ISSN: 2339-2541
JURNAL GAUSSIAN, Volume 14, Nomor 2, Tahun 2025, Halaman 356 - 365
Online di: https://ejournal3.undip.ac.id/index.php/gaussian/

COMPARATIVE ANALYSIS OF K-MEANS, K-MEDOIDS, AND FUZZY C-MEANS FOR CLUSTERING PROVINCES IN INDONESIA BASED ON RICE PRODUCTION IN 2024

Ilham Mujahidin[1*], Siti Hadijah Hasanah[2]
[1,2] Department of Statistics, Faculty of Science and Technology, Universitas Terbuka
***e-mail**: ilhammujahidin1708@gmail.com

DOI: 10.14710/j.gauss.14.2.356-365

Article Info:
Received: 2025-05-31
Accepted: 2025-10-09
Available Online: 2025-10-14

Keywords:
Clustering; K-Means; K-Medoids; Fuzzy C-Means; Davies-Bouldin Index

Abstract: Rice is one of the main commodities in Indonesia's agricultural sector that plays a crucial role in maintaining national food security. This study aims to cluster 38 provinces in Indonesia based on the similarity of rice production in order to understand the spatial variation of agricultural performance in different provinces. In this analysis, three clustering methods were used, namely K-Means, K-Medoids, and Fuzzy C-Means, by considering two main variables: the area of harvest and the amount of rice production in 2024, whose data were sourced from the Central Bureau of Statistics. Evaluation of cluster quality was conducted using the Davies-Bouldin Index (DBI). The results showed that the K-Means method produced the most optimal clustering with the lowest DBI value of 0.276 at the number of clusters $K = 8$, compared to K-Medoids (0.279 at $K = 8$) and Fuzzy C-Means (0.285 at $K = 2$). The clusters formed show a clear separation between provinces with high and low production levels. Provinces with high agricultural intensification and a large contribution to production belong to the main cluster, while areas with limited resources and low production form a separate cluster. This finding reflects the diversity of agricultural conditions influenced by infrastructure, intensification, and geographical and climatic factors.

1. INTRODUCTION

The agricultural sector, especially the rice commodity, plays an important role in maintaining national food security in Indonesia (Quirinno et al., 2024). Harvest area and rice production are the main indicators in assessing the performance of this sector. According to data from the Central Statistics Agency (BPS) in 2024, the rice harvest area is estimated to reach around 10.05 million hectares, a decrease of 167.25 thousand hectares or 1.64 percent compared to the previous year which recorded a harvest area of 10.21 million hectares. Along with the decrease in harvest area, rice production is also estimated to decrease to around 53.14 million tons of harvested dry grain, a decrease of 838.27 thousand tons of harvested dry grain (1.55 percent) compared to the production in 2023 which reached 53.98 million tons of harvested dry grain. This decline is likely influenced by various factors, such as climate change, agricultural policies, and the use of different technologies and infrastructure in each province (Malau et al., 2025).

The process of understanding the patterns and characteristics of the agricultural sector in various provinces requires a data-based analysis approach, one of which is through the clustering method. Clustering is one of the methods in data mining that aims to group data into several groups based on the similarity of certain characteristics or attributes (Zheng et

al., 2020). Clustering algorithms work by grouping data into several clusters whose members have a high degree of similarity with each other, while the differences between clusters remain significant (Ling et al., 2025). The main goal is to reveal the hidden structure in the data, so as to provide a deeper understanding of the categories or groups contained therein (Hendrastuty, 2024). Some clustering methods that are often used include K-Means, K-Medoids, and Fuzzy C-Means. The K-Means method has several advantages, including its simplicity and ability to cluster large amounts of data quickly and efficiently (Kousis & Tjortjis, 2021). K-Medoids is a cluster method that has better resistance to outliers, because it uses medoids as cluster centers (Soesmono et al., 2025). Meanwhile, Fuzzy C-Means is a clustering algorithm that works by distributing data to each cluster based on the principles of fuzzy theory. This algorithm measures the extent to which the level of probability of data is included as a member of a cluster (Turrahma et al., 2023).

Based on research by Andini et al. (2024), the K-Means algorithm proved to be the most effective in clustering purebred chicken eggs in South Sumatra Province. This is indicated by the Davies Bouldin Index (DBI) value of 0.193 on five clusters, which is lower than the DBI of 0.268 from the K-Medoids algorithm on three clusters. In contrast, Rudianto and Wijayanto (2023) found that the K-Medoids algorithm gave the best clustering results with five clusters, based on the evaluation of the Sw/Sb ratio, where K-Medoids had a smaller ratio than K-Means. On the other hand, Firdaus et al. (2021) found that the Fuzzy C-Means algorithm obtained a Silhouette Index (SI) value of 0.818, which is higher than the 0.569 value of the K-Means algorithm. This finding indicates that Fuzzy C-Means has superior performance compared to K-Means. This study was conducted to compare three clustering algorithms, namely K-Means, K-Medoids, and Fuzzy C-Means, in clustering provinces in Indonesia based on rice production data in 2024. Through this analysis, it is expected to obtain more precise information about the agricultural characteristics of each province, which can later be used as a reference in formulating more effective and targeted agricultural policies.

2. LITERATURE REVIEWS

Clustering is an unsupervised learning method that aims to group data into several clusters based on similar characteristics. Each cluster consists of members that have a high level of similarity. This similarity is measured by distance, where the distance between members in one cluster is kept as small as possible, while the distance between clusters is made as far as possible (Tandean and Purba, 2020). Three popular methods that are often used are K-Means, K-Medoids, and Fuzzy C-Means (FCM).

a. K-Means

K-Means algorithm is a data clustering method that classifies data based on its proximity to the cluster center (centroid) (Triayudi et al., 2024). K-Means divides data into c clusters by minimizing intra-cluster variation. The iterative process starts with the initialization of the cluster center v_k, then the data is grouped based on the closest Euclidean distance, then the cluster center is updated using:

$$v_k = \frac{1}{|C_K|} \sum_{x_i \epsilon C_k} x_i \tag{1}$$

where C_K is the set of data in the to cluster k, and $|C_K|$ is the number of data in the cluster. The minimized objective function is:

$$J = \sum_{k=1}^{c} \sum_{x_i \epsilon C_k} x_i \, \|\boldsymbol{x_i} - \boldsymbol{v_k}\|^2 \quad (2)$$

The distance between two points x_i and x_j uses the Euclidean distance:

$$d(x_i, x_j) = \sqrt{\sum_{j=1}^{p} (X_{il} - C_{jl})^2} \quad (3)$$

b. K-Medoids

K-Medoids is similar to K-Means, but uses a medoid as the cluster center, which is the actual data that minimizes the total distance to all points in the cluster (Supriyadi et al., 2021). Objective function:

$$J = \sum_{k=1}^{c} \sum_{x_i \epsilon C_k} d(x_i, m_k) \quad (4)$$

where m_k is the medoid (representative data) of the to cluster k, and $d(x_i, m_k)$ can be:

Euclidean: $d(x_i, m_k) = \sqrt{\sum_{j=1}^{p} (X_{il} - C_{jl})^2}$ (5)

Manhattan: $d(x_i, m_k) = \sum_{l=1}^{p} |x_{il} - m_{kl}|$ (6)

Medoid selection is done by finding point x_j in the minimizing cluster:

$$\sum_{x_i \epsilon C_k} d(x_i, x_j) \quad (7)$$

c. Fuzzy C-Means

Fuzzy C-Means is a clustering algorithm that determines the optimal cluster by using membership degrees as the basis for determining the extent to which data belongs to a particular cluster. In addition, Fuzzy C-Means is able to form clusters that are more varied and scalable (Nur et al., 2023). FCM uses a soft clustering approach, where data can be a member of several clusters with membership level $\mu_{ik} \epsilon [0,1]$ (Herijanto and Riti, 2024). Minimized objective function:

$$J_m = \sum_{i=1}^{n} \sum_{k=1}^{c} (\mu_{ik})^m \, \|\boldsymbol{x_i} - \boldsymbol{v_k}\|^2 \quad (8)$$

where:

μ_{ik}: degree of membership of data x_i to cluster k

m: fuzziness parameter ($m > 1$)

Cluster center update formula:

$$v_k = \frac{\sum_{i=1}^{n} (\mu_{ik})^m \, x_i}{\sum_{i=1}^{n} (\mu_{ik})^m} \quad (9)$$

Membership degree update formula:

$$\mu_{ik} = \left(\sum_{j=1}^{c} \left(\frac{\|\boldsymbol{x_i} - \boldsymbol{v_k}\|}{\|\boldsymbol{x_i} - \boldsymbol{v_j}\|} \right)^{\frac{2}{m-1}} \right)^{-1} \quad (10)$$

d. Evaluation with Davies-Bouldin Index (DBI)

The DBI method is used to evaluate the quality of clusters, by considering how compact and separate a cluster is. The DBI value is calculated by:

$$DBI = \frac{1}{c} \sum_{i=1}^{c} R_i \quad (11)$$

with:

$$R_i = \max_{j \neq 1} \left(\frac{S_i + S_j}{M_{ij}} \right) \quad (12)$$

where:

$S_i = \frac{1}{|C_i|} \sum_{x \epsilon C_i} ||x - v_i||$: average distance in the to cluster i

$M_{ij} = ||\boldsymbol{v_i} - \boldsymbol{v_j}||$: distance between cluster centers i and j

A smaller DBI value indicates better clustering results because the clusters are more compact and well separated (Singh et al., 2020).

3. RESEARCH METHOD

This study uses secondary data obtained from the publication of the Central Bureau of Statistics (BPS) in 2024 regarding the harvest area and rice production in Indonesia. The variables used in this study are harvest area (in hectares) and rice production (in tons). Data analysis was conducted using Python through the Google Collaboratory platform, with the following steps:

1. Data Preparation

   Data was imported into Google Collaboratory using the relevant Python libraries. Once the data was successfully loaded, an initial review of the structure and summary statistics was conducted to understand the general characteristics of the data.
2. Data Standardization

   Numerical data is standardized so that all variables are on a comparable scale. This process aims to prevent the dominance of certain variables in distance calculations during the clustering process.
3. Determination of the Number of Clusters

   The determination of the number of clusters is done exploratively by testing several K values, namely K=2 to K=9. The purpose of this step is to compare the quality of cluster results based on different number of clusters.
4. Clustering Process

   The clustering process is performed using three methods, namely:
   a. K-Means which groups data based on the average center (centroid).
   b. K-Medoids which forms clusters based on median points so that it is more resistant to outliers.
   c. Fuzzy C-Means which allows each data to have membership levels in more than one cluster simultaneously.
5. Cluster Evaluation with Davies-Bouldin Index (DBI)

The cluster results of each method and the number of clusters are evaluated using the Davies-Bouldin index. The lower the DBI value, the better the quality of the resulting clusters, as it reflects clear separation and high cohesiveness within each cluster.

6. Cluster Visualization

Visualization of cluster results is carried out to obtain an overview of the distribution patterns and separation between clusters, so as to support the interpretation of the analysis results visually.

4. RESULT AND DISCUSSION

The three clustering algorithms used in this study are K-Means, K-Medoids, and Fuzzy C-Means. Data from 38 provinces were analyzed and clustered using the three clustering methods. The clustering process begins with determining the number of clusters used. To achieve optimal results, it is necessary to experiment in determining the number of clusters The number of clusters selected are $K = 2, K = 3, K = 4, K = 5, K = 6, K = 7, K = 8$, and $K = 9$. Each K value was evaluated using DBI to determine the most optimal number of clusters.

Tabel 1. Experimental results of clustering with K-Means

Experiment	$K1$	$K2$	$K3$	$K4$	$K5$	$K6$	$K7$	$K8$	$K9$
$K = 2$	4	34	-	-	-	-	-	-	-
$K = 3$	3	27	8	-	-	-	-	-	-
$K = 4$	9	3	25	1	-	-	-	-	-
$K = 5$	3	23	3	1	8	-	-	-	-
$K = 6$	5	3	19	1	3	7	-	-	-
$K = 7$	3	12	3	4	1	11	4	-	-
$K = 8$	1	3	12	1	11	4	2	4	-
$K = 9$	11	3	2	4	1	4	4	1	8

Based on the results in Table 1, it shows the number of members of each cluster from the results of the K-Means method for various number of clusters ($K = 2$ to $K = 9$). From this table it can be seen how the distribution of data into each cluster changes according to the value of K. The selection of the optimal number of clusters will be determined in the evaluation section using the DBI index.

Table 2. Experimental results of clustering with K-Medoids

Experiment	$K1$	$K2$	$K3$	$K4$	$K5$	$K6$	$K7$	$K8$	$K9$
$K = 2$	4	34	-	-	-	-	-	-	-
$K = 3$	3	25	10	-	-	-	-	-	-
$K = 4$	6	3	4	25	-	-	-	-	-
$K = 5$	4	3	11	14	6	-	-	-	-
$K = 6$	11	3	1	6	3	14	-	-	-
$K = 7$	11	3	5	5	3	10	1	-	-
$K = 8$	11	3	1	10	2	5	5	1	-
$K = 9$	1	3	4	8	2	4	4	11	1

Table 2 presents the results of the number of cluster members of the K-Medoids method for each trial number of clusters (K). As in K-Means, this information is important to see the distribution of the number of cluster members and analyze the stability of the clustering results.

Table 3. Experimental results of clustering with Fuzzy C-Means

Experiment	$K1$	$K2$	$K3$	$K4$	$K5$	$K6$	$K7$	$K8$	$K9$
$K = 2$	34	4	-	-	-	-	-	-	-
$K = 3$	8	3	27	-	-	-	-	-	-
$K = 4$	9	3	23	3	-	-	-	-	-
$K = 5$	3	3	19	6	7	-	-	-	-
$K = 6$	1	3	5	3	7	19	-	-	-
$K = 7$	1	1	6	2	6	19	3	-	-
$K = 8$	1	1	6	1	6	19	2	2	-
$K = 9$	2	2	5	1	5	11	1	1	10

Table 4 shows the results of clustering using Fuzzy C-Means, namely the number of provinces that have dominant membership in each cluster. Fuzzy C-Means allows one object to belong to several clusters, but the dominant cluster is shown. The distribution of members for each K is also observed here. After clustering using the K-Means, K-Medoids, and Fuzzy C-Means algorithms, the next step is to evaluate the clusters using the DBI (Davies-Bouldin Index) method.

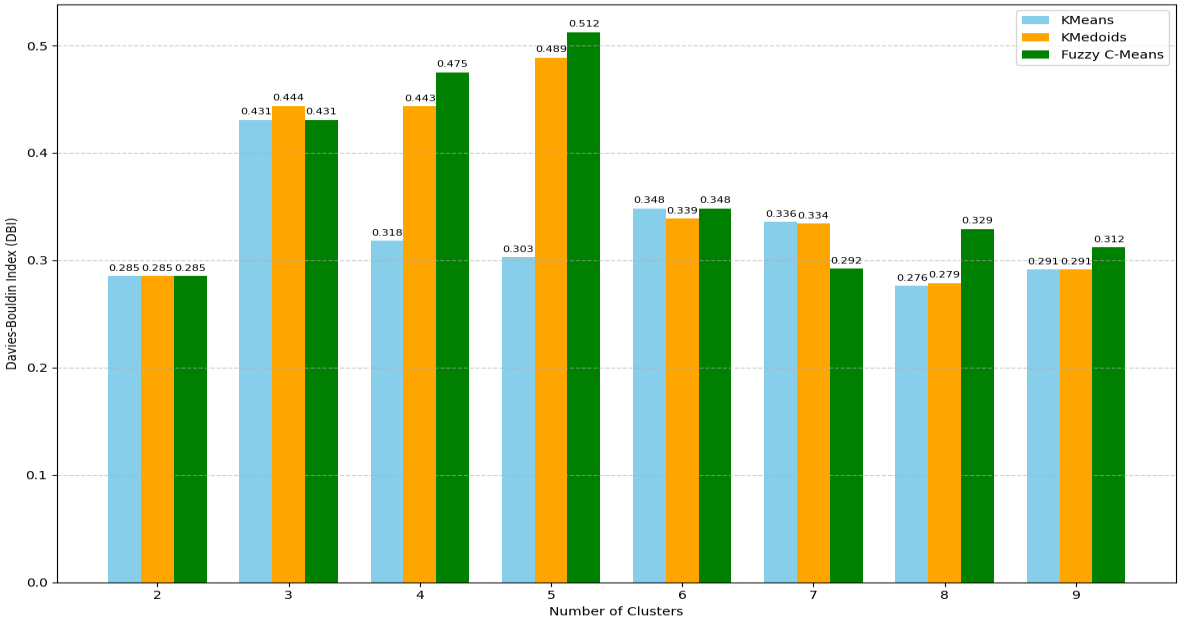

Figure 1. Comparison of DBI values for various clustering methods

Based on the comparison of DBI values (Figure 1), K-Means has the smallest index value when $K = 8$ with a value of 0.276. The K-Medoids method also obtained the smallest index value when $K = 8$ with a value of 0.279. Meanwhile, Fuzzy C-Means has the smallest index value when $K = 2$ with a value of 0.285. DBI using the K-Means algorithm of 0.276 has the smallest value of the other algorithms. According to Fathurrahman et al. (2023), the closer to 0 the DBI value, the more optimal the data grouping results. Thus, the K-Means method shows a more optimal performance than K-Medoids and Fuzzy C-Means in grouping data based on the harvest area and the amount of rice production. This result is consistent with the findings of Syukron et al. (2022), who concluded that the K-Means algorithm is able to produce a better cluster structure and performance than the other two methods.

Table 5. Clustering results with K-Means method when $K = 8$

Cluster	Province
Cluster 1	North Sumatra
Cluster 2	Central Java, East Java, West Java
Cluster 3	DI Yogyakarta, North Sulawesi, Gorontalo, East Kalimantan, West Sulawesi, Bali, Southeast Sulawesi, Central Kalimantan, Bengkulu, Jambi, Riau, South Papua
Cluster 4	South Sulawesi
Cluster 5	Riau Islands, Bangka Belitung Islands, West Papua, Southwest Papua, North Maluku, Central Papua, Maluku, Papua, North Kalimantan, DKI Jakarta, Papua Mountains
Cluster 6	Aceh, West Nusa Tenggara, Banten, West Sumatra
Cluster 7	Lampung, South Sumatra
Cluster 8	West Kalimantan, South Kalimantan, East Nusa Tenggara, Central Sulawesi

Clustering provinces in Indonesia using the K-Means method with a value of $K = 8$ (Table 5) produces a pattern that illustrates the differences in rice production characteristics between regions. Provinces with high rice production tend to belong to the same cluster, while provinces with low production are grouped separately.

a. Cluster 2 is the main group consisting of areas with high agricultural intensification, adequate infrastructure, and high land productivity. This cluster plays a major role in supporting rice production in Indonesia.
b. Cluster 5 includes areas with very low production levels. Regions in this cluster are generally located in eastern Indonesia, with significant geographical challenges and limited means of production.
c. Cluster 3 and Cluster 8 reflect groups of provinces with low to medium production characteristics, spread across different parts of Indonesia. Although their contribution to national production is not dominant, the provinces in these two clusters show similarities in their relatively limited but consistent production patterns.
d. Cluster 4 consists of one high-producing province outside Java. This province occupies a separate position due to its significant role as a regional rice granary, although its scale is not as large as the main cluster.
e. Cluster 1 and Cluster 7 reflect groups of provinces with fairly high production levels, although not comparable to the main region. The provinces in these two clusters have important contributions, especially outside Java.
f. Cluster 6 includes regions with medium production categories. The provinces in this cluster have good development potential and could increase through intensification and expansion of agricultural land.

Overall, the clustering results reflect spatial variations in rice production in Indonesia. These differences are influenced by the resource potential, level of agricultural intensification, agricultural infrastructure, and diverse geographical and climatic conditions in each region.

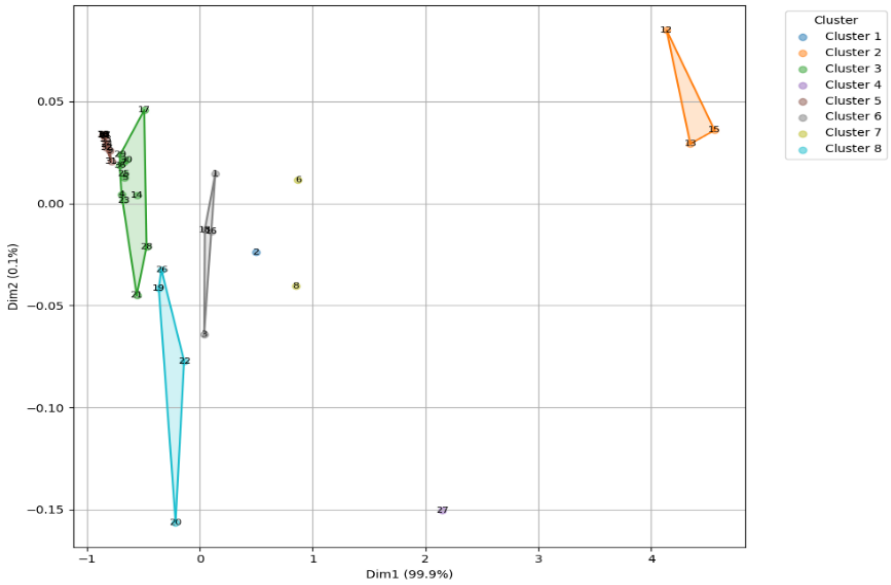

Figure 2. Visualization of clustering results with the K-Means method when $K = 8$

Visualization of the clustering results using the K-Means method (Figure 2) shows the results of grouping provinces based on rice production into eight clusters. It can be seen that provinces with high production characteristics tend to form their own groups that are clearly separated from other groups, reflecting the dominant contribution to rice production in Indonesia. Meanwhile, provinces with low production levels are scattered in several different groups, showing similarities in production patterns despite coming from diverse geographical areas. This cluster distribution pattern indicates that differences in resource potential, level of agricultural intensification, and geographical and infrastructural conditions affect rice production capacity between regions in Indonesia. Thus, the government can focus its production enhancement programs on provinces that are part of clusters with low production levels, for example through the provision of agricultural production facilities, farmer training, and improved access to irrigation and fertilizers.

5. CONCLUSION

Based on the analysis, the provinces in Indonesia were successfully clustered based on rice production data in 2024. Clustering using the K-Means algorithm resulted in the most optimal clustering with the lowest Davies-Bouldin Index (DBI) value of 0.276 at $K = 8$, better than K-Medoids (0.279 at $K = 8$) and Fuzzy C-Means (0.285 at $K = 2$). This result shows that the K-Means algorithm is able to form more separate and compact clusters, according to the principle that the smaller the DBI value, the better the quality of clustering. Visualization and further analysis of the clustering results revealed spatial patterns that reflect differences in agricultural characteristics between regions. Provinces with high production tend to belong to the same cluster, while regions with lower production levels are scattered in separate clusters. Overall, the clustering results reflect the diversity of potential, level of intensification, agricultural infrastructure, and geographical and climatic conditions that influence rice production patterns in Indonesia.

REFERENCES

Andini, R., Aminisari, S. T., Husin, V. A. F., & Jambak, M. I. (2024). Perbandingan algoritma K-Means dan K-Medoids untuk klasterisasi produksi telur ras ayam kampung di Provinsi Sumatera Selatan. *JATI (Jurnal Mahasiswa Teknik Informatika)*, *8*(3), 3910–3915.

Badan Pusat Statistik. (2025). *Ringkasan Eksekutif luas Panen, dan Produksi Padi di Indonesia 2024 (Angka Tetap)*. Badan Pusat Statistik.

Fathurrahman, Harini, S., & Kusumawati, R. (2023). Evaluasi Clustering K-Means dan K-Medoid Pada Persebaran Covid-19 di Indonesia dengan Metode Davies-Bouldin Index (DBI). In *Jurnal MNEMONIC, 6*(2), 117-128.

Firdaus, H. S., Nugraha, A. L., Sasmito, B., Awaluddin, M., & Nanda, C. A. (2021). Perbandingan metode Fuzzy C-Means dan K-Means untuk pemetaan daerah rawan kriminalitas di Kota Semarang. *Elipsoida*, *4*(1), 58–64.

Hendrastuty, N. (2024). Penerapan Data Mining Menggunakan Algoritma K-Means Clustering Dalam Evaluasi Hasil Pembelajaran Siswa. *Jurnal Ilmiah Informatika dan Ilmu Komputer (Jima-Ilkom)*, *3*(1), 46–56. https://doi.org/10.58602/jima-ilkom.v3i1.26

Kousis, A., & Tjortjis, C. (2021). Data Mining Algorithms for Smart Cities: A Bibliometric Analysis. In *Algorithms*, *14*(8), 1-35. https://doi.org/10.3390/a14080242

Ling, L. S., dan Weiling, C. T. (2025). Enhancing Segmentation: A Comparative Study of Clustering Methods. *Institute of Electrical and Electronics Engineers (IEEE)*, vol. 13, 47418-47439. https://doi.org/10.1109/ACCESS.2025.3550339

Malau, H., Saragih, J. R., & Purba, T. (2025). Strategi Perencanaan Wilayah dalam Menanggulangi Dampak Perubahan Iklim pada Kawasan Pertanian. *PESHUM: Jurnal Pendidikan, Sosial dan Humaniora, 4*(2), 2316-2323.

Nur, I. M., Syifa, A. N. L., Kharis, M., Permatasari, S. H. 2023. Implementasi Fuzzy C-Means dalam Pengelompokan Hasil Panen Padi di Provinsi Bali. *Variance: Journal of Statistics and its Applications, 5*(1), 13-24. https://doi.org/10.30598/variancevol5iss1page132-4

Quirinno, R. S., Murtiana, S., Asmoro, N. (2024). Peran Sektor Pertanian dalam Meningkatkan Ketahanan Pangan dan Ekonomi Nasional. NUSANTARA*: Jurnal Ilmu Pengetahuan Sosial, 11*(7), 2811-2822. https://doi.org/10.31604/jips.v11i7.2024

Rudianto, R. D., & Wijayanto, A. W. (2023). Analisis perbandingan K-Means dan K-Medoids dalam pengelompokan provinsi berdasarkan Indeks Demokrasi Indonesia 2021. *Komputika: Jurnal Sistem Komputer*, *13*(1), 19–27. https://doi.org/10.34010/komputika.v13i1.10812

Singh, A. K., Mittal, S., Malhotra, P., dan Srivastava, Y. V. (2020). Clustering Evaluation by Davies Bouldin Indeks (DBI) in Cereal data using K-Means. *Proc. 4th Int. Conf.*

Comput. Methodol. Commun. ICCMC 2020, pp. 306–310. https://doi.org/10.1109/ICCMC48092.2020.ICCMC-00057

Supriyadi, A., Triayudi, A., & Sholihati, I. D. (2021). Perbandingan Algoritma K-Means dengan K-Medoids pada Pengelompokan Armada Kendaraan Truk Berdasarkan Produktivitas. *JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)*, *6*(2), 229-240.

Soesmono, S., Pertiwi, R., Saputri, B., Putri, N., & Widodo, E. (2025). Pengelompokan Provinsi di Indonesia Berdasarkan Tingkat Pengangguran Tahun 2023 Menggunakan K-Medoids. *Emerging Statistics and Data Science Journal*, *3*(1), 498-515.

Syukron, H., Fauzi Fayyad, M., Junita Fauzan, F., Ikhsani, Y., & Gurning, U. R. (2022). Perbandingan K-Means K-Medoids dan Fuzzy C-Means untuk Pengelompokan Data Pelanggan dengan Model LRFM. *MALCOM: Indonesian Journal of Machine Learning and Computer Science 2*(2), 76–83.

Tendean, T., & Purba, W. (2020). Analisis Cluster Provinsi Indonesia Berdasarkan Produksi Bahan Pangan Menggunakan Algoritma K-Means. *Jurnal Sains dan Teknologi*, *1*(2), 5-11. https://doi.org/10.34013/saintek.v1i 2.31

Triayudi, A., Pinastawa, I. W. R., Pratama, J. D., Ningsih, S. (2024). *Clustering*. Yogyakarta: PT Penamudamedia.

Turrahma R. N., Putra, A. N. C., Alfajri, M. D., Gusmanto, R., dan Oktoeberza, W. K. Z. (2023). Implementasi *Fuzzy C-Means* untuk Clustering data Harga Saham Harian Pada PT. Astra International TBK. *Jurnal Rekursif, 11*(1), 64-69.

Zheng, A. Cai, J., Yang, H., and Zhao, X. (2020). CPGAN: Curve Clustering Architecture Based on Projected Latent Vector of Generative Adversarial Network. *Institute of Electrical and Electronics Engineers (IEEE).* https://doi.org /10.1109/access.2020.299