TIERS Information Technology Journal
Vol.6, No.1, June 2025, pp. 16~32
P-ISSN: 2723-4533 / E-ISSN: 2723-4541
DOI: https://doi.org/10.38043/tiers.v6i1.6244

Optimizing Socioeconomic Features for Poverty Prediction in South Sumatera

Terttiaavini[1], Agustina Heryati[2], Tedy Setiawan Saputra[3]

avini.saputra@uigm.ac.id[1], agustina.heryati@uigm.ac.id[2], tdyfaith@gmail.com[3]

[1] Master of Computer Science, Universitas Indo Global Mandiri, South Sumatera, Indonesia
[2] Information Systems, Universitas Indo Global Mandiri, South Sumatera, Indonesia
[3] Management, STIE APRIN, South Sumatera, Indonesia

ABSTRACT

Poverty in South Sumatera remains a complex challenge influenced by socioeconomic factors. Traditional methods often fail to capture nonlinear relationships critical for accurate prediction. This study enhances poverty prediction by optimizing feature engineering using 32-variable socioeconomic data from South Sumatra for the years 2019 to 2023. Data preprocessing included cleaning, imputation, normalization, and outlier handling. Feature aggregation created composite indices: Education Index (P1, P2, P3), Health Index (AH1–AH4), Economic Index (IE, GR, AI, EG), and Healthcare Workforce Index (HW1–HW9). Feature interaction derived ratios such as Income vs. Economy (AN/Education Index), Infrastructure vs. Health (road length/Healthcare Workforce Index), and Unemployment vs. Workforce (HI/AT), highlighting interdependencies. Dimensionality reduction (PCA) and Lasso Regression selected eight key predictors, including Year and Poverty Level. Among tested models, Random Forest performed best ($R^2=0.7244$, MAE=0.2489). SHAP analysis identified Education and Economic Indices as top predictors. Optimized feature engineering improved model accuracy and interpretability, supporting targeted poverty reduction strategies in South Sumatera.

Keywords: Poverty Prediction; Machine Learning; Feature Engineering; Random Forest; SHAP Analysis

Article Info

Received : 15-01-2025
Revised : 11-03-2025
Accepted : 22-05-2025

Correspondence Author:

Terttiaavini
Master of Computer Science,
Universitas Indo Global Mandiri,
Jendral Sudirman Street No 629 KM. 4 Palembang South Sumatera, 30129
Email: avini.saputra@uigm.ac.id

1. INTRODUCTION

Poverty remains a complex and multidimensional socio-economic issue, posing a significant challenge in developing countries, including Indonesia. South Sumatera Province, with its large population, faces difficulties in sustainably reducing poverty rates [1] Although data from the Central Bureau of Statistics (BPS) indicate a declining poverty trend in recent years, social inequality and economic factors continue to hinder poverty alleviation efforts [2]. Various aspects, including economic conditions, education levels, access to healthcare services, and infrastructure availability, contribute to poverty dynamics [3], [4]. However, these variables do not always have a linear relationship, making it difficult to identify patterns using traditional analytical approaches. Several previous studies have attempted to measure poverty using different approaches and indicators. However, many of these studies have limitations, particularly in terms of the indicators used and the analytical methods employed. A common drawback is the use of a limited set of indicators, which fails to capture the full scope of poverty's multidimensional nature. Furthermore, traditional methods often do not adequately account for complex, non-linear relationships between socio-economic factors, leading to suboptimal poverty prediction models [5],[6]. To address these issues, there is a need for a more comprehensive and systematic approach to poverty analysis.

This study aims to overcome the limitations of previous research by applying advanced feature engineering techniques to optimize the use of socio-economic indicators. Unlike earlier studies, which often relied on basic models and a subset of variables, this research integrates a wider range of socio-economic factors, including education, health, economy, and infrastructure, and utilizes a more structured feature optimization process. Feature Aggregation, Feature Interaction, and Dimensionality Reduction (PCA) are employed to generate composite indices and better capture complex relationships among variables. The SHAP analysis method is utilized for systematic feature selection, while K-Means clustering is applied to group regions based on socio-economic characteristics, enhancing the understanding of poverty patterns across different areas.

Furthermore, while traditional regression models, such as Linear Regression, have been used for poverty prediction, they fail to capture the non-linear interactions among socio-economic variables. This study introduces more sophisticated machine learning algorithms, including Random Forest, XGBoost, and Neural Networks (MLPRegressor), which are capable of identifying complex, non-linear relationships and improving the accuracy of poverty predictions [7]. By combining these advanced techniques with optimized feature engineering, this study aims to enhance both the accuracy and interpretability of poverty prediction models.

The novelty of this study lies in its comprehensive approach to feature engineering and model evaluation. Previous works have either used basic models or failed to optimize feature selection, which limits their ability to capture the full complexity of poverty dynamics. By adopting a systematic, feature-optimized methodology, this research provides a more reliable and actionable model for poverty prediction, offering valuable insights for policymakers to design data-driven strategies aimed at poverty alleviation.

To further illustrate the distinctive nature of this study, the following table compares the approaches, methods, and indicator selection of previous poverty-related research with the approach applied in this study. Table 1 highlights the differences in methodologies and the comprehensive approach adopted here, which underscores the unique contributions of this research to the field of poverty prediction.

Several previous studies have measured poverty using different approaches and indicators. However, many of these studies have limitations in terms of indicator coverage and analytical methods. To provide a clearer picture of the novelty and unique contributions of this study, the following table compares the approaches, methods, and indicators used in relevant poverty-related research with the approach applied in this study. Table 1 presents a comparison of various studies, highlighting the differences in approaches, methods, and indicator selection, which emphasize the unique aspects of the current study's methodology.

Table 1. Comparison of Methodologies and Approaches in Related Poverty Prediction Studies

Study	**Methodology**	**Dataset**	**Key Features/ Indicators**	**Limitations**	**Findings**
Lismana & Sumarsono (2022) [5]	Stata (Descriptive Analysis)	Population Growth, HDI, Unemployment Rate	Population Growth, HDI, Unemployment Rate	Limited to basic indicators, no feature engineering, Missed nonlinear relationships, no multidimensional analysis	Focused on basic indicators of poverty, missing comprehensive factors
Annas et al. (2022) [6]	K-Means Clustering	Poverty Gap Index, Poverty Severity Index	Poverty Gap, Severity Index, Limited Socio-economic indicators	Clustering only, no predictive modeling	Classified poverty but not predictive, limited factors considered
Rahman et al. (2021) [7]	Clustering Analysis	Education, Standard of Living, Employment	Education, Standard of Living, Employment	Narrow indicator selection, may not represent poverty fully	Poor representation of multidimensional poverty due to limited factors
Zixi H (2021) [8]	Gradient Boosting	56 features, feature-rich dataset	Multiple socio-economic features	Risk of overfitting, no feature selection	Identified Gradient Boosting as best model but lacked optimization
Qi Li (2022) [9]	Logistic Regression	DHS data from 8040 households	Demographic, Housing, Asset Ownership	Missing direct economic data such as income, expenditure	Limited accuracy due to missing key economic variables
Alsharkawi et al (2021) [10]	Machine learning (LightGBM)	Household expenditure and income survey data	Household income and expenditure	Survey data is costly and time-consuming	LightGBM achieved 81% F1 score, cost-effective and real-time poverty prediction

Maruejols et al (2022) [11]	ML comparison (Random Forest, SVM, LASSO)	Socioeconomic data from rural China	Household income, health, education, family demographics	Limited to rural China	Random forest: 85.29% accuracy, subjective poverty linked to income and non-income factors
Current Study (2025)	Machine Learning (Random Forest, XGBoost, Neural Networks)	Socio-economic data from South Sumatera (2019-2023)	Education, Health, Economy, Infrastructure	Systematic feature engineering, handling missing data, outliers	Provides accurate poverty prediction with optimized feature engineering

Some studies also predict poverty using machine learning, but the use of limited indicators can result in prediction models that are less accurate and do not fully reflect poverty holistically [12], [13]

Various international studies and applications have demonstrated the relevance and feasibility of poverty prediction by utilizing advanced data and machine learning methods supported by World Bank data. For instance, Lee and Braithwaite (2022) applied machine learning techniques to predict poverty in Sub-Saharan Africa and South Asia using satellite imagery and mobile phone data, enabling real-time poverty monitoring in these regions. Similarly, Abbas et al. (2022) conducted a multidimensional study across 59 developing countries in Asia and Africa using the World Bank's World Development Indicators (WDI), applying supervised machine learning to identify key determinants of energy poverty [14] . These global studies collectively highlight the value of multidimensional socioeconomic data and machine learning techniques, predominantly using World Bank open data, for enhancing poverty prediction accuracy across different contexts.

However, while these approaches have proven successful globally, their implementation in South Sumatera presents distinct challenges, including fragmented district-level socioeconomic data coverage and limited real-time data integration capabilities. This study adapts these proven methodologies to local conditions by leveraging available provincial data sources while overcoming data limitations through customized feature engineering techniques specifically designed for South Sumatra's socioeconomic landscape."

Data-driven and machine learning approaches are increasingly applied in poverty research to understand patterns and generate more accurate predictions. However, gaps remain in feature optimization, indicator selection, and handling of data imbalances, which need to be addressed to ensure that the models produce more robust and interpretable results [15], [16].

Feature engineering plays a crucial role in improving the accuracy of poverty prediction by optimizing the quality of data used in predictive models [17]. This process aims to create, transform, and select the most relevant features to enhance machine learning model performance. However, the primary challenge in this study lies in processing multidimensional socio-economic data, including handling missing values, outliers, and complex relationships between variables, which are often non-linear [18]. Furthermore, optimizing clustering techniques and dimensionality reduction remains an issue, as these approaches can help group regions based on similar socio-economic characteristics and reduce data complexity without losing significant information.

Several previous studies have applied machine learning for poverty prediction, yet they still face limitations in feature optimization and data analysis methods. Research utilizing linear regression has demonstrated its inability to capture non-linear relationships between socio-economic variables. Meanwhile, decision-tree-based algorithms such as Random Forest and Gradient Boosting have been employed in some studies, but optimal feature selection remains a challenge [19]. Additionally, earlier research tended to use variable subsets without a systematic feature optimization strategy, preventing predictive models from fully capturing the complex patterns within socio-economic data. Therefore, this study adopts a more structured feature engineering approach, incorporating Feature Aggregation, Feature Interaction, Feature Selection using SHAP Analysis, and Dimensionality Reduction with PCA to enhance poverty prediction model accuracy [20], [21].

Based on these challenges, this study aims to develop a more accurate poverty prediction model by optimizing feature engineering techniques and applying machine learning methods. To achieve this goal, this study utilizes socio-economic data from South Sumatera Province for the period 2019–2023, obtained from BPS and other official sources. The dataset includes key indicators such as education, health, economy, and infrastructure, with poverty rate as the target variable.

During data processing, this study applies various preprocessing techniques to improve data quality before using it in predictive models. Missing values are handled using mean-based imputation for variables with minimal missing data, while variables with a large number of missing values are addressed using the RandomForestRegressor method. Outliers are detected and corrected using the Interquartile Range (IQR) and Capping (Winsorization) methods to ensure a representative data distribution [22]. Additionally, MinMaxScaler normalization is applied to enhance machine learning model stability and maintain consistent variable scaling.

During the Feature Engineering stage, this study constructs new features that better represent the poverty phenomenon. Feature Aggregation is used to develop composite indices such as Education Index, Health Index, and Economic Index to summarize information from multiple related variables. Meanwhile, Feature Interaction is applied to capture more complex inter-variable relationships by creating ratios such as Income_vs_Economy and Infrastructure_vs_Health. To improve model efficiency, Feature Selection using SHAP Analysis is conducted to identify the most influential features in poverty prediction. Furthermore, Dimensionality Reduction with Principal Component Analysis (PCA) is applied to reduce the number of features without losing significant information. Finally, Clustering using K-Means is implemented to group regions based on poverty patterns, enabling a more targeted and data-driven analysis [23], [24].

In the modeling stage, this study evaluates the performance of several machine learning algorithms, including Linear Regression, Random Forest, XGBoost, and Neural Network (MLPRegressor). Model evaluation is conducted using Mean Absolute Error (MAE), Mean Squared Error (MSE), and R^2 Score to measure prediction accuracy and assess the effectiveness of each model in analyzing complex inter-variable relationships [25], [26].

This study presents novelty through a more systematic and comprehensive feature engineering approach in poverty analysis. Unlike previous studies that only used variable subsets without feature optimization, this study implements a structured method to capture complex relationships among socio-economic variables. Additionally, this study evaluates not only traditional regression models but also compares the performance of ensemble learning and deep learning models, providing broader insights into the effectiveness of different approaches in poverty prediction.

Current poverty prediction models exhibit three methodological shortcomings: (1) inadequate feature engineering, (2) improper handling of nonlinear relationships, and (3) suboptimal algorithm selection. This study examines two corresponding hypotheses: (H1) Feature engineering techniques incorporating aggregation and interaction terms yield significantly better predictive performance than basic feature sets. (H2) Ensemble machine learning methods demonstrate superior accuracy compared to traditional linear models when analyzing South Sumatra's multidimensional poverty data. These predictions follow from existing approaches' failure to capture known complex interactions among education, employment, and infrastructure indicators in the region.

This study contributes to the development of a more systematic feature engineering approach to improve poverty prediction accuracy based on multidimensional socio-economic data. By evaluating various machine learning algorithms, this study determines the best-performing model for poverty analysis. After selecting the optimal model, the SHAP (SHapley Additive Explanations) method is applied to identify key factors influencing poverty, enhancing model interpretability and transparency in inter-variable relationship analysis. This approach contributes to the development of more accurate predictive models and supports policymakers in designing data-driven strategies for more effective poverty alleviation.

2. RESEARCH METHOD

This study implements a systematic approach to data processing, variable analysis, and predictive model development in a structured manner. The stages of this research are as follows:

2.1. Data Collection

This study uses socioeconomic data from South Sumatra for the period 2019–2023, obtained from the Central Bureau of Statistics (BPS). This timeframe was chosen because, as of 2024, the poverty rate in South Sumatra remains high at 10.97%, while the national poverty rate has decreased to 9.03% (Badan Pusat Statistik Indonesia, 2024; BPS Provinsi Sumatera Selatan, 2024) [27], [28]. The data includes various key aspects, namely education, employment, access to healthcare services, the number of healthcare workers, as well as economic and infrastructure indicators.

In the education aspect, the collected variables include the number of people attending elementary school (ED1), junior high school (ED2), and senior high school (ED3). In the employment aspect, the data covers the unemployment rate based on education level, including unemployment at the elementary level (AJ1), junior high level (AJ2), and senior high level (AJ3). Additionally, the labor force participation rate (TPAK) is considered at each education level, namely elementary (LP1), junior high (LP2), and senior high (LP3).

In the health sector, this study collected data on access to healthcare services, such as the availability of general hospitals (AH1), community health centers (AH2), clinics or health posts (AH3), and maternal and child health centers (AH4). Moreover, the number of healthcare professionals was analyzed, including the number of doctors (HW1), dentists (HW2), nurses (HW3), midwives (HW4), pharmacists (HW5), nutritionists (HW6), and medical laboratory technicians (HW7).

Apart from these variables, this study also considers infrastructure and economic factors, such as road length based on surface type (LR), the number of motorized vehicles by type (MV), access to technology (AT), average monthly net income (AI), the growth rate of Gross Regional Domestic Product (EG), the percentage of households receiving KPS/KKS assistance (HH), and indicators of gender inequality and access to opportunities (V8). This study integrates various socioeconomic variables to provide a comprehensive analysis for predicting poverty levels in South Sumatera. It utilizes a total of 90 data samples with 32 features.

2.2. Descriptive Statistics Analysis

This analysis is conducted to understand the dynamics of socioeconomic variables in South Sumatra Province during the period of 2019–2023, before further processing and analysis are carried out. Descriptive statistics are used to provide a comprehensive overview of the distribution and characteristics of the data analyzed in this study. The data are presented in the form of a summary table that includes the mean, standard deviation, as well as the minimum and maximum values of each variable. Table 2 presents the summary and descriptive statistics of the poverty data.

Table 2. Summary and Descriptive Statistics of Poverty Data

fiture	mean	std	min	25%	50%	75%	max
P1 - P3	79.98	4.25	42.66	76.98	79.66	83	99.6
AJ1 - AJ3	79.98	4.25	42.66	76.98	79.66	83	99.6
LP1 - LP3	79.98	4.25	42.66	76.98	79.66	83	99.6
AN	1733821.6	423158.1	997280	1463891.5	1672149	1931359.7	2920519
IE	15054.4	30469	1509	4279.2	7513	10104.2	158294
GR	2.67	3.7	-24.1	2.52	3.29	3.91	8.53
AH1 - AH4	178.04	357.73	1	40.31	94.5	150.44	6740
HW1 - HW9	476.7	976.64	1	215.5	215.5	294.61	17292
LR	149.62	274.74	13.44	61.27	97.72	140.98	1781.02
MV	29037	64683.8	1256	3332	8718	11361.5	287760
HI	68.76	3.33	64.32	67.02	67.99	69.41	79.47
AT	9.01	4.64	3.02	5.76	7.83	11.24	22.41
AI	1511401.2	379766.4	787843	1246915.5	1459105	1674635.2	2487288
EG	2.67	3.70	-24.1	2.52	3.29	3.91	8.53
HH	11.6	4.67	5.21	8.07	10.64	13.9	26.08

The poverty data contains outliers and missing values in several variables, namely AN (1 missing value), IE (18 missing values), GR (5 missing values), AH3 (1 missing value), AH4 (1 missing value), HW1 (3 missing values), HW2 (4 missing values), HW6 (2 missing values), HW7 (2 missing values), HW9 (2 missing values), LR (13 missing values), HI (18 missing values), and EG (5 missing values). The presence of outliers and missing values may hinder the prediction process, therefore, they need to be addressed with appropriate methods such as imputation to handle missing values and outlier detection techniques to manage abnormal data.

2.3. Data Preprocessing

Data Cleaning: The data cleaning stage includes missing value imputation and outlier handling to ensure the optimal quality of the dataset [29]; *Missing Value Imputation:* Missing value imputation is performed with two methods based on the amount of missing data. Variables with a small number of missing values (less than 3), namely HW1, HW6, HW7, HW9, AN, and AH4, are imputed using mean imputation with `impleImputer (strategy= 'mean')`. This method is chosen because the data distribution is close to normal, so the mean can provide a reasonably accurate estimate, and the imputation process is simpler and faster. In contrast, variables with a larger number of missing values (more than 4), namely IE, HI, LR, GR, EG, and HW2, are imputed using Random Forest Imputer through : `IterativeImputer(estimator= RandomForestRegressor(random_state=42`. This method is better at capturing the relationships between variables compared to mean imputation, especially when the data is not normally distributed. After the imputation process is complete, a check is performed using `data.isnull().sum()` to ensure there are no remaining missing values [30] . *Outlier Detection and Handling:* Outlier detection is performed using the Interquartile Range (IQR) method. This method is chosen because it is effective in identifying extreme values without relying on assumptions about the data distribution, is more resistant to skewness, and detects outliers based on the lower bound (Q1 - 1.5 * IQR) and the upper bound (Q3 + 1.5 * IQR). Outlier handling is applied using Capping (Winsorization), where extreme values are replaced with the lower or upper bound. This method is chosen because it preserves important information in the dataset without removing data, while minimizing the impact of outliers on the analysis results and predictive models. After this process, the six variables with the highest number of outliers are visualized using histograms and Kernel Density Estimation (KDE), as shown in Figure 1.

.

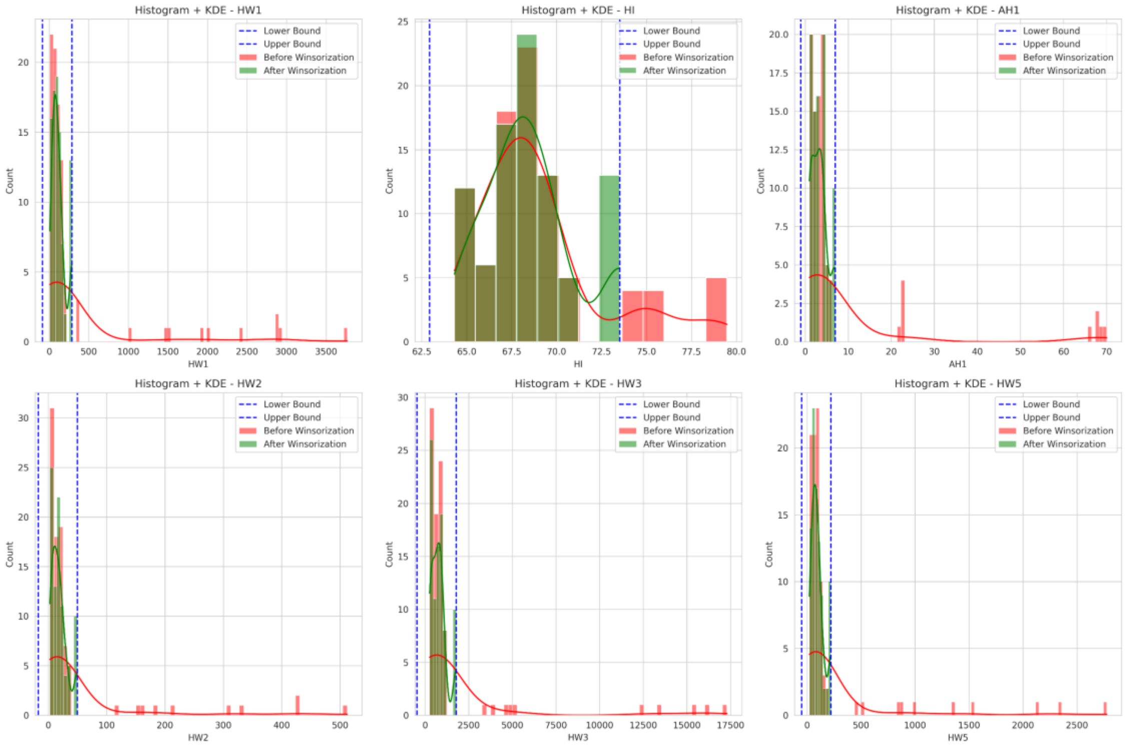

Figure. 1 Comparison of Data Distribution Before and After Outlier Handling Using Winsorization

Data Transformation: The data transformation stage is performed to align the variable scales so that analysis and machine learning models can function more optimally.Data normalization using MinMaxScaler is applied to align the variable scales within the range of 0 to 1. This technique is crucial to prevent variables with larger scales from dominating others [31] . MinMaxScaler is chosen because it preserves the original distribution shape of the data and is more effective for distance-based algorithms such as K-Nearest Neighbors (KNN), Support Vector Machine (SVM), and Neural Networks. Normalization is performed using the following formula:

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}} \tag{1}$$

where X' is the normalized value, X is the original value, X_{min} is the minimum value, and X_{max} is the maximum value in the feature. After the transformation is complete, the dataset is saved in a new file for use in further analysis. *Correlation Analysis and Heatmap Visualization:* Correlation analysis is performed to understand the relationships between variables in the dataset, which is an essential step in Exploratory Data Analysis. Correlation is measured using the Pearson correlation coefficient, with a value range between -1 and 1, where a value close to 1 indicates a strong positive relationship, a value close to -1 indicates a strong negative relationship, and a value close to 0 indicates a weak or insignificant relationship [32]. To visualize these relationships, a heatmap is used, providing an overview of the interrelationships between features in the dataset. Figure 2 displays the correlation matrix between all features used in the analysis.

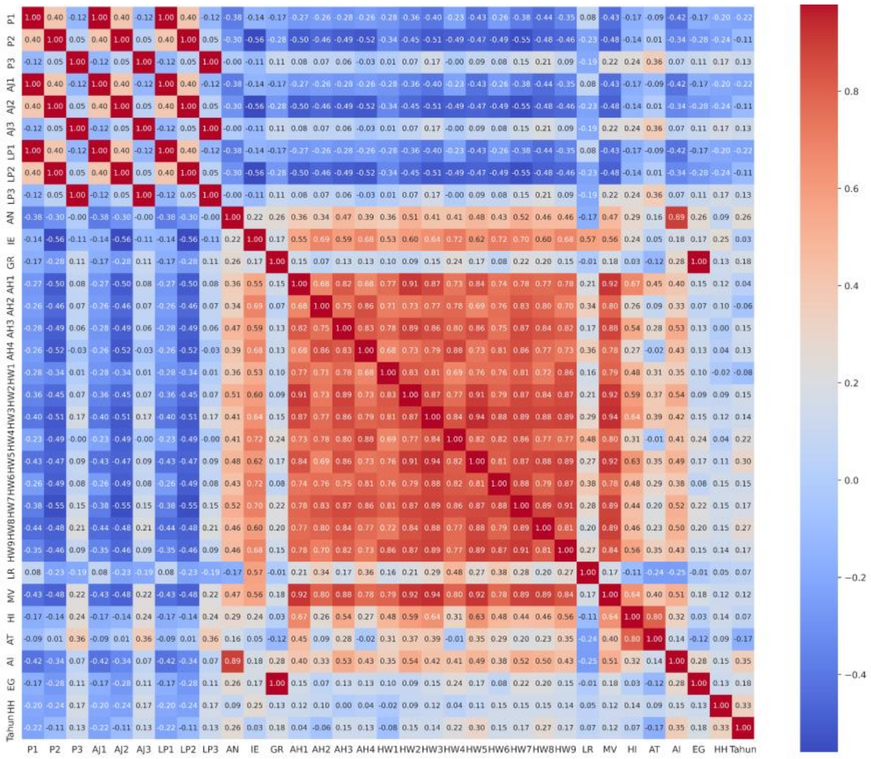

Figure. 2 Correlation Matrix of Variables in the Dataset

The correlation analysis results show several significant patterns of relationships between variables, including : *Education variables* (P1, P2, P3) have a positive correlation with labor force participation (LP1, LP2, LP3), which indicates that the higher the level of education, the greater the likelihood that an individual will enter the labor force; *Unemployment variables* (AJ1, AJ2, AJ3) have a negative correlation with education, which indicates that individuals with higher education levels tend to have lower unemployment rates; *Access to healthcare services* (AH1, AH2, AH3, AH4) is correlated with the number of healthcare workers (HW1 - HW9), which shows that the number of medical workers affects public access to healthcare services. Economic indicators have a relationship with unemployment and labor force participation, indicating a connection between economic conditions and social welfare.

The results of this correlation analysis serve as the basis for performing feature engineering, where new features are constructed to improve the representation of information in the dataset.

2.4. Feature Engineering

Feature engineering is an essential step in data analysis that involves transforming raw data into meaningful features to enhance the performance of predictive models. This process includes techniques like feature aggregation, variable interaction, dimensionality reduction, and feature selection, all of which aim to improve the model's predictive accuracy and interpretability. In this study, Feature Engineering is carried out through three main approaches : Feature Aggregation, Feature Interaction, and Dimensionality Reduction using PCA and Lasso Regression [33].

Feature Aggregation: This technique combines related variables into composite indices to provide a more informative representation of the data. The following indices were created: Education Index: The average of education indicators (P1, P2, P3); Health Index: The average of health indicators (AH1, AH2, AH3, AH4); Economic Index: The average of economic variables (IE, GR, AI, EG); Healthcare Workforce Index: The average of key healthcare workforce indicators (HW1-HW9). Implementation Code:

```
python
df['Education Index'] = df[['P1', 'P2', 'P3']].mean(axis=1)
df[' Health Index'] = df[['AH1', 'AH2', 'AH3', 'AH4']].mean(axis=1)
df[' Economic Index '] = df[['IE', 'GR', 'AI', 'EG']].mean(axis=1)
df['Healthcare Workforce Indeks'] = df[['HW1', 'HW2', 'HW3', 'HW4','HW5', 'HW6',
'HW7', 'HW8', 'HW9']].mean(axis=1)
```

The creation of new features is carried out using the feature ratio approach to capture hidden relationships between variables. The feature Income vs. Economy is derived from the ratio of income (AN) to the Education Index, reflecting the economic contribution of individuals relative to educational attainment. The

Infrastructure vs. Health feature represents the comparison between road length (LR) and the Healthcare Workforce Index, illustrating the relationship between infrastructure and the availability of health personnel. The Unemployment vs. Workforce feature is generated from the ratio of the unemployment rate (HI) to the labor force (AT), indicating the proportion of unemployment in the labor market. Meanwhile, the Transportation vs. Roads feature is obtained from the ratio of the number of motorized vehicles (MV) to road area (AI), representing traffic density. A small constant value of 1e^{-6} is added to the denominator to prevent division by zero.

Feature Interaction: Feature interaction aims to capture the relationships between variables by creating ratio-based features. These ratios help to reveal hidden connections between socio-economic indicators: Income vs. Economy: The ratio of income (AN) to the Education Index, indicating the economic contribution relative to education levels; Infrastructure vs. Health: The ratio of road length (LR) to the Healthcare Workforce Index, highlighting the relationship between infrastructure and healthcare availability; Unemployment vs. Workforce: The ratio of unemployment rate (HI) to the labor force (AT), indicating unemployment within the labor market; Transportation vs. Roads: The ratio of the number of motorized vehicles (MV) to road area (AI), representing traffic density. A small constant value of 1e^{-6} is added to the denominator to prevent division by zero. Implementation Code:

```
python
df['Income vs. Economy '] = df['AN'] / (df['Education Index '] + 1e-6)
df['Infrastructure vs. Health']=df['LR']/(df['Healthcare Workforce Indeks ']+1e-6)
df[' Unemployment vs. Workforce'] = df['HI'] / (df['AT'] + 1e-6)
df[' Transportation vs. Roads'] = df['MV'] / (df['AI'] + 1e-6)
```

The application of feature engineering techniques helps improve the performance of analytical and predictive models by enhancing the dataset with more representative new features.

Dimensionality Reduction (PCA): To optimize the created features, the Principal Component Analysis (PCA) technique is applied to reduce the number of features without losing significant information. PCA is used to extract five main components that represent the largest variation in the data. Besides PCA, Lasso Regression is also applied for feature selection to remove variables with low contributions to the prediction. This approach ensures that only truly relevant features are used in the model, thereby improving efficiency and prediction accuracy.

Removal of Constant Features: Further exploration of the generated features from Feature Engineering reveals that four features Basic LFPR_Ratio, Basic_Unemployment_ratio, Intermediate_ LFPR_Ratio, and Intermediate_unemployment_ratio, have constant values (1) across all observations. These four features lack variation and therefore do not provide useful information for the model. In data analysis, constant features do not contribute to data separation or classification and can increase computational complexity without significant benefits. Thus, these features are removed from the dataset to improve efficiency and avoid redundancy. After the Feature Engineering process, eight main features are produced, which are more representative and ready for modeling. Additionally, the dataset retains two other key variables: "Year" as the time variable and "Poverty Rate" as the target variable for prediction.

2.5. Clustering Analysis Based on Poverty Patterns

Clustering is employed to group regions based on poverty characteristics derived from the feature engineering process. This approach enables the identification of poverty patterns without relying on explicit regional labels, resulting in a more objective and data-driven analysis [34], [35].

The K-Means method is applied due to its computational efficiency, effectiveness in capturing numerical patterns, and flexibility in determining the optimal number of clusters. In addition, Principal Component Analysis (PCA) is utilized for dimensionality reduction, aiming to retain the most relevant information in the dataset while minimizing redundancy.

Through PCA, five principal components were selected, which collectively explain 87.4% of the total variance in the data. These components effectively capture the key socio-economic variations needed for the clustering process. The clustering results reveal three distinct clusters : Cluster 0 (red): Regions with high poverty levels, Cluster 1 (blue): Regions with stable socio-economic conditions, and Cluster 2 (green): Regions with moderate conditions. Figure 3 presents a visualization of the clustering results, illustrating the distribution of each region based on the grouped socio-economic characteristics. The color-coded representation helps to clearly understand the differences in poverty patterns across the clusters.

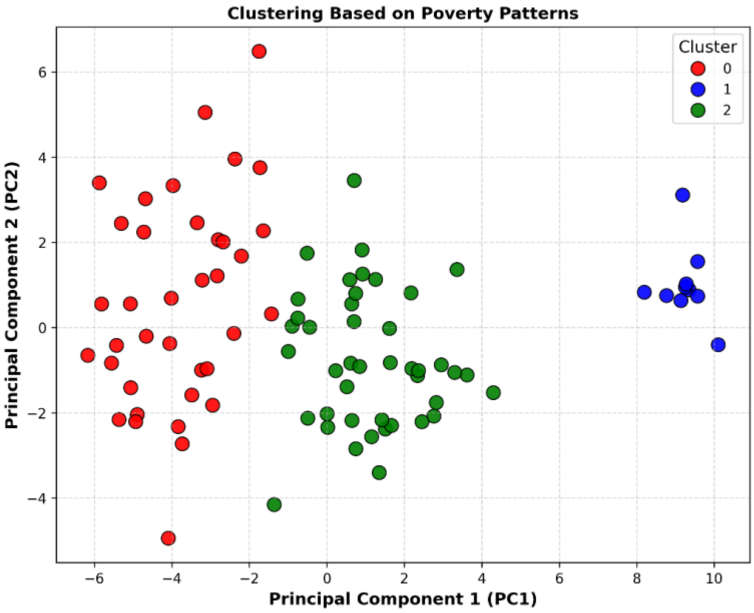

Figure. 3 Visualization of Clustering Results Based on Poverty Patterns

2.6. Model Evaluation

In this study, several models were applied to predict poverty levels using the processed dataset. The models used are Linear Regression (LR), Random Forest Regression (RF), and Gradient Boosting Regression (GBR). These models were selected for their ability to handle complex relationships and non-linearity in the data. Linear Regression was used as the baseline model, while Random Forest and Gradient Boosting were applied to capture more complex patterns and feature interactions. To evaluate the model's performance, three main metrics were used:

1. Mean Absolute Error (MAE)

   MAE measures the average absolute error between predictions and actual values, without considering the direction of the error. MAE is calculated as the average of the absolute differences between the predicted values and the actual values. MAE provides a clear indication of how accurate the model is in terms of how large the deviations are compared to the true values. The formula for MAE is:

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |\, y_i - \hat{y}_i \,| \tag{2}$$

   where y_i is the actual value and $\hat{y}_i$ is the predicted value

2. Mean Squared Error (MSE)

   MSE calculates the average of the squared differences between predicted and actual values. MSE places more weight on larger errors, making it more sensitive to outliers than MAE. The formula for MSE is:

$$MSE = \frac{1}{n}\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{3}$$

   This metric penalizes large errors more heavily and helps identify models with significant deviations in predictions.

3. R² Score

   R^2, or the coefficient of determination, measures the proportion of variance in the dependent variable that can be explained by the independent variables in the model. This metric indicates how well the model fits the data, with values ranging from 0 to 1. The higher the R^2 value, the better the model fits the data. The formula for R^2 is:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} \tag{4}$$

   where $\bar{y}_i$ is the mean of the actual values.

.

The test results of the models applied show different performances in predicting poverty levels. Table 3 presents the performance evaluation results of these models.

Table 3. Performance Evaluation of Models for Predicting Poverty Levels

Model	MAE	MSE	R2 Score
Linear Regression	0.550830	0.423618	0.474129
Random Forest	0.248889	0.423618	0.724428
XGBoost	0.371582	0.604624	0.249432
Neural Network	7.311705	54.323025	-66.435479

2.7. Evaluation of Feature Influence on Poverty Prediction

The model was trained and evaluated, followed by further analysis to identify the factors that have the greatest impact on poverty level prediction. To achieve this, this study utilizes the SHAP Summary Plot to determine the most contributing features in the model and the SHAP Dependence Plot to analyze the relationship between each feature and the prediction more transparently

The SHAP Summary Plot visualization results showed the contribution of features to poverty level prediction. This plot illustrates the importance level and direction of influence of each feature based on the distribution of SHAP values.

Significant Feature Influence on Poverty Prediction The Education Index had SHAP values ranging from approximately -0.6 to 0.7, indicating that education plays a crucial role in determining poverty levels. Negative SHAP values suggest that an increase in the education index contributes to poverty reduction. Additionally, the Health Index had SHAP values ranging from approximately -0.5 to 0.8, indicating that access to healthcare services is closely related to poverty levels. The higher the healthcare access, the lower the predicted

1. Poverty level.
   The Economic Index had SHAP values ranging from approximately -0.4 to 0.6, demonstrating that economic stability significantly affects poverty conditions. A more stable economy tends to reduce poverty, while economic instability can increase poverty levels. Meanwhile, the Unemployment Ratio to Workforce had SHAP values ranging from approximately -0.3 to 0.3, indicating that an increase in unemployment rates correlates with higher poverty levels.
2. Features with Limited Influence
   Several other features contributed less to poverty prediction than the primary features mentioned earlier. The Income-to-Economy Ratio had SHAP values ranging from approximately -0.2 to 0.2, indicating that while income is related to the economy, its impact on poverty levels is more stable compared to education or employment factors.
   Additionally, the Infrastructure-to-Health Ratio had SHAP values ranging from approximately -0.1 to 0.1, indicating that its contribution to the prediction model was smaller than that of the primary features. This finding suggests that although healthcare infrastructure is important, its impact on poverty prediction in this model is relatively limited. The Transportation-to-Road Ratio had SHAP values ranging from approximately -0.2 to 0.3, indicating that transportation accessibility plays a role in poverty levels, but its influence is not as significant as economic, health, and education factors.
3. Implications for Poverty Reduction Strategies
   Policy focus should be directed toward factors with the highest SHAP values, such as the Health Index (-0.5 to 0.8), Education Index (-0.6 to 0.7), Economic Index (-0.4 to 0.6), and Unemployment Ratio to Workforce (-0.3 to 0.3). These factors have a dominant influence on poverty and can serve as the foundation for designing poverty alleviation strategies.
   Enhancing access to healthcare services, strengthening the economic sector, and creating job opportunities should be top priorities in poverty reduction strategies. Better education will improve workforce skills and competitiveness, while a stable economy will create more job opportunities. Furthermore, although features with smaller SHAP values have a more limited impact than the primary factors, they remain relevant in poverty analysis and can be part of a comprehensive policy approach. Figure 4 is a visualization of the SHAP Summary Plot.

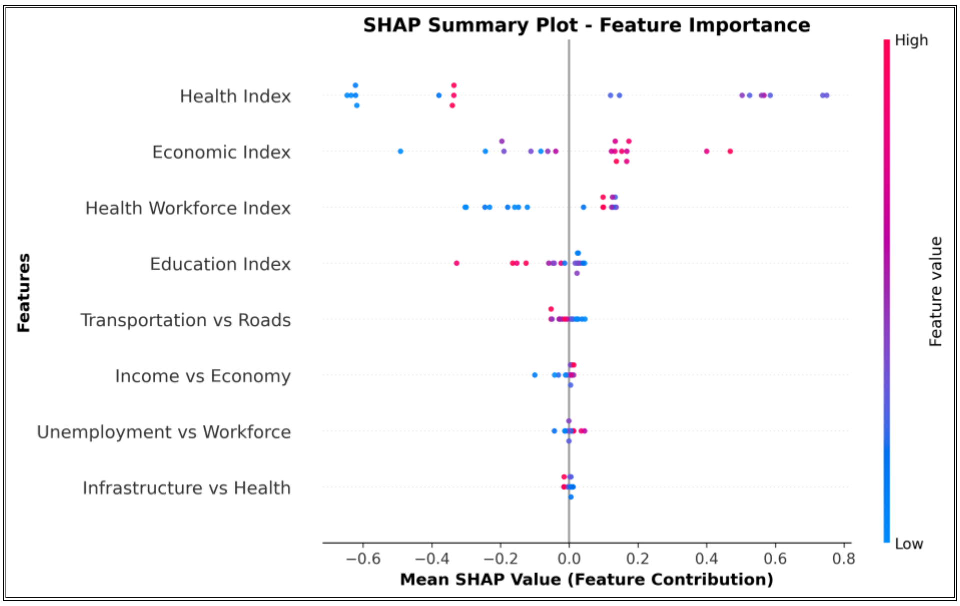

Figure. 4 SHAP Summary Plot – Feature Influence on Poverty Prediction

The SHAP Dependence Plot was used to analyze the relationship between a feature's value and its contribution to poverty level prediction. This visualization helps in understanding how changes in a specific feature's value affect the prediction outcome and in identifying potential interactions between features in the model. This technique enables a deeper interpretation of variable influence on predictions, providing clearer insights into the key factors that play a role in determining poverty levels. Figure 5 presents the SHAP Dependence Plot, which illustrates the relationship between features and their contributions to the model.

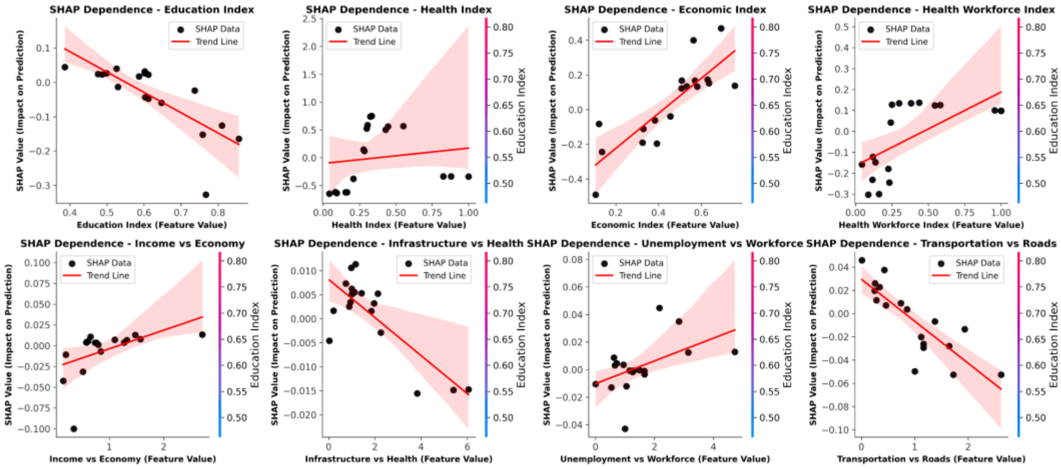

Figure. 5 SHAP Dependence Plot – Feature Relationship and Contribution to the Model

2.7. Analysis of Feature Influence Trends on Poverty Prediction

Based on the SHAP Dependence Plot analysis, two main patterns were identified in the influence of variables on poverty levels: a positive trend and a negative trend. Each pattern reflects how changes in a feature's value affect poverty prediction.

Features in this category indicate that as their values increase, their contribution to higher poverty predictions also increases. These features include the Health Index, Economic Index, Health Workforce Index, and the Unemployment-to-Workforce Ratio.

This positive trend suggests that while improvements in the Health Index and Economic Index are generally expected to reduce poverty levels, in some cases, an increase in these indices may also reflect social and economic disparities. For example, increased access to healthcare services does not always correlate directly with poverty reduction if there are still obstacles to equitable healthcare distribution. Similarly, a high Economic Index may indicate economic growth, but without fair distribution, certain societal groups may remain in poverty.

.

Features in this category indicate that as their values increase, they contribute to lowering poverty predictions. These features include the Education Index, Infrastructure-to-Health Ratio, Transportation-to-Road Ratio, and Income-to-Economy Ratio.

From the analysis results, the Education Index has the most significant negative impact on poverty levels, indicating that increased access to and quality of education directly contribute to poverty reduction. Infrastructure and transportation access also support economic development, although their impact is not as significant as education and healthcare. Good transportation access can improve workforce mobility and expand economic opportunities for communities in underdeveloped areas.

This study has evaluated the key factors contributing to poverty prediction using feature selection techniques. The next section will discuss the overall model performance, compare the effectiveness of various algorithms, and interpret the findings obtained in poverty prediction in South Sumatera

3. RESULTS AND DISCUSSION

3.1. Data Preprocessing

The results of the data preprocessing stage indicate a significant improvement in dataset quality, ensuring that the data is more structured, consistent, and suitable for machine learning models.

Missing Value Imputation Results: The mean imputation method and Random Forest Imputer successfully addressed missing values without distorting the overall data distribution. The verification using `data.isnull().sum()` confirmed that no missing values remained after the process, ensuring data completeness for model training without introducing artificial bias.

Outlier Handling Results: The IQR and Winsorization methods effectively stabilized the data distribution while preserving essential information. By reducing extreme values, the methods prevented outliers from disproportionately influencing model predictions, thereby enhancing robustness. Variables with a high number of outliers showed a more normalized and interpretable distribution, as illustrated in Figure 1.

Data Transformation Results: Normalization using MinMaxScaler successfully aligned all variable scales within the range of 0-1, which is critical for preventing models from being biased toward features with larger numerical ranges. This step ensured numerical stability and optimized performance across different machine learning algorithms.

Exploratory Data Analysis (EDA) Results: The correlation heatmap in Figure 2 revealed significant relationships among socio-economic variables, providing valuable insights into feature dependencies. These insights guided the feature engineering process by identifying key indicators that contribute most to poverty prediction. For instance, the positive correlation between education variables (P1, P2, P3) and labor force participation (LP1, LP2, LP3) suggested that higher education levels are associated with higher labor force participation, leading to the creation of the "Income vs. Economy" feature. Similarly, the correlation between healthcare access variables (AH1, AH2, AH3, AH4) and the number of healthcare workers (HW1-HW9) motivated the development of the "Health Index." This index aggregates healthcare access and workforce data to better reflect public well-being. These relationships led to the creation of aggregate and interaction features to better capture the socio-economic dynamics that influence poverty prediction.

Overall Impact of Data Preprocessing: The data preprocessing stage effectively reduced noise, handled inconsistencies, and enhanced dataset reliability. These improvements laid a strong foundation for feature engineering and predictive modeling, ensuring higher model accuracy and stability.

3.2. Feature Engineering

Feature engineering plays a crucial role in transforming raw data into meaningful representations, allowing machine learning models to extract deeper insights and improve predictive accuracy. This study applied advanced techniques to construct more informative features, refining the dataset for better performance in poverty prediction.

1. *Feature Engineering Summary:* The feature engineering process included : Feature Aggregation → Created composite indices such as Education Index, Health Index, Economic Index, and Health Workforce Index, which captured multi-dimensional aspects of poverty.

   Feature Interaction → Generated ratio-based features, including Income vs Economy, Infrastructure vs Health, Unemployment vs Workforce, and Transportation vs Road, which helped in understanding relative economic conditions rather than absolute values.

   Dimensionality Reduction (PCA & Lasso Regression) → Used to eliminate irrelevant features and enhance model efficiency.
2. *Feature Selection and Validation*: After generating 12 new features, selection was performed using PCA and Lasso Regression to retain the most impactful ones. Four features (Basic Unemployment Ratio, Intermediate Unemployment Ratio, Basic Labor Force Participation Rate Ratio, and Intermediate Labor

Force Participation Rate Ratio) had constant values of 1 across all observations. These were removed to avoid redundancy and improve computational efficiency.

Feature selection ensured that only the most informative variables were retained, leading to a more interpretable and efficient model.

3. *Final Feature Engineering Results:* After selection, eight key features remained, offering a more refined and representative dataset for predictive modeling. Additionally, two primary variables "Year" (time variable) and "Poverty Rate" (target variable) were maintained.

   This structured feature set significantly enhances the model's ability to detect poverty patterns, making the analysis more accurate and actionable for socio-economic policy interventions.

3.3. Model Development and Performance

The evaluation results indicate that model performance varies significantly in predicting poverty levels, highlighting the importance of selecting an appropriate model for socio-economic data analysis. Random Forest achieved an MAE of 0.2489, an MSE of 0.2219, and an R^2 of 0.7244, making it the best-performing model in this study**.** Its ensemble learning capability allows it to effectively capture complex, non-linear relationships between socio-economic indicators, resulting in higher accuracy and lower error rates compared to other models. This performance is further enhanced by the optimized feature set, which includes interaction terms and aggregated features derived from socio-economic indicators. These optimized features, such as "Income vs. Economy" and "Infrastructure vs. Health," are crucial for capturing the interdependencies between variables and directly contribute to the model's ability to predict poverty levels more accurately**.** Additionally, its robustness to outliers and feature importance analysis capability makes it particularly useful for interpreting key poverty determinants.

To optimize Random Forest performance, Grid Search Cross-Validation was applied for hyperparameter tuning. The parameter search included: n_estimators: [100, 200, 300], max_depth: [10, 20, None], min_samples_split: [2, 5, 10], min_samples_leaf: [1, 2, 4]. The optimal combination was selected based on cross-validation results to enhance both accuracy and generalizability.

Linear Regression showed weaker performance, with an MAE of 0.5508, an MSE of 0.4236, and an R^2 of 0.4741. This indicates that the model struggles to capture non-linear interactions between variables, leading to higher prediction errors. The assumption of linearity in poverty data, which is often influenced by multiple interacting socio-economic factors, likely limits its effectiveness. XGBoost had a lower MAE than Linear Regression (0.3715) but produced a higher MSE (0.6046) and a lower R^2 (0.2494).

While XGBoost is known for its powerful boosting mechanism, its suboptimal performance in this study suggests potential overfitting or insufficient hyperparameter tuning. Randomized Search CV was used for tuning, with parameter ranges such as: learning_rate: [0.01, 0.1, 0.2], max_depth: [3, 5, 7], n_estimators: [100, 200, 300]. Further optimization might be needed to improve generalization on this socio-economic dataset.

Neural Network performed the worst, with an MAE of 7.3117, an MSE of 54.3230, and a negative R^2 (-66.4354), indicating a failure to generalize the data. The poor performance could be attributed to the relatively small dataset size, which may not provide enough training samples for deep learning models. Additionally, the model's high complexity requires careful hyperparameter tuning. A grid search was conducted to explore combinations of: hidden_layer_sizes: [(50,), (100,), (50, 50)], activation: ['relu', 'tanh'], alpha: [0.0001, 0.001]. However, performance remained limited, suggesting the need for more training data and potentially deeper architectures with regularization.

Based on the evaluation results, Random Forest was selected as the best model for predicting poverty levels due to its higher accuracy, robustness, and ability to capture complex relationships between socio-economic indicators. Unlike other models, it balances predictive power with interpretability, making it a suitable choice for data-driven poverty analysis and policy formulation**.** The findings highlight the importance of using ensemble-based models in socio-economic studies, where relationships between variables are often non-linear and multidimensional.

3.4. Summary of Key Findings

This study systematically enhances the accuracy of poverty prediction through optimized feature engineering, with key findings at each methodological stage. Data preprocessing, which includes normalization, imputation, and handling of outliers, significantly improves the dataset quality for modeling. Feature engineering techniques generate eight representative predictors, including aggregation (creating composite indices for Education, Economy, Health, and Healthcare Workforce), feature interactions (capturing complex relationships), and feature selection using PCA and Lasso Regression, which identify the most influential predictors, such as education and economic indices. Cluster analysis reveals three distinct poverty patterns: Cluster 0 (high poverty with deficits in education/infrastructure, red), Cluster 1 (stable socio-

.

economic conditions, blue), and Cluster 2 (moderate conditions requiring targeted interventions, green), with clear spatial segregation shown in Figure 3. These findings collectively enable data-driven policy interventions that are precise and contextually relevant to the region.

Model Development and Evaluation: The poverty prediction model was evaluated using four algorithms, with results showing that Random Forest performed best in capturing complex relationships between socio-economic variables. However, the Neural Network model demonstrated a negative R^2 value (-66.43), indicating poor performance. This is likely due to overfitting caused by the small size of the dataset (90 samples). The limited data may not fully represent regional variations, reducing the model's generalizability. Overfitting occurs when a model learns noise or irrelevant patterns from the data, leading to poor performance on unseen data. To address this limitation: A larger dataset is recommended for future studies to improve model generalization, Synthetic data augmentation techniques may help generate additional representative data, Alternatively, simpler and more regularized model architectures can be considered to prevent overfitting.

Feature Importance Analysis: SHAP analysis identified the most influential features in the poverty prediction model. Notably: *Education Index:* Higher education index consistently correlates with lower poverty levels. For instance, districts with values above 0.75 on the education index showed a significant drop in predicted poverty probability. *Infrastructure-to-Health Ratio:* Areas with low health infrastructure access despite good general infrastructure tend to have higher predicted poverty. For example, a ratio below 0.3 increased poverty predictions by 0.2 units on the SHAP scale. *Transportation-to-Road Ratio:* Poor access to public transportation relative to road length was a strong poverty predictor, highlighting mobility as a determinant of socio-economic access. *Income-to-Economy Ratio:* This ratio, which compares average income to regional economic output, revealed that areas with low income contributions relative to their economic scale tend to have persistent poverty indicating inequality in distribution.

3.5. Policy Recommendations

Based on the model findings and SHAP analysis, several policy recommendations are proposed to address multidimensional poverty. First, increase the education budget allocation in regions with low education index values, prioritizing community-based education programs and scholarship schemes. Second, improve infrastructure development by expanding healthcare facilities in regions with high general infrastructure but limited healthcare services, particularly in remote sub-districts. Third, develop rural transport networks and subsidized public transportation in areas with low Transportation-to-Road Ratios to reduce economic isolation. Fourth, address income inequality through inclusive economic programs, considering the low Income-to-Economy Ratio in some regions. Lastly, encourage local governments to adopt data science tools like Random Forest and SHAP analysis to support evidence-based decision-making and more effective policy formulation.

4. CONCLUSION

This study provides empirical evidence on the effectiveness of machine learning techniques in predicting poverty levels, particularly within the context of South Sumatra Province. The data-driven approach applied in this research has proven to significantly enhance the identification of key socio-economic determinants, thereby supporting the formulation of more targeted and evidence-based policies.

By implementing context-specific feature engineering, clustering analysis, and comprehensive model evaluation, this study successfully improved both the accuracy and interpretability of the predictive models. The Random Forest algorithm outperformed other models, demonstrating superior capability in capturing complex relationships among socio-economic variables associated with poverty.

The feature importance analysis identified education, infrastructure, and economic indicators as the most significant predictors of poverty. These findings underscore the importance of policy interventions focused on improving educational quality, infrastructure development, and strengthening economic sectors as key strategies for sustainable poverty reduction.

The novelty of this study lies in the development of a region-specific feature engineering strategy and the comprehensive evaluation of predictive model performance tailored to local socio-economic characteristics. In contrast to prior studies that employed more general frameworks, this research reveals unique and contextually relevant feature configurations for the South Sumatra region. Moreover, the application of clustering analysis contributes to a more granular understanding of poverty profiles, enabling the formulation of location-based and adaptive policy measures.

Despite these contributions, this study has certain limitations, including potential data biases and the need for further optimization of feature selection techniques. Future research should consider integrating larger and more diverse datasets, incorporating real-time economic indicators, and applying explainable AI approaches to enhance the transparency and trustworthiness of the predictive models.

Overall, this study contributes to the growing body of literature on data-driven poverty assessment by integrating machine learning with socio-economic analysis in a contextualized manner. It offers a relevant and adaptive predictive framework to support sustainable development efforts and data-informed poverty alleviation strategies.

CONFLICT OF INTEREST STATEMENT

The Authors state no conflict of interest.

REFERENCES

[1] R. Riangga and E. Desmamora, “Jumlah Penduduk Miskin Kota Palembang masih Terbanyak di Sumsel, Daya Neli Berkurang, Pentingkan Rokok,” Sumeks.com. Accessed: Mar. 13, 2025. [Online]. Available: https://sumeks.disway.id/read/740341/jumlah-penduduk-miskin-kota-palembang-masih-terbanyak-di-sumsel-daya-beli-berkurang-pentingkan-rokok

[2] E. Saputra and D. Setiawan, “0,38 Persen Penduduk Palembang Berada di Garis Kemiskinan Ekstrim,” Pal TV. Co.id. Accessed: Mar. 13, 2025. [Online]. Available: https://paltv.disway.id/read/31710/038-persen-penduduk-palembang-berada-di-garis-kemiskinan-ekstrim

[3] Y. Wang, Y. Jiang, D. Yin, C. Liang, and F. Duan, “Examining Multilevel Poverty-Causing Factors in Poor Villages: a Hierarchical Spatial Regression Model linear model,” *Appl Spat Anal Policy*, vol. 14, no. 14, pp. 969–998, Aug. 2021, doi: 10.1007/s12061-021-09388-1.

[4] Y. Jiang, Y. Wang, W. Qi, B. Cai, C. Huang, and C. Liang, “Detecting Multilevel Poverty-Causing Factors of Farmer Households in Fugong County: A Hierarchical Spatial–Temporal Regressive Model,” *Agriculture*, vol. 12, no. 11, p. 1844, Nov. 2022, doi: 10.3390/agriculture12111844.

[5] A. I. Lismana and H. Sumarsono, “Analysis of the Effect of Population Growth, Human Development Index and Unemployment Rate on Poverty in West Java Province 2017-2020,” *Jurnal Ekonomi Pembangunan*, vol. 20, no. 01, pp. 88–97, Jun. 2022, doi: 10.22219/JEP.V20I01.20286.

[6] S. Annas, B. Poerwanto, S. Sapriani, and M. F. S, “Implementation of K-Means Clustering on Poverty Indicators in Indonesia,” *MATRIK : Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer*, vol. 21, no. 2, pp. 257–266, Mar. 2022, doi: 10.30812/matrik.v21i2.1289.

[7] M. Rahman and V. Kumar, “Machine Learning Based Customer Churn Prediction in Banking,” in *Proceedings of the 4th International Conference on Electronics, Communication and Aerospace Technology, ICECA 2020*, 2020. doi: 10.1109/Iceca49313.2020.9297529.

[8] H. Zixi, “Poverty Prediction through Machine Learning,” in *Proceedings - 2nd International Conference on E-Commerce and Internet Technology, ECIT 2021*, Institute of Electrical and Electronics Engineers Inc., Mar. 2021, pp. 314–324. doi: 10.1109/ecit52743.2021.00073.

[9] Q. Li, S. Yu, D. Échevin, and M. Fan, “Is poverty predictable with machine learning? A study of DHS data from Kyrgyzstan,” *Socioecon Plann Sci*, vol. 81, p. 101195, Jun. 2022, doi: 10.1016/j.seps.2021.101195.

[10] A. Alsharkawi, M. Al-Fetyani, M. Dawas, H. Saadeh, and M. Alyaman, “Poverty classification using machine learning: The case of Jordan,” *Sustainability (Switzerland)*, vol. 13, no. 3, pp. 1–16, Feb. 2021, doi: 10.3390/su13031412.

[11] L. Maruejols, H. Wang, Q. Zhao, Y. Bai, and L. Zhang, “Comparison of machine learning predictions of subjective poverty in rural China,” *China Agricultural Economic Review*, vol. 15, no. 2, pp. 379–399, May 2023, doi: 10.1108/caer-03-2022-0051.

[12] A. A. Hassan, A. H. Muse, and C. Chesneau, “Machine Learning Study Using 2020 SDHS Data to Determine Poverty Determinants in Somalia,” *Scientific Reports 2024 14:1*, vol. 14, no. 1, pp. 1–19, Mar. 2024, doi: 10.1038/s41598-024-56466-8.

[13] S. K. Satapathy, S. Saravanan, S. Mishra, and S. N. Mohanty, “A Comparative Analysis of Multidimensional COVID-19 Poverty Determinants: An Observational Machine Learning Approach,” *New Gener Comput*, vol. 41, no. 1, pp. 155–184, Mar. 2023, doi: 0.1007/s00354-023-00203-8.

[14] K. Abbas *et al.*, “Measurements and determinants of extreme multidimensional energy poverty using machine learning,” *Energy*, vol. 251, p. 123977, Jul. 2022, doi: 10.1016/J.ENERGY.2022.123977.

[15] W. Sosa-Escudero, M. V. Anauati, and W. Brau, “Poverty, Inequality and Development Studies with Machine Learning,” *Advanced Studies in Theoretical and Applied Econometrics*, vol. 53, pp. 291–335, 2022, doi: 10.1007/978-3-031-15149-1_9.

[16] A. Nachev, “Exploring Poverty Factors Through Predictive Modeling,” pp. 329–342, 2025, doi: 10.1007/978-3-031-85628-0_24.

[17] M. Kuhn and K. Johnson, “Feature Engineering and Selection: A Practical Approach for Predictive Models,” *Feature Engineering and Selection: A Practical Approach for Predictive Models*, pp. 1–297, Jan. 2019, doi: 10.1201/9781315108230.

[18] A. Karim, “Perbandingan Prediksi Kemiskinan di Indonesia Menggunakan Support Vector Machine (SVM) dengan Regresi Linear,” *Jurnal Sains Matematika dan Statistika*, vol. 6, no. 1, pp. 107–113, Jan. 2020, doi: doi:10.24014/jsms.v6i1.9259.

[19] I. P. Putri, T. Terttiaavini, and N. Arminarahmah, “Analisis Perbandingan Algoritma Machine Learning untuk Prediksi Stunting pada Anak,” *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, vol. 4, no. 1, 2024, doi: 10.57152/malcom.v4i1.1078.

[20] D. Antoni, T. Avini, A. Heryati, and H. Syaputra, *Business Process Reengineering*, 1st ed. Perkumpulan Rumah Cemerlang Indonesia, 2023. Accessed: Apr. 25, 2025. [Online]. Available: https://www.rcipress.rcipublisher.org/index.php/rcipress/catalog/book/862

[21] C. S. Kumar, M. N. S. Choudary, V. B. Bommineni, G. Tarun, and T. Anjali, “Dimensionality Reduction based on SHAP Analysis: A Simple and Trustworthy Approach,” *Proceedings of the 2020 IEEE International Conference on Communication and Signal Processing, ICCSP 2020*, pp. 558–560, Jul. 2020, doi: 10.1109/iccsp48568.2020.9182109.

[22] K. Cheng and D. S. Young, “An Approach for Specifying Trimming and Winsorization Cutoffs,” *J Agric Biol Environ Stat*, vol. 28, no. 2, pp. 299–323, Jun. 2023, doi: 10.1007/s13253-023-00527-4.

[23] H. Wang, Q. Liang, J. T. Hancock, and T. M. Khoshgoftaar, “Feature selection strategies: a comparative analysis of SHAP-value and importance-based methods”, doi: 10.1186/s40537-024-00905-w.

[24] W. E. Marcilio and D. M. Eler, “From explanations to feature selection: Assessing SHAP values as feature selection mechanism,” *Proceedings - 2020 33rd SIBGRAPI Conference on Graphics, Patterns and Images, SIBGRAPI 2020*, pp. 340–347, Nov. 2020, doi: 10.1109/sibgrapi51738.2020.00053.

[25] D. Chicco, M. J. Warrens, and G. Jurman, “The Coefficient of Determination R-Squared is More Informative than SMAPE, MAE, MAPE, MSE and RMSE in Regression Analysis evaluation,” *PeerJ Comput Sci*, vol. 7, pp. 1–24, Jul. 2021, doi: 10.7717/peerj-cs.623.

[26] T. Verdonck, B. Baesens, M. Óskarsdóttir, and S. vanden Broucke, “Special issue on feature Engineering Editorial,” *Mach Learn*, vol. 113, no. 7, pp. 3917–3928, Jul. 2024, doi: 10.1007/s10994-021-06042-2.

.

[27] BPS Provinsi Sumatera Selatan, “Persentase Penduduk Miskin Provinsi Sumatera Selatan Maret 2024 Sebesar 10,97 Persen - Badan Pusat Statistik Provinsi Sumatera Selatan.” Accessed: Mar. 16, 2025. [Online]. Available: https://sumsel.bps.go.id/id/pressrelease/2024/07/01/810/persentase-penduduk-miskin-provinsi-sumatera-selatan-maret-2024-sebesar-10-97-persen-.html

[28] Badan Pusat Statistik Indonesia, “Persentase Penduduk Miskin Maret 2024 turun menjadi 9,03 persen,” Badan Pusat Statistik Indonesia. Accessed: Mar. 16, 2025. [Online]. Available: https://www.bps.go.id/id/pressrelease/2024/07/01/2370/persentase-penduduk-miskin-maret-2024-turun-menjadi-9-03-persen-.html

[29] “Data Cleaning - Venkatesh Ganti, Anish Das Sarma - Google Buku.” Accessed: Mar. 13, 2025. [Online]. Available: https://books.google.co.id/books?hl=id&lr=&id=qYdyEAAAQBAJ&oi=fnd&pg=PP1&dq=1)%09Data+Cleaning&ots=1mHWRrOPVN&sig=R9C88poXk7TjrU2k8MQ9cadSlkY&redir_esc=y#v=onepage&q=1)%09Data%20Cleaning&f=false

[30] M. M. Mijwil, A. W. Abdulqader, S. M. Ali, and A. T. Sadiq, “Null-values Imputation Using Different Modification Random Forest Algorithm,” *IAES International Journal of Artificial Intelligence*, vol. 12, no. 1, pp. 374–383, Mar. 2023, doi: 10.11591/ijai.v12.i1.pp374-383.

[31] V. N. G. Raju, K. P. Lakshmi, V. M. Jain, A. Kalidindi, and V. Padma, “Study the Influence of Normalization/Transformation process on the Accuracy of Supervised Classification,” *Proceedings of the 3rd International Conference on Smart Systems and Inventive Technology, ICSSIT 2020*, pp. 729–735, Aug. 2020, doi: 10.1109/icssit48917.2020.9214160.

[32] K. U. Singh, S. K. Pandey, D. P. Yadav, T. Singh, G. Kumar, and A. Kumar, “Data Science - A Compendious Study on Statistical Methods and Visualization Techniques,” *Proceedings of International Conference on Computational Intelligence and Sustainable Engineering Solution, CISES 2023*, pp. 227–232, 2023, doi: 10.1109/cises58720.2023.10183429.

[33] F. Rahmat *et al.*, “Supervised feature selection using principal component analysis,” *Knowl Inf Syst*, vol. 66, no. 3, 2024, doi: 10.1007/s10115-023-01993-5.

[34] K. P. Sinaga and M.-S. Yang, “Unsupervised K-Means Clustering Algorithm,” *IEEE Access*, vol. 8, pp. 80716–80727, 2020, doi: 10.1109/access.2020.2988796.

[35] T. Terttiaavini *et al.*, “Clustering Analysis of Premier Research Fields,” *International Journal of Engineering & Technology*, vol. 7, no. 4.44, 2018, doi: 10.14419/ijet.v7i4.44.26860.

BIOGRAPHIES OF AUTHORS

Dr. Terttiaavini, S.Kom., M.Kom is a Lecturer in the Master of Computer Science at Universitas Indo Global Mandiri, South Sumatera, Indonesia. She holds a Doctoral degree (Dr.) in Computer Engineering with a specialization in Data Science. Her research interests include Machine Learning, Data Analysis, Data Mining, and Decision Support Systems. From 2016 to 2024, she has received four research grants and four community service grants from the Ministry of Higher Education, Science, and Technology of the Republic of Indonesia (Kemdiktisaintek RI). In addition, she has authored and published six reference books in her field of expertise. She is actively involved in various research and community service activities and is open to research collaborations with international scholars. Terttiaavini can be contacted via email: avini.saputra@uigm.ac.id

Agustina Heryati, S.Kom., M.M., M.Kom is a lecturer, researcher, and writer from Palembang, South Sumatera. She has a high dedication in the field of Informatics Engineering, especially in the sciences that include Information System, Decision Support System (DSS), Data Mining, and Artificial Intelligence (AI), Intelligent Systems. He is currently pursuing his doctoral degree at Sriwijaya University in Informatics Engineering. This dissertation research discusses green transportation route optimization in the manufacturing industry, which will contribute to determining the most optimal and environmentally friendly route. Being a lecturer and writer is a life calling that provides an opportunity to share knowledge while creating a meaningful impact on the future. She can be contacted at email: agustina.heryati@uigm.ac.id.

Tedy Setiawan Saputra, S.E., M.M., CRMPA, CACP, is a lecturer at STIE Aprin Palembang, Indonesia. He currently serves as the Head of the Institute for Research and Community Service at the same institution. In addition to his academic role, he is also a member of the Audit Committee at PT Petromuba (Perseroda), reflecting his active involvement in both academic and corporate governance sectors.
He is currently pursuing a Doctoral Program in Management Science at the Islamic University of Indonesia, Yogyakarta, with a concentration in Strategic Human Resources Management. His research interests include sustainability, green concepts, and human resources management. Mr. Saputra was awarded a competitive research grant by the Ministry of Research, Technology, and Higher Education (Kemenristekdikti) in 2021. Among his notable scholarly contributions is the publication “Innovative Leadership in the Digital Age: A Preliminary Research on Digital Leadership in Contemporary Management” (Springer Nature Singapore, 2024). He can be contacted via email at tdyfaith@gmail.com

.