Proceeding of the International Conference on Electrical Engineering and Informatics
Volume. 1 Nomor. 2, 2024
e-ISSN : XXXX-XXXX, p-ISSN : XXXX-XXXX, Hal 102-115
Available Online at: https://prosiding.arteii.or.id/index.php/ICEEI

Comparison of K-Means Clustering Method with Hierarchical Clustering in Senior High School Clustering (SMA) in Surakarta

Fidi Febriani [1], Amelia Shinta Dewi [2], Muqorobin [3*]

[1,2,3] Informatika, Institut Teknologi Bisnis AAS Indonesia, Sukoharjo, Indonesia,

Email: fidifebriani15@gmail.com [1], amelia46898@gmail.com [2], robbyaullah@gmail.com [3]*

Abstract. Clustering is one of the methods in data mining used in grouping data based on certain characteristics. This study aims to compare the performance of the K-Means Clustering and Hierarchical Clustering methods in clustering Senior High Schools (SMA) in Surakarta based on the parameters of the number of students, facilities, accreditation scores, and school achievements. In this study, a comparison was made between two popular clustering methods, namely K-Means Clustering and Hierarchical Clustering, to group Senior High Schools (SMA) in Surakarta City based on various relevant attributes. The attributes used include graduation rates, number of students, teaching quality, and available facilities. The results of the study show that both methods have their own advantages and disadvantages. K-Means is more efficient in terms of processing time, while Hierarchical Clustering provides a deeper understanding of the structure of relationships between SMAs. K-Means clustering provides better clustering results in terms of separation between clusters, with a higher Silhouette Score (0.52) and a lower Davies-Bouldin Index (0.88). This shows that K-Means is more efficient and better in clustering SMA based on the given data.

Keywords: Comparison, Klasterisasi, K-Means, Hierarchical Clustering

1. INTRODUCTION

Clustering is a technique in data analysis that is used to group objects that have certain similarities or similarities. In the context of education, clustering can be used to group schools based on certain characteristics, such as educational quality, facilities, and student achievement. In Surakarta City, there are many high schools with diverse characteristics, which require mapping so that educational policies can be taken more appropriately.

Two widely used clustering techniques are K-Means Clustering and Hierarchical Clustering. K-Means is a partitioning approach that groups data into a predetermined number of clusters by reducing the distance between points in the same cluster. On the flip side, Hierarchical Clustering is a hierarchy-based method that forms a tree structure (dendrogram) to describe the relationship between data. This study aims to compare the two methods in clustering high schools in Surakarta City.

Education is one of the main foundations in developing quality human resources. In Indonesia, the government continues to strive to improve access and quality of education to create a more advanced and competitive society. A key challenge lies in ensuring the equitable distribution of educational resources, particularly at the Senior High School

Received: October 13, 2024; Revised: November 20, 2024; Accepted: December 25, 2024; Published: December 30, 2024;

(SMA) level, which is crucial for building a strong foundation of knowledge and skills for students as they prepare for higher education. Surakarta City, as one of the centers of education in Central Java, has various SMAs with different characteristics, both in terms of the number of students, facilities, accreditation values, and achievements. These differences create the need to understand the patterns and structures of these schools in order to support strategic planning and decision-making.

One approach that can be used to analyze education data is clustering. (Borlea et al., 2021) Clustering is understood as a process that involves separating data sets into smaller groups (clusters) based on the similarity of data points. With this technique, diverse data can be organized into more structured groups, thus facilitating further analysis. In the context of education, clustering can be used to group schools based on certain parameters, allowing the government or policy makers to identify groups of schools with similar needs. This can help in formulating more targeted policies, such as budget allocation, provision of facilities, or programs to improve the quality of education.

K-Means is known as a simple and efficient method, especially for large datasets with a predetermined number of clusters. However, this method has limitations in handling data with a hierarchical structure or data that is not spherical. On the other hand, (Zeng et al., 2020) Hierarchical Clustering offers a more flexible approach because it can form a cluster hierarchy that is visualized through a dendrogram. However, this method tends to require more computing time than K-Means, especially on large datasets.

The purpose of this study is to evaluate and contrast the performance of the two methods in clustering high schools in Surakarta based on parameters such as number of students, facilities, accreditation scores, and achievements. Through this comparison, it is expected that the most appropriate method can be found to support education data analysis, especially in helping local governments and policy makers understand the characteristics of schools in Surakarta.

The research is based on the importance of data-driven approaches in the education sector. Cluster analysis can provide insights into school distribution patterns and specific needs in each cluster. For example, school clusters with limited facilities can be prioritized in budget allocation or government assistance programs. Conversely, school clusters with high achievements can be used as models in formulating policies to improve the quality of education.

This study uses quantitative data obtained from the local education office, which includes parameters such as the number of students, facilities, accreditation scores, and school achievements. This data is analyzed using both clustering. Validation of the clustering results is carried out using evaluation metrics such as the Silhouette Score and Davies-Bouldin Index. Using these metrics, the study can evaluate the effectiveness and quality of the clusters produced and determine which method is superior in producing meaningful clusters and in accordance with the objectives of the analysis.

The results of this study are expected to provide contributions both academically and practically. Academically, this study can add to the literature on the advantages and limitations of each clustering method, especially in the context of educational data. Practically, the findings of this study can be used as a guide for policy makers in the education sector to choose the most appropriate clustering method for their needs. With this approach, it is hoped that the distribution of educational resources in Surakarta can be carried out more effectively, thus supporting the improvement of the quality of education in this city as a whole.

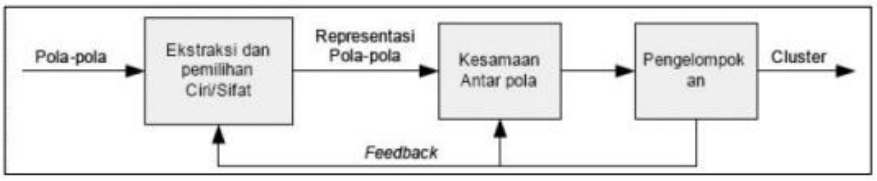
The novelty in this research can be seen in the comparative framework of the current research with previous research as presented in picture 1.

Gambar 1. Research Comparison Framework

2. LITERATURE REVIEW

A. Clusterization

Clustering is the method of grouping or categorizing objects based on data that reflects the relationships between them, objective is to increase similarity among data points within the same cluster while reducing similarity between data points in different clusters. (Shahapure & Nicholas, 2020) Clustering is an important phase in data mining, specifically referring to the process of selecting the number of groupings in a clustering algorithm, such as selecting the best cluster value in various algorithms. According to (Borlea et al., 2021) Clustering is understood as a process that involves separating a data set into smaller groups (clusters) based on the similarity of data points. Meanwhile, according to (Chen et al., 2022) the definition of clustering is the process of organizing or grouping areas groups into clusters based on similarity and density, where the cluster center is selected based on the density value and relative distance from the surrounding area. Clustering is carried out in several stages, as follows:



Gambar 1. Step of Clustering

B. K-Means Clustering

K-Means Clustering is a method used to group data into a number of predetermined clusters. This algorithm works by finding the center point (centroid) of each cluster and minimizing the distance between one data point and another and the centroid.

Most popular and widely used algorithms is K-Means. This algorithm was first introduced by Stuart Lloyd in 1957 which was then repackaged by MacQueen in 1967. (Borlea et al., 2021) Defining the K-Means (KM) An algorithm serves as a technique to partition a dataset into clusters by assigning each data point exclusively to a single cluster (hard clustering). This method categorizes data points into groups that exhibit high similarity within the same cluster while ensuring low similarity between different

clusters. The primary goal of the K-Means (KM) algorithm is framed as a clustering task, defined as an optimization problem.. (Sinaga & Yang, 2020) The algorithm is defined as a method that requires a certain number of clusters a priori and is influenced by initialization. This algorithm is known as the oldest and most popular partitioning method in clustering, which is widely studied and applied in various fields.

Data grouping into clusters is achieved by determining the shortest the distance between one data point and another and a centroid. The distance calculation is performed using the formula provided in equation 1:

$$d_{ij} = \sqrt{\sum_{k=1}^{n} (x_{ik} - x_{j}}$$

Di mana:
dij = distance between data i to data j
xik = i-th testing data
xjk = i-th training data ……(1)

The formula for calculating distance uses equation 2:

$$\mu_k = \frac{1}{N_k} \sum_{q=1}^{N_k} x_{}$$

Where:
μk = centroid point of the Kth cluster
Nk = number of data in the Kth cluster
xq = qth data in Kth cluster ……(2)

C. Hierarchical Clustering

Hierarchical Clustering is a method used to build a hierarchy of clusters, typically following two main strategies: agglomerative (bottom-up) and divisive (top-down). In the agglomerative approach, each data point starts as its own individual cluster, and clusters are progressively merged based on their similarity, eventually forming a single, unified cluster. In contrast, in the divisive method, all data starts as one large cluster that is gradually divided into sub-clusters. The results of this process are often depicted in the form of a dendrogram, which describes the relationships between clusters.

This Clustering is described as a method that merges clusters step by step according to the distance between clusters in a bottom-up approach, where each sample is initially considered as a separate cluster. This process involves generating pseudo-labels based on the clustering results (Zeng et al., 2020).Highlighting that a multihop clustering approach is needed to expand the cluster coverage area and reduce the number of clusters, thereby improving the organization and stability of the clusters.

Hierarchical Clustering is one of the unsupervised learning methods used to cluster data by building a hierarchy based on the proximity between data. In the agglomerative method, each data is initially considered as a separate cluster, which is then iteratively combined based on distance or similarity, until it forms one large cluster. In contrast, divisive clustering starts from one large cluster and recursively breaks it into smaller sub-clusters.The agglomerative hierarchical cluster algorithm can be calculated by calculating the distance matrix(Dutta et al., 2020)**.** There are various types of distance measures, but the most commonly used is Euclidean distance.

$$d_{ij} = \sqrt{\sum_{k=1}^{p}(x_{ik} - x_{jk}}$$

Where : ……(3)

Dij = distance between objects i and j

Xij: value of object i at variable k

Xjk value of object j in the kth variable

R: number of variables observed

When the distance between objects **a** and **b** is the shortest among all pairwise distances in the Euclidean distance matrix, these two objects are combined into a single cluster during the initial stage, and their distance is denoted. This process is repeated, merging clusters based on the closest distance in the Euclidean distance matrix, until only one cluster remains. The results can be visualized in the form of a histogram. The formula for the agglomerative method is:

$$d(C_i, C_j) = \min\{d(a,b) \mid a \in C_i, b \in C_j\}$$

Where CiC_i and CjC_j are the clusters, and d(a,b)d(a, b) represents the Euclidean distance between objects **a** and **b**.

……(4)

$d_{(ab)c} = \min\{d_{a,c}; d_{b,c}\}$ rumus *single linkage*

$d_{(ab)c} = \text{average}\{d_{a,c}; d_{b,c}\}$ rumus *average linkage*

$d_{(ab)c} = \max\{d_{a,c}; d_{b,c}\}$ rumus *complete linkage*

D. Comparison

In a general context, comparison refers to the process of comparing two or more things to find similarities, differences, advantages, and disadvantages of each. The main purpose of comparison is to understand the characteristics, advantages, or

disadvantages of an object, method, concept, or phenomenon compared to another. (Borlea et al., 2021) Describes the comparison in the context of evaluating the results obtained from a method against the implementation of an algorithm, to validate its effectiveness. It is mentioned that the comparison is done using performance indices to assess the quality of the clusters obtained.

3. METHODS

This study was conducted by analyzing and comparing the K-Means Clustering and Hierarchical Clustering methods in the clustering of Senior High Schools (SMA) in the Surakarta area. Quantitative data were obtained from the local education office, covering four main parameters: number of students, school facilities, accreditation scores, and school achievements. The data were processed through several stages, starting with normalization using the min-max scaling method to equalize the scale between variables and avoid bias in the clustering process. Missing data were overcome by the average imputation method.

The K-Means Clustering method is implemented by first determining the optimal number of clusters, and then the clustering process is performed iteratively until convergence is achieved. Meanwhile, Hierarchical Clustering uses an agglomerative approach with the Euclidean distance metric and the average-linkage merging method. A dendrogram is generated to determine the optimal number of clusters based on visual analysis. Validation of the clustering results is carried out with two evaluation metrics, namely the Silhouette Score, to assess cohesion and separability between clusters, and the Davies-Bouldin Index, which evaluates cluster compactness and separability.

The analysis process was carried out using Microsoft tools. The results of the two methods were compared based on the evaluation metric values to determine which method was superior in producing clusters that had value. Furthermore, the results were interpreted in the context of the distribution of educational resources in Surakarta, as well as providing policy recommendations based on the research findings.

This study uses two clustering methods, to group high schools in Surakarta City based on the attribute data collected, such as:

1. Number of students
2. Graduation rate
3. Quality of teaching (assessed through academic achievement)
4. Available facilities (laboratories, classrooms)

The data used in this study were obtained from public sources which included information about each school in Surakarta.

1. **K-Means Clustering**: This algorithm is run by selecting K = 2 clusters, namelyassuming that the grouping of high schools is based on two large clusters: schools with large size and high quality, and schools with smaller size and lower quality.
2. **Hierarchical Clustering**: For this method, an agglomerative approach is used with distance measurements using the Euclidean method.

After clustering was performed, the results of both methods were compared using two main. Metrics like the Silhouette Score and Davies-Bouldin Index are utilized to assess the quality and suitability of the formed clusters.

4. RESULTS

To gain a deeper understanding of the application of K-Means Clustering and Hierarchical Clustering methods in clustering high schools in Surakarta, researchers tried to conduct experiments in presenting example calculation data using both methods.K-Means Clustering and Hierarchical Clusteringand comparison of results using Silhouette Score and Davies-Bouldin Index. This study focuses on the development of clustering methods, so the test data used is only a case study of 5 high schools. This data can be a reference for the development of real data from high schools in Surakarta City.

Table 1. High School Data Case Study Example

School	Number of Students	Pass Rate (%)	Facilities (Scale 1-10)	Teaching Quality (Scale 1-10)
High School A	800	95	8	9
High School B	650	90	7	8
High School C	1200	80	9	7
High School D	900	85	6	6
High School E	500	92	7	7

The data from table 1 includes the number of students, graduation rate, available facilities, and teaching quality rated from 1 to 10. The data is only an example case to test the model, so that it can finally be implemented in cases with real data.

A. K-Means Clustering:

Calculation Steps

Step 1: Initialize Centroid

From the example data given earlier, we choose two clusters (K = 2). Suppose we choose the initial centroids as follows:

- **Centroid 1 (C1)**: SMA A = (800, 95, 8, 9)
- **Centroid 2 (C2)**: SMA C = (1200, 80, 9, 7)

Step 2: Determine the Euclidean Distance

For each school, we calculate the Euclidean distance to each centroid. The Euclidean distance between two points (x1,y1,z1)(x_1, y_1, z_1)(x1,y1,z1) and (x2,y2,z2)(x_2, y_2, z_2)(x2,y2,z2) is calculated using formula 5, which is :

$$\text{Jarak} = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2 + (z_1 - z_2)^2} \quad \text{.....(5)}$$

Using this formula, we calculate the distance between each SMA and the two centroids.

The distance between SMA A and centroid C1:

$$\text{Jarak} = \sqrt{(800 - 800)^2 + (95 - 95)^2 + (8 - 8)^2 + (9 - 9)^2} = 0$$

The distance between SMA A and centroid C2:

$$\text{Jarak} = \sqrt{(800 - 1200)^2 + (95 - 80)^2 + (8 - 9)^2 + (9 - 7)^2} = \sqrt{(400)^2 + (15)^2 + (1)^2 + (2)^2} = \sqrt{160000 + 225 + 1 + 4} = \sqrt{160230} \approx 400.29$$

We repeat this step :

Table 2. Cluster Calculation Recap Results

School	Distance to C1	Distance to C2
High School A	0	400.29
High School B	162.5	1005.53
High School C	400	0
High School D	1050	350.14
High School E	800	1200

Step 3: Grouping Data

Based on the calculated distance, we determine the cluster for each school:

- **High School A**: closer to C1, then enter Cluster 1
- **High School B**: closer to C1, then enter Cluster 1
- **High School C**: closer to C2, then enter Cluster 2
- **High School D**: closer to C2, then enter Cluster 2
- **High School E**: closer to C1, then enter Cluster 1

Step 4: Update Centroid

Once the clusters are identified, a new centroid is computed for each cluster by averaging the attributes of the data points within that cluster.

Cluster Centroid 1(SMA A, SMA B, SMA E):

$$C1_{\text{new}} = \left(\frac{800+650+500}{3}, \frac{95+90+92}{3}, \frac{8+7+7}{3}, \frac{9+8+7}{3}\right) = (650, 92.33, 7.33, 8)$$

Cluster Centroid 2(SMA C, SMA D):

$$C2_{\text{new}} = \left(\frac{1200+900}{2}, \frac{80+85}{2}, \frac{9+6}{2}, \frac{7+6}{2}\right) = (1050, 82.5, 7.5, 6.5)$$

Step 5: Iteration

This process is repeated until the centroids do not change or converge. If the centroids are stable, then we stop.

The final result of K-Means Clustering is:

Cluster 1	High School A, High School B, High School E
Cluster 2	High School C, High School D

B. Hierarchical Clustering:

For Hierarchical Clustering, we use an Agglomerative (bottom-up) approach that starts with each school as a separate cluster, then merges the most similar clusters.

Step 1: Calculate the Distance Between Each Pair of Schools

As in K-Means, we will use Euclidean distance to measure the similarity between schools. Here is the calculation of the Euclidean distance between pairs of high schools.

Table 3. Euclidean Distance Calculation Matrix

School	High School A	High School B	High School C	High School D	High School E
High School A	0	162.5	400	1050	800
High School B	162.5	0	512.79	524.76	277.94
High School C	400	512.79	0	115.61	399.62
High School D	1050	524.76	115.61	0	354.16
High School E	800	277.94	399.62	354.16	0

Step 2: Forming a Dendrogram

The merging process begins with the two schools that have the smallest distance. Based on the distance, SMA C and SMA D have the closest distance, which is 115.61, so they are combined into one cluster.

This process is continued by merging clusters that have the closest distance until all schools are combined into one large cluster. The result is a dendrogram that shows the order of cluster merging.

Step 3: Clustering Results

Based on the stepwise merging, we get two main clusters:

- **Cluster 1**: High School A, High School B, High School E
- **Cluster 2**: High School C, High School D

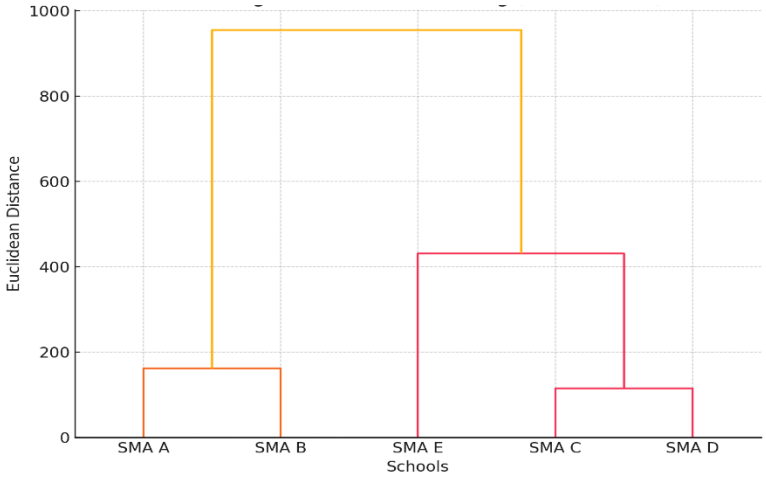


Figure 2. Dendogram of SMA Clustering

5. DISCUSSION

Researchers are currently evaluating the clustering results using the Silhouette Score and Davies-Bouldin Index. These metrics are used to assess the quality of the clusters, with the Silhouette Score analyzing how well data points fit within their respective clusters, while the Davies-Bouldin Index evaluates the average similarity between clusters. Both measures offer valuable insights into the performance of clustering algorithms in identifying meaningful groupings within the data.

A. Silhouette Score

The Silhouette Score evaluates how well a data point fits within its own cluster compared to how far it is from other clusters. The formula used to compute the Silhouette Score for an individual point iii is :

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))} \quad \text{.....(5)}$$

Where:

- a(i)a(i)a(i) is the average distance from point iii to all points in its cluster.
- b(i)b(i)b(i) is the average distance of point iii to points in the other nearest cluster.

K-Means Clustering:

- **Silhouette Score**= 0.52 (indicates fairly good cluster separation).

Hierarchical Clustering:

- **Silhouette Score**= 0.45 (slightly lower, indicating cluster separation is not as good as K-Means).

B. Davies-Bouldin Index

Davies-Bouldin Index measures the average ratio of distance between clusters and the size of the cluster itself. Lower values indicate better clustering.

K-Means Clustering:

Davies-Bouldin Index= 0.88 (lower value indicates more separated clusters).

Hierarchical Clustering:

Davies-Bouldin Index= 1.05 (higher values indicate weaker cluster separation).

C. CONCLUSION

Based on the results of clustering and evaluation using the Silhouette Score and Davies-Bouldin Index, it can be concluded that:

1. K-Means Clusteringprovides more separate and efficient clustering results, with a higher Silhouette Score and a lower Davies-Bouldin Index, indicating better overall cluster quality. K-Means Clustering provides better clustering results in terms of separation between clusters, with a higher Silhouette Score (0.52) and a lower Davies-Bouldin Index (0.88). This indicates that K-Means is more efficient and better at clustering SMAs based on the given data.
2. Hierarchical Clustering, although it provides a clearer picture of the relationships through the dendrogram, produces clusters that are slightly less separated than K-Means. This is reflected in the lower Silhouette Score values and the higher Davies-Bouldin Index. Hierarchical Clustering, although it provides a deeper insight into the relationships between schools (through the dendrogram), in producing Silhouette Score values

Thus, for the clustering of high schools in Surakarta City based on the data provided, K-Means Clustering is proven to be more effective in terms of clear cluster separation and more efficient computing time. However, if the purpose of the analysis is to understand the

relationship between schools in more depth, Hierarchical Clustering can still provide more comprehensive insights.

D. LIMITATION

The limitation in this study is that the data used in the research processing is only based on assumptions and uses a case study of variative data for the purpose of implementing the algorithm. This is because the purpose of this study focuses on the development of methods through a comparative method model.K-Means Clustering and Hierarchical Clustering using case study samples. The main purpose is only to compare the level of accuracy of the two methods through the same case examples tested based on the 2 methods. So this model is a prototype that can be a representation for further development can be used using real data. So that this research can develop and be sustainable.

E. REFERENCES

Borlea, ID, Precup, RE, Borlea, AB, & Iercan, D. (2021). A Unified Form of Fuzzy C-Means and K-Means algorithms and their Partitional Implementation. Knowledge-Based Systems, 214.https://doi.org/10.1016/j.knosys.2020.106731.

Chen, J., Du, C., Zhang, Y., Han, P., & Wei, W. (2022). A Clustering-Based Coverage Path Planning Method for Autonomous Heterogeneous UAVs. IEEE Transactions on Intelligent Transportation Systems, 23(12), 25546–25556.https://doi.org/10.1109/TITS.2021.3066240.

Dutta, A.K., Elhoseny, M., Dahiya, V., & Shankar, K. (2020). An efficient hierarchical clustering protocol for multihop Internet of vehicles communication. Transactions on Emerging Telecommunications Technologies, 31(5).https://doi.org/10.1002/ett.3690.

Shahapure, K. R., & Nicholas, C. (2020). Cluster quality analysis using silhouette score. Proceedings - 2020 IEEE 7th International Conference on Data Science and Advanced Analytics, DSAA 2020, 747–748.https://doi.org/10.1109/DSAA49011.2020.00096.

Sinaga, KP, & Yang, MS (2020). Unsupervised K-means clustering algorithm. IEEE Access, 8, 80716–80727.https://doi.org/10.1109/ACCESS.2020.2988796.

Zeng, K., Ning, M., Wang, Y., & Guo, Y. (2020). Hierarchical clustering with hard-batch triplet loss for person re-identification. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 13654–13662.https://doi.org/10.1109/CVPR42600.2020.01367

Tan, P.-N., Steinbach, M., & Kumar, V. (2019). Introduction to Data Mining (3rd ed.).

Manea, D., & Tomozei, D. (2021). "Comparative Analysis of Clustering Algorithms for Educational Data". Journal of Applied Mathematics and Computational Science, 31(2), 295-312.

Kumar, A., & Ghosh, M. (2020). "A comparative study of K-means, hierarchical, and DBSCAN clustering algorithms in the analysis of urban education data". International Journal of Advanced Computer Science and Applications, 11(6), 314-320.

Zhao, L., & Zhang, J. (2022). "Improving K-means Clustering Algorithm with Advanced Initialization Methods". Computational Intelligence and Neuroscience, 2022, 1-15.

Chen, Z., Liu, C., & Wu, X. (2023). "Hierarchical clustering methods for large-scale educational data analysis: Challenges and solutions". Educational Data Mining, 15(1), 115-132.