population	0,397150
Average Length of Schooling	0,277641

To evaluate the relationships among the independent variables, a Pearson correlation analysis was conducted and visualized in the form of a heatmap (Figure 4). This correlation analysis is exploratory in nature and important within the context of inferential logistic regression, as high correlations between variables may lead to multicollinearity, which can reduce the reliability of the regression coefficient estimates.

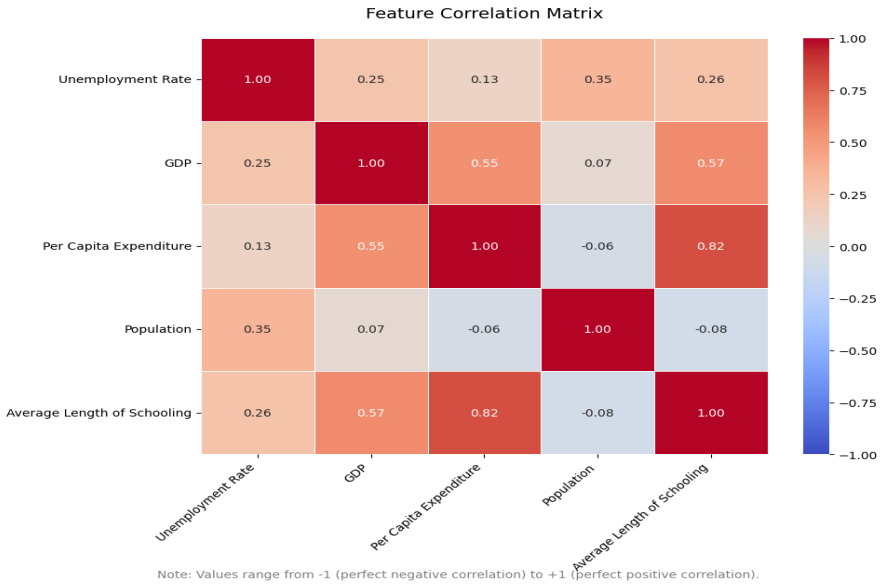

Fig 4. Correlation Heatmap.

Figure 4 illustrates the heatmap visualization shows that the variable *Per Capita Expenditure* has a very strong correlation with *Average Length of Schooling* (r = 0.82), and a moderate correlation with *GDP* (r = 0.55). Additionally, *GDP* also shows a moderate correlation with *Average Length of Schooling* (r = 0.57). Correlations of 0.7 or higher indicate a potential for multicollinearity, which can lead to instability in the estimation of logistic regression coefficients and affect p-values. Therefore, these findings serve as the basis for conducting a Variance Inflation Factor (VIF) test in the subsequent stage, in order to confirm and address potential multicollinearity within the model.

B. Logistic Regression Model Training Results

The logistic regression model was employed to identify the influence of socioeconomic variables on poverty status across districts and cities in Aceh Province. Parameter estimation was conducted using an inferential approach based on Maximum Likelihood Estimation (MLE) with the support of the *statsmodels* library.

The model was constructed using the full set of observational data from 2019 to 2023, comprising 115 observations. Based on the model training results, the log-likelihood value was –61.668, with a likelihood ratio test p-value of 9.459e-07, indicating that the model is statistically significant at the 95% confidence level. The Pseudo R-squared value of 0.2260 suggests that approximately 22.60% of the variation in poverty status can be explained by the combination of predictor variables included in the model.

The estimated parameters of the logistic regression model including coefficients, standard errors, z-values, p-values, and 95% confidence intervals are presented in Table 4.

TABLE IV
LOGISTIC REGRESSION COEFFICIENTS (FINAL MODEL).

Variabel	Coefficient	Std. Error	Z	P >|z|	[0.025	0.975]
Intercept	-0.0020	0.222	-0.009	0.993	-0.437	0.433
Open Unemployment Rate (TPT)	-0.2225	0.256	-0.870	0.384	-0.724	0.279
Gross Regional Domestic Product (GRDP)	0.4304	0.263	1.634	0.102	-0.086	0.946
Expenditure per Capita	0.0134	0.349	0.038	0.969	-0.670	0.697
Total Population	-0.8797	0.293	-3.007	0.003	-1.453	-0.306
Number of Poor People	0.3829	0.260	1.474	0.141	-0.126	0.892
Average Length of Schooling	-1.3297	0.385	-3.452	0.001	-2.085	-0.575

Based on the results, only two variables show a statistically significant influence on poverty status, namely *Average Length of Schooling* (p = 0.001) and *Total Population* (p = 0.007).

- The negative coefficient of *Average Length of Schooling* (–1.3936) indicates that an increase of one year in average schooling duration significantly reduces the log-odds of poverty.
- The coefficient for *Total Population* is also negative (–4.953e–06) and statistically significant, suggesting that an increase in population is associated with a lower likelihood of a region being classified as poor, although the numerical effect is relatively small.

The other three variables *Unemployment Rate*, *GRDP*, and *Per Capita Expenditure* have p-values greater than 0.05 and 95% confidence intervals that include zero, and are therefore considered statistically insignificant in this model. This lack of significance is suspected to be related to potential multicollinearity among the independent variables, which will be further examined through the VIF analysis in the following section.

To facilitate a clearer interpretation of the model results, Table 5 presents the odds ratios (OR) for each predictor variable along with their respective 95% confidence intervals. These values allow a practical understanding of how each independent variable affects the likelihood of a region being categorized as poor, assuming other variables remain constant, as shown in Table 5.

TABLE V
ODDS RATIOS AND 95% CONFIDENCE INTERVALS

Variabel	Odds Ratio (OR)	CI 95% (Lower)	CI 95% (Upper)	Brief Interpretation
Intercept (const)	371.980,82	1.621,76	85.320.880,00	Not interpreted; model constant only
Open Unemployment Rate (TPT)	0,876	0,697	1,102	Not significant; CI includes 1
Gross Regional Domestic Product (GRDP)	1,000046	0,999988	1,000103	Not significant; very small effect
Expenditure per Capita	1,000083	0,999628	1,000537	Not significant; small effect and CI includes 1
Total Population	0,999995	0,999991	0,999999	Significant; population increase → slightly decreases poverty likelihood
Number of Poor People	0,248	0,111	0,556	Significant; each additional year of schooling → reduces poverty likelihood by ~75%

Table 5 presents the Odds Ratio (OR), 95% Confidence Interval (CI), and the practical interpretation of each variable in the model. The OR indicates the direction and magnitude of a variable's effect on poverty status—values above 1 imply an increased likelihood of poverty, while values below 1 imply a decrease. The CI reflects the reliability of the estimate; if it includes 1, the effect is not statistically significant. These values offer practical insights into how each predictor contributes to poverty classification, summarized as follows:

- *Average Length of Schooling* has an OR of 0.248 (CI: 0.111 – 0.556), indicating that each additional year of education reduces the likelihood of poverty by approximately 75.2%, and this result is statistically significant.
- *Total Population* has an OR close to 1 (0.999995), but with a very narrow confidence interval that does not include 1, indicating a statistically significant effect despite its small numerical magnitude.
- *Unemployment Rate (TPT)*, *Gross Regional Domestic Product (GRDP)*, and *Per Capita Expenditure* have confidence intervals that include 1, suggesting that these three variables do not have a statistically significant effect on poverty status in this model.

To prevent bias in coefficient estimation due to multicollinearity, a diagnostic assessment using the Variance Inflation Factor (VIF) was conducted. High multicollinearity can inflate standard errors and compromise the statistical significance of predictors. Prior to the analysis, the variable *number of poor people* was excluded due to its strong conceptual association with the dependent variable, which could introduce redundancy and distort interpretation. The remaining VIF values were then evaluated to ensure that no independent variable exhibited excessive collinearity. Detailed results of the VIF analysis are presented in Table 6.

TABLE VI
VARIANCE INFLATION FACTOR (VIF).

Variable	VIF	Interpretation
Open Unemployment Rate (TPT)	10.41	Moderate to high multicollinearity
Gross Regional Domestic Product (GRDP)	17.83	High multicollinearity
Expenditure per Capita	108.13	Very high multicollinearity
Total Population	5.99	Safe (significant)
Number of Poor People	2.52	Safe (not significant)
Average Length of Schooling	119.81	Very high multicollinearity

The results of the Variance Inflation Factor (VIF) analysis indicate a very high degree of multicollinearity in the variables *Average Length of Schooling* (118.65), *Per Capita Expenditure* (107.23), and *GRDP* (17.79), which may potentially affect the statistical significance of these variables.

Nevertheless, these three variables were retained in the model due to their strong empirical and theoretical relevance to the poverty phenomenon. Previous studies have demonstrated that the average length of schooling has a significant negative effect on poverty levels, as evidenced by Safrina and Hasanah [1], Salong et al. [2], and Ratnasari et al. [6], who emphasize that education is a key component in poverty alleviation.

Likewise, per capita expenditure and GRDP have been recognized as essential indicators in reflecting a region's economic capacity and households' access to basic needs. Excluding these variables from the model may compromise conceptual integrity and omit critical information necessary for evidence-based policy formulation.

Although methods such as Principal Component Analysis (PCA) or Backward Elimination can be employed to reduce multicollinearity, this study does not adopt such approaches in order to maintain the interpretability of the model, particularly in the context of socio-economic policy. As a future alternative, such variable selection techniques may be considered to obtain more stable parameter estimates. Meanwhile, the variable *Total Population* exhibits a lower VIF value (4.36), indicating a clearer independent contribution to the model.

C. Model Performance Evaluation

To evaluate the performance of the logistic regression model in classifying poverty status, testing was conducted using 2023 test data through a classification approach based on scikit-learn. The model's predictions were assessed using several evaluation metrics: accuracy, confusion matrix, precision, recall, F1-score, and Area Under the Curve (AUC), as shown in Table 7.

TABLE VII
MODEL EVALUATION RESULT ON TEST DATA (YEARS 2023)

Evaluation Metric	Value
Accuracy	0.70
Precision	0.81
Recall	0.70
F1-Score	0.66
Area Under the Curve (AUC)	0.69

Table 7 presents the evaluation results of the logistic regression model on the 2023 test data using five key metrics: accuracy, precision, recall, F1-score, and AUC. The results indicate a reasonably good classification performance, with an accuracy of 70% and a balanced trade-off between precision and recall. The AUC value of 0.69 reflects a moderate ability to distinguish between poor and non-poor regions. This evaluation provides a comprehensive overview of the model's effectiveness in a socio-economic context, further supported by the confusion matrix visualization in Figure 5.

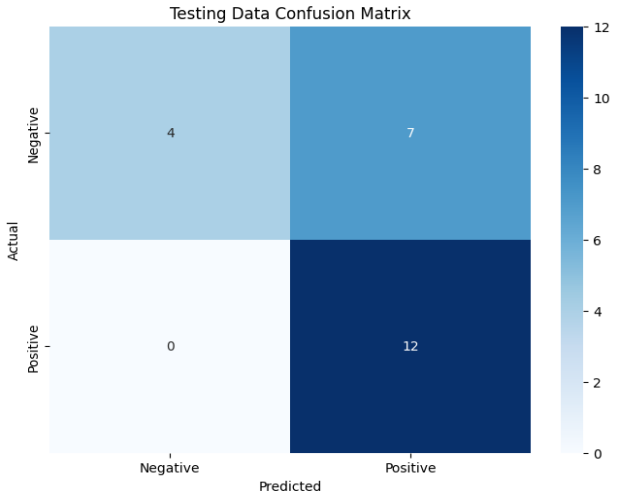

Fig 5. Confusion Matrix of the Test Data.

Based on the confusion matrix, the model successfully identified all poor regions in the test data correctly (True Positives), with no False Negatives. However, there were 7 non-poor regions misclassified as poor (False Positives), and only 4 non-poor regions were correctly identified (True Negatives). To compare performance, a baseline classifier using a majority class strategy (DummyClassifier, strategy='most_frequent') was employed, which achieved an accuracy of only 0.48. These results indicate that the model demonstrates better classification performance than a naive approach based solely on class distribution.

In addition, the model's performance is also visualized through the Receiver Operating Characteristic (ROC) curve, as presented in Figure 6 below.

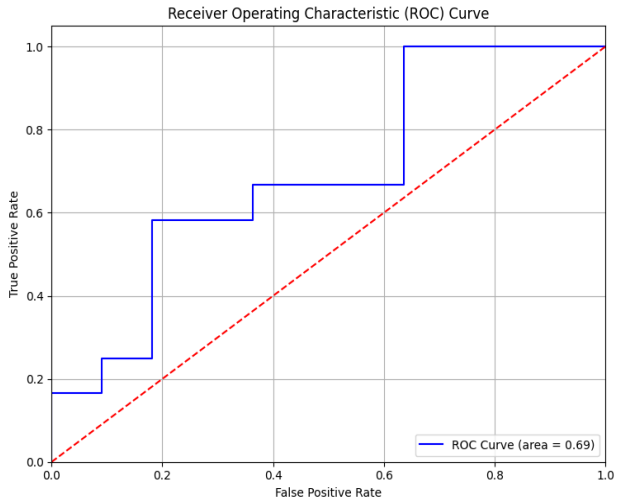

Fig 6. ROC and Precision-Recall Curves.

The AUC value of 0.69 indicates a moderate classification ability of the model in distinguishing between the two target classes. In the ROC curve, the blue line represents the actual performance of the model, while the red line serves as the random baseline. The baseline accuracy of 0.48 was calculated based on majority class predictions from the training data. Although logistic regression is generally inferential in nature, these results demonstrate its potential as a classification tool for socio-economic data, with room for improvement through non-linear algorithms such as Random Forest, SVM, or Gradient Boosting.

D. Discussion

The objective of this study is to analyze the socio-economic factors influencing poverty status in the districts and cities of Aceh Province using a binary logistic regression approach. The inferential model based on the statsmodels library reveals that, among the five independent variables examined, only two show statistically significant effects on poverty status: Average Years of Schooling and Total Population.

The variable *Average Years of Schooling* exhibits a statistically significant negative logit coefficient at the 1% level (p = 0.001), with a value of –1.3936. This implies that each additional year of average education among the population reduces the log-odds of a district or city being classified as poor. When converted to an odds ratio, the value becomes 0.248, indicating that the likelihood of an area being categorized as poor decreases by approximately 75.2%. Empirically, this suggests that education plays a direct role in enhancing the community's capacity to escape the cycle of poverty. In Aceh, regions such as Aceh Singkil and Pidie Jaya, which record relatively low average years of schooling, generally exhibit higher percentages of people living in poverty. This finding is consistent with the study by Salong et al. (2024), which emphasizes the importance of education in strengthening regional socio-economic resilience [2].

The *Total Population* variable exhibits a statistically significant negative logit coefficient at the 1% level (p = 0.007), with a value of –0.000004953. This indicates that each additional unit increase in population reduces the log-odds of a district or city being classified as poor. When converted to an odds ratio, the value becomes 0.999995, meaning that a one-unit increase in population decreases the likelihood of poverty by approximately 0.0005%. Although the numerical effect is small, the negative direction of the coefficient suggests that districts or cities with larger populations tend to have a lower probability of being categorized as poor. This may be associated with potential economies of scale, better access to infrastructure, and broader distribution of fiscal resources. Nevertheless, as the odds ratio is very close to 1, this effect should be interpreted with caution.

Conversely, the other three variables—Open Unemployment Rate (TPT), GRDP, and Expenditure per Capita—were not statistically significant in this model. One of the main factors that may explain this result is the presence of multicollinearity among the independent variables, particularly between Average Years of Schooling and Expenditure per Capita (VIF > 100), as well as between GRDP and Expenditure per Capita (VIF > 10). When two or more variables are highly correlated, the stability of coefficient estimates decreases, and the associated p-values tend to increase. This phenomenon is also discussed in the study by Aini et al. (2023), which states that multicollinearity can cause theoretically important variables to appear statistically insignificant in logistic regression models [4].

Although some variables were found to be statistically insignificant, this study opted not to remove them from the model. The rationale was to maintain theoretical coherence, as all included variables are grounded in development theory and prior poverty studies. This strategy aligns with the approach recommended by Doherty and Wells (2025), who emphasize the importance of retaining substantively relevant variables even when they are not statistically significant particularly in the context of complex socio-economic phenomena [5].

In line with this approach, the discussion does not aim to formulate new policies but rather to relate the research findings to existing programs and policies at both national and regional levels. The objective is to strengthen the justification for data-driven interventions

and to support the development of more contextually appropriate policies. Therefore, the policy implications derived from the regression results are presented as a conceptual contribution to reinforce programs that are relevant to the socio-economic conditions of Aceh Province.

Based on the research findings, the average length of schooling has been proven to be the most significant factor in reducing the likelihood of poverty. Therefore, the policy implication of this result is the importance of strengthening the quality and effectiveness of existing educational interventions through region-specific, data-driven adjustments. The Aceh Provincial Government and its regencies/cities need to:

- Identify regions with an average years of schooling below 9 years as priority targets.
- Encourage the development of more targeted scholarship programs for poor families, utilizing welfare data such as the *Integrated Social Welfare Data (DTKS)* to improve accuracy in targeting.
- Strengthen public education efforts through schools and village institutions to reduce the risk of school dropouts, particularly among vulnerable groups.

In addition, the finding that population size has a negative effect on poverty status albeit with a small effect suggests that more populous regions tend to possess better socio-economic capacity. Therefore, local governments may:

- Adjust the development budget allocation approach by considering population size, ensuring that highly populated districts such as Bireuen, North Aceh, and Banda Aceh receive proportionate access to public services.
- Expand the coverage of education, healthcare, and social protection services in densely populated or high-growth areas to strengthen local socio-economic resilience.

Thus, these findings provide an empirical foundation to refine education policies and the distribution of social services based on evidence, rather than relying solely on macro-level indicators. Overall, the results of this study affirm that an inferential statistical approach can be effectively utilized to identify the determinants of poverty at the regional level, and that the outcomes can support the formulation of evidence-based socio-economic policies in Aceh Province.

In addition to strengthening existing strategies, these findings also present an opportunity to evaluate the effectiveness of interventions that have not shown significant impact. For instance, the insignificance of per capita expenditure and GRDP in the model may serve as a basis to assess whether economic development policies in Aceh have truly contributed to poverty reduction, or whether they remain trapped in patterns of non-inclusive growth.

V. CONCLUSION

This study aims to analyze the socio-economic factors influencing poverty status in Aceh Province using a binary logistic regression approach. The inferential analysis conducted with the statsmodels framework reveals that out of the five independent variables tested, only two have a statistically significant effect on poverty status at the district/city level: Average Years of Schooling and Total Population.

- The Average Years of Schooling variable has a significant negative effect on poverty status, with an odds ratio of 0.248. This indicates that an increase of one year in average schooling can reduce the likelihood of a region being classified as poor by 75.2%.
- The Population Size variable also shows a significant negative effect on poverty status, with an odds ratio of 0.999995. This means that an increase in population slightly decreases the likelihood of a region being classified as poor by approximately 0.0005%, although the numerical effect is minimal. This finding suggests that regions with larger populations tend to have lower poverty levels, possibly due to economies of scale and better access to infrastructure.

Meanwhile, the variables Open Unemployment Rate (OUR), Gross Regional Domestic Product (GRDP), and Expenditure per Capita were found to be statistically insignificant in the model, likely due to high multicollinearity among the independent variables. However, these variables were retained in the model due to their relevance in poverty literature and to preserve the conceptual integrity of the model.

As a complement to the inferential approach, the predictive model implemented using scikit-learn on the 2023 test data achieved an accuracy of 70%, with an F1-score of 0.66 and an AUC of 0.69. Although not the main focus, these results indicate that the model has a reasonably good classification capability in identifying impoverished regions based on socioeconomic data.

These findings reinforce the importance of the education sector as a key instrument for poverty alleviation and highlight the potential of population-based approaches in social policy allocation. The binary logistic regression model has proven to provide valuable inferential insights for the formulation of data-driven socioeconomic policies at the regional level.

REFERENCES

[1] U. Hasanah and L. Safrina, "Determinan tingkat kemiskinan di Provinsi Aceh," *I-Finance: A Research Journal on Islamic Finance*, vol. 10, no. 1, pp. 138–154, 2024.

[2] E. Salong, M. Bugis, I. T. Matitaputty, and F. Ramly, "The effect of education, health and per capita income on poverty rate in Maluku Province," *Daengku: Journal of Humanities and Social Sciences Innovation*, vol. 4, no. 3, pp. 457–464, 2024, doi: 10.35877/454RI.daengku2567.

[3] I. Azis, I. M. Sumertajaya, S. S. Purwaningsih, and S. S. Tjahjawati, "Penentuan faktor kemiskinan Indonesia menggunakan regresi logistik," *Journal of Mathematics, Computations, and Statistics*, vol. 6, no. 1, pp. 61–65, 2023.

[4] A. N. Aini, A. U. O. Ashar, T. I. Lestari, I. M. Nur, and R. Wasono, "Pemodelan tingkat kemiskinan di Papua Barat dengan pendekatan binary logistic regression," *Square: Journal of Mathematics and Mathematics Education*, vol. 5, no. 2, pp. 113–120, 2023, doi: 10.21580/square.2023.5.2.17169.

[5] I. A. Doherty and M. E. Wells, "Sleep disparities in the United States: Comparison of logistic and linear regression with stratification by race," *Sleep Epidemiology*, vol. 5, Art. no. 100106, 2025, doi: 10.1016/j.sleepe.2025.100106.

[6] V. Ratnasari, Purhadi, M. Rifada, and A. T. R. Dani, "Explore poverty with statistical modeling: The bivariate polynomial binary logit regression (BPBLR)," *MethodsX*, vol. 14, Art. no. 103099, 2025, doi: 10.1016/j.mex.2024.103099.

[7] D. F. Silaban, F. A. Simanjuntak, I. O. Lumbantobing, R. Napitupulu, and Arnita, "Analisis faktor-faktor yang mempengaruhi peluang penerimaan pelamaran kerja menggunakan metode regresi logistik biner," *Trigonometri: Jurnal Matematika dan Ilmu Pengetahuan Alam*, vol. 5, no. 1, pp. 1–10, Dec. 2024, doi: 10.8734/trigo.v1i2.365.

[8] B. W. Kusuma and A. S. Widawati, "Analisis determinasi tingkat kemiskinan di Provinsi Aceh," *JIMEA: Jurnal Ilmiah MEA (Manajemen, Ekonomi, dan Akuntansi)*, vol. 8, no. 1, pp. 2211–2228, Apr. 2024.

[9] I. Hartika, "Analisis kemiskinan di Provinsi Aceh," *Brilliant: Journal of Islamic Economics and Finance*, vol. 2, no. 1, pp. 14–29, Jun. 2024.

[10] W. S. Nainggolan, Z. M. Pulungan, A. Faradhila, and G. T. Sinaga, "Analisis klasifikasi dan regresi logistik biner untuk menilai faktor-faktor yang memengaruhi kepuasan pelanggan," *Trigonometri: Jurnal Matematika dan Ilmu Pengetahuan Alam*, vol. 5, no. 2, pp. 50–60, Dec. 2024, doi: 10.8734/trigo.v1i2.365.

[11] R. K. Dinata, Fajriana, Z. Zulfa, and N. Hasdyna, "Klasifikasi Sekolah Menengah Pertama/sederajat wilayah Bireuen menggunakan algoritma K-Nearest Neighbors berbasis web," *CESS (Journal of Computer Engineering System and Science)*, vol. 5, no. 1, pp. 33–37, Jan. 2020.

[12] M. Maulita and N. Nurdin, "Pendekatan data mining untuk analisa curah hujan menggunakan metode regresi linear berganda (studi kasus: Kabupaten Aceh Utara)," *IDEALIS: Indonesian Journal of Information System*, vol. 6, no. 2, pp. 99–106, Jul. 2023.

[13] D. R. Sahputra et al., "Model regresi logistik pada indeks kedalaman kemiskinan di Provinsi Jawa Timur tahun 2021," in *Prosiding Seminar Nasional Matematika, Statistika, dan Aplikasinya*, Terbitan III, Samarinda, Indonesia, pp. 1–9, Aug. 2023.

[14] A. I. Nurrizqi, Erfiani, Indahwati, A. Fitrianto, and R. Amelia, "Pemodelan regresi logistik berbasis backward elimination untuk mengetahui faktor yang memengaruhi tingkat kemiskinan di Indonesia tahun 2021," *Jurnal Statistika dan Aplikasinya*, vol. 6, no. 2, pp. 160–169, Dec. 2022.

[15] N. P. N. Hendayanti and M. Nurhidayati, "Regresi logistik biner dalam penentuan ketepatan klasifikasi tingkat kedalaman kemiskinan provinsi-provinsi di Indonesia," *Sainstek: Jurnal Sains dan Teknologi*, vol. 12, no. 2, pp. 63–70, Dec. 2020.