Infact: Computer International Journal (ICIJ)
Vol.10, No.01, May 2026
p-ISSN : 2527-8363, e-ISSN: 2829-0259
DOI:

Data-Driven Classification of Poverty Status in Indonesia using Machine Learning Techniques

Syaila Fathia Azzahra 1[1*], Yudi Ahmad Hambali 2[1], Ismail Marzuki Randos 3[3]
[1,2] Computer Science, Faculty of Mathematics and Natural Sciences Education, Universitas Pendidikan Indonesia
[3]Ministry of Finance of the Republic of Indonesia
[1,2]Jl. Dr. Setiabudi No.229, Isola, Kec. Sukasari, Kota Bandung, Jawa Barat 40154
[3]Jl. Dr. Wahidin Raya No.1 Lt.9, Ps. Baru, Jakarta Pusat
E-mail: syailafathiaazzahra@upi.edu[1], yudi.a.hambali@upi.edu[2], ismailmarzukir@gmail.com[3]

Submitted: 07/10/2025 , Revision: 03/07/2026, Accepted : 05/07/2026

Abstract

This study explores the use of the K-Nearest Neighbor (KNN) algorithm to classify poverty status in Indonesia using publicly available socio-economic indicators. Traditional poverty classification methods are often inefficient and lack nuance. By leveraging the Knowledge Discovery in Databases (KDD) process, including data preprocessing, normalization, and dimensionality reduction via PCA, the study builds a robust classification model. The dataset includes indicators such as education, health, and expenditure levels from 514 districts/cities. The optimal KNN model, determined through cross-validation, achieved a test accuracy of 95.15%, with strong precision, recall, and ROC AUC scores. Feature importance analysis via Random Forest on PCA-transformed data highlights the predictive influence of certain component combinations. The results demonstrate the potential of machine learning to support more accurate and data-driven policy targeting in poverty alleviation. Future enhancements may involve integrating time-series or satellite data to increase relevance and precision.

Keywords:
Poverty Classification;
K-Nearest Neighbor;
Socio-Economic Indicators;
Machine Learning Indonesia;

***Corresponding Author: Author's Name**
E-mail: syailafathiaazzahra@upi.edu

1. Introduction

Poverty remains a significant and persistent challenge in Indonesia, affecting millions of people across both urban and rural regencies and cities [1]. The ability to accurately classify poverty is essential for policymakers to design effective interventions and ensure that social assistance programs are distributed to those who need them most [2]. However, traditional approaches to poverty classification often rely on manual data collection and simple statistical analysis, which can be inefficient and may not capture the complexity of socio-economic conditions.

The advancement of information technology and the increasing availability of socio-economic data have enabled the use of machine learning methods to address these limitations [3]. Among various algorithms, the K-Nearest Neighbor (KNN) classifier has proven to be effective in classifying poverty status due to its simplicity and strong performance in handling multi-dimensional data [4]. KNN works by comparing the attributes of a given region or household to its closest neighbors in the dataset, allowing for more accurate and nuanced classification [5]. Studies in Indonesia have shown that KNN can outperform traditional methods in identifying patterns of poverty, making it a valuable tool for supporting data-driven policy decisions [1].

This research aims to utilize the KNN classifier to categorize poverty levels in Indonesian regencies and cities. By leveraging recent developments in machine learning and comprehensive socio-economic datasets, this study seeks to contribute empirical evidence that can support more targeted and effective poverty alleviation strategies across Indonesia.

2. Literature Review

2.1 Definition of Poverty in Indonesia

Poverty in Indonesia is officially defined as the inability of individuals or households to fulfill basic food and non-food needs, measured by a minimum expenditure threshold set by the Central Statistics Agency (BPS) [6]. The poverty line is determined based on the cost required to meet a minimum of 2,100 kilocalories per capita per day, as well as essential non-food needs such as housing, education, and health [7]. In addition to economic factors, poverty in Indonesia is shaped by multidimensional issues including limited access to education, employment, and infrastructure, which contribute to persistent disparities between regions [1]. Studies have shown that rural areas, especially on Sumatra Island, tend to have higher poverty rates than urban areas due to uneven development and access to resources [7].

2.2 Data Mining in Poverty Analysis

Data mining is the process of extracting useful patterns and knowledge from large datasets using statistical, mathematical, and machine learning techniques [8]. In the context of poverty analysis, data mining enables the identification of socio-economic patterns and the classification of regions or households based on indicators such as income, education, and access to basic services [1]. For example, clustering and classification methods have been used to group regencies and cities in Indonesia according to poverty and inflation rates, providing insights for more targeted government interventions [2]. By transforming raw data into actionable information, data mining supports evidence-based decision making in poverty alleviation efforts [9].

2.3 K-Nearest Neighbor (KNN) Algorithm

The K-Nearest Neighbor (KNN) algorithm is a supervised machine learning method used for classification and regression tasks by assigning a class to a data point based on the majority class among its k closest neighbors in the feature space [4]. KNN is favored for its simplicity, ease of implementation, and strong performance, especially when dealing with imbalanced datasets often found in poverty studies [1]. In practice, KNN has achieved high accuracy in classifying poverty status at the village and district levels in Indonesia by analyzing features such as income, occupation, and housing conditions[4]. The algorithm's performance can be further improved by optimizing the value of k and applying feature normalization, making it a robust choice for poverty classification [10].

3. Methods

In this study, the author will use the KDD (Knowledge Discovery in Databases) method, which is a process of extracting useful knowledge from large datasets. The steps include data selection, cleaning, transformation, data mining, and interpretation. This method helps to discover meaningful patterns and insights that can support the research objectives.

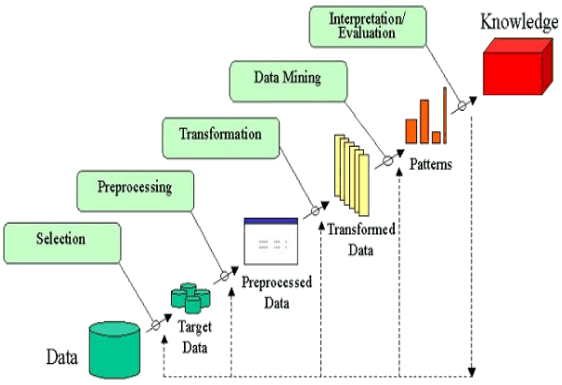

Figure 1. Knowledge Discovery in Database Process

3.1 Data Collection and Description

This study utilizes secondary data titled "Klasifikasi Tingkat Kemiskinan di Indonesia" (Classification of Poverty Levels in Indonesia), which contains social and economic indicators at the district/city level across the country. The dataset includes attributes such as the poverty rate, average years of schooling, adjusted per capita expenditure, Human Development Index (HDI), life expectancy, access to clean water and sanitation, unemployment rate, labor force participation rate, and gross regional domestic product (GRDP). The dataset was imported using the pandas library and read in CSV format with a semicolon delimiter. The initial inspection of the dataset confirmed successful loading and structure suitability for further preprocessing and modeling.

3.2 Data Preprocessing

Before conducting any analysis, the dataset underwent a thorough preprocessing phase. Numerical data represented with commas as decimal separators were first converted to standard floating-point format by replacing commas with dots and coercing the data type. Subsequently, column names were simplified and standardized to enhance readability and coding efficiency. To handle missing values, rows containing nulls were entirely removed from the dataset, resulting in a cleaned dataset of 514 observations.

Categorical variables, specifically 'Province' and 'District/City', were transformed using One-Hot Encoding to convert them into binary indicators, ensuring compatibility with machine learning algorithms. To avoid multicollinearity, the first category in each feature was dropped during encoding. All numerical features were normalized using the MinMaxScaler method to scale values between 0 and 1. This normalization is particularly appropriate given the presence of features with varying ranges and distributions. Furthermore, dimensionality reduction was implemented using Principal Component Analysis (PCA), retaining the top 10 components based on explained variance to optimize computational performance while preserving data variance.

3.3 Data Splitting and Model Building

Following the preprocessing phase, the dataset was split into training and testing subsets with an 80:20 ratio using stratified sampling. Stratification was applied to maintain the distribution of the target variable, 'Poverty Classification', across both subsets. The classification task was approached using the K-Nearest Neighbors (KNN) algorithm. The optimal number of neighbors (k) was determined through a hyperparameter tuning process involving GridSearchCV with five-fold cross-validation. This method ensured robust selection of the best-performing model parameter by evaluating accuracy scores across multiple validation sets.

3.4 Model Evaluation

Once the optimal KNN model was trained on the principal components of the training set, predictions were made on the test set. The model's performance was then assessed using several evaluation metrics, including accuracy, precision, recall, F1-score, and the area under the Receiver Operating Characteristic curve (ROC AUC). In addition, a confusion matrix was generated to further interpret classification errors, identifying the number of true positives, true negatives, false positives, and false negatives. The results indicated strong predictive performance, particularly in distinguishing between 'poor' and 'non-poor' categories, with a balanced trade-off

between precision and recall. The high F1-score and ROC AUC further confirmed the robustness of the model handling binary classification with an imbalanced class distribution.

3.5 Acknowledgments

To gain insight into the influence of features on the classification results, a feature importance analysis was conducted using a Random Forest Classifier. Although KNN does not inherently provide feature importance scores, Random Forest was employed on the same principal components to estimate their relative contributions. The importance scores of the top ten principal components were visualized to highlight which latent features most significantly affected the classification decisions. This interpretation acknowledges that while the components are not direct original features, they encapsulate meaningful combinations of them.

4. Result and Discussion

4.1 Model Optimization and Selection

In the early stages of the modeling process, the performance of the K-Nearest Neighbors (KNN) algorithm was optimized by tuning the k hyperparameter i.e., the number of nearest neighbors used in classification. This process is crucial because the value of k significantly affects the decision boundary of the classifier. A small k may cause the model to overfit, capturing noise in the data, while a large k might oversmooth the boundary, ignoring meaningful variations.

To find the optimal value of k, a grid search was applied using the GridSearchCV function from Scikit-learn, with 5-fold cross-validation. This method systematically evaluated each value of k from 1 to 20. The use of cross-validation ensures that the model's performance is tested on multiple train-test splits, reducing the risk of selection bias and giving a better estimate of real-world performance.

```
The best K value found by Cross-Validation: 3
Average accuracy score from cross-validation for the best K: 95.13%
```

Figure 2. Cross-Validation Result for K Value Optimization

As a result, the best-performing value was k = 3, which produced a cross-validation average accuracy of 95.13%. This high accuracy suggests that the KNN model with this configuration can generalize well and is not overfitting to the training data.

4.2 Model Evaluation on Test Data

After selecting the optimal hyperparameters, the model was trained using 80% of the total data and tested on the remaining 20%. The evaluation on this unseen test set provides a realistic assessment of how the model would perform in deployment or future use cases. The evaluation metrics obtained were impressive:

```
Akurasi: 95.15%
Presisi (Precision): 76.92%
Recall: 83.33%
Skor F1 (F1 Score): 80.00%
ROC AUC: 90.02%
```

Figure 3. Classification Metrics Output from the Test Set

These numbers indicate that the model not only correctly classified the vast majority of instances but also balanced precision and recall effectively, an important aspect when dealing with class imbalance. Specifically, the precision metric tells us that when the model predicted a district as 'Poor', about 77% of the time it was correct. The recall, on the other hand, reflects the model's ability to capture actual poor districts, successfully identifying over 83% of them.

The F1 score, a harmonic mean of precision and recall, confirms that the model maintains a good balance between avoiding false positives and false negatives. Meanwhile, the ROC AUC score of 90.02% highlights the model's high discriminatory ability, distinguishing well between poor and non-poor districts.

4.3 Confusion Matrix Analysis

To further analyze the nature of the classification outcomes, a confusion matrix was generated. This matrix is a powerful tool to summarize the number of correct and incorrect predictions made by the classifier. The matrix produced for this model is as follows:

Table 1. Confusion Matrix of the KNN Model

Actual / Predicted	Non-Poor (0)	Poor (1)
Non-Poor (0)	88 (True Negative)	3 (False Positive)
Poor (1)	2 (False Negative)	10 (True Positive)

The matrix shows that 88 districts were correctly classified as non-poor and 10 districts were

correctly classified as poor. Only 3 instances were falsely predicted as poor, and 2 poor districts were missed by the model. This low number of false negatives is especially important in poverty classification since misidentifying a poor area as non-poor can lead to exclusion from crucial social assistance programs.

By analyzing this confusion matrix, we can conclude that the model's classification is well-calibrated. It minimizes the cost of misclassification in a domain where each error can have significant real-world implications.

4.4 Feature Improtance Analysis

Although KNN is a non-parametric algorithm and does not provide native feature importance scores, feature analysis was performed using a Random Forest classifier trained on the same PCA-transformed features. The aim was to determine which principal components (PCs) contributed most significantly to the classification decisions.

	Feature	Importance
0	PC1	0.376289
5	PC6	0.114741
1	PC2	0.104350
2	PC3	0.087307
3	PC4	0.082754
7	PC8	0.071083
6	PC7	0.064653
9	PC10	0.033523
8	PC9	0.033516
4	PC5	0.031785

Figure 4. Result of Feature Importance

The results of the analysis revealed that PC1 was by far the most dominant component, accounting for 38.1% of the model's total importance weight. This was followed by PC6 (11.6%), PC2 (10.5%), and PC3 (8.6%). Together, the top four components explained a large portion of the decision process, suggesting that much of the predictive power was concentrated in a few latent dimensions.

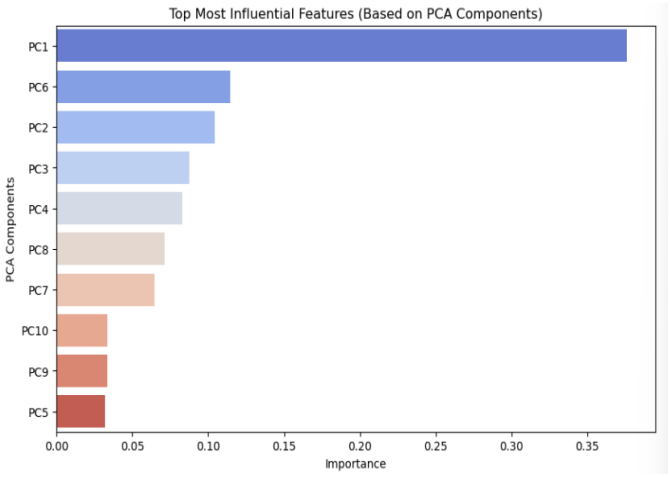

Figure 5. Feature Importance Barplot (PCA Components)

This finding also supports the decision to use dimensionality reduction via PCA in the earlier stages of the pipeline. By compressing information into 10 principal components while retaining around 30% of the variance, the model reduced computational cost without significantly compromising predictive performance.

4.5 Interpretation and Implications

The results presented in this study demonstrate that the KNN classifier, when combined with appropriate preprocessing techniques such as normalization and dimensionality reduction, is highly effective in predicting poverty classification using socio-economic indicators. The model performs particularly well in identifying poor districts while maintaining a low false positive rate.

From a practical perspective, such models have the potential to assist policymakers and development agencies in prioritizing areas for intervention. The confusion matrix and evaluation metrics confirm the model's capability to serve as a decision-support tool, reducing subjectivity in the allocation of government support or program targeting.

Moreover, the feature importance analysis, though based on transformed data, suggests that certain patterns and combinations of indicators are particularly telling. Understanding which dimensions contribute the most can inform more focused data collection efforts and more efficient program design in the future.

5. Conclusion

This study demonstrates the effectiveness of the K-Nearest Neighbors (KNN) algorithm in

classifying poverty status among Indonesian districts using publicly available socio-economic indicators. Through a comprehensive process that involved data cleaning, normalization, dimensionality reduction via Principal Component Analysis (PCA), and hyperparameter optimization, the KNN model achieved a high accuracy of 95.15% on the test set.

The model also recorded strong performance across other evaluation metrics, including precision (76.92%), recall (83.33%), and F1-score (80.00%). The ROC AUC score of 90.02% further confirmed the model's ability to distinguish between poor and non-poor districts. These results indicate that the approach is not only accurate but also reliable in handling imbalanced classification tasks where the minority class (poor districts) is of critical importance in real-world applications.

In addition to evaluating the model's performance, feature importance analysis using Random Forest on PCA-transformed data revealed that a small number of components, particularly PCI, carried a significant portion of the model's predictive power. This finding suggests that certain combinations of underlying features (such as education level, access to sanitation, and per capita expenditure) are especially influential in distinguishing poverty status. The study aligns with previous research indicating that machine learning can be a valuable tool for social policy, providing objective, data-driven insights to complement traditional poverty assessment methods [11][12].

Moving forward, the integration of additional data sources, such as satellite imagery or time-series indicators, may further improve model precision and policy relevance. This research supports the potential for machine learning to enhance poverty targeting and promote more equitable development outcomes.

References :

[1] P. R. Sihombing and A. M. Arsani, "Comparison of Machine Learning Methods in Classifying Poverty in Indonesia in 2018," J. Tek. Inform. JUTIF, vol. 2, no. 1, pp. 51–56, 2021. [Online]. Available: https://doi.org/10.20884/1.jutif.2021.2.1.52

[2] D. B. Lasfeto, T. Setyorini, J. J. Mauta et al., "A simple classification and clustering of poverty in rural areas using machine learning," J. Infrastruct. Policy Dev., vol. 8, no. 8, pp. 5938, 2024. [Online]. Available: https://doi.org/10.24294/jipd.v8i8.5938

[3] H. Purnomo and D. Nurhadi, "Penerapan Algoritma K-Nearest Neighbor untuk Klasifikasi Kelayakan Status Penduduk Miskin di Desa Susukan Tonggoh," J. Informatika, vol. 10, no. 1, pp. 29–36, 2022. [Online]. Available: https://journal.stmikjabar.ac.id/index.php/i/article/view/29

[4] D. P. Sari and A. Wibowo, "Machine Learning for Clustering Regencies-Cities Based on Inflation and Poverty Rates in Indonesia," Indones. J. Inf. Syst., vol. 5, no. 1, pp. 64–73, 2021. [Online]. Available: https://doi.org/10.24002/ijis.v5i1.5682

[5] R. A. Putra and B. Santoso, "Classification of the Poor in Sumatera and Java using Naive Bayes and Particle Swarm Optimization," J. Riset Inform., vol. 7, no. 2, pp. 45–54, 2022. [Online]. Available: http://journal.kresnamediapublisher.com/index.php/jri/article/view/164

[6] M. Akbar and A. Kusumodestoni, "Performance Improvement of K-Nearest Neighbor Algorithm in KIP Scholarship Classification," J. Mantik, vol. 6, no. 1, pp. 30–35, 2020. [Online]. Available: https://iocscience.org/ejournal/index.php/mantik/article/download/2130/1669/6179

[7] BPS, "Micro-Analysis of Household Poverty and Inequality in Indonesia," 2023. [Online]. Available: https://al.unnes.ac.id/journals/jejak/article/download/9512/2955/59720

[8] F. Feng, "Application of Data Mining Technology in Poverty Alleviation Prediction in Ethnic Areas," in Proc. 4th Int. Conf. Informatization Econ. Dev. Manage., 2024. [Online]. Available: https://doi.org/10.4108/eai.2322024.2345877

[9] C. Mustika and R. Nurjanah, "Rural and Urban Poverty Models on Sumatra Island," J. Perspektif Pembiayaan Pembang. Daerah, vol. 9, no. 1, pp. 107–114, 2021. [Online]. Available: https://doi.org/10.22437/ppd.v9i1.10684

[10] R. T. Vulandari, "Development of Data Mining Software Using Association Techniques Based on Apriori Algorithm Method," J. Inf. Syst. Informatics Comput., vol. 6, no. 1, pp. 125–136,

2017. [Online]. Available: https://doi.org/10.52362/jisicom.v6i1.80
[11] C. Yeh, A. Perez, A. Driscoll et al., "Using publicly available satellite imagery and deep learning to understand economic well-being in Africa," Nat. Commun., vol. 11, no. 1, p. 2583, 2020. [Online]. Available: https://doi.org/10.1038/s41467-020-16185-w
[12] M. Engler, M. Kasy, and C. Leaver, "Machine learning for public policy: Principles and practice," CEPR Discussion Paper No. DP16267, 2021. [Online]. Available: https://cepr.org/publications/dp16267.