Performance Evaluation of Clustering Algorithms for Grouping Regencies and Cities in Kalimantan Based on Socio-Demographics

Khusnia Nurul Khikmah[1*] and A'yunin Sofro[2]

[1]Mathematics Study Program, Faculty of Mathematics and Natural Sciences, Universitas Palangka Raya, Palangka Raya, Central Kalimantan, 74874 Indonesia

[2]Actuarial Sciences Study Program, Faculty of Mathematics and Natural Sciences, Universitas Negeri Surabaya, Surabaya, East Java, 60231 Indonesia

*Corresponding author: khusniank@gmail.com

Received: 11 March 2026
Revised: 12 August 2026
Accepted: 18 August 2026

ABSTRACT – **Cluster analysis is an important data exploration technique that aims to identify patterns or characteristics of data, which, in practice, are used to support evidence-based policy making, one of which is used in population and regional development. However, choosing the wrong clustering method can produce useless analysis results. Therefore, this study proposes a comparative study to determine the performance of various clustering algorithms, comparing the performance of hierarchical clustering algorithms (single linkage, complete linkage, average linkage, centroid, ward) and non-hierarchical algorithms (k-means) in regencies and cities in Kalimantan. The study uses secondary data taken from the official websites of the Central Statistics Agency of the five provinces in Kalimantan, where the data to be clustered consists of 56 regencies and cities with eight socio-demographic variables. The analysis results show that all hierarchical methods produce robust and consistent clusters compared to non-hierarchical methods. This decision is based on the silhouette score value 0.701, which is 0.308 better than k-means. This shows that hierarchical methods can clearly identify regencies and cities in Kalimantan based on their socio-demographic factors.**

Keywords– **Analysis, Hierarchical Clustering, Non-Hierarchical Clustering, Silhouette Score, Socio-Demographic.**

I. INTRODUCTION

Clustering is a data analysis technique in multivariate statistics that has various approaches. This technique aims to group objects into several groups based on the similarity of their characteristics [1]. In general, this analysis technique works by calculating the level of closeness or similarity between objects and grouping them into one or another. This technique calculates the level of closeness, which can be grouped into hierarchical and non-hierarchical. The hierarchical clustering technique clearly shows the hierarchical relationship between sub-clusters and clusters [2]. Meanwhile, the non-hierarchical clustering technique divides the data so that each point in the data is only in one cluster at a certain level [3]. Both techniques have advantages and disadvantages in data clustering, so this study compares the two techniques.

This study compares the performance of four methods for clustering analysis using hierarchical techniques, namely single linkage, average linkage, complete linkage, centroid, and Ward's method. These four methods were selected based on previous studies on [4], which showed that hierarchical clustering produces clear dendrogram visualizations and does not require cluster number initialization. Meanwhile, for the study of non-hierarchical techniques, the proposed method is k-means. The method is proposed considering the efficiency of k-means described in [5] and its robustness to noise as described in previous research on [6].

Based on this background, this study aims to apply six clustering methods to the social demographic data of regencies and cities in Kalimantan. The selection of regencies and cities in Kalimantan generally took into account various aspects such as the size of the area and the problems that arise in Kalimantan, as was done in previous research on [7], which conducted modeling and prediction using basic statistical models and machine learning. The results show that various aspects, including socio-demographics, influence poverty in Kalimantan. These aspects will be statistically analyzed using clustering to identify groups of regencies and cities in Kalimantan based on their socio-demographic conditions.

In general, this study provides the latest perspective on the comparative performance of the proposed methods, namely single linkage, average linkage, complete linkage, centroid, Ward's method, and k-means, for clustering the socio-demographic problems of regencies and cities in Kalimantan. The results of this study are statistically capable of comparing the performance of clustering methods and providing new insights into the groups of regencies and cities in Kalimantan based on their social demographic characteristics.

II. LITERATURE REVIEW

A. Min-Max Normalization

Min-Max normalization is a data pre-processing technique that performs a linear transformation of numerical data into a specific interval. Based on previous research on [8], this method can improve numerical stability in computations, such as cluster analysis, by ensuring that all variables contribute equally to cluster formation. Suppose the original value of a data is x, with a minimum value of $\min(x)$ and a maximum value of $\max(x)$. In that case, the normalization result of

x, symbolized by x' in the range [0,1], follows the following equation [9], [10].

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)} \quad (1)$$

B. Euclidean Distance

Euclidean distance is a measure of inequality that measures the distance between two observations, which is usually represented as a straight line between the two points. If two points with $p = (p_1, p_2, \dots, p_n)$ and q $q = (q_1, q_2, \dots, q_n)$ are in n-dimensional space. Then the Euclidean distance can be calculated based on the following equation [11].

$$d(p,q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \cdots + (p_n - q_n)^2} \quad (2)$$

C. Single Linkage

Single linkage is an agglomerative hierarchical clustering method in which the distance between two clusters is defined as the minimum distance between one point and another in a cluster [12]. This technique is often called the nearest neighbor technique, which can capture complex cluster shapes. If point $p = (p_1, p_2, \dots, p_n)$ is in cluster C_i and point $q = (q_1, q_2, \dots, q_n)$ is in cluster C_j, with the distance between them symbolized by $D(C_i, C_j)$, then this minimum distance can generally be calculated using the following equation [13].

$$D(C_i, C_j) = \min\{d(p,q) | p \in C_i, q \in C_j\} \quad (3)$$

D. Complete Linkage

Complete linkage is a hierarchical clustering method that defines the distance between two clusters as the maximum distance between two points in each cluster, a technique commonly referred to as the furthest neighbor, where, in general, all clusters are characterized by the least similar cluster members [4]. If point $p = (p_1, p_2, \dots, p_n)$ is in cluster C_i and point $q = (q_1, q_2, \dots, q_n)$ is in cluster C_j, with the distance between them symbolized by $D(C_i, C_j)$, then this maximum distance can generally be calculated using the following equation [14].

$$D(C_i, C_j) = \max\{d(p,q) | p \in C_i, q \in C_j\} \quad (4)$$

E. Average Linkage

Average linkage is the midpoint between the single linkage and complete linkage methods, where the distance between two clusters is calculated based on the average distance between all pairs of points in two different clusters. Based on previous research on GH, it is mentioned that average linkage can produce well-separated clusters and a more balanced hierarchy. If point $p = (p_1, p_2, \dots, p_n)$ is in cluster C_i and point $q = (q_1, q_2, \dots, q_n)$ is in cluster C_j with a distance symbolized by $D(C_i, C_j)$, then this average distance can generally be calculated using the following equation [15].

$$D(C_i, C_j) = \frac{1}{|C_i||C_j|} \sum_{p \in C_i} \sum_{q \in C_j} d(p,q) \quad (5)$$

F. Centroid

Centroid is a hierarchical clustering method that defines the distance between two clusters as the distance between their centroids or average vectors. If point $p = (p_1, p_2, \dots, p_n)$ is in cluster C_i where the centroid is $\mu_i = \frac{1}{|C_i|} \sum_{p \in C_i} p$, the centroid distance can generally be calculated using the following equation [16].

$$D(C_i, C_j) = d(\mu_i, \mu_j) \quad (6)$$

G. Ward's Method

Ward's method is an agglomerative technique that aims to minimize the total variance within clusters by minimizing the sum of squared differences (ESS) in all clusters. Based on previous research at [17], this method is one of the most reliable hierarchical methods because it optimizes the best criteria with cluster homogeneity. If point $p = (p_1, p_2, \dots, p_n)$ is in cluster C_i, then the value of $ESS(C) = \sum_{p \in C} \|p - \mu C\|^2$, where $\|p - \mu C\|^2$ is the Euclidean distance squared between point p and centroid μC. Thus, mathematically, the distance between two clusters increases the total within-cluster sum of squares, which can be calculated using the following equation [12].

$$D(C_i, C_j) = ESS(C_i \cup C_j) - [ESS(C_i) + ESS(C_j)] \quad (7)$$

H. K-Means Clustering

K-means is a non-hierarchical clustering algorithm method that aims to partition n observations of data into k clusters, where each observation that falls within the cluster with the nearest centroid acts as a prototype, thereby minimizing the within-cluster sum of squares (WCSS) [18]. This method has advantages in computational efficiency and ease of implementation. However, it has limitations in its sensitivity to determining the number of clusters k and sensitivity to the placement of initial random centroids. Suppose point $p = (p_1, p_2, \dots, p_n)$ is in cluster C_i, where μC is the centroid of cluster C_i and $\|p - \mu C\|^2$ is the Euclidean distance squared between point p and centroid μC. Mathematically, the distance between two clusters can be calculated using the following equation to minimize the within-cluster sum of squares [3].

$$WCSS = \sum_{i=1}^{k} \sum_{p \in C_i} \|p - \mu C\|^2 \quad (8)$$

I. Silhouette Score

Silhouette score is a clustering evaluation matrix that measures how well each object fits into its cluster [19]. This matrix combines the concepts of cohesion, which is the closeness of objects to other objects in the same cluster, and separation, which is how far objects are from other clusters. Silhouette score $(S_i) \in [-1,1]$, where if the silhouette score value is closer to 1 $(S_i \approx 1)$, then the clustering accuracy is good or the cohesion is high and the separation is high. If $(s_i \approx 0)$, then the cohesion and separation are balanced. If $(S_i \approx -1)$, the point of i may be placed in the wrong cluster. If cohesion $a(p) = \frac{1}{|C_i|-1}\sum_{p \in C_i, p \neq q} d(p,q)$ and separation $b(p) = \min_{k \neq c}\left(\frac{1}{C_k}\sum_{q \in C_i} d(p,q)\right)$. Then, mathematically, the silhouette score equation can be calculated based on the following equation [20].

$$S_i = \frac{b(p) - a(p)}{\max\{b(p), a(p)\}} \tag{9}$$

III. METHODOLOGY

The data used in this study is secondary data taken from the official websites of the Central Statistics Agency of Central Kalimantan, East Kalimantan, South Kalimantan, West Kalimantan, and North Kalimantan provinces. In general, the variables used in this study are the eight variables presented in Table 1.

Table 1 Data Description

Variable	Symbol	Explanation
Poverty Rate	X_1	%
Total Population	X_2	In Thousands
Population Growth Rate	X_3	%
Percentage of Population	X_4	%
Population Density	X_5	/km^2
Length of Time Spent in School	X_6	In Years
Per Capita Expenditure on Food	X_7	%
Total Working Population	X_8	In Thousands

This study has data analysis steps that begin with data collection, followed by data exploration. Missing data in the total working population variable was handled as in the previous study on HJK. Then, clustering was performed and the silhouette score was calculated. Finally, the results were evaluated and interpreted.

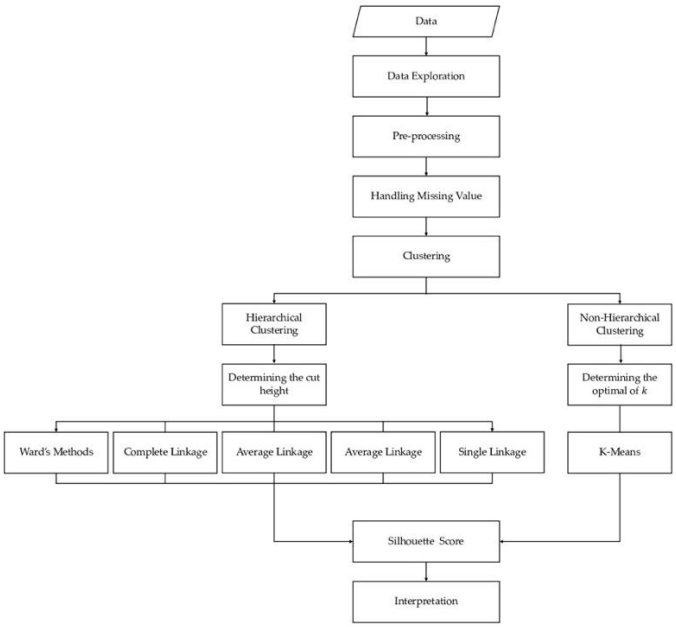


Figure 1 Analysis Flow Chart

IV. RESULTS AND DISCUSSIONS

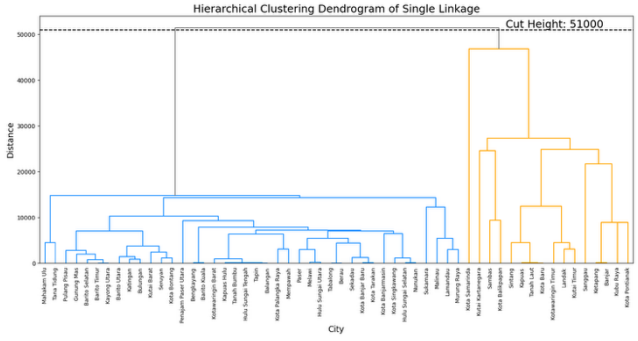
The analysis results began with data collection in Table 1 from 56 regencies and cities in Kalimantan, where the descriptive statistics of the eight variables are presented in Table 2. The descriptive statistics conducted by this study aim to understand the distribution, data center, and spread of each variable used to provide richer information. The descriptive statistics results for the population density variable (X_5) show an extreme data distribution, which allows the data to have outliers, as indicated by the high standard deviation and variance values.

Table 2 Descriptive Statistics of The Data

Measure	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8
Mean	1.71	320.56	8.54	8.93	382.41	8.66	52.51	150734.84
Median	1.36	269.67	5.96	6.36	37.00	8.39	53.32	130071.50
Standard Deviation	2.32	204.32	6.01	6.45	1200.45	1.17	4.78	89660.36
Variance	5.38	41746.58	36.16	41.58	1441070.86	1.37	22.86	8038979826.43
Range	18.10	836.50	25.34	33.44	6817.00	5.17	20.55	388498.00
Minimum	0.39	28.80	2.23	0.81	2.00	6.54	41.43	14640.00
Maximum	18.49	865.30	27.57	34.25	6819.00	11.71	61.98	403138.00

The first analysis was conducted using a hierarchical clustering approach, where the data submitted was standardized or scaled to provide a balanced contribution to the analysis process using the min-max method, see equation [1]. Next, the standardized data of the eight variables submitted, namely poverty rate, total population, population growth rate, population percentage, population density, length of schooling, per capita expenditure on food, and total working population, were analyzed using the single, average, complete linkage, and Ward's methods as clustering features. The cluster formation process is carried out by calculating distances, where this study uses Euclidean distances, see equation [2]. The results of these distance calculations form a square and symmetric matrix ($\boldsymbol{n} \times \boldsymbol{n}$). The resulting clusters are obtained by calculating the smallest distance matrix for the single linkage method. In contrast, the closest distance between the first two objects is calculated based on the equation in equation [3]. This agglomerative process is repeated until the number of clusters is obtained, which in this study is $\boldsymbol{k} = \mathbf{2}$.

The second method is calculated using the average value, or average linkage. This method pairs the closest clusters based on the average calculation of the Euclidean distance matrix of two objects. Then, agglomerative clustering is repeated until $\boldsymbol{k} = \mathbf{2}$ clusters are obtained. The third method calculates the farthest distance matrix to determine the two closest objects, then agglomerative clustering is repeated until $\boldsymbol{k} = \mathbf{2}$ clusters are obtained. The fourth is the centroid method, which calculates the distance between two clusters as the Euclidean distance between their mean vectors (centroids), merging the pair of clusters that results in the smallest increase in the total within-cluster variance at each step of the hierarchical agglomerative clustering process. The last method in the hierarchical clustering approach is Ward's method, which focuses on minimizing the variance within each cluster. The algorithm is repeated by calculating the total sum of squares, which will increase with each possible combination of two clusters. If combining two clusters increases the sum of squares, the smallest value of this total increase in the sum of squares is selected for combination (principle of minimal increase of variance). The clustering results of the four hierarchical clustering methods in regencies and cities in Kalimantan with eight socio-demographic variables are shown in the dendrogram in Figure 2. Figure 2. Dendrogram of clustering results



(a)

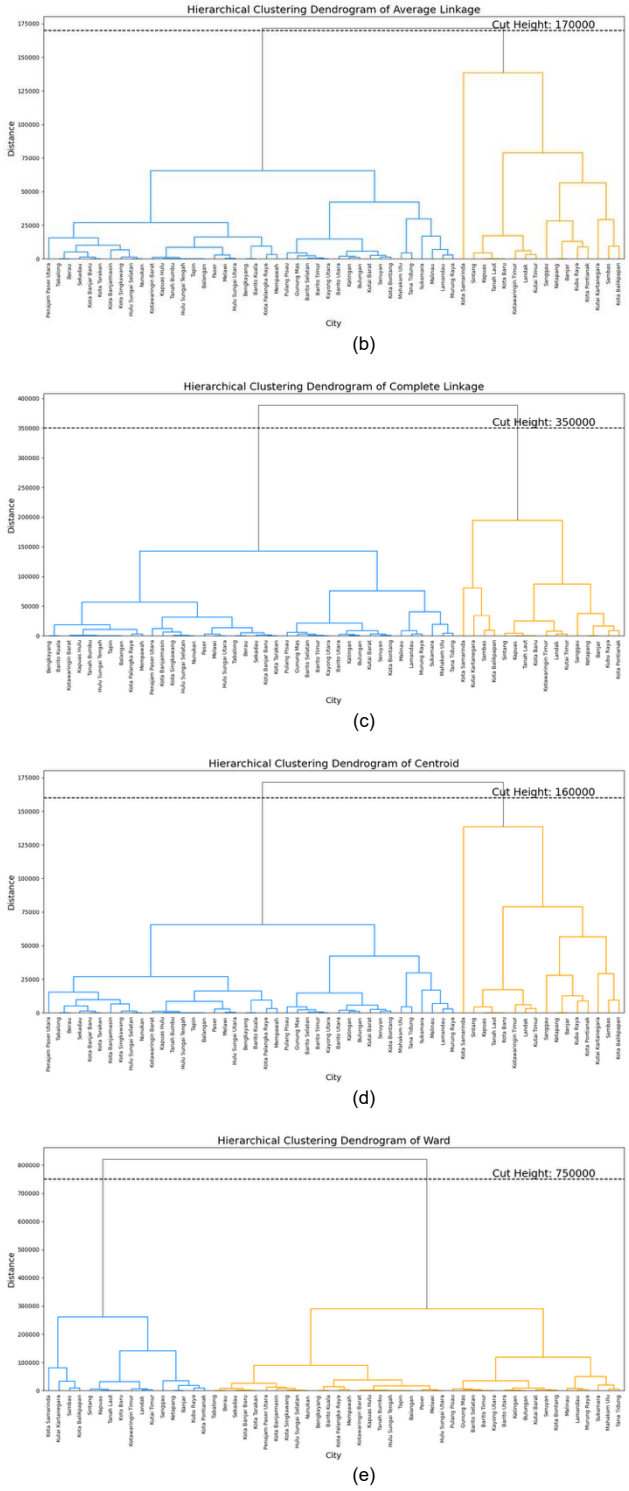


Figure 1 Dendrogram of Clustering Results. (a) Single Linkage, (b) Average Linkage, (c) Complete Linkage, (d) Centroid, and (e) Ward'd Methods

The clustering results shown in the dendrogram using the hierarchical clustering approach reveal a clear separation into two clusters. The first cluster consists of regencies and cities in Kalimantan with homogeneous eight socio-demographic variables. In contrast, the second cluster shows a different pattern, where the regencies and cities in Kalimantan in this cluster have specific socio-demographic characteristics.

This is reinforced by the dendrogram in Figure 2 above, showing good consistency across all different hierarchical clustering methods used. This means that the data has a strong inherent structure, so comparisons of clustering methods in the hierarchical clustering approach do not have a significant effect. The dendrogram results also indicate that of all the methods used in the hierarchical clustering approach for 56 regencies and cities in Kalimantan, there are 13 regencies and cities with specific socio-demographic conditions, namely East Kotawaringin, Kapuas, Sambas, Landak, Ketapang, Sintang, Kapuas Hulu, Kayong Utara (North Kayong), Kubu Raya, Pontianak, Singkawang, Berau, and Penajam Paser Utara (North Penajam Paser).

The second approach was non-hierarchical clustering, namely k-means clustering, determining the $\boldsymbol{k}$ value using the elbow method, as shown in Figure 3. This value shows a significant decrease in WCSS (within-cluster sum of squares) from $\boldsymbol{k} = \mathbf{1}$ to $\boldsymbol{k} = \mathbf{2}$, then the graph flattens after $\boldsymbol{k} = \mathbf{2}$. In general, this indicates that $\boldsymbol{k} = \mathbf{2}$ is the optimal cluster. Furthermore, this study also determined the optimal $\boldsymbol{k}$ value using the Mcclain methods, which also showed that the optimal $\boldsymbol{k}$ value was $\boldsymbol{k} = \mathbf{2}$. The k-means clustering results showed that of the 56 regencies and cities in Kalimantan, there were six regencies and cities with specific socio-demographic conditions, namely Pontianak, Banjarmasin, Kutai Kertanegara, Samarinda, Balikpapan, and Tarakan. Meanwhile, the other 50 regencies and cities fall under general socio-demographic conditions.

The goodness of these clustering results can generally be compared using the silhouette score to validate that objects within clusters have high similarity and evident dissimilarity with objects outside the clusters. The silhouette clusters of the six methods proposed by this study are presented in Table 3.

Table 3 Silhouette Score

Matrix Evaluation	Single Linkage	Average Linkage	Complete Linkage	Centroid	Ward's Methods	K-Means
Silhouette Score	0.701	0.701	0.701	0.701	0.701	0.393

The silhouette score in Table 3 above shows that the hierarchical clustering approach produces consistent results compared to the k-means (non-hierarchical clustering approach), namely 0.701. This indicates that the cluster structure produced has good or robust differences in clustering the socio-demographic data of regencies and cities in Kalimantan. This result is reasonable considering that the data contains outliers, as shown in Table 1, where the population density variable ($\boldsymbol{X_5}$) has outliers. Previous research on [21] shows that k-means is unsuitable for data with outliers.

Cluster 1 (43 regencies and cities) requires policies that focus on equitable development, improving access to education, and economic development based on local potential. Programs to improve the quality of human resources and basic infrastructure must be prioritized. Cluster 2 (13 regencies and cities) requires policies that focus on controlling population density, providing urban infrastructure, and alleviating urban poverty. Programs to expand formal employment opportunities and control food prices should be prioritized. In addition, these results provide insight into the importance of comprehensive comparative studies for achieving robust clustering. The hierarchical clustering approach also shows that it can capture the natural structure of the socio-demographic data of regencies and cities in Kalimantan.

This study used the Min-Max Scaling method to standardize the data prior to cluster analysis. This method is known to have limitations when the data contains severe outliers. For future research, it is recommended to use normalization methods that are more robust to outliers, such as Robust Scaling, Log-Transformation, or Z-Score after data transformation, in order to produce more stable and accurate clusters.

V. CONCLUSIONS AND SUGGESTIONS

Clustering analysis of 56 regencies and cities in Kalimantan based on the socio-demographic conditions of eight variables representing them shows that it is possible to identify two natural clusters from the data. Whereas for the hierarchical clustering approach with five proposed methods, single, average, complete linkage, centroid, and Ward's methods showed good clustering results compared to the k-means non-hierarchical clustering approach, which was based on a silhouette score of 0.71 or higher by 0.308. These results also show that 13 of the 56 regencies and cities in the minority cluster have different socio-demographic characteristics, requiring different regional development policies to be on target. A comprehensive statistical understanding of the comparative study of clustering methods shows that data structure is important in determining the analysis method.

ACKNOWLEDGEMENT

The author expresses sincere gratitude to Badan Pusat Statistik Provinsi Kalimantan Tengah for providing the statistical data used in the completion of this research.

REFERENCES

[1] G. J. Oyewole and G. A. Thopil, "Data clustering: application and trends," *Artif Intell Rev*, vol. 56, no. 7, pp. 6439–6475, 2023.

[2] X. Ran, Y. Xi, Y. Lu, X. Wang, and Z. Lu, "Comprehensive survey on hierarchical clustering algorithms and the recent developments," *Artif Intell Rev*, vol. 56, no. 8, pp. 8219–8264, 2023.

[3] P. Kudal, A. Patnaik, S. Dawar, R. K. Satankar, and P. Dawar, "Segmentation of OECD countries on the basis of selected global environmental indicators using k-means non-hierarchical clustering," *Environmental Science and Pollution Research*, vol. 31, no. 7, pp. 10334–10345, 2024.

[4] K. N. Khikmah and A. Sofro, "Clustering Regency and City in East Java Based on Population Density and Cumulative Confirmed COVID-19 Cases," *ComTech: Computer, Mathematics and Engineering Applications*, vol. 12, no. 2, pp. 111–121, 2021.

[5] F. Nie, Z. Li, R. Wang, and X. Li, "An effective and efficient algorithm for K-means clustering with new formulation," *IEEE Trans Knowl Data Eng*, vol. 35, no. 4, pp. 3433–3443, 2022.

[6] X. Ran, X. Zhou, M. Lei, W. Tepsan, and W. Deng, "A novel k-means clustering algorithm with a noise algorithm for capturing urban hotspots," *Applied Sciences*, vol. 11, no. 23, p. 11202, 2021.

[7] K. N. Khikmah and A. S. Sofro, "Comparison of Basic Statistics and Machine Learning Classification Algorithms in Kalimantan Poverty Prediction with Handling Missing Data," *Jurnal Matematika, Statistika dan Komputasi*, vol. 22, no. 1, pp. 90–101, 2025.

[8] M. Shantal, Z. Othman, and A. A. Bakar, "A novel approach for data feature weighting using correlation coefficients and min–max normalization," *Symmetry (Basel)*, vol. 15, no. 12, p. 2185, 2023.

[9] N. Singh and P. Singh, "Exploring the effect of normalization on medical data classification," in *2021 International Conference on Artificial Intelligence and Machine Vision (AIMV)*, IEEE, 2021, pp. 1–5.

[10] K. Cabello-Solorzano, I. Ortigosa de Araujo, M. Peña, L. Correia, and A. J. Tallón-Ballesteros, "The impact of data normalization on the accuracy of machine learning algorithms: A comparative analysis," in *International conference on soft computing models in industrial and environmental applications*, Springer, 2023, pp. 344–353.

[11] H. Zhao, "Design and Implementation of an Improved K-Means Clustering Algorithm," *Mobile Information Systems*, vol. 2022, no. 1, p. 6041484, 2022.

[12] S. Miyamoto, *Theory of agglomerative hierarchical clustering*, vol. 15. Springer, 2022.

[13] D. Bakkelund, "Order preserving hierarchical agglomerative clustering," *Mach Learn*, vol. 111, no. 5, pp. 1851–1901, 2022.

[14] E. Oti and M. Olusola, "Overview of agglomerative hierarchical clustering methods," *British Journal of Computer, Networking and Information Technology*, vol. 7, no. 2, pp. 14–23, 2024.

[15] L. R. Emmendorfer and A. M. de Paula Canuto, "A generalized average linkage criterion for hierarchical agglomerative clustering," *Appl Soft Comput*, vol. 100, p. 106990, 2021.

[16] A. Dogan and D. Birant, "K-centroid link: a novel hierarchical clustering linkage method," *Applied Intelligence*, vol. 52, no. 5, pp. 5537–5560, 2022.

[17] N. Randriamihamison, N. Vialaneix, and P. Neuvial, "Applicability and interpretability of Ward's hierarchical agglomerative clustering with or without contiguity constraints," *J Classif*, vol. 38, no. 2, pp. 363–389, 2021.

[18] M. Annas and S. N. Wahab, "Data mining methods: K-means clustering algorithms," *International Journal of Cyber and IT Service Management*, vol. 3, no. 1, pp. 40–47, 2023.

[19] S. K. Anand and S. Kumar, "Experimental comparisons of clustering approaches for data representation," *ACM Computing Surveys (CSUR)*, vol. 55, no. 3, pp. 1–33, 2022.

[20] M. Shutaywi and N. N. Kachouie, "Silhouette analysis for performance evaluation in machine learning with applications to clustering," *Entropy*, vol. 23, no. 6, p. 759, 2021.

[21] C. Grunau and V. Rozhoň, "Adapting k-means algorithms for outliers," in *International Conference on Machine Learning*, PMLR, 2022, pp. 7845–7886.

© 2026 by the authors. This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License (http://creativecommons.org/licenses/by-sa/4.0/).