Journal of Soft Computing Exploration

Homepage: htttps://shmpublisher.com/index.php/joscex

p-ISSN: 2746-7686
e-ISSN: 2746-0991

.

Comparison of clustering analysis of K-means, K-medoids, and fuzzy C-means methods: case study of school accreditation in west java

Yunia Hasnataeni[1], M Rizky Nurhambali[2], Rizky Ardhani[3], Siti Hafsah[4], Agus M Soleh[5]
[1,2,3,4,5]Department of Statistics, IPB University, Indonesia

Article Info

Article history:

Received May 22, 2025
Revised June 16, 2025
Accepted June 22, 2025

Keywords:

Clustering
Education quality
Fuzzy C-means
K-means
K-medoids

ABSTRACT

This research aims to analyze school accreditation data in West Java using clustering methods: K-Means, K-Medoids, and Fuzzy C-Means, to identify patterns and groups of schools based on similar characteristics. K-Means, known for its simplicity, suggests an optimal two-cluster solution based on silhouette values but employs three clusters for detailed analysis. K-Medoids, noted for its robustness against outliers, achieves the best clustering with a lowest Davies-Bouldin Index (DBI) of 0.8 and the highest Silhouette Information (SI) value of 0.46. Fuzzy C-Means, which assigns membership degrees to each data point across clusters, performs reasonably well with a DBI of 0.87 and an SI value of 0.40, while K-Means shows the highest DBI of 0.9 and the lowest SI value of 0.39. The findings highlight K-Medoids as the superior method for clustering. Regions with lower educational quality, such as Bekasi and Cianjur regions, require priority interventions, whereas areas with better quality, like Bandung and Bekasi regions, can serve as models. Data-driven approaches, inter-regional collaboration, and continuous monitoring and evaluation are recommended to optimize educational policies and enhance overall educational quality in West Java.

Corresponding Author:

Yunia Hasnataeni,
Department of Statistics, IPB University, Indonesia
Jalan Meranti Wing 22 Level 4, Dramaga, Babakan,
Kec. Dramaga, Kabupaten Bogor, Jawa Barat 16680
Email: yunia200698@gmail.com
https://doi.org/10.52465/joscex.v6i2.575

1. INTRODUCTION

Clustering is a multivariate analysis technique that aims to group objects based on the similarity of their characteristics. Objects in one cluster have a high degree of similarity, while between clusters have significant differences [1]. The application of this method has been widespread in various fields such as market segmentation, prediction of business problems, and image processing in computer vision [2]. Clustering-based clustering allows the identification of latent structures in data without dependence on labels, making it useful in the exploration of complex data. Some commonly used methods in clustering analysis include K-Means, K-Medoids, and Fuzzy C-Means, each of which has algorithmic characteristics, advantages, and limitations [3].

K-Means is a partition-based method that works by minimizing the Euclidean distance between data points and the cluster center (centroid). An iterative procedure is performed until the centroid position reaches

convergence [4]. Although computationally efficient, K-Means has a high sensitivity to outliers and is strongly influenced by centroid initialization, thus potentially trapping local optimum solutions [5]. K-Medoids, or Partitioning Around Medoids (PAM), is an alternative to K-Means that uses medoids, which are actual data points, as cluster centers [6]. The robustness of this method to outliers is higher as it does not depend on the mean value. The clustering results are more stable when there is noise in the data [7]. Fuzzy C-Means adopts a soft clustering approach by assigning membership degrees between 0 and 1 to each data to all clusters. This approach allows one object to be partially associated with more than one cluster, making it suitable for data that has indistinct or overlapping cluster boundaries [8].

In the context of education, clustering methods are relevant for grouping schools based on quality characteristics measured through accreditation data. School accreditation reflects the level of fulfillment of the eight national standards of education, such as content standards, learning processes, graduate competencies, educators and education personnel, infrastructure facilities, management, financing, and educational assessment [9], [10]. West Java Province shows significant variations in conditions between regions, both in terms of the quality of teachers, facilities, and the academic achievements of students. The quality gap demands special attention so that policy allocations are more targeted [11].

Previous studies have generally only applied one clustering method without comparing the performance of various algorithms in the context of primary and secondary education. There is no study that systematically conducts a comparative evaluation between K-Means, K-Medoids, and Fuzzy C-Means using school accreditation data at the provincial level. This study fills this gap by providing an empirical analysis of the three clustering methods in the context of education quality in West Java Province.

This study aims to compare the performance of three clustering algorithms, namely K-Means, K-Medoids, and Fuzzy C-Means, in forming structurally uniform school groups. The performance evaluation is conducted using two internal metrics, namely Silhouette Score as an indicator of cluster cohesion and separation, and Davies-Bouldin Index as a measure of efficiency and statistical validity of cluster formation. The final results of this study are expected to provide an empirical basis for strategic decision-making in data-driven education quality improvement.

2. METHOD

In this research, the analysis process is carried out through several stages that have been systematically designed to achieve the research objectives. The series of stages can be seen more clearly in Figure 1 below, which presents the overall research workflow.

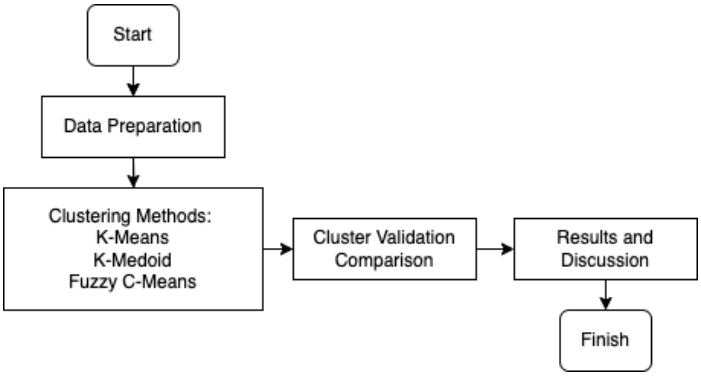

Figure 1. Flowchart of data analysis procedure

In this research, the analysis process was carried out through several systematically designed stages to achieve the research objectives. The overall stages are illustrated in Figure 1, which presents the data analysis workflow. The initial stage was data preprocessing, which involved selecting relevant variables based on theoretical considerations and domain knowledge, handling missing data, and transforming categorical variables into numerical form using one-hot encoding. The data was then normalized using Min-Max scaling so that all features were within a comparable range of values.

After preprocessing, clustering analysis was conducted using three algorithms: K-Means, K-Medoids, and Fuzzy C-Means. K-Means and K-Medoids are partitioning-based algorithms, while Fuzzy C-Means allows

soft clustering where data points can belong to more than one cluster with varying degrees of membership. To evaluate the clustering performance, two internal validation indices were used: the Silhouette Coefficient and the Davies-Bouldin Index, which measure cluster compactness and separation. All stages of the analysis were performed using RStudio, with the support of packages such as cluster, fclust, factoextra, readr, and ggplot2.

Data

The data used in this research were obtained from accreditation documents that must be prepared by schools/madrasasah in West Java. These documents cover the various standards assessed in the accreditation process, including standards on content, process, graduate competencies, educators and education, facilities and infrastructure, management, financing, and assessment. These documents reflect the level of fulfillment of each standard by schools. The variables that will be used in this study are described in Table 1 below:

Tabel 1. Data variables

Variable	Type	Description
Province Name	Categorical	Name of the province where the school is located
City Name	Categorical	Name of the city/regency where the school is located
Education Level	Categorical	Level of education (e.g., SD/MI, SMP/MTS, SMA/MA, SMK)
Education Level Name	Categorical	Specific name of the education level
School Type	Categorical	Type of school (public or private)
Content Standard	Numerical	Score reflecting fulfillment of the content standard
Process Standard	Numerical	Score reflecting fulfillment of the process standard
Graduate Competency Standard	Numerical	Score reflecting fulfillment of graduate competency standard
Teacher Standard	Numerical	Score reflecting fulfillment of teacher and education personnel standard
Facility Standard	Numerical	Score reflecting fulfillment of infrastructure and facility standard
Management Standard	Numerical	Score reflecting fulfillment of management standard
Funding Standard	Numerical	Score reflecting fulfillment of financing standard
Assessment Standard	Numerical	Score reflecting fulfillment of assessment standard

K-Means Clustering

K-Means is one of the algorithms in data mining that can be used to perform data grouping or clustering. This algorithm belongs to the category of distance-based clustering algorithms, where data is divided into a number of clusters based on the similarity between its members. There are various approaches to forming clusters, one of which is to define rules that dictate membership in the same group based on the degree of similarity among its members. K-Means classifies data based on distance, and this algorithm only works on numeric attributes [12]. In calculating the distance between data and points at the cluster center using the euclidian distance equation. The euclidian distance equation is as follows [13]:

$$\mathrm{D(a,b)} = \sqrt{(\mathrm{X1p} - \mathrm{X1q})^2 + (\mathrm{X2p} - \mathrm{X2q})^2 + \cdots + (\mathrm{Xnp} - \mathrm{Xnq})^2} \quad \textbf{(1)}$$

$(\mathrm{a,b})$: Distance from data point p to cluster center q
Xnp : Value of the n-th attribute for data point p
Xnq : Value of the n-th attribute for cluster center q

K-Medoid Cluestering

The K-Medoids algorithm, also known as Partitioning Around Medoids (PAM), is a partition clustering method used to group a set of objects into clusters. Unlike K-Means, the K-Medoids algorithm uses objects in the data set as representatives of each cluster. Objects selected to represent a cluster are called medoids [14]. K-Medoids aims to minimize the total distance between points in the cluster and their medoids, thus providing better robustness against outliers compared to K-Means. In this algorithm, data is clustered with a partitioning system, where the distance between non-medoid and medoid data is calculated. To allocate data to the nearest cluster, K-Medoids uses distance calculation with the euclidean distance method [13].

Fuzzy C- Means Clustering

Fuzzy C-Means Clustering (FCM), also known as Fuzzy Isodata, is one of the clustering methods that is part of the Hard K-Means method. FCM uses a fuzzy clustering model, where data can be a member of all the clusters formed with different degrees or levels of membership, ranging from 0 to 1. This degree of membership determines the extent to which a data belongs to a cluster. FCM allows data to have partial membership in multiple clusters. This means that each data has a certain membership value for each cluster, which reflects its level of presence in that cluster [15]. To find a cluster, you can use the following equation [16]:

$$\mu_{ik} = \frac{[\sum_{j=1}^{m}(x_{ij} - V_{kj})^2]^{\frac{-1}{w-1}}}{\sum_{k=1}^{c}[\sum_{j=1}^{m}(x_{ij} - V_{kj})^2]^{\frac{-1}{w-1}}} \quad (2)$$

μ_{ik} : Random number (or random value)
V_{kj} : Cluster center

Silhouette Score

Silhouette score is a method for evaluating cluster quality by measuring how well the objects in a cluster are grouped. The stages of calculating the silhouette score involve several steps, and the equation can be seen in the equation [17]:

$$S_i = \frac{b(x_i) - a(x_i)}{\max(b(x_i) - a(x_i))} \quad (3)$$

S_i : Silhouette Score
$b(x_i)$: Minimum average distance between the object and all other clusters (i.e., inter-cluster distance)
$a(x_i)$: Average distance between the object and other objects in the same cluster (i.e., intra-cluster distance)

Davies-Bouldin Index

Davies-Bouldin Index (DBI) is a method introduced by David L. Davies and Donald W. Bouldin in 1979. The DBI method assesses cluster results based on the closeness between the data in the cluster. DBI measurement aims to minimize the intra-cluster distance and maximize the inter-cluster distance. The smaller the DBI value, the better the quality of the cluster. The DBI equation formula is as follows [13]:

$$DBI = \frac{1}{M}\sum_{j=1}^{M} \max R_j.k \quad (4)$$

3. RESULTS AND DISCUSSIONS

The following are the results and discussion of clustering in education data analysis is particularly relevant for identifying patterns and groups that have similar characteristics with the K-Means, K-Medoids, and Fuzzy C-Means methods.

K-Means Clustering

Figure 2 below shows the relationship between the number of clusters and the average silhouette value for each number of clusters tested.

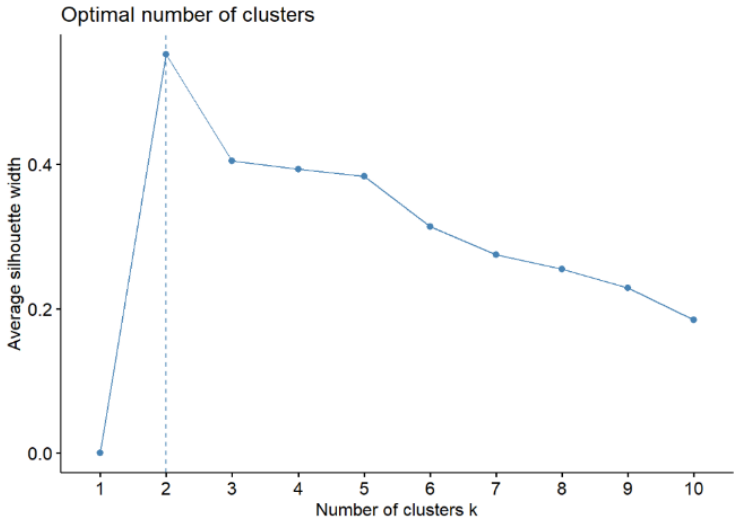

Figure 2. Optimal K-means cluster

The silhouette graph displayed shows the average silhouette value for various numbers of clusters. The silhouette value measures how similar an object is to their own cluster compared to other clusters. This value ranges from -1 to 1, with higher values indicating better clustering. From the graph, the optimal number of clusters is two, as it has the highest average silhouette value among the options considered. After two clusters, the silhouette value decreases as the number of clusters increases, indicating that adding clusters does not significantly improve the clustering quality. A plot of the cluster results is shown in Figure 3 below:

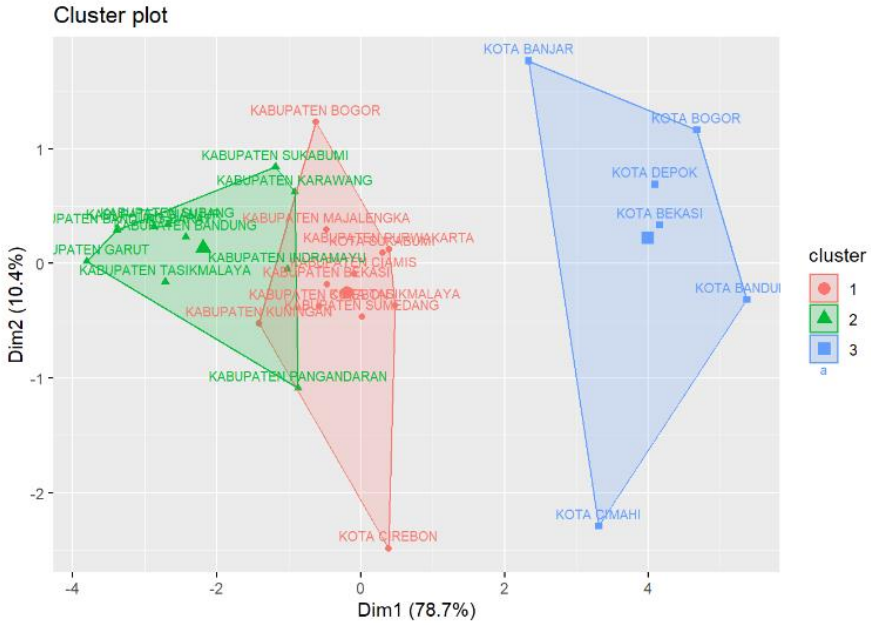

Figure 3. K-means cluster plot

Although the silhouette graph shows that two clusters is the optimal number based on the highest average silhouette value, we chose to use three clusters in this analysis. This decision was based on the consideration that adding one more cluster allows for a more detailed identification of the variation in school accreditation data in West Java. With three clusters, we can better capture the differences in education quality across regions, as well as identify areas that require special attention or that can serve as examples for other regions.

The clustering results using the K-Means method on school accreditation data in West Java show the division into three clusters. The first cluster (red) includes schools in areas such as Bekasi Regency, Cianjur Regency, and Sumedang Regency, which show relatively low education quality and require more attention for quality improvement. The second cluster (green) includes schools from regions such as Bandung Regency,

Garut Regency, Tasikmalaya Regency, Kuningan Regency, and Majalengka Regency. The schools in this cluster have varying education quality but tend to be average, with some good standards and some that need improvement. The third cluster (blue) includes schools from big regions such as Bandung City, Bogor City, Depok City, Bekasi City and Cimahi City, which show better education quality and can be used as examples for other regions.

K-Medoids Clustering

Figure 4 below shows the relationship between the number of clusters and the average silhouette value for each number of clusters tested.

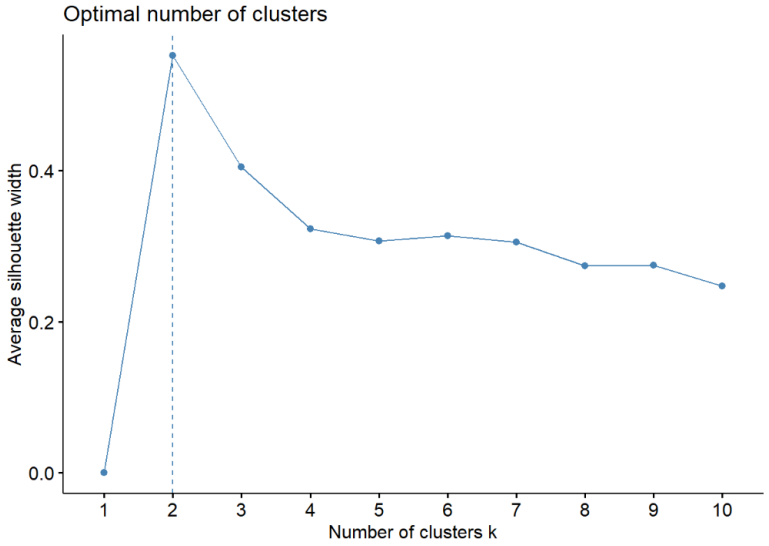

Figure 4. Optimal K-medoids cluster

Based on the results of the analysis using the K-Medoids algorithm and the silhouette method, it was found that the optimal number of clusters for school accreditation data in West Java is two clusters. This indicates that schools in West Java can be divided into two main groups that have quite different characteristics in terms of their accreditation. This clustering can help in identifying factors that affect the quality of education and provide policymakers with deeper insights to improve the quality of education in the area. However, to enable a more detailed identification of the variation in school accreditation data in West Java we decided to add one more cluster. The plot of the cluster results is shown in Figure 5 below:

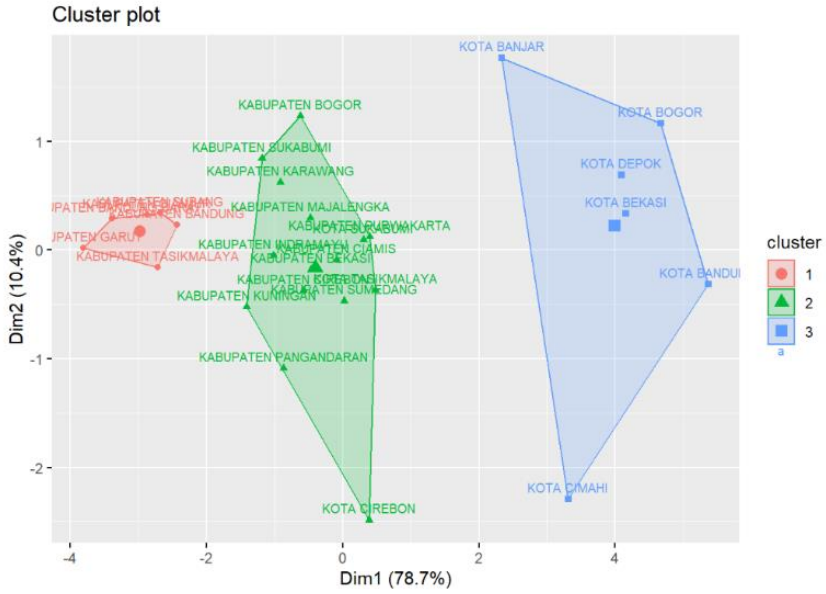

Figure 5. K-medoids cluster plot

Based on analysis using the K-Medoids algorithm, school accreditation data in West Java is grouped into three main clusters. These clusters help in identifying patterns and characteristics underlying the quality of education in different regions and regions in West Java. The resulting medoids represent the center of each cluster with values for various education standards. The first cluster includes Bandung, West Bandung and Cianjur regions that have some lower standards and may need more attention in education quality improvement efforts. The second cluster includes Majalengka district, Bogor district and Cirebon city which show better results but there are still areas that need improvement. The third cluster includes Bekasi City, Cimahi City and Depok City which showed very good results in almost all standards, suggesting that education policies and practices in these areas can be used as examples for other areas.

Fuzzy C-Means Clustering

Figure 6 below shows the relationship between the number of clusters and the average silhouette value for each number of clusters tested.

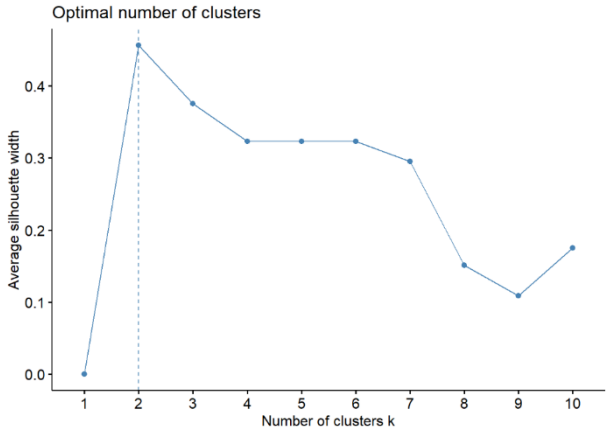

Figure 6. Optimal fuzzy C-means cluster

Based on the results of the analysis using the Fuzzy C-Means algorithm and the silhouette method, it was found that the optimal number of clusters for school accreditation data in West Java is two clusters. The plot of cluster results is shown in Figure 7 below:

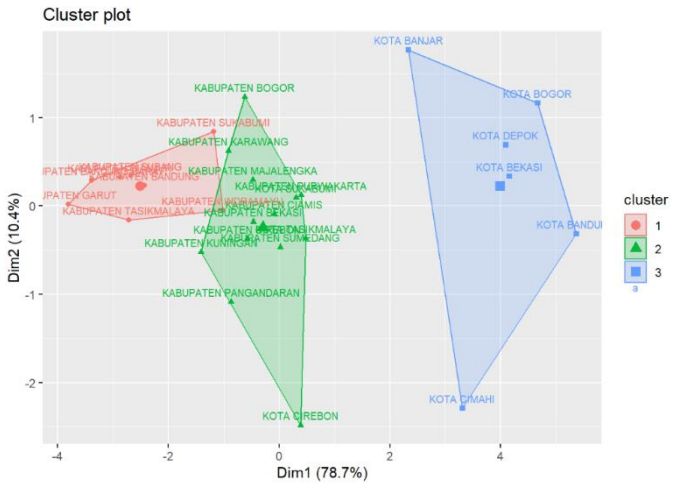

Figure 5. Fuzzy C-means cluster plot

The results of clustering with the Fuzzy C-Means method on school accreditation data in West Java are divided into 3 clusters. The first cluster (red) consists of schools in areas such as Bandung Regency and Tasikmalaya Regency that have lower education quality. The second cluster (green) includes areas such as Bogor Regency and Karawang Regency with varying but average education quality. The third cluster (blue) includes regions such as Bandung City and Bekasi City that show the best quality of education.

Silhouette Score

Figure 8 below is the result of the comparison of Silhouette Score of K-Means, K-Medoids, and Fuzzy C-Means methods:

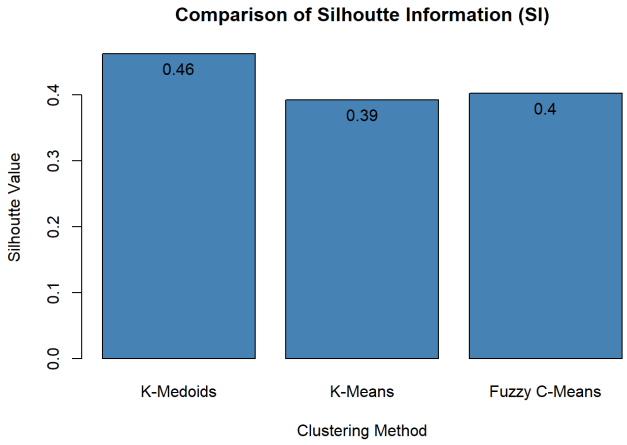

Figure 8. Comparison of silhouette score

The graph above shows the comparison of Silhouette Information values for K-Medoids, K-Means, and Fuzzy C-Means clustering methods. K-Medoids has the highest Silhouette value of 0.46, showing the best clustering results among the three methods. K-Means has the lowest value of 0.39, indicating poor clustering performance, possibly caused by sensitivity to outliers and initial centroid selection. Fuzzy C-Means has a Silhouette value of 0.40, better than K-Means but still below K-Medoids. Thus, K-Medoids proved to be more reliable in producing better clustering for school accreditation data in West Java.

Davies-Bouldin Index

Figure 9 below is the Davies-Bouldin Index comparison result of K-Means, K-Medoids, and Fuzzy C-Means methods:

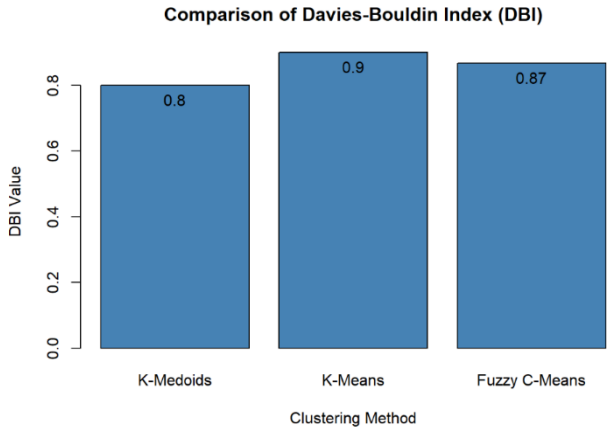

Figure 9. Comparison of davies-bouldin index

The graph above compares the Davies-Bouldin Index (DBI) values for three clustering methods: K-Medoids, K-Means, and Fuzzy C-Means. From the comparison of Davies-Bouldin Index (DBI) values, it can be concluded that the K-Medoids method produces the best quality clustering among the three methods tested, as it has the lowest DBI value (0.8). Fuzzy C-Means shows fairly good results with a DBI value of 0.87, while K-Means has the highest DBI value (0.9), indicating slightly less good clustering quality than the other two methods.

4. CONCLUSION

Clustering analysis of school accreditation data in West Java using K-Means, K-Medoids, and Fuzzy C-Means shows that the three clusters provide a more detailed understanding of the quality of education in different regions. Based on the DBI and SI values, the K-Medoids method is the best clustering method followed by the Fuzzy C-Means and K-Means methods. K-Medoids has the lowest DBI value of 0.80 and the highest SI of 0.46. To improve the quality of education, low-quality areas such as Bekasi and Cianjur regions need priority intervention, while good-quality areas such as Bandung and Bekasi can be modeled. Data-driven approaches, inter-regional collaboration and continuous monitoring and evaluation are essential to optimize education policies and improve the overall quality of education in West Java. Therefore, the research objectives have been successfully achieved, as the study was able to compare the performance of clustering methods and identify meaningful regional groupings to support policy recommendations.

REFERENCES

[1] A. Aditya, I. Jovian, and B. N. Sari, "Implementasi K-Means Clustering Ujian Nasional Sekolah Menengah Pertama di Indonesia Tahun 2018/2019," *J. MEDIA Inform. BUDIDARMA*, vol. 4, no. 1, p. 51, Jan. 2020, doi: 10.30865/mib.v4i1.1784.

[2] A. F. Mohamed Nafuri, N. S. Sani, N. F. A. Zainudin, A. H. A. Rahman, and M. Aliff, "Clustering Analysis for Classifying Student Academic Performance in Higher Education," *Appl. Sci.*, vol. 12, no. 19, p. 9467, Sep. 2022, doi: 10.3390/app12199467.

[3] Annisa Nadaa Shabrina, M. Afdal, and Siti Monalisa, "Comparison Of K-Means, K-Medoids, and Fuzzy C-Means Algorithms for Clustering Drug User's Addiction Levels," *J. Sist. Cerdas*, vol. 6, no. 2, pp. 113–122, Aug. 2023, doi: 10.37396/jsc.v6i2.313.

[4] M. A. Syakur, B. K. Khotimah, E. M. S. Rochman, and B. D. Satoto, "Integration K-Means Clustering Method and Elbow Method For Identification of The Best Customer Profile Cluster," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 336, p. 012017, Apr. 2018, doi: 10.1088/1757-899X/336/1/012017.

[5] M. E. Celebi, H. A. Kingravi, and P. A. Vela, "A comparative study of efficient initialization methods for the k-means clustering algorithm," *Expert Syst. Appl.*, vol. 40, no. 1, pp. 200–210, Jan. 2013, doi: 10.1016/j.eswa.2012.07.021.

[6] E. Herman, K.-E. Zsido, and V. Fenyves, "Cluster Analysis with K-Mean versus K-Medoid in Financial Performance Evaluation," *Appl. Sci.*, vol. 12, no. 16, p. 7985, Aug. 2022, doi: 10.3390/app12167985.

[7] E. Schubert and P. J. Rousseeuw, "Fast and Eager k-Medoids Clustering: O(k) Runtime Improvement of the PAM, CLARA, and CLARANS Algorithms," *Inf. Syst.*, vol. 101, p. 101804, Nov. 2021, doi: 10.1016/j.is.2021.101804.

[8] S. Sarbaini, W. Saputri, Nazaruddin, and F. Muttakin, "Cluster Analysis Menggunakan Algoritma Fuzzy K-Means Untuk Tingkat Pengangguran Di Provinsi Riau," *J. Teknol. dan Manaj. Ind. Terap.*, vol. 1, no. 2, pp. 78–84, Jun. 2022, doi: 10.55826/tmit.v1iII.30.

[9] M. A. E. Rahadianto, A. D. Sakti, and K. Wikantika, "Evaluasi Kualitas Infrastruktur Fasilitas Pendidikan Dasar di Provinsi Jawa Barat Indonesia Menggunakan Pendekatan Berbasis Model Multi-Hazard dan Aksesibilitas," in *Seminar Nasional Geomatika 2020: Informasi Geospasial untuk Inovasi Percepatan Pembangunan Berkelanjutan*, 2020.

[10] S. A. Nurfatimah, S. Hasna, and D. Rostika, "Membangun Kualitas Pendidikan di Indonesia dalam Mewujudkan Program Sustainable Development Goals (SDGs)," *J. Basicedu*, vol. 6, no. 4, pp. 6145–6154, May 2022, doi:

10.31004/basicedu.v6i4.3183.
[11] C. Mongi and C. Montolalu, "Penggerombolan Sekolah Menengah Atas Berdasarkan Nilai Ujian Nasional Di Kota Manado," *d'CARTESIAN*, vol. 6, no. 2, p. 80, Aug. 2017, doi: 10.35799/dc.6.2.2017.17969.
[12] W. M. P. Dhuhita, "Clustering menggunakan metode K-Means untuk menentukan status gizi balita," *J. Inform.*, vol. 15, no. 2, 2015.
[13] A. A. Harahap, M. Raihan, N. Amani, and P. R. Andini, "Comparison of Unsupervised Learning Techniques for Clustering Data on the Number of Villages in Indonesia," in *SENTIMAS: Seminar Nasional Penelitian dan Pengabdian Masyarakat*, 2023, pp. 163–170.
[14] S. Nurlaela, A. Primajaya, and T. N. Padilah, "Algoritma K-Medoids Untuk Clustering Penyakit Maag Di Kabupaten Karawang," *INFORMATIKA*, vol. 12, no. 2, 2020.
[15] D. L. Rahakbauw, V. Y. I. Ilwaru, and M. H. Hahury, "IMPLEMENTASI FUZZY C-MEANS CLUSTERING DALAM PENENTUAN BEASISWA," *BAREKENG J. Ilmu Mat. dan Terap.*, vol. 11, no. 1, pp. 1–12, Mar. 2017, doi: 10.30598/barekengvol11iss1pp1-12.
[16] R. D. L. N. Karisma, T. S. Arinda, H. Widayani, and A. Kusumastuti, "Clustering of COVID-19 Provinces in Indonesia Using Fuzzy Means Cluster Methods," 2023, pp. 394–406. doi: 10.2991/978-94-6463-148-7_39.
[17] 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, Jun. 2021, doi: 10.3390/e23060759.

.