Scientific Committee

- Prof. Hideki Kuwahara, Ph.D. (Sophia University, Japan)
- Prof. Dr. Willy Susilo ((Wollongong University, Australia)
- Prof. Dr. rer. nat. Herry Pribawanto Suryawan (Universitas Sanata Dharma - Indonesia)
- Prof. Ir. Sudi Mungkasi, Ph.D. (Universitas Sanata Dharma - Indonesia)
- Prof. Bambang Soelistijanto, Ph.D. (Universitas Sanata Dharma - Indonesia)
- Prof. Dr. Ir. Feri Yusivar, M.Eng. (Universitas Indonesia, Indonesia)
- Prof. Dr. Eng. Ir. Deendarlianto, S.T., M.Eng. (Gadjah Mada University, Indonesia)
- Dr. Uday Pandit Khot (St. Francis Institute of Technology, University of Mumbai, India)
- Dr. Farah Suraya Md Nasrudin (Universiti Teknologi MARA, Malaysia)
- Assoc. Prof. Dr. Shi Bai (Florida Atlantic University, U.S)
- Parkpoom Phetpradap, Ph.D. (Chiang Mai University, Thailand)
- Tri Hieu Le, Ph.D. (Ho Chi Minh University of Technology (HUTECH), Vietnam)
- Alex Baenaorama, Ph.D. (St. Paul University Dumaguete, Philipine)
- Ignasia Handipta Mahardika, Ph.D. (Sogang University, South Korea)
- Dr. Irma Saraswati, M.Sc. (Universitas Sultan Ageng Tirtayasa, Indonesia)
- Dr. Imamul Muttakin (Universitas Sultan Ageng Tirtayasa, Indonesia)
- Dyonisius Dony Ariananda, S.T., M.Sc., Ph.D. (Gadjah Mada University, Indonesia)
- Dr. I Gusti Ngurah Bagus Catrawedarma, S.T., M.Eng. (Politeknik Negeri Banyuwangi, Indonesia)

Steering Committee

- Chairperson: Ir. Drs. Haris Sriwindono, M.Kom., Ph.D. (Universitas Sanata Dharma - Indonesia)
- Member:
    - Dr. Ir. I Gusti Ketut Puja (Universitas Sanata Dharma - Indonesia)
    - Prof. Dr. Frans Susilo, S.J. (Universitas Sanata Dharma - Indonesia)
    - Dr. Ir. Bernadeta Wuri Harini (Universitas Sanata Dharma - Indonesia)
    - Prof. Dr. Ir. Anastasia Rita Widiarti (Universitas Sanata Dharma - Indonesia)
    - Dr. Lusia Krismiyati Budiasih (Universitas Sanata Dharma - Indonesia)

Organizing Committee

- Chairman: Dr. Ir. Achilleus Hermawan Astyanto (Universitas Sanata Dharma - Indonesia)
- Member:
    - Dr. Sri Hartati Wijono, M.Kom. (Universitas Sanata Dharma - Indonesia)
    - Ir. Kartono Pinaryanto, S.T., M.Cs. (Universitas Sanata Dharma - Indonesia)
    - Regina Chelinia Erianda Putri, M.T. (Universitas Sanata Dharma - Indonesia)
    - Ir. Rosalia Arum Kumalasanti, M.T. (Universitas Sanata Dharma - Indonesia)
    - Michael Seen, M.Eng. (Universitas Sanata Dharma - Indonesia)
    - Ir. Robertus Adi Nugroho, M.Eng. (Universitas Sanata Dharma - Indonesia)
    - Eduardus Hardika Sandy Atmaja, Ph.D. (Universitas Sanata Dharma - Indonesia)
    - Gilang Arga Dyaksa, M.Eng. (Universitas Sanata Dharma - Indonesia)
    - Heryoga Winarbawa, M.Eng. (Universitas Sanata Dharma - Indonesia)

Editors

- Prof. Peerapong Uthansakul, Ph.D. (Suranaree University of Technology - Thailand)
- Assoc. Prof. Dr. Eng. Ir. I Made Wicaksana Ekaputra (Universitas Sanata Dharma - Indonesia)
- Assoc. Prof. Ir. Damar Widjaja, Ph.D. (Universitas Sanata Dharma - Indonesia)

InCASST
International Conference on Applied Sciences and Smart Technologies

Innovating for Sustainability: Digitalization, Green Energy, and Achieving Energy Independence for a Greener Future

Yogyakarta, 15 October 2025 (Hybrid)

KEYNOTE SPEAKERS

Prof. Dr. Thomas Götz
University of Koblenz, Germany

Prof. Dr. Kavita Sonawane
SFIT, University of Mumbai, India

Ishak H.P. Tnunay, Ph.D.
Beehive Drones, PT. Aerotek Global Inovasi, Indonesia

Andreas Prasetyadi, Ph.D.
Universitas Sanata Dharma, Indonesia

Call for Papers

The 2nd International Conference on Applied Sciences and Smart Technologies (InCASST 2025)

SCOPES

Environmental Developments
- Environmental impact assessment and management
- Waste management and recycling
- Environmental biotechnology and microbiology
- Carbon capture and sequestration

Green Technologies
- Renewable energy technologies and systems
- Climate change and global warming
- Sustainable agriculture and land use practices
- Clean energy
- Energy Efficiency
- Water-Energy Nexus
- Green Materials

Digital Society
- IoT and AI-based sustainability solutions
- Data analysis and predictive modeling for environmental sustainability
- Sustainable transportation and mobility solutions
- Data and distributing computing

IMPORTANT DATES

- Early bird registration deadline : 14 July 2025
- Full paper submission deadline : 29 Aug 2025
- Acceptance notification : 12 Sept 2025
- Late registration deadline : 15 Sept 2025
- Conference day : 15 Oct 2025

CONTACT PERSON

Eduardus Hardika Sandy Atmaja, Ph.D. +62 895 3817 54488
Dr. Sri Hartati Wijono, M.Kom. +62 811 2646 471

REGISTRATION FEE

Offline Presenter
- Early bird : IDR 2,650,000 (USD 185)
- Late : IDR 2,900,000 (USD 205)

Online Presenter
- Early bird : IDR 2,500,000 (USD 175)
- Late : IDR 2,750,000 (USD 195)

Offline Non Presenter : IDR 150,000 (USD 15)

PAYMENT METHODS

- Bank Transfer
  Acc. Number : 8466045484
  Acc. Name : Michael Seen
  Bank Name : Bank Central Asia (BCA)
- Paypal : incasst@usd.ac.id
  Swift Code : CENAIDJA

Scan here

Co Hosted by

UBAYA
UNIVERSITAS PIGNATELLI TRIPUTRA

Sponsored by

wkc wijaya kusuma contractors

Published by

E3S Web of Conferences
IJASST International Journal of Applied Sciences and Smart Technologies

incasst_usd
incasst@usd.ac.id
s.usd.ac.id/incasst

Free Access to the whole issue

E3S Web of Conferences

Volume 687 (2026)

The 2nd International Conference on Applied Sciences and Smart Technologies (InCASST 2025)

Yogyakarta, Indonesia, October 15, 2025

P. Uthansakul, I M.W. Ekaputra and D. Widjaja (Eds.)

Export the citation of the selected articles Export

Select all

- *Environmental Developments & Sustainable Systems*

- Green Technologies & Digital Society

Published online: 15 January 2026

DOI: https://doi.org/10.1051/e3sconf/202668702014

Abstract | PDF (683.5 KB) | References | NASA ADS Abstract Service

E3S Web of Conferences

eISSN: 2267-1242

Open Access

Issue	E3S Web Conf. **Volume** 687, 2026 The 2nd International Conference on Applied Sciences and Smart Technologies (InCASST 2025)
Article Number	02008
Number of page(s)	12
Section	Green Technologies & Digital Society
DOI	https://doi.org/10.1051/e3sconf/202668702008
Published online	15 January 2026

E3S Web of Conferences **687**, 02008 (2026)

Evaluation of the Performance of K-means and Bisecting K-means in Clustering Indonesian Regions using Poverty Data

Dhea Cindera Bantala* **and Paulina Heruningsih Prima Rosa**

Department of Informatics, Faculty of Science and Technology, Sanata Dharma University, Indonesia

* Corresponding author: cinderadhea@gmail.com

Table of Contents

Article contents

Abstract PDF (1.044 MB) References

Database links

NASA ADS Abstract Service

Metrics

Show article metrics

Services

Same authors

- Google Scholar
- EDP Sciences database

Recommend this article
Download citation
Alert me if this article is corrected
Alert me if this article is cited

Related Articles

Optimizing MSMEs clusters using type-2 Fuzzy K-Medoids method in Sampang Madura District
EPJ Web of Conferences 344, 01057 (2025)

Abstract

Clustering regions in Indonesia based on poverty data is essential for developing targeted policies. A key challenge is determining the optimal number of clusters to accurately reflect regional disparities. This study compares the Bisecting K-means and K-means algorithms, evaluating them with Sum Squared Error (SSE), David-Bouldin Index (DBI), and Silhouette Coefficient (SC). Evaluation results identified K-means with six clusters (k= 6) as the most optimal model. It achieved a low SSE (2129.37), a relatively low DBI (0.82), and a moderate SC (0.3625). This model successfully maps regions from the most prosperous to the poorest, providing a clear basis for poverty alleviation strategies. For future work, a deeper analysis of cluster members using heatmaps or boxplots is recommended. Visualizing the results on a map would also help stakeholders easily understand the spatial distribution of poverty levels. Furthermore, comparing clustering results year-on-year would be valuable for tracking regional progress.

E3S Web of Conferences

eISSN: 2267-1242

Evaluation of the Performance of K-means and Bisecting K-means in Clustering Indonesian Regions using Poverty Data

Dhea Cindera Bantala[1]*, and *Paulina Heruningsih Prima* Rosa[1]

[1]Department of Informatics, Faculty of Science and Technology, Sanata Dharma University, Indonesia

Abstract. Clustering regions in Indonesia based on poverty data is essential for developing targeted policies. A key challenge is determining the optimal number of clusters to accurately reflect regional disparities. This study compares the Bisecting K-means and K-means algorithms, evaluating them with Sum Squared Error (SSE), David-Bouldin Index (DBI), and Silhouette Coefficient (SC). Evaluation results identified K-means with six clusters (k = 6) as the most optimal model. It achieved a low SSE (2129.37), a relatively low DBI (0.82), and a moderate SC (0.3625). This model successfully maps regions from the most prosperous to the poorest, providing a clear basis for poverty alleviation strategies. For future work, a deeper analysis of cluster members using heatmaps or boxplots is recommended. Visualizing the results on a map would also help stakeholders easily understand the spatial distribution of poverty levels. Furthermore, comparing clustering results year-on-year would be valuable for tracking regional progress.

1 Introduction

According to data from the Central Statistics Agency (BPS), the percentage of poverty in Indonesia as of March 2023 was 9.36%, while as of March 2024 it was 8.47% with the number of poor people reaching 23.85 million [1]. Major challenges for the central government in addressing poverty include formulating national poverty alleviation policies, determining more targeted budget allocations, and identifying regions that fall into the extreme, middle, and low poverty categories. For local governments, for example, the challenge is understanding their region's position relative to other regions and determining development priorities (education, health, infrastructure, and economic empowerment). The first step in overcoming the problem of poverty in Indonesia is to understand the pattern and description of poverty in each region using the clustering method.

Clustering and mapping regions in Indonesia based on poverty data is useful for identifying areas with similar poverty characteristics, enabling more targeted policies and programs. Central and regional governments can use it to formulate poverty alleviation strategies and allocate budgets efficiently; academics and researchers can utilize it for data-driven studies and recommendations; NGOs and international organizations can use it as a

* Corresponding author: cinderadhea@gmail.com

basis for targeted social interventions; and the public and media can gain transparency into regional conditions and play a role in policy oversight.

One of the obstacles in applying clustering to poverty data is determining the most appropriate number of clusters to represent the differences in poverty levels between regions more accurately. Several studies on clustering indicators and poverty rates using clustering methods have been conducted. Luchia et al. conducted research on comparing K-means and K-Medoids on clustering poverty data in Indonesia [2]. The results showed that the K-means algorithm is superior to K-Medoids in clustering poverty data in Indonesia. This is evidenced by the best K-means DBI value of 0.041 with K = 8 trials.

Munandar implemented clustering algorithm for grouping poverty level in Banten Province using K-Medoids and K-means [3]. The clustering results show an equal distribution of 3 clusters in K-medoid and K-means using RStudio, namely 3 regencies/cities with low poverty, 3 regencies/cities with moderate poverty, and 2 regencies/cities with high poverty. However, the DBI value in the K-Medoid algorithm is lower than K-means with results of 0.582 and 0.602, respectively. Rahman et al. performed a clustering approach to identify multidimensional poverty indicators for the bottom 40 percent group. K-means algorithm was tested with four different distance measure: Euclidean Distance (ED), Correlation Similarity (CrS), Cosine Similarity (CS) and Dice Similarity (DS) to choose the best distance measure [4]. Wardani et al. conducted a research on implementation of the K-means method for clustering Regency/City in North Sumatra based on poverty indicators [5].

In this paper, the author conduct research on the comparison of K-means Clustering and Bisecting K-means Clustering algorithms in clustering poverty information in Indonesia. Three different evaluation metrics, namely Sum Squared Error (SSE), Davies-Bouldin Index (DBI), and Silhouette Coefficient (SC) are applied to evaluate the clustering results using the K-means and Bisecting K-means algorithms and to determine the most appropriate number of clusters for poverty data in Indonesia.

2 Research Methods

The data used in this study was taken from the Indonesian Central Statistics Agency (BPS) website (https://www.bps.go.id). The data contains 22 attributes that serve as poverty indicators in 514 regencies/municipalities in Indonesia in 2023. The following are the attributes:

1. REG: Regency/city
2. JPM: Number of poor people
3. PPM: Percentage of poor population
4. P1: Poverty depth level
5. P2: Poverty severity index
6. PL: Poverty line (Rp/cap/month)
7. < SD: Percentage of poor population aged 15 years and above, who have completed less than elementary school
8. TMT SD/SMP: Percentage of poor population aged 15 years and above who have completed elementary/middle school
9. >SMA: Percentage of poor people aged 15 years and above who have completed high school or above
10. 7-12: School participation rates of poor population by regency/city and age group 7-12
11. 13-15: School participation rate of poor population by regency/city and age group 13-15
12. TB: Percentage of poor population aged 15 years and above who are unemployed
13. BSI: Percentage of poor population aged 15 years and above working in informal sector
14. BSF: Percentage of poor population aged 15 years and above working in formal sector

15. TB1: Percentage of poor population aged 15 years and above who are unemployed
16. BSP: Percentage of poor population aged 15 years and above working in the agricultural sector
17. BBSP: Percentage of poor population aged 15 years and above working in non-agricultural sectors
18. PM: Percentage of per capita expenditure on food and having poor status
19. PTM: Percentage of per capita expenditure on food and having non-poor status
20. PMPTM: Percentage of per capita expenditure on food and having poor and non-poor status
21. AL: Percentage of poor households using adequate water
22. JS: Percentage of poor households using their own toilet

Figure 1 shows the research stages. The explanations of each stage are elaborated in the following sections.

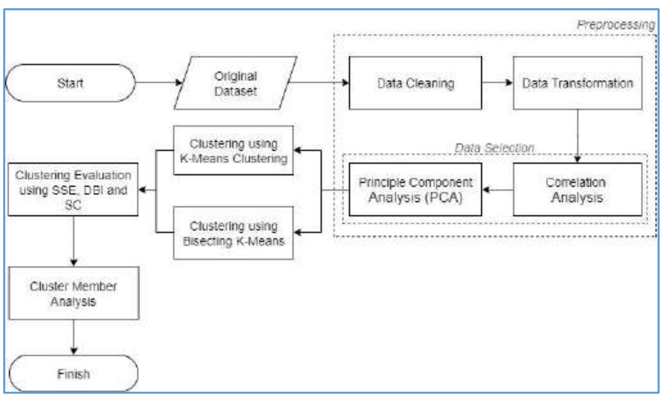


Fig. 1. Research stages.

2.1 Data Preprocessing

Data preprocessing carried out in this study consisted of data cleaning, data transformation, and data selection. Data cleaning process was performed to ensure no missing data. The results showed 21 missing values represented by "-" in the dataset. This prevented the clustering process from being performed due to the presence of string values. To address this issue, the data was replaced with NaN values. Then, imputation using average values were applied toward the NaN values of the following attributes: < SD, BSF, BSP, BBSP. The next step was checking for outliers by examining the descriptive statistics of the dataset and calculating outliers using the Interquartile Range (IQR). After the outliers were found, the Robust Scaler method was applied to the dataset to reduce the influence of outliers on the data scale. Next, data transformation was carried out. Data transformation was changing the data type to the appropriate format and simplifying the data [5]. The dataset contains object-type attributes, namely < SD, BSF, and BSP. Those attributes should be float data types. Therefore, a data transformation process was performed to change the data type from object to float.

The clustering process requires data with relevant variables for analysis. Therefore, a selection process was carried out. This study involved two stages of selection: manual and Principal Component Analysis (PCA). In the manual selection process, three attributes were removed because the values were already represented by other attributes, namely PL, TB1, and PMPTM. Thus, the manual selection results in 19 remaining attributes.

The next stage of data selection was examined the relationships between variables by implementing correlation analysis. Figure 2 shows the results of the correlation analysis.

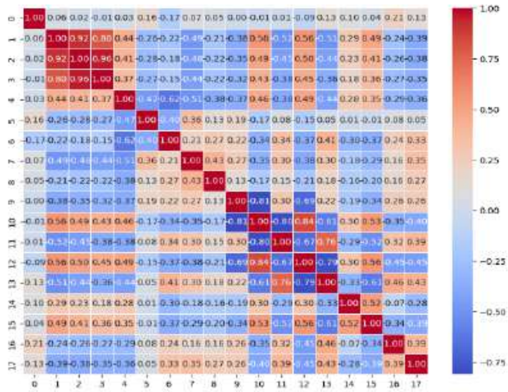


Fig. 2. Correlation analysis of numerical attributes.

The correlation analysis results indicate that several variables have a significant correlation, which can lead to suboptimal clustering results. The attributes with the highest correlation are the first index attribute and the second index attribute, with a correlation of 92%. Next, the second index attribute and the third index attribute have a correlation of 96%. To address this issue, the solution was not to remove one of the highly correlated variables. The problem was solved using the PCA method.

From 19 variables, the data was reduced to three main variables using PCA with a cumulative explained variance value of 0.8265. This value indicates that data reduction to three main components still retains 82.65% of the information or total variation in the original data. This process can help improve the quality and performance of the clustering model and can produce quite significant evaluation values. Figure 3 shows the results of the PCA process.

	PCA1	PCA2	PCA3
0	-0.541426	1.631035	0.213172
1	0.460395	1.830319	-0.038424
2	-0.855202	-0.758805	0.094231
3	1.067232	0.698512	-0.939763
4	1.194284	0.937345	0.384838
...	...	...	...
509	-27.836898	-2.741172	-1.086964
510	-12.963176	-1.632001	-1.808152
511	-23.208380	5.965640	4.692103
512	-27.675657	-10.163288	1.466124
513	-1.402541	-1.874007	2.990290

514 rows × 3 columns

Fig. 3. Result of principle component analysis.

2.2 K-means clustering

Clustering is a grouping of objects in each group or cluster that has the same characteristics between objects in a cluster but not with objects from other clusters [6]. K-means Clustering is one of the clustering methods that determines the number k and sets the initial centroid as the basis for cluster formation [7]. The following are the stages of K-means Clustering:

1) Determine the value of k.
2) Determine the initial centroid.
3) Calculate the centroid distance with each data using the euclidean distance formula as follows to find the minimum distance to the centroid:

$$d(x_i, \mu_j) = \sqrt{\sum (x_i - \mu_j)^2} \quad (1)$$

where :
x_i = i-th data
μ_j = center data of each cluster

4) Enter the data that has the smallest distance from the centroid.
5) Create a new centroid from the cluster average calculation with the following equation:

$$\mu_k = \frac{1}{N_k} \sum X_q \quad (2)$$

where :
μ_k = centroid point of the K cluster
N_k = the number of data in the K cluster
X_q = is the q-th data in the K cluster

6) Repeat steps 3-5 until the data in the cluster does not change.

The implementation of the clustering process with the K-means algorithm was carried out using the sklearn.cluster KMeans library in Python. Next, clustering experiments were conducted with k = 2 to k = 10.

2.3 Bisecting K-means clustering

Bisecting K-means clustering is a variation of the K-means clustering algorithm. The main concept is to divide one cluster into two sub-clusters at each iteration until the number of clusters reaches a predetermined k value [8]. The following are the steps of Bisecting K-means:

1) Set k to determine the number of clusters.
2) Assign one cluster to all data.
3) Randomly determine the initial centroid and use K-means clustering with k = 2 to divide the clusters.
4) Calculate the Sum of Squared Error (SSE) value of the two clusters using the following equation [8]:

$$\mathrm{SSE} = \sum_{j=1}^{K} \sum_{x_i \in C_k} \| x_i - \mu_j \|^2 \quad (3)$$

where:
K = number of clusters
C_k = the k-th cluster
μ_k = centroid for the cluster C_k
x_i = the i-th data point in the cluster
$\sum_{j=1}^{K}$ = the sum of the SSE values of all clusters
$\sum_{x_i \in C_k}$ = the sum of the squared errors of all data points in a particular cluster

$\| x_i - \mu_j \|^2$ = the squared error formula. The euclidean distance is squared

5) Select the cluster that has the higher SSE value and divide the cluster into 2 clusters using K-means clustering.
6) Repeat steps 3 - 5 until K clusters are formed.

The implementation of the clustering process with the Bisecting K-means algorithm was carried out using the sklearn.cluster Bisecting KMeans library in Python. Furthermore, experiments were also conducted with varying values of k = 2 to k = 10.

2.4 Clustering evaluation

Several metrics are commonly used to evaluate clustering results, including Sum Squared Error (SSE), Davies-Bouldin Index (DBI), and Silhouette Coefficient (SC). The SSE measures the sum of the squared distances of each data point from its cluster center. A smaller value indicates a more compact cluster, although the SSE tends to decrease as the number of clusters increases. The DBI compares the ratio between intra-cluster distances and inter-cluster distances; a lower DBI value indicates better clustering quality, as the clusters are denser and more clearly separated. Meanwhile, the SC evaluates how similar a point is to its own cluster compared to other clusters, with values ranging from -1 to 1; the closer to 1, the better the clustering quality. In this study, these three metrics were applied to evaluate the clustering results of the K-means and Bisecting K-means algorithms. Clustering experiments were conducted for each algorithm, varying from k = 2 to k = 10, and the results were then compared. A comparative analysis of the experimental results for both algorithms was then performed to draw conclusions.

2.5 Clustering evaluation

Analysis of the recommended cluster members was carried out by describing the characteristics of the cluster members based on the average value of each attribute.

3 Results and analysis

3.1 Evaluation of the number of clusters

By using 514 pre-processed data, the clustering process was implemented using module sklearn.cluster Bisecting KMeans and module sklearn.cluster K-means which are available in scikit-learn Python machine learning library. The experiment was carried out with k = 2 to k = 10. Evaluation of clustering results was carried out using 3 clustering evaluation metrics, namely SSE, DBI, and SC. Table 1 describes the comparison of SSE, DBI, and SC values of K-means and Bisecting K-means. Visualization of the relationship between k and each type of clustering evaluation metric is depicted in Figures 4, 5, and 6.

Based on Elbow method, Figure 4 shows that for both algorithm (K-means and Bisecting K-means), the elbow point is seen at k = 5 since the SSE (Sum of Squared Errors) decreases sharply from k = 2 to k = 5 and after that, the decline persists but tends to be more gradual. This means that with 5 clusters, the model is good enough to minimize the SSE, and adding clusters after that does not reduce the SSE significantly.

Figure 5 depicts the 3D visualization of clustering results using K-means and Bisecting K-means at k = 5 as the results of the Elbow method above. There is one very large cluster (289 points) and several small clusters (9, 26, and 29 points). This means the clustering tends to be unbalanced, with one cluster dominating. This could indicate that K-means is not optimal for separating small groups. Means, on the other hand, bisecting K-means (k = 5) produces a more even cluster distribution, making it more representative.

Table 1. Comparison of SSE, DBI, and SC values of K-means and Bisecting K-means.

K	SSE		DBI		SC	
	K-means	Bisecting K-means	K-means	Bisecting K-means	K-means	Bisecting K-means
2	5519.9221	5519.9221	0.6639	0.6639	0.7973	0.7973
3	4143.836	4358.3972	0.8559	1.0929	0.4402	0.5661
4	3337.8056	3549.4262	0.9534	1.1345	0.444	0.3201
5	2547.5861	2917.5976	0.8586	1.0934	0.3424	0.3641
6	2129.3691	2466.0845	0.82	1.0445	0.3625	0.3774
7	1858.8423	2055.7257	0.8592	0.9266	0.3655	0.3623
8	1602.1058	1758.7802	0.8608	0.9357	0.3585	0.3609
9	1452.2105	1635.3366	0.9037	1.0201	0.2694	0.3134
10	1332.8769	1487.1604	0.9018	1.0003	0.2683	0.3118

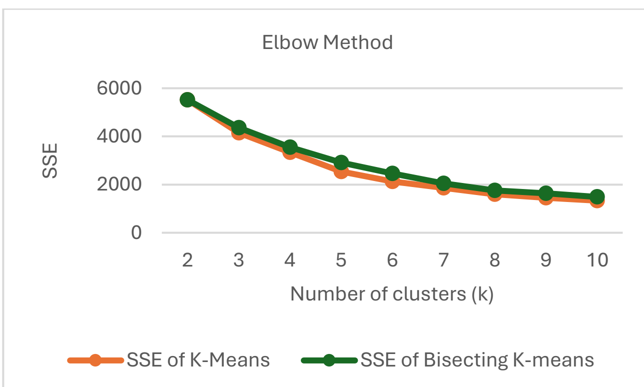


Fig. 4. The relationship between the k value and SSE.

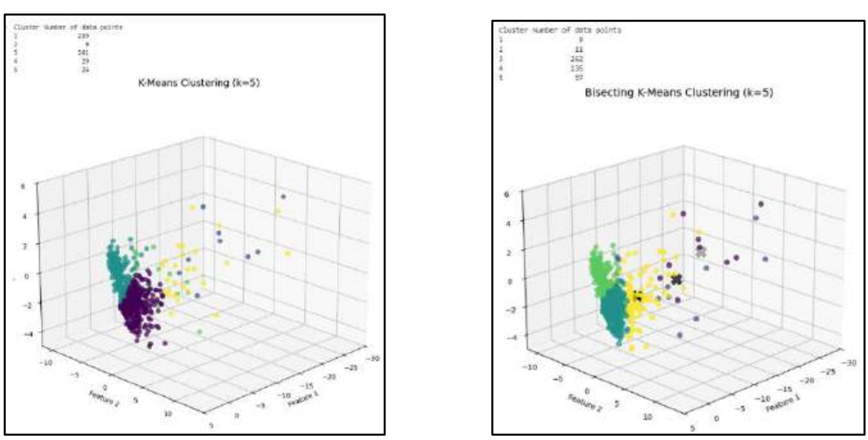


Fig. 5. 3D visualization of clustering results using K-means (k = 5) and Bisecting K-means (k = 5).

Using Davies-Bouldin Index, a lower DBI value is better because it indicates more separate and compact clusters. As can be seen in Figure 6, for K-means, the lowest DBI value is 0.6639 (k = 2), but k = 2 is too simple since it is only divided the data in two large clusters. After that, the DBI value is relatively stable and the next best is around k = 6 (0.82), which is smaller than the values around other k values. For Bisecting K-means, the lowest DBI value is around k = 7 (0.9266).

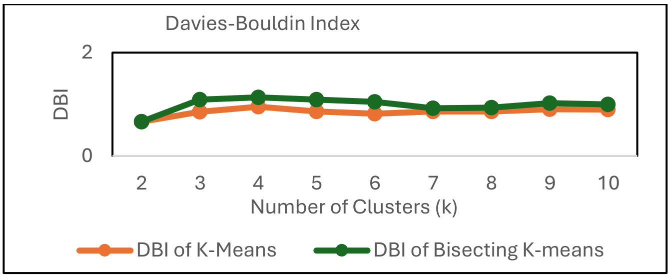


Fig. 6. The relationship between the k value and DBI.

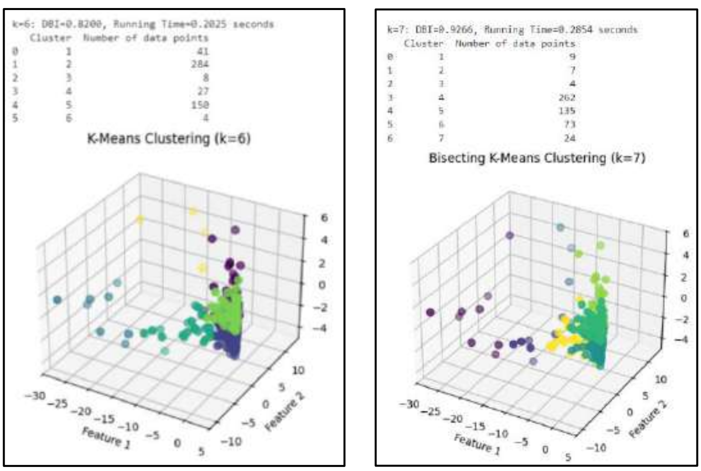


Fig. 7. 3D visualization of clustering results using K-means (k = 6) and Bisecting K-means (k = 7).

Figure 7 depicts a 3D visualization of the results of K-means clustering at k = 6 and Bisecting K-means at k = 7, according to the best results using the DBI metric. In K-means (k = 6), the clusters appear quite clearly separated. However, there is an imbalance in size (some clusters are very small). The DBI value is lower than Bisecting K-means (0.82), indicating better inter-cluster separation than Bisecting.

In Bisecting K-means (k = 7), the clusters appear more "broken" into smaller parts. Many small clusters result in a more fragmented distribution. Compared to K-means, the DBI value is higher (0.93), indicating poorer cluster quality (less compact/less separated clusters). In other words, K-means produces better clustering (lower DBI) than bisecting K-means in this dataset.

Using Silhouette Coefficient as evaluation metric, a higher SC value is better because it indicates that objects are in the appropriate clusters (clear separation between clusters). As shown in Table 2 and Figure 8, the highest Silhouette Coefficient value of K-means and Bisecting K-means are the same, namely 0.7973, which is obtained at k = 2. According to Kaufmann & Rousseeauw, an SC value of 0.7973 indicates that the cluster formed has a strong structure [9]. However, in this dataset, clustering in two clusters is too simplistic.

After k = 2, the SC value drops sharply, but stabilizes around k = 6–7 (0.36–0.37), indicating that the cluster structure is starting to balance. For Bisecting K-means, a similar pattern occurs. The highest SC of Bisecting K-means is at k = 6 (0.3774).

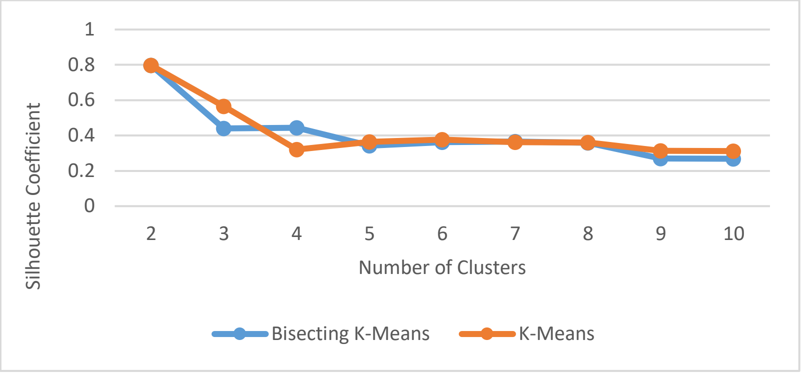
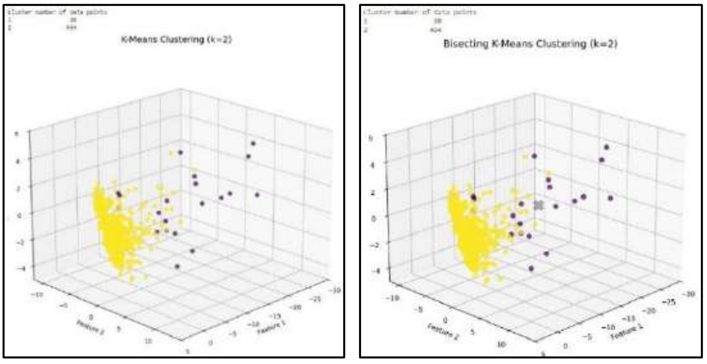


Fig. 8. Comparison of Silhouette Coefficient of K-means Clustering and Bisecting K-means Clustering.

Figure 9 is a 3D visualization of K-means and Bisecting K-means clustering, both are results from the k = 2. The figure also shows the SC values for each cluster resulting from each clustering method. It is clear from both figures that the first cluster, whose points are colored yellow, appears to be clustered in close proximity, while the second cluster, containing points colored other than yellow, tends to be scattered. This indicates that the first cluster is very similar to each other, while the second cluster is actually not very similar. The SC values for both clusters, both from K-means and from Bisecting K-means, also reflect the same thing. The first cluster has a strong structure, while the second cluster has a weak structure, or even almost no structure.



(a) (b)

Fig. 9. 3D visualization of K-means Clustering results (k = 2) and Bisecting K-means (k = 2).

Based on the analysis of the tables and graphs above, several general conclusions are drawn. K-means tends to produce more compact and separate clusters than bisecting K-means, as evidenced by its better (lower) SSE and DBI values. Bisecting K-means' performance is not significantly different, but tends to be more stable at medium cluster sizes (k = 6–8), as evidenced by its slightly higher SC.

Although k = 2 yields the highest SC and lowest DBI values for both the K-means and Bisecting K-means methods, the number of clusters = 2 is considered to oversimplify regional grouping. While it may be useful as a general clustering technique, it is less able to capture more detailed clusters.

Based on the balance of the three metrics (low SSE, low DBI, and moderate SC), the optimal k value is around k = 6. At this point, K-means has a low SSE (2129.37), a relatively low DBI (0.82), and a moderate SC (0.3625).

3.2 Cluster member analysis

A more in-depth analysis was conducted based on the description of cluster members at k = 6. The average value data for each attribute at k = 6 was used as the basis for the description of cluster members. Table 2 shows the number of members of each cluster. Table 2 elaborates the average value of each attribute at k = 6. In this paper, only clusters formed from the K-means method are analyzed further. Table 4 describes the characteristic of each cluster based on the data from Table 3. The 6 clusters that were formed have diverse characteristics, ranging from clusters containing very poor residents to clusters containing relatively prosperous residents.

Table 2. Number of cluster members.

Cluster	Number of Cluster Member	
	K-means	Bisecting K-means
1	41	9
2	284	11
3	8	262
4	27	135
5	150	73
6	4	24

Table 3. Average value of each atribute.

ATTRIBUTE / CLUSTER	JPM	PPM	P1	P2	<SD	TMT SD/SMP	>SMA	7-12	13-15
1	43.7	24.6	5.5	1.7	25.0	50.8	24.1	97.1	93.2
2	54.3	10.9	1.6	0.4	21.0	55.4	23.7	99.4	91.7
3	43.1	34.2	6.8	2.1	59.0	30.3	10.7	59.1	52.4
4	20.3	12.4	2.1	0.6	28.0	50.7	21.3	86.9	88.2
5	50.6	7.0	1.0	0.2	13.4	52.1	34.7	99.4	95.2
6	48.2	35.7	12.1	5.8	50.5	34.8	14.8	81.6	73.7

ATTRIBUTE / CLUSTER	TB	BSI	BSF	BSP	BBSP	PM	PTM	AL	JS
1	38.2	52.2	9.6	47.2	14.7	63.5	54.6	73.8	74.2
2	41.5	44.5	14.1	37.0	21.5	63.1	54.0	79.6	81.6
3	24.1	74.4	1.4	74.4	1.4	68.7	64.9	68.9	41.6
4	42.9	46.7	10.4	40.0	17.1	62.2	53.2	75.6	74.3
5	47.4	25.1	27.5	14.6	41.9	58.5	46.6	92.5	93.9
6	17.7	81.4	0.8	81.4	7.4	68.1	64.4	50.4	59.6

Table 4. Characteristics of each cluster.

Cluster	Description
1	Poverty is moderate, but the community is relatively well-off in terms of education and basic access.
2	Large number of poor people but low poverty rate; good education and access.
3	Severe poverty with low levels of education and infrastructure, dominated by the agricultural sector.
4	Moderate poverty, education and infrastructure.
5	The most prosperous region, lowest poverty, best education and access.
6	The poorest and most underdeveloped region, low level of education, predominantly agricultural, minimal infrastructure.

4 Conclusion

K-means tends to produce more compact and separate clusters than Bisecting K-means, as evidenced by its better (lower) SSE and DBI values. Bisecting K-means' performance is not significantly different, but tends to be more stable at medium cluster sizes (k = 6–8), as evidenced by its slightly higher SC.

Based on the evaluation results of three metrics (SSE, DBI, and SC), the K-means method with k = 6 is the most optimal clustering model for poverty data with a low SSE (2129.37), a relatively low DBI (0.82), and a moderate SC (0.3625). The results of this clustering are able to map areas from the most prosperous to the poorest, so that they can become the basis for formulating more targeted development and poverty alleviation policies.

Future works need to be done are analyzing the member of each cluster by using heatmap or boxplot for each attribute, and visualizing the clustering results in the form of a map such that relevant parties might be easier to understand the picture and pattern of the distribution of poverty levels in various regions of Indonesia. Comparing clustering results from year to year is also interesting to do to see progress in each region.

References

1. Badan Pusat Statistik, Profil Kemiskinan di Indonesia, (2025). https://www.bps.go.id/id/pressrelease/2025/07/25/2518/persentase-penduduk-miskin-maret-2025-turun-menjadi-8-47-persen-.html?utm_source=chatgpt.com
2. N.T. Luchia, H. Handayani, F.S. Hamdi, D. Erlangga, S.F. Octavia, Comparison of K-means and K-medoids on poor data clustering in Indonesia. MAILCOM: Indonesian Journal of Machine Learning and Computer Science **2**, 35-41 (2022). https://doi.org/10.57152/malcom.v2i2.422
3. T.A. Munandar, Penerapan algoritma clustering untuk pengelompokan tingkat kemiskinan provinsi Banten. JSiI (Jurnal Sistem Informasi) **9**, 109-114 (2022). https://doi.org/10.30656/jsii.v9i2.5099
4. M.A. Rahman, N.S. Sani, R. Hamdan, Z.A. Othman, A.A. Bakar, A clustering approach to identify multidimensional poverty indicators for the bottom 40 percent group. Plos One, (2021). https://doi.org/10.1371/journal.pone.0255312
5. S.E. Wardani, S.Z. Harahap, R. Muti'ah, Implementation of the K-means method for clustering regency/city in North Sumatra based on poverty indicators. Jurnal dan Penelitian Teknik Informatika **8**, 1429-1442 (2024). https://doi.org/10.33395/sinkron.v8i3.13720
6. J. Han, M. Kamber, J. Pei, Data mining concepts and techniques, 3rd edition (Morgan Kaufmann, San Francisco, 2011)

7. I.M.A.W. Putra, G. Indrawan, K.Y.E. Aryanto, Recommendation system based on clustering of Tabanan regional library using K-means clustering. Indonesian Journal of Computer Science (JIK) **3**, 18-22 (2018).
8. A. Halim, H. Gohzali, M.D. Panjaitan, I. Maulana, Movie recommendation by using bisecting k-means and collaborative filtering. CITISEE 2017, 37-41 (2017).
9. L. Kaufman, P.J. Rousseeuw, Finding groups in data: an introduction to cluster analysis (John Wiley and Sons, New Jersey, 2005)