Jurnal Varian
Vol. 7, No. 2, April 2024, Hal. 191–198
e-ISSN: 2581-2017

An Implementation of K-Medoids Method in Provincial Clustering Based on Education Indicator Data

Annisa Zuhri Apridayanti[1], M. Fathurahman[1], Surya Prangga[1]
[1]Department of Statistics, Mulawarman University, Indonesia

Article Info

Article history:

Received : 07-23-2023
Revised : 06-30-2024
Accepted : 06-30-2024

Keywords:

Data Mining;
Clustering;
Outlier;
K-Medoids;
Davies-Bouldin index; Education.

ABSTRACT

K-Medoids is an essential method in data mining clustering, which have the outliers. This research aims to apply the K-Medoids method to grouping provinces in Indonesia based on education indicator data in 2021. The optimum number of clusters obtained from the Davies-Bouldin index, whereas similarity measure used the Euclidean distance. The number of groups proposed were K=2,3,,10, and the optimal number of clusters were determined by the lowest of the Davies-Bouldin index. The result shows that the optimal number of clusters is five. These five clusters have characteristics that differentiate one cluster from other clusters.

Accredited by Kemenristekdikti, Decree No: 200/M/KPT/2020
DOI: https://doi.org/10.30812/varian.v7i2.3205

Corresponding Author:

M. Fathurahman,
Department of Statistics, Mulawarman University Email: fathur@fmipa.unmul.ac.id

How to Cite:
Annisa Zuhri Apridayanti, M. Fathurahman, Surya Prangga. (2024). An Implementation of K-Medoids Method in Provincial Clustering Based on Education Indicator Data. Jurnal Varian, 7(2), 191-198.

Journal Homepage: https://journal.universitasbumigora.ac.id/index.php/Varian

A. INTRODUCTION

Clustering is a data mining method that divides data into groups that have objects with the same characteristics. The K-Medoids method is widely used in data mining clustering based on the partitioning clustering framework (Han et al., 2022). The K-Medoid method is also called the Partition Around Medoids (PAM) method, which was discovered by Leonard Kaufman and Peter J. Rousseeuw in 1990 (Satoto et al., 2015).

The K-Medoids method has several advantages, namely being robust against outlier data and having good performance for large datasets. (Kaur et al., 2014), (Arora et al., 2016). Several studies have examined and developed the K-Medoids method. (Syukra et al., 2019) developed and implemented the K-Medoids method with the FP-Growth algorithm. Sureja et al. (2022) developed the K-Medoids method using the crow search algorithm. Effendie and Kariyam (2023) used the deviation ratio index to determine the optimum number of clusters in the K-Medoids method.

The K-Medoids method in this research is applied to grouping provinces in Indonesia based on education indicator data in 2021.

Education is an indicator of human development in a province. If a province has good- quality education, then human development in that province will be advanced and of high quality, which will have an impact on the progress and improvement of the quality of human development in Indonesia. The existence of disparities in the quality of education in several provinces that are not evenly distributed is an obstacle and challenge to human development in Indonesia (Badan Pusat Statistik, 2021).

Many studies have been carried out applying the K-Medoids method to educational data. Grouping of districts and cities in South Sulawesi and West Sulawesi Provinces based on high school and equivalent education participation rates (AS et al., 2019). Grouping of study program selection patterns for new students at Kanjuruhan University, Malang (Wira et al., 2019). (Qona'ah et al., 2020) grouping laboratories. (Sinaga et al., 2022) Grouping the ratio of students to teachers, the ratio of students to study groups, the ratio of study groups to elementary school education classes, and the ratio of study groups to junior high school education classes by province. Grouping of sub-districts in Bojonegoro Regency based on educational supporting factors (Kartini and Husen, 2023). Grouping of independent campus learning program (MBKM) survey results (Mayadi et al., 2023). Grouping of districts and cities in South Sulawesi Province based on education indicators (Raja, 2020).

Meanwhile, research applying the K-Medoids method to the grouping of provinces in Indonesia based on education indicator data has been carried out by using the Silhoutte index to determine the number of clusters and Euclidean distance to measure the closeness between objects (Fialine et al., 2021). The results of this research obtained an optimum number of clusters of 3 clusters, namely the first cluster consisting of 6 provinces with high education indicator characteristics; the second cluster consists of 13 provinces with medium education indicator characteristics; and cluster 3 consists of 15 provinces with low education indicator characteristics. Research conducted by (Septian and Darnah, 2023) also used the Silhoutte index and Euclidean distance to determine the optimum number of clusters and measure the proximity between objects. Based on the results of grouping using the K-Medoids method, the optimum number of clusters was obtained as 2 clusters, namely the first cluster consisting of 14 provinces and cluster 2 consisting of 20 provinces.

Furthermore, (Rahmawati and Fauzan, 2024) apply the K-Medoids method to handle outlier data using three methods: the K-Medoids method without treatment, K-Medoids with mean trimming, and K-Medoids with min-max trimming. Optimum number of clusters using the Silhoutte index. This research resulted in the three proposed methods having relatively similar Silhoutte index values, and referring to the parsimony principle, the K-Medoids without treatment method is recommended for grouping provinces in Indonesia based on education indicator data. The optimum number of clusters formed was 2 clusters, namely cluster 1 consisting of 10 provinces with high education indicator characteristics and cluster 2 consisting of 24 provinces with low education indicator characteristics.

This research aims to apply the K-Medoids method to grouping provinces in Indonesia based on education indicator data by determining the optimum number of clusters using the Davies-Bouldin index. This index was still rarely used in previous studies using the K-Medoids method. The distance used is the Euclidean distance, which has been widely used in previous research. Most of the educational indicator data used in this research refers to previous studies, but there are additional educational indicators that are different from previous studies, namely the percentage of school libraries, the percentage of school laboratories, the ratio of students per teacher, and the percentage of well-owned classrooms.

B. RESEARCH METHOD

1. K-Medoids

The K-Medoids method is a method that represents clusters formed using medoids. Clusters are formed by calculating the proximity between medoids and non-medoid objects using a distance measure. So, the partition method can still be carried out based on the principle of minimizing the number of dissimilarities between each object and its corresponding medoid. The stages of the K-Medoids algorithm are as follows (Purba, Saifullah, & Dewi, 2019):

1. Randomly initialize the cluster center with as many K objects as representative objects o_m (medoids).
2. Calculating the Euclidean distance for each object against each medoid using the following equation:

$$d(x_{ij}, o_{mj}) = \sqrt{(x_{i1}, o_{m1})^2 + (x_{i2}, o_{m2})^2 + \ldots (x_{iq}, o_{mq})^2} \quad (1)$$

where $d(x_{ij}, o_{mj})$ is the distance between the data at ith observation and j-th observation on the m-th medoid for $i = 1, 2, , n$; $j = 1, 2, , q$ and $m = 1, 2, , k$.

3. Assigns each object to the cluster closest to its medoids and calculates the objective function, which is the sum of the proximity of all objects to the nearest medoids based on the minimum distance between objects to each medoid.

4. Selecting cluster centers randomly as many as K objects as non-representative objects o_h (non-medoids).
5. Calculating the Euclidean distance for each object against each of the non-medoids using the following formula:

$$d(x_{ij}, o_{hj}) = \sqrt{(x_{i1}, o_{h1})^2 + (x_{i2}, o_{h2})^2 + \ldots (x_{iq}, o_{hq})^2} \tag{2}$$

where $d(x_{ij}, o_{hj})$ is the distance between the data at ith observation and j-th observation on the h-th medoid for $i = 1, 2, , n$; $j = 1, 2, , q$ and $h = 1, 2, , k$.
6. Place each object into the cluster closest to its non-medoids and calculate the objective function, which is the sum of the proximity of all objects to the nearest non-medoids based on the minimum distance between objects to each non-medoids.
7. Compute the difference between the objective function and the non-medoid objective function and the non-medoid objective function by subtracting the non-medoid objective function from the medoid objective function.
8. Replace medoids with non-medoids if the objective function value is < 0.
9. Repeat steps d and g until there are no more medoids changes.
10. After clustering, there is no change in the representative object.
11. Determine the optimum cluster using the Davies-Bouldin index

2. Davies-Bouldin Index

The Davies-Bouldin index is a method that can be used to maximize the distance between one cluster and another and, at the same time, try to minimize the distance between objects in a cluster (Badruttamam et al., 2020). The clustering with the best number of clusters is the clustering that has the lowest Davies-Bouldin index value. The Davies-Bouldin index value is formulated as follows:

$$DBI = \frac{1}{K} \sum_{k=1}^{K} R_K \tag{3}$$

with

$$R_K = \max_{l=1,2,\ldots,K, k \neq 1} R_{kl}, \quad R_{kl} = \frac{S_k + S_l}{D(T_k, T_l)} \tag{4}$$

where K is the number of clusters; R_{kl} is a measure of similarity between the k-th cluster and the l-th cluster; and S_k is a measure of dispersion of the k-th cluster-k, $k = 1, 2, \ldots, K$, and defined as follows:

$$S_k = \left[\frac{1}{n_k} \sum_{U_i \in V_k, i=1}^{n_k} D^2(U_i, T_k) \right]^{\frac{1}{2}}, \quad D^2(U_i, T_k) = (D(U_i, T_k))^2 \tag{5}$$

where n_k is the number of k-th member cluster, $k = 1, 2, \ldots, K$; V_k is the k-th cluster; U_i is the i-th member of the k-th cluster; $D(U_i, T_k)$ is the distance from the i-th member of the k-th cluster (U_i) to the k-th cluster centroid (T_k) which can be obtained with a simple matching dissimilarity measure as follows:

$$D(U_i, T_k) = \sum_{m=1}^{M} \delta(x_{im}, t_{km}) \tag{6}$$

with

$$\delta(x_{im}, t_{km}) = \begin{cases} 0 & , x_{im} = t_{km} \\ 1 & , x_{im} \neq t_{km} \end{cases} \tag{7}$$

where x_{im} is the m-th value at i-th variable of U; t_{km} is the m-th value at the centroid cluster; and M is the number of cluster.

$D(T_k, T_l)$ is the distance between the k-th cluster centroid (T_k) and the l-th cluster centroid (T_l) can be obtained by:

$$D(T_k, T_i) = \sum_{m=1}^{M} \delta(t_{km}, t_{lm}) \tag{8}$$

DOI: https://doi.org/10.30812/varian.v7i2.3205

with

$$\delta(t_{km}, t_{lm}) = \begin{cases} 0 & , t_{km} = t_{lm} \\ 1 & , t_{km} \neq t_{lm} \end{cases} \tag{9}$$

where t_{lm} is the m-th value at the l-th cluster centroid.

3. Data Sources and Research Variables

The data used in this research was secondary data obtained from the Ministry of Education, Culture, Research, and Technology of the Republic of Indonesia. The research unit was in 34 provinces in Indonesia in 2021, whereas the research variables are presented in Table 1.

Table 1. Research Variables

Variable	Description	Measurement Scale
X_1	Percentage of school libraries	Ratio
X_2	Percentage of school laboratories	Ratio
X_3	School participation rate	Ratio
X_4	Pure participation rate	Ratio
X_5	Gross participation rate	Ratio
X_6	Repeating rate	Ratio
X_7	Drop-out rate	Ratio
X_8	The average length of school	Ratio
X_9	The ratio student per teacher	Ratio
X_{10}	Percentage of well-owned classrooms	Ratio

4. Data Analysis Techniques

The techniques of data analysis in this research are as follows:

1. Detecting the outlier data using a boxplot.
2. Transforming data using z-score normalization.
3. Detecting multicollinearity between the variables using the Variance Inflation Factor (VIF) criteria.
4. Clustering data using the K-Medoids method.
5. Getting to the conclusion.

C. RESULT AND DISCUSSION

This section begins by detecting outliers in the research variable data using a boxplot. The result was displayed in Figure 1. In Figure 1, it appears that the research variable data contains outliers, as in the variables: percentage of school laboratories (X_2), pure participation rate (X_4), and percentage of well-owned classrooms (X_{10}).

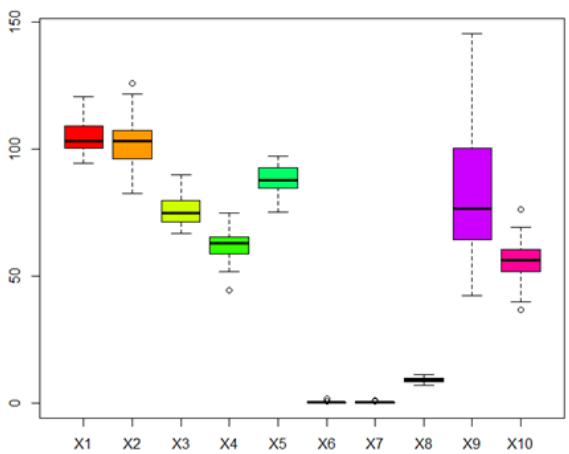

Figure 1. The distribution of research variable data

The next step is detecting multicollinearity between the variables. One of the assumptions in clustering using the K-Medoids method is that there is no multicollinearity between variables. The results of multicollinearity detection are presented in Table 2.

Table 2. The VIF Value of Variables

Variable	Description	Measurement Scale
X_1	Percentage of school libraries	1.51
X_2	Percentage of school laboratories	2.56
X_3	School participation rate	4.77
X_4	Pure participation rate	7.48
X_5	Gross participation rate	4.45
X_6	Repeating rate	2.13
X_7	Drop-out rate	1.41
X_8	The average length of school	3.02
X_9	The ratio student per teacher	2.46
X_{10}	Percentage of well-owned classrooms	2.75

Based on Table 2, all variables have a VIF value less than 10. It shows that there is no multicollinearity between variables. Therefore, all variables can be used for the process of clustering provinces in Indonesia based on the educational indicators using the K-Medoids method.

Furthermore, transform the research variable data using z-score normalization and determine the optimal number of clusters. The Davies-Bouldin index was used to determine and validate the optimal number of clusters, and which is shown in Table 3.

Table 3. DBI Values of the Optimal Number of Cluster

K	DBI
2	1.78
3	1.91
4	1.5
5	1.31
6	1.57
7	1.36
8	1.32
9	1.56
10	1.77

Based on Table 3, the DBI value for validating data from the results of clustering provinces in Indonesia based on education indicators using the K-Medoids method has different values. The smallest value is the clustering of 5 clusters, namely 1.31. Therefore, it can be concluded that the most optimal clustering using the K-Medoids method is at 5 clusters. Clustering the provincial data based on education indicators in Indonesia in 2021 using the K-Medoids method is shown in Table 4

Table 4. The Number of Cluster and The Member of Cluster

Cluster	The Number of Clusters	The Member of the Cluster
1	1	DKI Jakarta
2	15	West Java, Central Java, East Java, Jambi, South Sumatra, Lampung, West Kalimantan, Central Kalimantan, South Kalimantan, South Sulawesi, East Nusa Tenggara, Banten, Bangka Belitung Islands, Gorontalo and West Sulawesi
3	9	DI Yogyakarta, Aceh, West Sumatra, Riau, East Kalimantan, Central Sulawesi, Bali, Bengkulu and the Riau Island
4	8	North Sumatra, North Sulawesi, Southeast Sulawesi, Maluku, West Nusa Tenggara, North Maluku, West Papua and North Kalimantan
5	1	Papua

Based on Table 4, Cluster 1 consists of 1 province, namely DKI Jakarta Province, which has a higher average dropout rate, average length of schooling, and well-owned classrooms compared to other clusters. Cluster 2 consists of 15 provinces, namely West Java, Central Java, East Java, Jambi, South Sumatra, Lampung, West Kalimantan, Central Kalimantan, South Kalimantan, South Sulawesi, East Nusa Tenggara, Banten, Bangka Belitung Islands, Gorontalo, and West Sulawesi. It has an average percentage

DOI: https://doi.org/10.30812/varian.v7i2.3205

of school libraries lower than clusters 3 and 4 and higher than clusters 1 and 5. Cluster 3 consists of 9 provinces, namely DI Yogyakarta, Aceh, West Sumatra, Riau, East Kalimantan, Central Sulawesi, Bali, Bengkulu, and the Riau Islands, which have an average percentage of school libraries, school laboratories, net enrollment rates, and school enrollment rates that are higher than the other clusters. Cluster 4 consists of 8 provinces, namely North Sumatra, North Sulawesi, Southeast Sulawesi, Maluku, West Nusa Tenggara, North Maluku, West Papua, and North Kalimantan, which have an average gross enrollment rate that is higher than the other clusters. Cluster 5 consists of 1 province, namely Papua, which has a higher average repetition rate and student-per-teacher ratio than other clusters. Spatial mapping of clustering the provinces based on educational indicator data using K-Medoids is displayed in Figure 2.

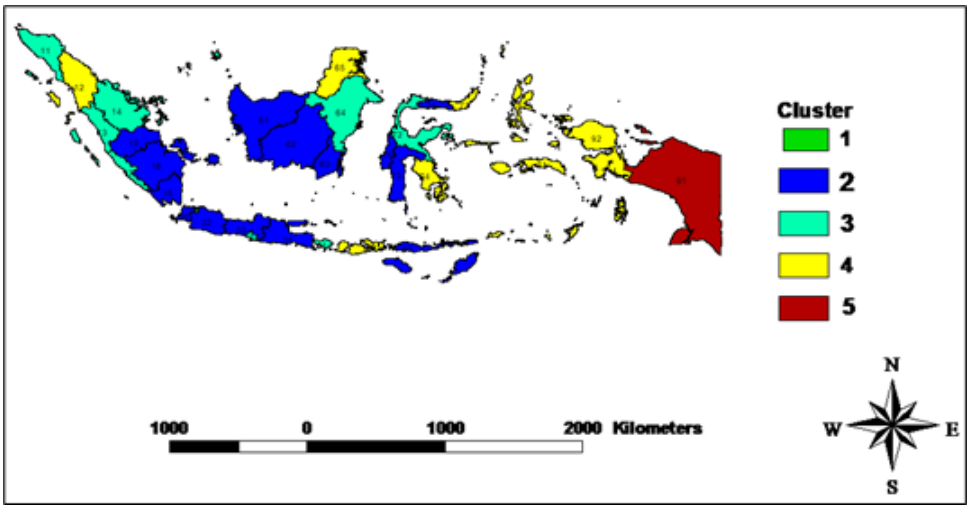

Figure 2. Spatial mapping of provincial clustering based on education indicators in Indonesia in 2021

D. CONCLUSION AND SUGGESTION

Based on the results and discussion, the conclusions and suggestions that are the education indicator data in Indonesia in 2021 contains outlier data, so it can be analyzed using the K-Medoids method. The optimum number of clusters obtained from the Davies-Bouldin validation results was five clusters. The number of members in cluster 1 is 1 province, cluster 2 is 15 provinces, cluster 3 is 9 provinces, cluster 4 is 8 provinces, and cluster 5 is 1 province. This research can still be continued by adding other variables as education indicators so that more optimal results can be obtained. Apart from that, further research can use other methods, for example, CLARA as an alternative method.

E. DECLARATIONS

AUTHOR CONTIBUTION

All authors contributed to the writing of this article.

FUNDING STATEMENT

- COMPETING INTEREST

All authors involved in completing this paper are not in a conflict of interest that results in mutual downfall. We, the authors, are verysupportive of each other. Towards the journal editor, we have no interest whatsoever. We declare clean from conflicts of interest.

REFERENCES

Arora, P., Varshney, S., et al. (2016). Analysis of k-means and k-medoids algorithm for big data. *Procedia Computer Science*, 78:507–512. https://doi.org/10.1016/j.procs.2016.02.095.

AS, W., Aidid, M. K., and Nusrang, M. (2019). Pengelompokan kabupaten/kota provinsi sulawesi selatan dan barat berdasarkan angka partisipasi pendidikan sma/smk/ma menggunakan k-medoid dan clara. *VARIANSI: Journal of Statistics and Its application on Teaching and Research*, 1(3):48.

Badan Pusat Statistik (2021). Indeks pembangunan manusia 2021.

Badruttamam, A., Sudarno, S., and Di Asih, I. M. (2020). Penerapan analisis klaster k-modes dengan validasi davies bouldin index dalam menentukan karakteristik kanal youtube di indonesia (studi kasus: 250 kanal youtube indonesia teratas menurut socialblade). *Jurnal Gaussian*, 9(3):263–272. https://doi.org/10.14710/j.gauss.9.3.263-272.

Effendie, A. R. and Kariyam, A. (2023). A medoid-based deviation ratio index to determine the number of clusters in a dataset. *MethodsX*, 10:102084. https://doi.org/10.1016/j.mex.2023.102084.

Fialine, A., Alodia, D., Endriani, D., and Widodo, E. (2021). Implementasi metode k-medoids clustering untuk pengelompokan provinsi di indonesia berdasarkan indikator pendidikan. *Journal of Mathematics Education and Applied*, 2(2).

Han, J., Pei, J., and Tong, H. (2022). *Data mining: concepts and techniques*. Morgan kaufmann.

Kartini, A. Y. K. and Husen, S. (2023). Comparison of k-means and k-medoids clustering for grouping the sub-districts in bojonegoro regency based on educational supporting factors. *J Statistika: Jurnal Ilmiah Teori dan Aplikasi Statistika*, 16(2):514–523. https://doi.org/10.36456/jstat.vol16.no2.a8079.

Kaur, N. K., Kaur, U., and Singh, D. (2014). K-medoid clustering algorithm-a review. *Int. J. Comput. Appl. Technol*, 1(1):42–45.

Mayadi, M., Setiawati, S., and Priatna, W. (2023). Pengelompokan hasil survei mbkm menggunakan k-mean dan k-medoids clustering. *JURNAL MEDIA INFORMATIKA BUDIDARMA*, 7(1):426–435. http://dx.doi.org/10.30865/mib.v7i1.5003.

Qona'ah, N., Devi, A. R., and Dana, I. M. G. M. (2020). Laboratory clustering using k-means, k-medoids, and model-based clustering. *Indonesian Journal of Applied Statistics*, 3(1):64–77. https://doi.org/10.13057/ijas.v3i1.40823.

Rahmawati, O. and Fauzan, A. (2024). Provincial clustering based on education indicators: K-medoids application and k-medoids outlier handling. *BAREKENG: Jurnal Ilmu Matematika dan Terapan*, 18(2):1167–1178. https://doi.org/10.30598/barekengvol18iss2pp1167-1178.

Raja, N. A. (2020). *IMPLEMENTASI ALGORITMA CENTROID LINKAGE SAN K-MEDOIDS DALAM MENGELOMPOKKAN KABUPATEN/KOTA DI SULAWESI SELATAN BERDASARKAN INDIKATOR PENDIDIKAN*. PhD thesis, Universitas Hasanuddin.

Satoto, B. D., Khotimah, B. K., and Iswati, I. (2015). Pengelompokan wilayah madura berdasar indikator pemerataan pendidikan menggunakan partition around medoids dan validasi adjusted random index. *J. Inf. Syst. Eng. Bus. Intell*, 1(1):17–24.

Septian, R. and Darnah, D. (2023). Penerapan algoritma k-medoids pada pengelompokan wilayah provinsi di indonesia berdasarkan indikator pendidikan. *EKSPONENSIAL*, 14(2):85–90. https://doi.org/10.30872/eksponensial.v14i2.1150.

Sinaga, D. M., Windarto, A. P., and Hartama, D. (2022). Analisis k-medoids dalam pengelompokkan rasio murid dengan guru, murid dengan rombel, dan rasio rombel dengan kelas jenjang pendidikan sd dan smp menurut provinsi. *Jurnal Riset Teknik Informatika dan Data Sains*, 1(1):1–6.

Sureja, N., Chawda, B., and Vasant, A. (2022). An improved k-medoids clustering approach based on the crow search algorithm. *Journal of Computational Mathematics and Data Science*, 3:100034. https://doi.org/10.1016/j.jcmds.2022.100034.

Syukra, I., Hidayat, A., and Fauzi, M. Z. (2019). Implementation of k-medoids and fp-growth algorithms for grouping and product offering recommendations. *Indonesian Journal of Artificial Intelligence and Data Mining*, 2(2):107–115.

Wira, B., Budianto, A. E., and Wiguna, A. S. (2019). Implementasi metode k-medoids clustering untuk mengetahui pola pemilihan program studi mahasiswa baru tahun 2018 di universitas kanjuruhan malang. *Rainstek: Jurnal Terapan Sains & Teknologi*, 1(3):53–68. https://doi.org/10.21067/jtst.v1i3.3046.

.

Vol. 7, No. 2, April 2024, Hal. 191–198
DOI: https://doi.org/10.30812/varian.v7i2.3205