Vol. 3 No. (2) (December 2025) 83-91

Journal of Data Insights

http://journalnew.unimus.ac.id/index.php/jodi

Analysis of Data Mining in Predicting Poverty Levels in Indonesia Using the Decision Tree Method

Zaehol Fatah [1], **Ahsin Ilallah**[2]
[1]Sistem Informasi, Sains & Teknologi, Universitas Ibrahimy
[2]Sistem Informasi, Sains & Teknologi, Universitas Ibrahimy

DOI: https://doi.org/10.26714/jodi.v3i2.878

Info Artikel

Sejarah Artikel:
Disubmit 29 October 2025
Direvisi 20 November 2025
Disetujui 20 December 2025

Keywords:
Data Mining; Decision Tree; RapidMiner; Poverty; Organizational Interoperability.

Abstract

This study aims to examine the application of the Decision Tree method in predicting poverty levels in Indonesia using the RapidMiner software. Poverty is a complex issue influenced by social, economic, and educational factors. Through a data mining approach, this research seeks to identify patterns within poverty data to support more accurate decision-making. The research data were obtained from the public platform Kaggle and include key variables such as individual expenditure, the Human Development Index (HDI), average study time, access to proper sanitation and safe drinking water, as well as the open unemployment rate. The results show that the Decision Tree model achieved an accuracy of 94.90%, with a precision of 95.24% and a recall of 93.75%, based on the confusion matrix. The use of RapidMiner also facilitates the analysis, as the results are presented visually and are easy to understand. This model is recommended for implementation in government information systems as a data-driven decision-making tool to help reduce national poverty levels.

✉ Alamat Korespondensi:
E-mail : zaeholfatah@gmail.com

e-ISSN: 2988 - 2109

INTRODUCTION

Decision Tree is a method used in the decision-making process that applies a classification approach, where each characteristic is mapped into predefined categories. A Decision Tree can also be understood as a specialized framework designed to break down a large dataset into several more manageable subsets through the systematic application of a series of decision-making protocols (Werdiningsih, et al., 2022). In the context of Indonesia, the application of this method is highly relevant because the country has a very high diversity of ethnic groups. The archipelago within the territory of the Unitary State of the Republic of Indonesia reflects a nation rich in ethnic and cultural diversity (Pitoyo, et al., 2017). Several previous studies have demonstrated the effectiveness of data mining approaches in poverty analysis. Poverty has become a complex and multidimensional issue in Indonesia, even though the poverty rate has declined over the past few decades, particularly regarding the persistent problem of inaccurate targeting in economic distribution (Danil, et al., 2022). Poverty is a problem experienced by individuals or groups who struggle to meet their basic needs. This situation arises from weak human resource quality, the mismatch between minimum wages and the standard cost of living, as well as rapid population growth, which reduces competitiveness across various sectors, particularly in gaining employment (Sari, 2021).

Although the decline in national poverty statistics indicates positive progress, the overall data often masks more complex realities at the local level. Many regions continue to face significant levels of deprivation, with varying characteristics and challenges (Prasetyo, et al., 2025). A method is needed to predict the extent of decline or increase in the poor population in Indonesia in future periods. Changes in poverty levels in the following year can be estimated by utilizing previous years' poverty statistics as the basis for calculation (Mukmin, et al., 2021). Information technology–based approaches, such as data mining, are increasingly used in efforts to map poverty. Data exploration is the process of extracting models or hidden information from large datasets with the aim of assisting in the decision-making process (Setyawan, et al., 2025). Sehingga bigdata dan kecerdasan buatan (AI) menjadi alat utama dalam mengumpulkan, mengolah dan menganalisis data secara masif (Fatah, 2025).

This study will utilize the Decision Tree method, which is a classification technique that works by dividing data into smaller subsets based on the most informative attributes. At each branch, the algorithm selects the attribute with the highest information gain, namely the attribute that has the greatest ability to distinguish between classes. The method also involves analyzing large-scale data patterns to identify complex relationships between poverty factors and the potential rise in poverty rates in Indonesia. With the application of this method, it is expected to provide a positive contribution to reducing poverty levels in Indonesia (Sinaga, et al., 2024).

RESEARCH METHOD

The dataset used in this study contains several variables related to the social and economic conditions of the population. These variables consist of input variables (predictors) and an output variable (target). The input variables include per capita expenditure, the Human Development Index (HDI), average years of schooling, access to proper sanitation, access to safe drinking water, and the open unemployment rate. These variables were selected because they are considered capable of representing the factors that influence poverty conditions in Indonesia.

Meanwhile, the target variable predicted in this study is the poverty rate (P0). This variable is then converted into a categorical form so that it can be processed using the Decision Tree method. With clear variable specifications, the data analysis process becomes more structured and the classification results are easier to interpret. This study employs a Data Mining approach using the Decision Tree method to predict poverty levels through the RapidMiner application. All stages of the research are designed systematically to ensure the successful implementation of the method and the accuracy of the testing results.

DATA COLLECTION

One of the crucial elements in supporting poverty-alleviation planning is the availability of accurate poverty information. With such information, the government can formulate appropriate policies and actions to address poverty issues. In addition, this information also helps the government analyze year-to-year comparisons of poverty levels (Ferezagia, 2018). Therefore, the researchers used secondary information obtained from the public data-sharing platform Kaggle (www.kaggle.com). The dataset was selected because it contains attributes relevant to the factors causing poverty and has been widely used in studies related to poverty-level classification using machine learning.

The dependent variable used in this study is the percentage of the poor population by regency/city (Y), as presented in Table 1. The independent variables, as detailed in Table 1, include Percentage of Poor Population (P0) by District/City (X1); the average years of schooling of the population aged 15 years and above, measured in years (X2); adjusted per capita expenditure in thousand rupiah per person per year (X3); the Human Development Index (HDI) (X4); life expectancy at birth measured in years (X5); the percentage of households with access to improved sanitation (X6); the percentage of households with access to improved drinking water (X7); the open unemployment rate (X8); the labor force participation rate (X9); and gross regional domestic product (GRDP) at constant prices by expenditure, measured in rupiah (X10). In addition, as shown in Table 1, this study employs a poverty classification variable as a categorical variable to represent the overall level of regional poverty.

Table 1. Dataset

X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	Y
18,98	9,48	7148	66,41	65,28	71,56	87,45	5,71	71,15	1648096	0
20,36	8,68	8776	69,22	67,43	69,56	78,58	8,36	62,85	1780419	1
13,18	8,88	8180	67,44	64,4	62,55	79,65	6,46	60,85	4345784	0
13,41	9,67	8030	69,44	68,22	62,71	86,71	6,43	69,62	3487157	0
14,45	8,21	8577	67,83	68,74	66,75	83,16	7,13	59,48	8433526	0
15,26	9,86	10780	73,37	68,86	90,58	90,1	2,61	76,3	5953118	0
18,81	9,55	9593	71,67	67,99	89,6	94,22	7,09	60,05	7485861	0
14,05	10,33	9644	73,58	69,79	87,4	82,36	7,7	61,67	10261585	0
19,59	9	9860	70,7	66,95	54,1	89,24	7,28	60,29	7975099	0
13,25	9,29	8867	72,33	71,26	81,89	93,53	4,32	65,91	10374480	0
17,43	8,64	8201	69,46	68,81	79,97	91,09	8,31	58,47	16924103	0
16,34	8,67	8428	66,99	65,06	65,71	95,34	4,04	57,91	3069805	0
19,64	8,4	8856	67,56	65,53	47,63	84,68	1,84	78,99	1981879	0
13,34	8,91	8367	69,48	69,63	87,45	83,12	5,87	66,43	6062520	0

Data Mining

Data mining began to gain popularity in the 1990s, along with the increasing use of data as an essential element in various fields such as marketing and business, sciences and robotics, as well as arts and entertainment. Data mining is a process that utilizes one or more machine learning methods to automatically analyze and extract information (Eska, 2016). Based on various explanations, data explorati on is a technique applied to large datasets to identify hidden patterns by using an integrated approach that combines statistical analysis, machine learning algorithms, and data management technologies. Its main objective is to uncover previously unknown patterns, and once these patterns are identified, the results can be used to solve various types of problems (Derisma, 2020).

Decision Tree Method

A data processing technique used to predict future conditions by constructing a classification and estimation model represented in a tree format. The Decision Tree also serves as a visual representation that facilitates understanding of the decision-making process in a systematic, step-by-step, and logical manner (Nurani, et al., 2023). The Decision Tree method converts large datasets into a decision-tree structure that represents various rules or specific conditions. The resulting rules can be understood in human language and then translated into database systems, such as Structured Query Language, to trace records within certain data. A decision tree is a structure that functions to break down big data into several subsets of sample data through the application of a series of decision rules. (Pralarsya, et al., 2020).

The stages of the Decision Tree Algorithm are as follows :

a. Prepare the training dataset.
b. Determine the root node.
c. Calculate the information gain

$$Gain(S, A) = Entropy(S) - \sum_{i=1} n \frac{|Si|}{|S|} \times Entropy\ (Si)$$

d. Calculate the entropy value

$$Entropy(S) = \sum_{i-1} N - \pi \times log2\pi$$

e. The node splitting process stops when the node becomes pure.

Evaluation of the Decision Tree Model

To assess the performance of the Decision Tree model in predicting poverty levels, an evaluation method based on the confusion matrix was used. This evaluation is important to determine how accurately and consistently the model can recognize poverty categories.

1. Confusion Matrix
   The model evaluation table is presented as follows:

	Predicted: Poor	**Predicted: Not Poor**
Actual: Poor	15 (True Positive)	1 (False Negative)
Actual: Not Poor	2 (False Positive)	18 (True Negative)

2. Accracy
   Measures the percentage of correct predictions out of all test data.
   Accuracy=94.90%
3. Precission
   Indicates the level of accuracy when the model predicts the "Poor" category.
   Precision=95.24%
4. Recall
   Measures the model's ability to correctly identify data that truly belong to the poor category.
   Recall=93.75%

   The evaluation results show that the model has excellent performance in the classification process. The high precision and recall values indicate that the model is able to minimize prediction errors and accurately detect regions categorized as poor. Thus, the Decision Tree model is reliable for use as a data-driven decision support tool in poverty analysis.

RapidMiner

A software application used to process data using data-mining algorithm principles. This software can discover patterns from big data through a combination of statistical methods, AI, and database technology, and it helps users manage data by utilizing various available operators. These operators function to adjust the data by connecting them to each operator's node, and the processing results can be viewed by linking them to the output node (Novianti, 2019).

RESULTS AND DISCUSSION

Based on the existing problems, a system capable of providing information to predict poverty levels is needed, so that it can assist the government and the public in preventing the potential rise in poverty rates. With this prediction system, it is expected that fluctuations in poverty levels in Indonesia can be identified. This will enable the government or the general public to take strategic actions to improve societal welfare through well-targeted policies. Prediction is a systematic process used to forecast events that may occur in the future based on past evidence, existing information, and current conditions, with the aim of minimizing errors so that the results are more accurate. Prediction does not seek to provide completely certain answers about future events; rather, it aims to produce estimates that are as close as possible to reality. The main function of prediction is to identify hidden patterns contained within the available data (Kafil, 2019).

After certain patterns are discovered, these patterns can be used to predict the potential increase in poverty levels in a particular region, even before the impact is visibly apparent in the field. In this data-processing procedure, a data mining approach is used, which serves as a system for discovering information from big data. This approach combines statistical methods, mathematics, AI, and machine learning to perform data separation and identify useful and meaningful information that was previously hidden within large-scale databases. Through the application of data mining in poverty analysis, important patterns such as the main contributing factors to poverty, trends in the rise or decline of community welfare, and regions with high risk can be identified. Therefore, the prediction results can serve as a reference for the government in formulating preventive and strategic measures to reduce poverty rates more effectively.

Data Transformation

This stage plays a crucial role in improving the effectiveness of the data-mining algorithm that will be applied during the classification phase. Data transformation involves converting data into a more suitable format. The goal of this process is to make the data more structured and easier for the algorithm to process, thereby enhancing the accuracy of the classification results. Therefore, this stage serves as an essential foundation to ensure that the resulting data-mining model achieves high levels of accuracy, reliability, and validity in the decision-making process. Data transformation presented in Figure 1.

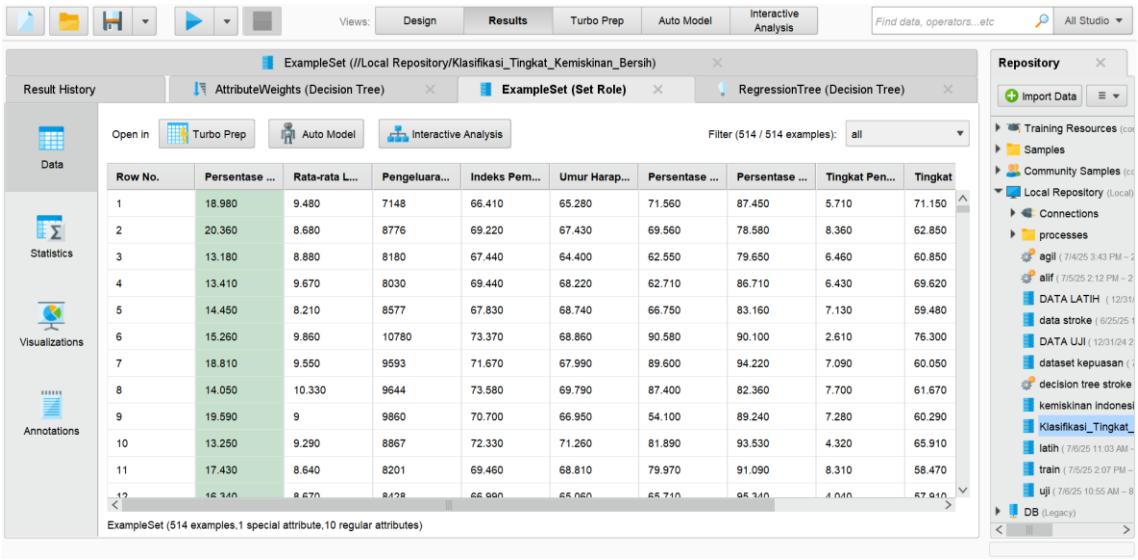

Figure 1. Data Transformation

Processing

After the data has been successfully transferred and transformed, the next stage is processing. This stage plays a very important role in ensuring that the developed data-mining model can function optimally not only on training data but also on new data that has never been processed or recognized before. Processing aims to prepare the data so that it fits the requirements of the algorithm being used, enabling the resulting model to accurately recognize patterns and relationships between variables. Processing presented in Figure 2.

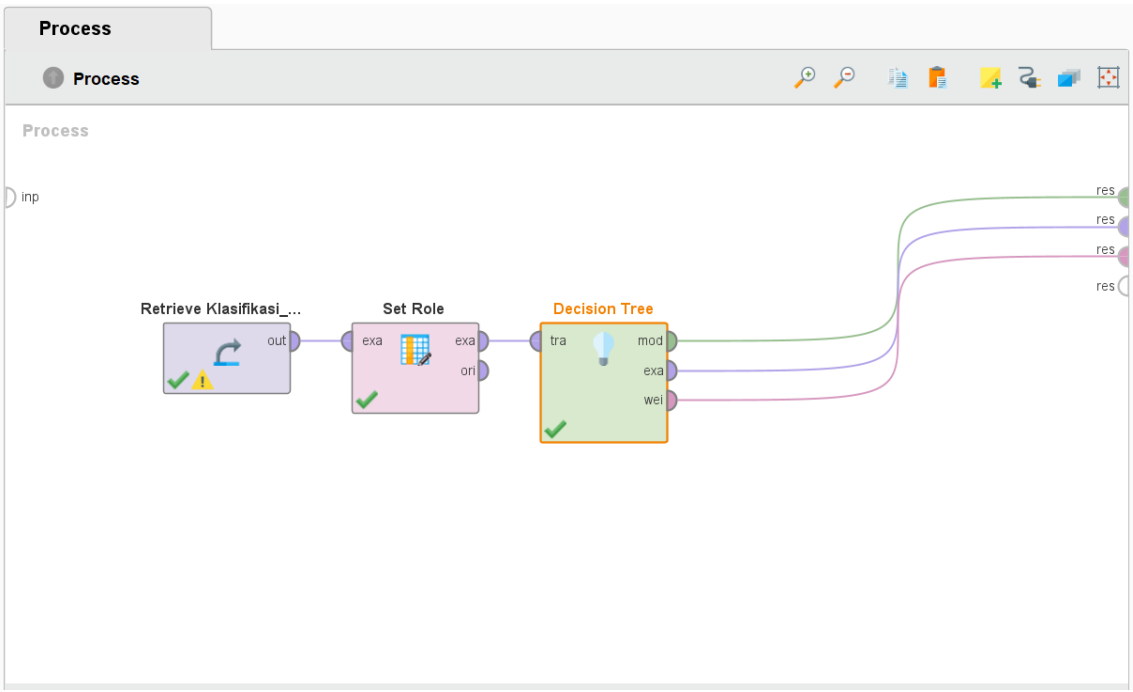

Figure 2. Processing

Decision Tree Modeling Results

The final stage represents the outcome of data that has undergone processing and analysis. At this stage, a model in the form of a Decision Tree is produced, which is a graphical representation illustrating various possible outcomes based on a series of interconnected data conditions or attributes. This model helps reveal patterns of relationships among variables that influence poverty levels, such as access to sanitation, availability of safe drinking water, the human development index, regional GDP, and the average length of schooling. The Data Mining model applied in this study is the Decision Tree method. The results presented in Figure 3.

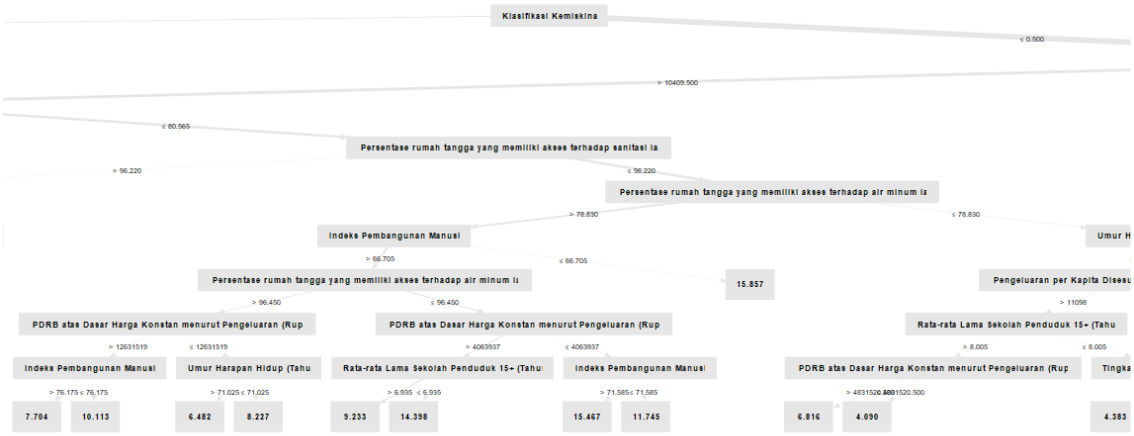

Figure 3. Decision Tree Modeling Results

The Decision Tree model indicates that the variables with the greatest influence on poverty levels are access to proper sanitation and safe drinking water. Areas with low sanitation access tend to fall into the poverty category. The model then considers other indicators such as the Human Development Index (HDI), per capita expenditure, and average years of schooling. The combination of these variables forms a set of rules that guides the data toward specific poverty categories. The results show that poverty is strongly influenced by the quality of basic public services and the economic capacity of the population. This stage is part of the evaluation and interpretation of the model in the data-mining process. At this stage, the system has built a regression decision-tree model (Regression Tree), which predicts numerical values (not just categories) of the target variable in this case, the poverty level based on the attributes contained in the dataset.

RegressionTree

```
Klasifikasi Kemiskinan > 0.500
|   Pengeluaran per Kapita Disesuaikan (Ribu Rupiah/Orang/Tahun) > 5722
|   |   Pengeluaran per Kapita Disesuaikan (Ribu Rupiah/Orang/Tahun) > 9781.500
|   |   |   Rata-rata Lama Sekolah Penduduk 15+ (Tahun) > 8.535: 11.145 {count=2}
|   |   |   Rata-rata Lama Sekolah Penduduk 15+ (Tahun) ≤ 8.535: 20.335 {count=2}
|   |   Pengeluaran per Kapita Disesuaikan (Ribu Rupiah/Orang/Tahun) ≤ 9781.500
|   |   |   Tingkat Pengangguran Terbuka > 2.175
|   |   |   |   Umur Harapan Hidup (Tahun) > 61.315
|   |   |   |   |   Persentase rumah tangga yang memiliki akses terhadap sanitasi layak > 84
|   |   |   |   |   |   Indeks Pembangunan Manusia > 65.715: 23.563 {count=3}
|   |   |   |   |   |   Indeks Pembangunan Manusia ≤ 65.715: 27.852 {count=4}
|   |   |   |   |   Persentase rumah tangga yang memiliki akses terhadap sanitasi layak ≤ 84
|   |   |   |   |   |   Pengeluaran per Kapita Disesuaikan (Ribu Rupiah/Orang/Tahun) > 9177: 27.065 {count=2}
|   |   |   |   |   |   Pengeluaran per Kapita Disesuaikan (Ribu Rupiah/Orang/Tahun) ≤ 9177
|   |   |   |   |   |   |   Pengeluaran per Kapita Disesuaikan (Ribu Rupiah/Orang/Tahun) > 7831
|   |   |   |   |   |   |   |   Tingkat Partisipasi Angkatan Kerja > 67.545: 22.777 {count=4}
|   |   |   |   |   |   |   |   Tingkat Partisipasi Angkatan Kerja ≤ 67.545: 18.930 {count=2}
|   |   |   |   |   |   |   Pengeluaran per Kapita Disesuaikan (Ribu Rupiah/Orang/Tahun) ≤ 7831
|   |   |   |   |   |   |   |   Rata-rata Lama Sekolah Penduduk 15+ (Tahun) > 7.080: 22.623 {count=9}
|   |   |   |   |   |   |   |   Rata-rata Lama Sekolah Penduduk 15+ (Tahun) ≤ 7.080: 26.117 {count=3}
|   |   |   |   Umur Harapan Hidup (Tahun) ≤ 61.315: 28.217 {count=3}
|   |   |   Tingkat Pengangguran Terbuka ≤ 2.175
```

Figure 4. Regression Tree

The structure of the Regression Tree shows that per capita expenditure is the first and most determining variable in the classification of poverty. Regions with low per capita expenditure particularly below approximately Rp 5,722,000 per person per year tend to fall into the poor category. In addition, the model also identifies that factors such as average years of schooling, life expectancy, and access to proper sanitation play important roles as subsequent splitting variables. For example, in certain branches, regions with an average schooling duration below 8 years, even if their expenditure is relatively high, are still at risk of being categorized as poor. This indicates that education remains a crucial factor influencing overall welfare.

The Human Development Index (HDI) also appears as a distinguishing factor in several branches of the tree. HDI values below the range of 65 – 66 frequently appear in nodes associated with poverty. This reinforces the notion that the overall quality of human development is closely related to poverty levels. Lastly, variables such as labor force participation rate and open unemployment rate emerge as additional splitting factors in some branches. This suggests that macroeconomic factors related to employment opportunities also influence the classification of poverty.

CONCLUSION

From the findings of the research conducted, it can be concluded that the application of the Decision Tree method through the RapidMiner application can provide relatively accurate predictions in assessing poverty levels in Indonesia. Through the data-mining process, which includes data collection, transformation, processing, and modeling, a classification model was produced that is able to identify the relationships between various social and economic factors and poverty levels. The factors that have the greatest impact in this model include expenditure per capita, average years of education, the human development index, access to sanitation and clean water, as well as the open unemployment rate. The resulting decision-tree model also provides a hierarchical overview of the factors causing poverty, which can serve as a reference in formulating poverty-reduction policies. The application of a data-mining approach based on the Decision Tree method demonstrates that integrating data from various institutions and sectors aligned with Gottschalk & Saether's concept of organizational interoperability can enhance the effectiveness of government analysis and decision-making in addressing poverty in a more targeted, accountable, and evidence-based manner. As an implication, the findings of this study can be used as a foundation for developing smarter poverty-prediction and monitoring systems that are capable of adapting to future changes.

REFERENCES

Werdiningsih, I., et al. (2022) Data Mining Management with Matlab Programming. Surabaya: Airlangga University Press, p. 65.

Pitoyo, A.J., et al. (2017) 'The dynamics of ethnic development in Indonesia in the context of national unity', Journal of Population and Policy, 25(1).

Danil, S., et al. (2022) 'Improving classification models in regencies/cities in Indonesia using the decision tree method', JITET (Journal of Informatics and Applied Electrical Engineering), 13(2). doi:10.23960/jitet.v13i2.6336.

Sari, A.A. (2021) 'The influence of minimum wage, open unemployment rate, and population size on poverty in Central Java Province', Equilibrium: Jurnal Ilmiah Ekonomi, Manajemen, dan Akuntansi, 10(1).

Prasetyo, T.L., et al. (2025) 'Clustering poverty data in Indonesia from 2007–2017 using K-Means and decision tree in Python', RIGGS (Journal of Artificial Intelligence and Digital Business), 4(2).

Mukmin, D.A., et al. (2021) 'Application of the moving average method in a poverty rate prediction information system', Al-Mantiq: Journal of Multidisciplinary Applications of Quantum Information Science, 1(1).

Setyawan, A., et al. (2025) 'Poverty classification in Indonesia using decision tree with RapidMiner', JMA (Academic Media Journal), 3(7).

Fatah, Z. (2025) Computers and Society. Jakarta: PT Penamuda Media, p. 114.

Sinaga, L.M., et al. (2024) 'Analysis and prediction of poverty rate percentage in Indonesia using multiple linear regression', KAKIFIKOM (Kumpulan Artikel Ilmiah Fakultas Ilmu Komputer), 6(2).

Ferezagia, D.V. (2018) 'Analysis of poverty levels in Indonesia', Journal of Applied Social Humanities, 1(1).

Eska, J. (2016) 'Application of data mining for wallpaper sales prediction using the C4.5 algorithm', JURTEKSI (Jurnal Teknologi dan Sistem Informasi), 2(2).

Derisma (2020) 'Comparison of algorithm performance for heart disease prediction using data mining techniques', JAIC (Journal of Applied Informatics and Computing), 4(1).

Nurani, A.T., et al. (2023) 'Comparison of decision tree regression and multiple linear regression performance for BMI prediction on the asthma dataset', Journal of Science and Science Education, 6(1).

Prakarsya, A., et al. (2020) 'Implementation of data mining for predicting the spread of HIV/AIDS in Bandar Lampung using the decision tree technique', Jurnal Siskomti, 3(2).

Novianti, D. (2019) 'Implementation of the Naive Bayes algorithm on the hepatitis dataset using RapidMiner', Jurnal Komputer dan Informatika, 21(1).

Kafil, M. (2019) 'Application of the K-nearest neighbors method for web-based sales prediction at Boutiq Dealove Bondowoso', JATI (Jurnal Mahasiswa Teknik Informatika), 3(2).