Bulletin of Applied Mathematics and Mathematics Education
Vol. 6, No. 1, April 2026, pp. 27-36
e-ISSN: 2776-1029 p-ISSN: 2776-1002 DOI: 10.12928/bamme.v6i1.15817

APPLIED MATHEMATICS

Clustering Indonesian provinces based on welfare level using several validity indices

Yudi Setyawan [1,*], Maria Kristina Yolanda Hawa [1], Kris Suryowati [1]

[1] Universitas AKPRIND Indonesia, Jl. Kalisahak 28, Komplek Balapan, Yogyakarta 55222, Indonesia
*Corresponding author

Abstract

One of the national development goals is to increase the level of community welfare. There are several aspects that influence the level of welfare, namely population, health, education, housing, social, employment, consumption, and poverty. This research aims to group provinces in Indonesia based on their level of welfare so that the government can determine appropriate policies in the context of economic recovery and improving the welfare of the Indonesian people. The data used are indicators of provincial welfare levels in Indonesia in 2022 from the Central Statistics Agency. Data is grouped into 3 clusters based on welfare level, namely high (C1), medium (C2), and low (C3) using the K-Means and Fuzzy C-Means methods. Based on the results of the validity test, it is known that the best method is the K-Means method with Euclidean distance using the parameter k = 3, the resulting DBI value is 0.989 and the C-Index is 0.076, where this value is better than those of the Fuzzy C-Means method. It is hoped that the results can provide information regarding the characteristics of provinces in Indonesia based on welfare level indicators and become a reference for the government in improving welfare in Indonesia.

Keywords: clustering, Fuzzy C-Means, K-Means, level of welfare, validity

How to cite: Setyawan, Y., Hawa, M. K. Y., & Suryowati, K. (2026). Clustering Indonesian provinces based on welfare level using several validity indices. *Bulletin of Applied Mathematics and Mathematics Education, 6*(1), 27-36. https://doi.org/10.12928/bamme.v6i1.15817

Article history: Received 10/02/2025, Accepted 20/05/2026, Published 19/06/2026

Correspondence address: Universitas AKPRIND Indonesia, Jl. Kalisahak 28, Komplek Balapan, Yogyakarta 55222, Indonesia. E-mail: setyawan@akprind.ac.id

INTRODUCTION

The primary objective of national development is to improve the welfare of its people. There are several aspects related to the level of community welfare, such as population, health, education, employment, consumption patterns, housing, poverty, and other social aspects (Nugraha et al., 2021). As an archipelagic country divided into many provinces, there is a noticeable disparity concerning the welfare levels of the provinces in Indonesia. This study aims to classify the provinces of Indonesia based on various related aspects using several variables, namely population density, life expectancy, average years of schooling, open unemployment rate, percentage of per capita expenditure on food, percentage of home ownership, percentage of poor population, and percentage of mobile phone ownership.

In non-hierarchical clustering, the number of clusters is determined first based on several considerations, both theoretical using certain methods and practical considerations according to the established goals. There are many methods that can be used in non-hierarchical clustering such as K-means, Fuzzy C-means clustering, and Mixture Modeling (Indah & Octaviana, 2025; Suraya & Wijayanto, 2022; Kembaren et al., 2022). The K-Means algorithm is a non-hierarchical algorithm that starts by dividing the data into several groups, where data with

almost similar characteristics are grouped into the same cluster, and data with different characteristics are grouped into different clusters. As a measure of similarity between data, distance concepts such as Euclidean, Manhattan, and Minkowski distances are used (Nishom, 2019). The aim of the K-Means method is to minimize the objective function generated during the clustering process (Dahnial, 2023). Meanwhile, Fuzzy C-Means is a data clustering method where the presence of each data point in a cluster is determined based on its degree of membership (Mustakim et al., 2023). The comparison of clustering results can be done using several types of validity indices such as the Davies-Bouldin Index, Dunn Index, and C-Index (Ikotun et al., 2025). The Davies-Bouldin Index (DBI) is a method used to evaluate clusters in general based on the quantity and relationship among cluster members. The smaller the DBI value, the better the resulting clusters (Alfarera, 2024). The Dunn Index (DI) is the ratio of the smallest distance between objects in different clusters to the largest distance between objects in the same cluster (Sary et al., 2024). A larger DI value indicates that the number of formed clusters is more optimal. This index calculates the validity of a grouping using minimum cluster distance (separation) and maximum cluster size (cohesion). The C-Index, discovered by Hubert and Levin in 1976, is one of the validity indices for clustering results with internal criteria. In the C-Index, clusters are optimal when its value is minimum, approaching zero (Saidah et al., 2022).

The final step in clustering is profiling, which involves finding the important characteristics of each group/cluster based on the studied variables to explain how these objects significantly differ in each of their dimensions. Profiling is the process of interpreting a cluster using the average or centroid of the initial data of all objects, so that the characteristics of each object can be known for comparing or distinguishing objects from different clusters. Based on the explanation above, in this study, clustering of provinces in Indonesia is carried out based on aspects that influence the welfare level of its population, which include population density, life expectancy, average length of schooling, open unemployment rate, percentage of per capita monthly expenditure on food, percentage of households according to home ownership status, percentage of poor population, and the percentage of mobile phone ownership per individual using the K-Means and Fuzzy C-Means methods. The data used is secondary data sourced from the Central Statistics Agency, both at the national and provincial levels, which is published and can be accessed openly through their respective official websites. This research is expected to find the best method for clustering the provinces in Indonesia based on their welfare levels and provide an overview of the clustering results so that the government can make appropriate policies for economic recovery post-COVID-19 for regions or provinces based on the best clustering results. For comparison, studies with similar methods or subjects have been conducted by (Febrianto & Palasara, 2019; Fitriani et al., 2021; Garini et al., 2022).

RESEARCH METHOD

In this study, secondary data related to welfare level obtained from the Central Bureau of Statistics (BPS) of Indonesia provinces in 2022 was used. The following are the stages of data analysis.

First, we collected the secondary data from the Central Bureau of Statistics. Then, we conducted descriptive analysis to elaborate the data. We standardized the data values using the z-score formula as follows.

$$z_i = \frac{x_i - \mu}{\sigma} \quad (1)$$

where x_i is the i^{th} value of a variable, μ and σ are the mean the standard deviation of the data respectively.

Furthermore, we drew assumption testing on cluster analysis, namely multicollinearity test, and performed clustering using the K-Means method with the following steps: (1) Determine the optimum K value using the Silhouette method, (2) randomly determine the cluster center points, (3) randomly determine the cluster center points, and (4) calculate the distance using Euclidean and Manhattan distances as follows.

The following is the formula for calculating Euclidean distance.

$$d_{euc}(x_i, x_j) = \sqrt{\sum_{k=1}^{p} (x_{ik} - x_{jk})^2} \quad (2)$$

while the formula for calculating Manhattan distance is

$$d_{man}(x_i, x_j) = \sum_{k=1}^{p} \left| x_{ik} - x_{jk} \right| \quad (3)$$

where,

$d(x_i, x_j)$: the distance between the i^{th} and the j^{th} objects

x_{ik} : the i^{th} object data in the k^{th} variable

x_{jk} : the j^{th} object data in the k^{th} variable

Then, we created the cluster plot, and conducted validity tests for the Davies-Bouldin Index (DBI), Dunn Index, and C-Index. The following is the calculation of the DBI validity index using the formula (Prihandoko et al., 2024).

$$DBI = \frac{1}{K} \sum_{i=1}^{K} \max_{j \neq i} \frac{s_i + s_j}{d_{i,j}} \quad (4)$$

where,

K : number of clusters

s_i : average distance between each point in cluster i and its centroid

$d_{i,j}$: distance between the centroids of cluster *i* and cluster *j*.

The Dunn Index formula is as follows (Malikhatin et al. 2021).

$$\text{Dunn Index} = \frac{\min\limits_{1 \le i < j \le K} d(C_i, C_j)}{\max\limits_{1 \le l \le K} (diam(C_l))} \quad (5)$$

where,

$min\, d(C_i, C_j)$: minimum distance between points in C_i and C_j clusters

$max(diam(C_k))$: maximum *clusters* diameter.

The formula of *C-Index* is given by (Cunningham, 2008):

$$C\text{-}Index = \frac{S_w - S_{min}}{S_{max} - S_{min}}, where\ S_{min} \neq S_{max}\ \ and\ C\text{-}Index\ \epsilon(0,1) \quad (6)$$

where,

S_w : the average distance of objects in the same cluster.

S_{min} : the smallest distance between all pairs of data objects within a cluster.

S_{max} : the largest distance between all pairs of data objects within a cluster.

Perform clustering using the Fuzzy C-Means method with the following steps: (1) determine the number of clusters, weights, maximum iterations, minimum error, initial objective function, and initial iterations; (2) Determine the initial random matrix; (3) Determine the degree of membership; (4) Determine the cluster centers using the following formula

$$V_{kj} = \frac{\sum_{i=1}^{n}((U_{ik})^w \times X_{ij})}{\sum_{i=1}^{n}(U_{ik})^w} \quad (7)$$

where,

V_{kj} : Center of cluster k and attribute j

U_{ik} : The membership degree of object i in cluster k

w : weight

Then, determine the objective function using Euclidean distance (2) and Manhattan (3), update the partition matrix, check the stop condition: if $t > maximum\ iteration$ then the calculation stops. Otherwise, $t = t + 1$, then a recalculation is done from step c; create a plot of the formed clusters; conduct validity tests for the DBI, $Dunn\ Index$, and $C - Index$; compare the best results by observing the best validity values, perform cluster profiling, and draw conclusions.

RESULTS AND DISCUSSION

This section discusses an overview of indicators of welfare levels and grouping using K-Means and Fuzzy C-Means methods, then conducts a comparison to determine the best method based on validity tests, determines the results of clustering, and profiling based on the best method.

Table 1. Descriptive Analysis

No	Variable	N	Minimum	Maximum	*Mean*	Standard Deviation
1	Population Density (People/km^2)	34	10.00	16084	750.7647	2739.173
2	Life Expectancy (Years)	34	65.63	75.08	70.42	2.451699
3	Average Length of Schooling (Years)	34	7.02	11.31	8.839118	0.92299
4	Open Unemployment Rate (%)	34	2.725	8.33	5.044559	1.536644
5	Percentage of Per Capita Expenditure per Month for Food (%)	34	37.75	54.85	46.76441	3.584195
6	Percentage of Households by Home Ownership Status (%)	34	56.13	92.51	83.18765	7.947134
7	Percentage of Poor Population (%)	34	4.53	26.68	10.27118	5.266617
8	Percentage of Individuals Who Own/ Have a Mobile Phone (%)	34	35.33	82.37	67.80529	8.464215

Overview of welfare in each province in Indonesia

Here is an overview of the welfare in each province in Indonesia in 2022. It can be seen from Table 1 which explains the descriptive statistics of 8 variables used to illustrate the Welfare in 34 Provinces in Indonesia. Where for the population density variable, it can be observed that out of the 34 provinces in Indonesia, the area with the highest population density reaches 16.084 people/km2, and the lowest population density is 10 people/km2, with an average population density of 750.7647 people/km2 and a standard deviation of 2,739.173 people/km2.

Clustering using the K-Means clustering method

The results of clustering using the K-Means clustering method and using two distance measurements, namely Euclidean distance and Manhattan distance are as follows.

K-Means Clustering using Euclidean Distance

The results of determining the optimal cluster using the silhouette width method can be seen in the following graph.

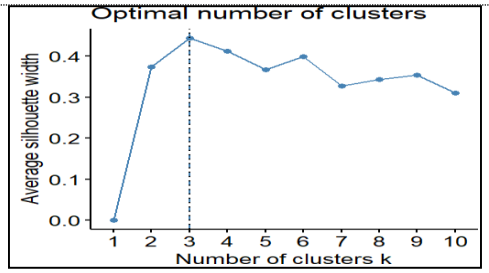

Figure 1. Silhouette Width using Euclidean and Manhattan distance

From Figure 1 it can be seen that the highest Silhouette value is located at the second point, which means the optimal number of clusters is 3 clusters. Then, the cluster members formed using the value k = 3 are obtained as shown in Table 2 below.

Table 2. K-Means Clustering Results with Euclidean Distance

Cluster	1	2	3
Number of provinces	3	25	6

It can be seen in Table 2 that cluster 1 consists of 3 provinces, cluster 2 consists of 25 provinces, and cluster 3 consists of 6 provinces. After the grouping of each object is identified, a plot of the K-Means clustering results is created. This plot presents the dispersion pattern of each member contained in each cluster as shown in Figure 2.

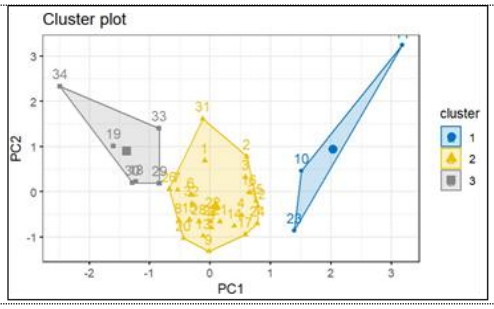

Figure 2. K-Means Clustering Plot with Euclidean Distance

e-ISSN: 2722-2691|p-ISSN: 2722-2683

As seen in Figure 2, it is known that the distribution pattern produces 3 clusters, consisting of cluster 1 in blue, cluster 2 in yellow, and cluster 3 in gray.

K-Means Clustering using Manhattan distance

The results of determining the optimal cluster using the silhouette width method show that the optimal k = 3. Then the members of the formed clusters with k = 3 are obtained as shown in the following Table 3.

Table 3. K-Means Clustering result using Manhattan distance

Cluster	1	2	3
Number of provinces	22	10	2

It can be seen in Table 3 that cluster 1 consists of 22 provinces, cluster 2 consists of 10 provinces, and cluster 3 consists of 2 provinces.

After the grouping of each object is identified, a plot of the K-Means clustering results is created. The plot presents the distribution pattern of each member found in each cluster as shown in Figure 3.

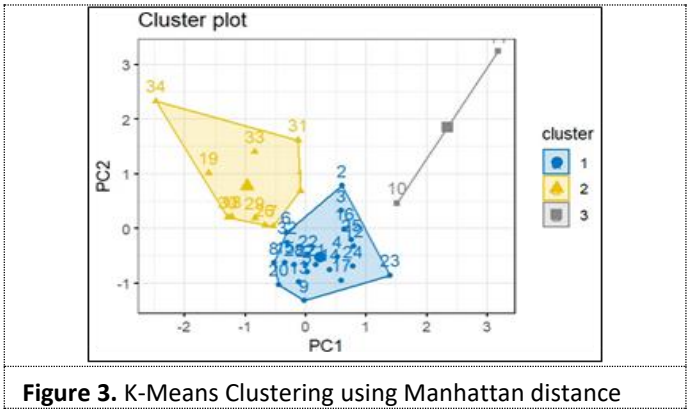

Figure 3. K-Means Clustering using Manhattan distance

As seen in Figure 3, it is known that the distribution pattern produces 3 clusters, consisting of cluster 1 in blue, cluster 2 in yellow, and cluster 3 in gray.

Clustering using Fuzzy C-Means clustering method

The results of clustering using the Fuzzy C-Means Clustering method and employing two distance measurements, namely Euclidean distance and Manhattan distance are as follows.

Clustering with the Fuzzy C-Means method using Euclidean Distance

For the Fuzzy C-Means method with Euclidean Distance, the optimal number of clusters based on the silhouette width method can be seen in the following figure.

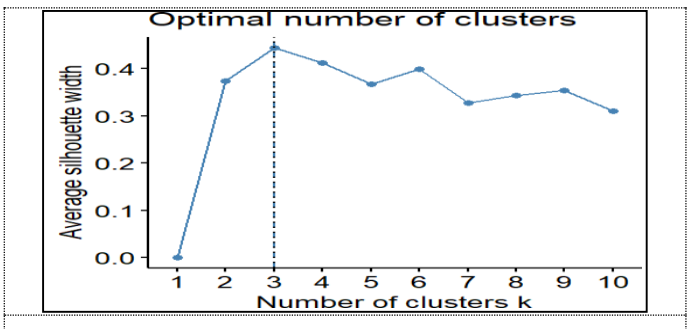

Figure 3. K-Means Clustering using Manhattan distance

From Figure 4 it can be seen that the highest Silhouette value is located at the second point, which means the optimal number of clusters is 3 clusters. Then, the cluster members formed using the value k = 3 are obtained as shown in Table 4 below.

Table 4. Result of Fuzzy C-Means clustering using Euclidean distance

Cluster	1	2	3
Number of provinces	10	12	12

It can be seen in Table 4 that cluster 1 consists of 10 provinces, cluster 2 consists of 12 provinces, and cluster 3 consists of 12 provinces.

After the grouping for each object has been identified, a plot of the results of the Fuzzy C-Means clustering is then conducted. The plot presents the distribution pattern of each member found in each cluster as shown in Figure 5.

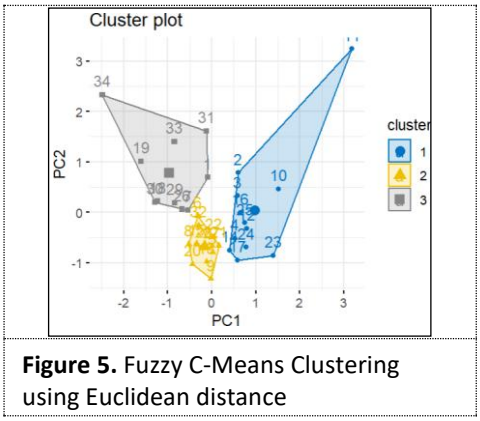

Figure 5. Fuzzy C-Means Clustering using Euclidean distance

As seen from Figure 5, it is observed that the distribution pattern produced three clusters, consisting of cluster 1 in blue, cluster 2 in yellow, and cluster 3 in gray.

Clustering with Fuzzy C-Means method using Manhattan distance

For the Fuzzy C-Means method with Manhattan Distance, the optimal number of clusters using the silhouette width method is k = 3. Then, the members of the formed clusters using k = 3 are shown in the following Table 5.

Table 5. Result of Fuzzy C-Means clustering using Manhattan distance

Cluster	1	2	3
Number of provinces	12	11	11

It can be observed in Table 5 that cluster 1 consists of 12 provinces, cluster 2 consists of 11 provinces, and cluster 3 consists of 11 provinces. After the grouping of each object is determined, the results of the Fuzzy C-Means clustering are plotted. The plot presents the distribution patterns of each member within the respective clusters as shown in Figure 6.

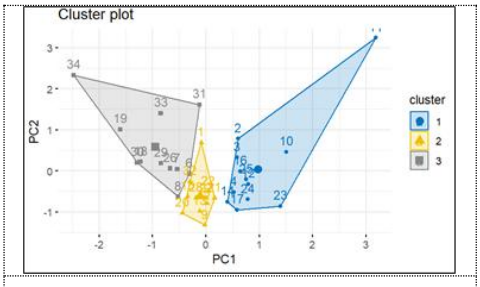

Figure 6. Fuzzy C-Means Clustering using Manhattan distance

e-ISSN: 2722-2691|p-ISSN: 2722-2683

As seen in Figure 6, it is known that the distribution pattern produces 3 clusters, consisting of cluster 1 in blue, cluster 2 in yellow, and cluster 3 in gray.

Comparison the clustering results using K-Means and Fuzzy C-Means

After obtaining the results of the province clustering in Indonesia using the K-Means and Fuzzy C-Means methods, a comparison of those results is then performed with the validity tests of the Davies-Bouldin Index, Dunn Index, and C-Index. The results are as follow.

Table 6. Results of Validity Test

No	Method	Distance	Parameter	Number of Clusters	Result of Validity Test		
					DBI	*Dunn Index*	*C-Index*
1	*K-Means*	*Euclidean*	*k* = 3	3 *clusters*	0.989	0.062	0.076
	K-Means	*Manhattan*	*k* = 3	3 *clusters*	1.134	0.086	0.104
2	*Fuzzy C-Means*	*Euclidean*	*k* = 3	3 *clusters*	1.340	0.065	0.163
	Fuzzy C-Means	*Manhattan*	*k* = 3	3 *clusters*	1.599	0.022	0.199

Based on Table 6, it can be observed that the K-Means method outperforms the Fuzzy C-Means method, as indicated by the K-Means method using Euclidean distance yielding the smallest DBI and C-Index values of 0.989 and 0.076, respectively. Therefore, it can be concluded that the best method in this study is the K-Means method using Euclidean distance with the parameter k = 3.

Cluster profile for the best method

Based on the best clustering results using the K-Means method with Manhattan distance and parameter k = 3. Then, categories were created for the value of each variable, namely low and high, as shown in Table 7.

Table 7. Cluster Profile

No	Variable	*Cluster* 1	*Cluster* 2	*Cluster* 3
1	Population Density (People/km^2)	5460	341	104
2	Life Expectancy (Years)	72.8	71.0	66.9
3	Average Length of Schooling (Years)	10.5	8.91	7.71
4	Open Unemployment Rate (%)	7.32	5.13	3.54
5	Percentage of Monthly Per Capita Expenditure for Food (%)	42.1	47.4	46.5
6	Percentage of Households by Home Ownership Status (%)	66.5	84.1	87.6
7	Percentage of Poor Population (%)	5.72	8.91	18.2
8	Percentage of Individuals who Own/Have a Mobile Phone (%)	81.5	68.6	57.7

Based on the average of each variable from each cluster compared to the average of each variable from all clusters, it can be determined that cluster 1 has high welfare characteristics in each province in Indonesia. Cluster 2 has moderate welfare characteristics in each province in Indonesia. Meanwhile, cluster 3 has low welfare characteristics in each province in Indonesia.

CONCLUSION

This study demonstrates that eight welfare indicators across Indonesian provinces exhibit substantial variation, as reflected in the five-number summary, with population density ranging from 10 to 16,084 people per km². Cluster analysis using K-Means and Fuzzy C-Means with both Euclidean and Manhattan distances consistently produced three clusters, though with differing provincial compositions. Evaluation of clustering validity revealed that K-Means with Euclidean distance yielded the most robust results, indicated by the lowest Davies-Bouldin Index (0.989)

and C-Index (0.076), confirming it as the optimal method with k=3. Cluster profiling further identified that provinces in Cluster 1 represent relatively high welfare, Cluster 2 moderate welfare, and Cluster 3 low welfare. These findings imply that policy interventions should prioritize provinces in Cluster 3 through improved access to education, healthcare, and basic infrastructure, while provinces in Cluster 2 require strategies to strengthen local economic capacity. Meanwhile, provinces in Cluster 1 may serve as benchmarks for best practices in welfare-oriented.

ACKNOWLEDGEMENT

Authors express gratitude to the AKPRIND University for providing full support and access during this research and the peer reviewers for proofreading and giving constructive feedbacks.

DECLARATION

Author contribution

All authors contribute in the research and/or writing the paper, and approved the final manuscript.

Yudi Setyawan	Conceptualization, Methodology, and leading the Investigation.
Maria Kristina Yolanda Hawa	Assisting the investigation, analyzing the data, and writing the original draft.
Kris Suryowati	Assisting the investigation, reviewing the paper, enriching the data analysis, and translating the paper into English.

Funding

This research did not receive any funding.

Conflict of interest

All authors declare that they have no competing interests.

Ethics declaration

We as authors acknowledge that this work has been written based on ethical research that conforms with the regulations of our institutions and that we have obtained the permission from the relevant institutes when collecting data. We support the Bulletin of Applied Mathematics and Mathematics Education (BAMME) in maintaining the high standards of personal conduct, practicing honesty in all our professional practices and endeavors.

The use of artificial intelligence

We do not use any generative AI tools to write any part of this paper.

Additional information

Not available.

REFERENCES

Alfarera, M. A. (2024). Analysis of Malang University student achievement grouping using the K-Means clustering method. *Journal of Electrical Engineering and Computer Sciences*, *9*(2), 159-172.

Cunningham, P. (2008). Unsupervised learning and clustering. In D. Greene, P. Cunningham, & R. Mayer, *Machine Learning Techniques for Multimedia* (pp. 51-90). Springer.

Dahnial. (2023). Implementation of K-Means clustering method to lecturers based on publications af national journals and accredited sinta. *Journal of Electrical Engineering and Computer Sciences*, *8*(1), 27-40.

Febrianto, N. I., & Palasara, N. (2019). Analisis clustering K-Means pada data informasi kemiskinan di Jawa Barat Tahun 2018. *Jurnal Sisfokom*, 8(2), 130-140.

Fitriani, D., Padilah, T. N., & Sari, B. N. (2021). Penerapan algoritma K-Means dalam pengelompokan kesejahteraan rakyat berdasarkan kecamatan di Kabupaten Karawang. *Progresif: Jurnal Ilmu Komputer*, 17(2).

Garini, F. C., Anbiya, W., & Purwandari, P. (2022). Optimalisasi pengelompokan provinsi di indonesia berdasarkan indikator kesejahteraan rakyat. *Seminar Nasional Statistika Aktuaria I* (pp. 1-10). Departemen Statistika FMIPA Universitas Padjadjaran.

Ikotun, A. M., Habyarimana, F., & Ezugwu, A. E. (2025). Benchmarking validity indices for evolutionary K-means clustering performance. *Nature Portfolio*.

Indah, Y. M., & Octaviana, S. (2025). Analisis perbandingan metode klaster hierarki dan non-hierarki terhadap tingkat pengangguran di pulau Jawa Tahun 2023. *BIAStatistics: Journal Of Statistics Theory and Aplications*, *19*(1), 92-104.

Kembaren, R. C., Sitompul, O. S., & Sawaluddin. (2022). Analysis clustering using normalized cross correlation in Fuzzy C-Means clustering algorithm. *Sinkron :Jurnal dan Penelitian Teknik Informatika, 6*(4), 2262-2271.

Malikhatin, H., Rusgiyono, A., & Maruddani, D. A. (2021). Penerapan K-Modes clustering dengan validasi dunn index pada pengelompokan karakteristik calon TKI Menggunakan R-GUI. *Jurnal Gaussian*, *10*(3), 359-366.

Mustakim, Aini, D. N., Batubara, A. U., Erkamim, M., & Legito. (2023). Fuzzy clustering-based grouping for mapping the distribution of student success data. *MALCOM: Indonesian Journal of Machine Learning and Computer Science, 3*(2), 366-372.

Nishom, M. (2019). Perbandingan akurasi euclidean distance, minkowski distance, dan manhattan distance pada algoritma K-Means clustering berbasis Chi-Square. *Jurnal Informatika: Jurnal Pengembangan IT (JPIT)*, *4*(1), 20-24.

Nugraha, A., Asnawi, M., & Purwandari, T. (2021). Analisis klaster hirarki untuk mengelompokkan provinsi di indonesia berdasarkan indikator kesejahteraan rakyat. *Seminar Nasional Statistika X*. Dept. Statistika FMIPA Universitas Padjadjaran.

Prihandoko, Jollyta, D., Gusrianty, Siddik, M., & Johan. (2024). Cluster validity for optimizing classification model: davies bouldin index – random forest algorithm. *Matrik: Jurnal Manajemen, Teknik Informatika, dan Rekayasa Komputer*, *24*(1), 61-72.

Saidah, D. A., Santoso, R., & Widiharih, T. (2022). Pengelompokan provinsi di indonesia berdasarkan indikator kesehatan lingkungan menggunakan metode partitioning around medoids dengan validasi indeks internal. *Journal Gaussian*, *11*(2), 302-312.

Sary, R. A., Satyahadewi, N., & Andani, W. (2024). Application of K-Means++ with dunn index validation of grouping West Kalimantan region based on crime vulnerability. *BAREKENG: Journal of Mathematics and Its Applications*, *18*(4), 2283–2292.

Suraya, G., & Wijayanto, A. W. (2022). Comparison of hierarchical clustering, K-Means, K-Medoids, and Fuzzy C-Means Methods in grouping provinces in indonesia according to the special index for handling stunting. *Indonesian Journal of Statistics and Its Applications*, *6*(2), 180-201.