BAREKENG: Journal of Mathematics and Its Applications
June 2024 Volume 18 Issue 2 Page 1167–1178
P-ISSN: 1978-7227 E-ISSN: 2615-3017
https://doi.org/10.30598/barekengvol18iss2pp1167-1178

PROVINCIAL CLUSTERING BASED ON EDUCATION INDICATORS: K-MEDOIDS APPLICATION AND K-MEDOIDS OUTLIER HANDLING

Octavia Rahmawati [1], Achmad Fauzan[2*]

[1,2]*Statistics Study Program, Faculty of Mathematics and Natural Sciences, Universitas Islam Indonesia*
Jl. Kaliurang KM 14.5, Sleman, 55584, Indonesia

*Corresponding author's e-mail: * **achmadfauzan@uii.ac.id***

Article History:

Received: 3rd January 2024
Revised: 6th February 2024
Accepted: 28th March 2024
Published: 1st June 2024

Keywords:

Education;
K-Medoids;
Outlier Imputation.

ABSTRACT

K-Medoids is a clustering algorithm that is often used because of its robustness against outliers. In this research, the focus is to cluster provinces based on educational level through several assessment indicators. This is in line with improving the quality of education in point 4 of the National Sustainable Development Goals (SDGs), namely "Quality Education". One of the points of the National Sustainable Development Goals (SDGs) that will still be improved is "Quality Education" which is the 4th point. This is because the success of a country is determined by the quality of good education. The condition of education in Indonesia still overlaps, so it is necessary to do equal distribution of education through clustering. The purpose of this research is to provide the best cluster results according to the Silhouette Index, which then the results of the clustering can be used as a consideration for advancing education in areas that still need attention, through policies or programs that can be developed by educational observers. This research was conducted in 34 provinces in Indonesia. The data source is from Statistical Publications by BPS RI. The method used is K-Medoids, because in this study there were outliers found. In addition to natural K-Medoids, the researcher also wants to compare methods by implementing K-Medoids with outlier handling in the form of imputed mean values and K-Medoids with imputed min-max values. The Silhouette Index results and cluster formation for the three comparators were 0.24 with 2 clusters, 0.26 with 8 clusters and 0.25 with 9 clusters, respectively. This research differs from previous work in its approach to outlier handling. While K-Medoids is a straightforward clustering method and generally indifferent to outliers, its effectiveness can be reduced by local outliers and random initial medoid selection.

This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution-ShareAlike 4.0 International License.

How to cite this article:

O. Rahmawati and A. Fauzan., "PROVINCIAL CLUSTERING BASED ON EDUCATION INDICATORS: K-MEDOIDS APPLICATION AND K-MEDOIDS OUTLIER HANDLING," *BAREKENG: J. Math. & App.*, vol. 18, iss. 2, pp. 1167-1178, June, 2024.

Copyright © 2024 Author(s)
Journal homepage: *https://ojs3.unpatti.ac.id/index.php/barekeng/*
Journal e-mail: *barekeng.math@yahoo.com; barekeng.journal@mail.unpatti.ac.id*

Research Article · Open Access

1. INTRODUCTION

K-Medoids is a clustering algorithm that aims to group data into clusters so as to spread the distance between points in the cluster. Generally, K-Medoids is used because of their robustness to outliers [1]- [4]. Apart from that, the use of medoids as cluster centers has a good impact because the cluster division is based on medoids, which are the actual data points in the cluster so that the resulting clusters have a more understanding meaning in data analysis. What differentiates the application of K-Medoids in this research from previous research is the handling of outliers. Some studies choose to use K-Medoids because of outliers in the data, but these outliers were handled by trimming of outlier values with mean values, and also trimming with min-max values [5] – [6]. This research aims to show whether there will be differences when handling outliers. What differentiates this research from previous research is the type of outlier handling. Generally, K-Medoids is very indifferent to the existence of outliers. K-Medoids is a widely recognized and straightforward clustering approach. Nevertheless, the algorithm's effectiveness might occasionally decline as a result of local outliers and the random selection of beginning medoids [7] – [9]. However, this research shows how researchers want to know the difference between natural K-Medoids using outlier and K-Medoids using outlier’s handling in research using educational data. This is based on the level of quality of education which is still a crucial problem in Indonesia[10] – [11].

Education is important in advancing the quality of a country’s human resources [12]. Because education itself is a series of processes of changing behavior, increasing knowledge, and life experience so that a person’s mind and characteristics can mature [13]. Quality education is key to establishing a great country [14], which is defined in terms of the quality of education, curriculum development, the quality of teaching staff, the quality of educational facilities, and also educational infrastructure. Based on the ranking results by the World Population Review on its website, it is recorded that Indonesia is ranked 67th in 2024 [15]. This figure is still not too superior when compared to other countries in Southeast Asia, as Singapore managed to occupy the 11th position. The hope is that in the following year, Indonesia will continue to create changes to advance and improve the quality of education in the eyes of the world. Based on the educational inequality that occurs in Indonesia, the K-medoids analysis is deemed appropriate for use as an educational clustering. The hope is that the government can group provinces based on the results of their indicator analysis.

To achieve quality education and education equality in accordance with the objectives of the Sustainable Development Goals (SDGs), Indonesia must have the courage to take steps to make breakthroughs. Indonesia itself still experiences disparity between one province and another. In several big cities such as DKI Jakarta and D.I Yogyakarta, the level of education is able to compete on the global stage. On the other hand, there are Papuans who have low literacy rates. According to Rosmana [16] it is difficult to find teaching staff in remote areas due to the inequality of education. This statement is supported by significant differences in the measurement indices between one province and another. Based on the 2021 Education Statistics Publication released by BPS, it is stated that the percentage of qualified teachers at elementary school level in DKI Jakarta and D.I Yogyakarta is 96.17% and 96.95% respectively, while in Papua it is only 76.05%. Using the same indicators, in the 2022 Education Statistics Publication, it is stated that in DKI Jakarta and D.I Yogyakarta, it is 96.44% and 97.51%, respectively, while in Papua, it is only 77.03% [17].

Therefore, to overcome the inequality of education that occurs in Indonesia, grouping or clustering is a viable solution [18]. By using existing education indicators, clustering can be done. The aim of applying clustering in this case study is to help the government and education stakeholders find out the regions that most need improved educational assistance and find out the education patterns in each province. Hopefully, in the future, new policies and effective education strategies that will have a big impact on education in Indonesia can be developed. Research on the clustering of provinces based on education indicators was previously conducted by D.A. Alodia et al. [19], the clustering algorithm used is K-Medoids, using eight variables as indicators for determining education, which was implemented across 34 provinces in Indonesia. The results obtained were three clusters, each cluster consisting of six provinces with the "High" class, 13 provinces with the "Medium" class, and 15 provinces with the "Low" class. Research on the clustering of provinces based on education indicators using K-Medoids was also conducted by R.A. Chandra [20]. This research uses a Clustering Data Mining methodology approach using K-Medoids to obtain a segmentation of the characteristics of Unjani University students' learning patterns. Two clusters were formed, where cluster 1 had a silhouette value close to 1, and the value for cluster 2 approached 0.80. Similar research was also conducted by D.M. Sinaga, et al [21]. The aim of this research is to create a grouping model using K-

Medoids. Similar research has also been carried out by D.A.I.C. Dewi and D.A. Paramita [22] using the K-Means algorithm to identify villages that are considered to have poor education. Eight clusters were obtained from a large amount of data, consisting of 229 in 33 villages in Kapuas Regency, Central Kalimantan. In the same year, similar research was also carried out by M.A. Putri, et al. [23] with the subjects of the research being students at state and private universities who were selected to receive Bank Indonesia Scholarship using K-Medoids.

This research refers to the research above that has been carried out previously. The K-Medoids algorithm is appropriate for data that contains outlier values. However, this study aims to compare the implementation of the K-Medoids algorithm without outlier handling and with various outlier handling. The objective of this research is to provide more understanding regarding the impact of employing the K-medoids approach to handle outlier data on the quality of the produced clusters. While outliers are sometimes disregarded as math mistakes or noise, they might really offer valuable insights. Hence, it is crucial to identify them before to the process of modeling and analysis [24]. The various outlier treatments applied include trimming of outlier values with mean values and also with min-max values.

2. RESEARCH METHODS

2.1 Research Design

The mapping of provinces into clusters was carried out using K-Medoids cluster analysis. K-Medoids algorithm was chosen because the data used contains outlier values that fall far from other data characteristics (outliers) [25]. K-Medoids is a clustering algorithm that is robust to outliers [26]. The research method used is descriptive analysis and K-Medoids clustering. The descriptive analysis produced is intended to provide an overview of the characteristics of each variable used in education indicator data, whereas cluster analysis is used to obtain mapping results of provinces in Indonesia based on educational indicator values attained by each province. The K-Medoids analysis used is expanded through some additions in the form of K-Medoids without outlier handling and with outlier handling in the form of trimming of mean values and trimming of min-max values.

2.2 Population and Sample

The population in this study is the education indicator for all provinces in Indonesia, which consists of 34 provinces. The sample in the study used ten provincial-level education indicator variables from 2022. Data comes from the 2022 Education Statistics Publication, which is available on the Indonesian Central Bureau of Statistics (BPS RI) website and accessed on March 4, 2023, and is defined as follows: Highest Educational Attainment (X_1), IT skills (X_2), School Enrollment Rates/ SER (X_3), Literacy Rate/ LR (X_4), Gross Participation Figures/ GPF (X_5), School Completion Rate (X_6), School Dropout Rate (X_7), Percentage of Qualified High School Teachers (X_8), Percentage of Qualified Vocational High School Teachers (X_9), Pure Participation Figures / PPF(X_{10}).

2.3 Data

Data was obtained through secondary means from the 2022 Education Statistics Publication from BPS RI. Table 1 is an illustration of the data used.

Table 1. Preview of Data

Province	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
Aceh	32.44	83.10	64.89	98.25	92.53	70.12	14.30	99.25	98.71	30.50
Sumatera Utara	40.53	78.66	74.44	99.11	97.23	76.48	12.41	98.16	95.61	22.58
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
Papua	24.51	65.93	29.82	81.19	77.06	38.74	34.87	97.63	93.88	15.28

2.4 Research Flowchart

The stages of this research are depicted in the research flow diagram as follows.

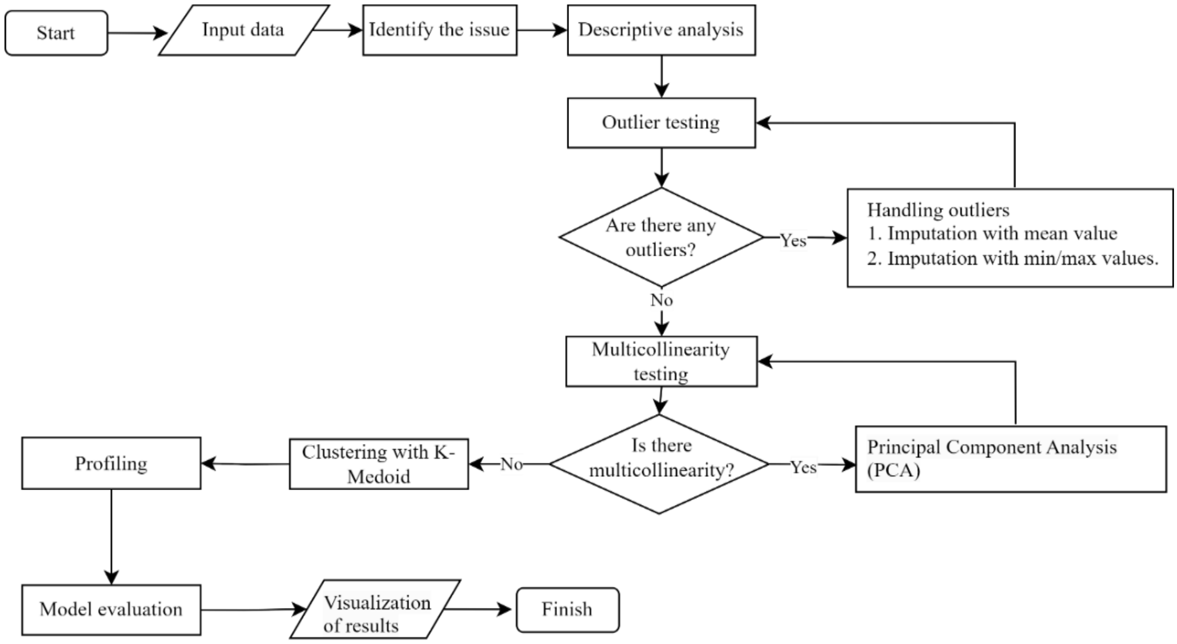


Figure 1. Research Flowchart

Figure 1 provides a visualization of the stages of implementing this research. The three clustering methods of analysis in this research are K-Medoids without outlier handling, K-Medoids with outlier handling in the form of trimming of mean values, and K-Medoids with outlier handling in the form of trimming of min-max values. This research began by identifying the problem and then collecting data by taking data from BPS publications. Subsequently, descriptive analysis was performed to determine the characteristics of the data. To obtain the final process, the clustering accuracy results were measured by looking at the comparison of the silhouette values of the three clusters method. The first step in performing clustering is to detect outliers, which is done using **Equation (1) [27]**.

$$\boldsymbol{D_M(x) = \sqrt{(x-\mu)^T \Sigma^{-1} (x-\mu)}} \tag{1}$$

Outliers are obtained from $d > X_p^2(1-\alpha)$ with the level of significance $\alpha = 5\%$. $\sqrt{(\boldsymbol{x}-\mu)^{\boldsymbol{T}}\Sigma(\boldsymbol{x}-\mu)}$ will be approximated to a chi-square distribution with a degree of p. $\boldsymbol{x}_{\mathrm{i}}$ is a vector that contains the data with which the distance from the mean is calculated and $\boldsymbol{\mu}$ is the mean value. The formula for calculating the number of clusters formed using the Silhouette Coefficient is as **Equation (2) [25]**.

$$\boldsymbol{SC = \frac{\sum_{i=1}^{n} \frac{b_i - a_i}{max\{a_i, b_i\}}}{n}} \tag{2}$$

$\frac{b_i - a_i}{max\{a_i, b_i\}}$ is the formula of the Silhoutte Index (SI). The values of SC are interval (-1, 1). The closer the silhouette Score value is to 1, the better the accuracy of the resulting cluster. In the K-Medoids algorithm, the cluster center is the medoid value, serving as a representative of the cluster centroid **[28]** .

$$\boldsymbol{s'_i(X,d,M) = \frac{b'_i - a'_i}{max(a'_i - b'_i)}} \tag{3}$$

Where a'_i mean $d(x_i, M_{l_i})$ and $b'_i = \min_{k \neq l_i} d(x_i, M_k)$. This formula is used for both K-Means and K-Medoids. To obtain the optimal number of clusters, the Silhouette Index is calculated. The silhouette index value is generally in the range of -1 to 1, the closer to 1, the better the cluster [29]. The formula used is presented in Equation (4) [30].

$$\boldsymbol{Silhouette\ index} = \frac{\boldsymbol{b_i - a_i}}{\boldsymbol{max\{a_i - b_i\}}} \tag{4}$$

a_i is average dissimilarity i to all other objects of A, mean $\left\{d(x_i, x_j) |\ I_{j=I_j}, i \neq j\right\}$. b is minimum $d(i, C)$, $C \neq A$, min mean $\left\{d(x_i, x_j) |\ I_j = k\right\}$.

3. RESULTS AND DISCUSSION

3.1 Descriptive Analysis

Each data certainly has characteristics that will describe its general condition. In this research, to determine the characteristics of each province, descriptive statistics were used to see the minimum and maximum values.

Table 2. Descriptive Statistics

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
Min	19.6	65.93	29.82	81.19	75.49	37.71	10.26	87.28	93.25	12.79
Mean	30.32	75.14	72.36	96.72	87.82	63.82	20.85	98.57	96.41	23.8
Max	48.56	89.95	92.36	99.81	97.71	87.68	34.87	99.59	98.71	52.3

Based on Table 2, nine out of ten variables have a positive influence on education indicators, meaning that when these variables increase, the quality of education in that province will also improve. Meanwhile, one variable (X_7) has a negative influence on the quality of education indicators, which means that an increase in variable X_7 will actually decrease the quality of education. Based on the data, it was found that the province with a high education indicator is DI Yogyakarta. This is evidenced by the fact that out of ten variables, three variables with the highest values are found in the province of DI Yogyakarta, namely X_1, X_6, and X_{10}. In addition, DI Yogyakarta also has the lowest value for variable X_7. This is because Yogyakarta is known as a city of education, where there are several universities such as Gadjah Mada University, Indonesian Institute of the Arts, Sunan Kalijaga State Islamic University, Yogyakarta State University, and 101 private colleges in DI Yogyakarta itself [31]. Moreover, based on Yogyakarta's role as a city of education [32] – [33], several characteristics can be found in Yogyakarta such as the heterogeneity of the students, a complete education level from elementary to university (*hierarchy*), and dynamic [34]. From the nine variables, 54.71% of provinces in Indonesia have education support indicators above average. Similarly, for the average variable (X_7), 47.06% of provinces in Indonesia are below the average for this variable.

3.2 Cluster with K-Medoid and Various K-Medoid Outlier Treatments

Checking for outliers is the first step of K-Medoids analysis in this research. If an outlier is found, then the K-Medoids algorithm is appropriate to implement, but if no outliers are detected, it is better to use the K-Means algorithm. Figure 2 presents the boxplot analysis for the K-Medoid clustering method, delineated into scenarios without treatment (a) and with treatment (b and c). The median is depicted as a bold line within the box, allocating approximately half of the observed values above this threshold and the remainder below, thereby suggesting a uniform distribution of data. The depicted box illustrates the interquartile range, with the lower boundary signifying the first quartile (bottom 25% of values) and the upper boundary representing the third quartile (bottom 75% of values). The X-axis enumerates the variables examined in the research (X_1 to X_{10}), whereas the Y-axis delineates the measured variable or the data distribution for each indicator under investigation. Observations from Figure 2 reveal numerous data points positioned beyond the upper and lower whiskers, signifying the identification of outliers or values that deviate markedly from the primary dataset. This observation substantiates the implementation of outlier treatment methodologies within this study. The treatment of outliers constitutes a singular phase within the study, resulting in the persistence of a limited number of outliers, albeit in a reduced capacity compared to their prevalence prior to the application of said treatment.

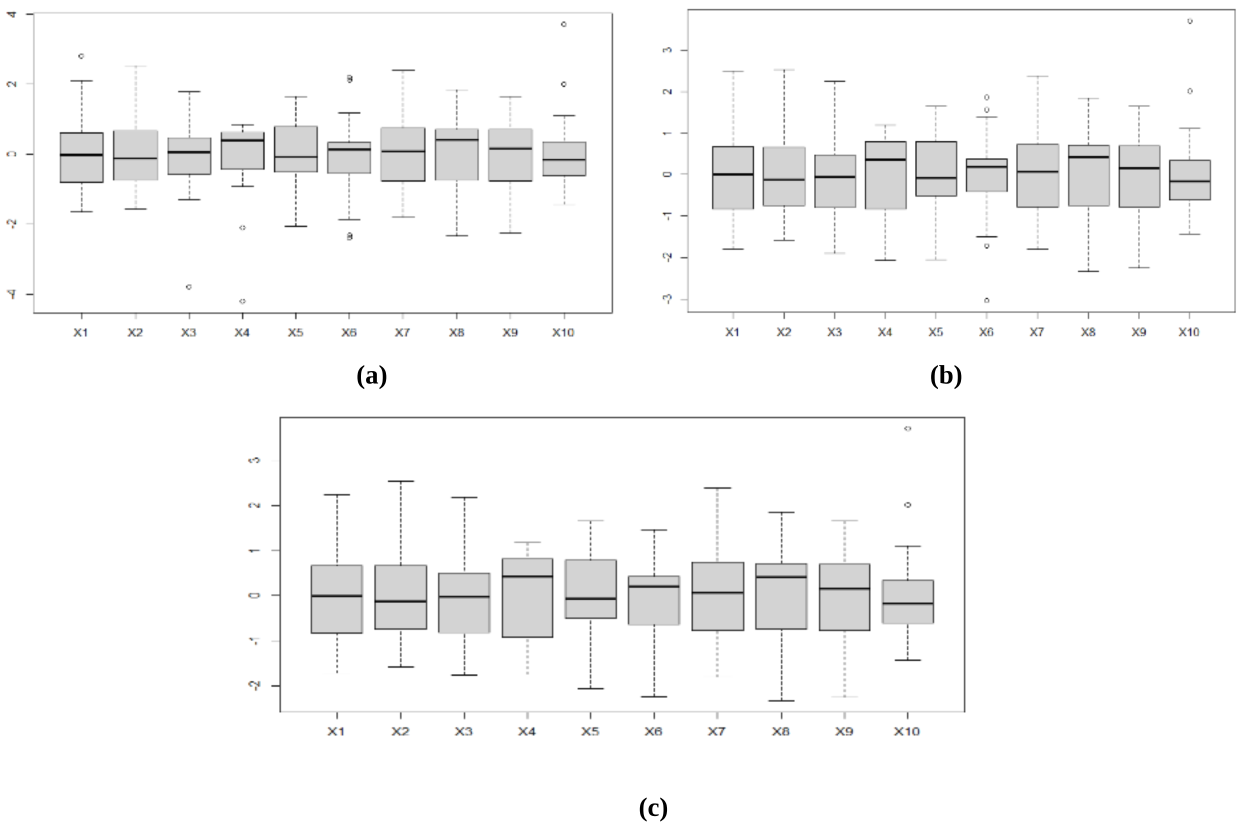


Figure 2. Outlier Visualization,
(a) K-Medoids Outlier without Treatment, (b) K-Medoids Outlier with Mean Trimming, (c) K-Medoids Outlier with Min-Max Trimming

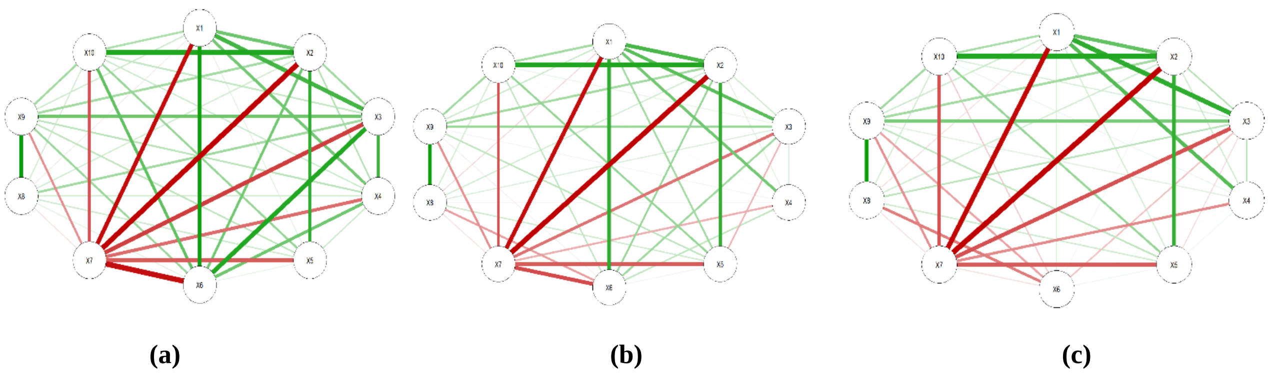
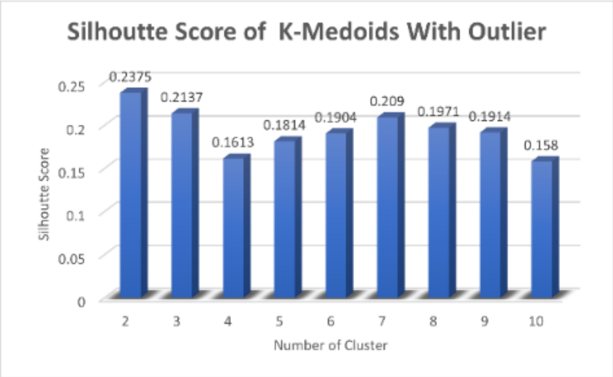
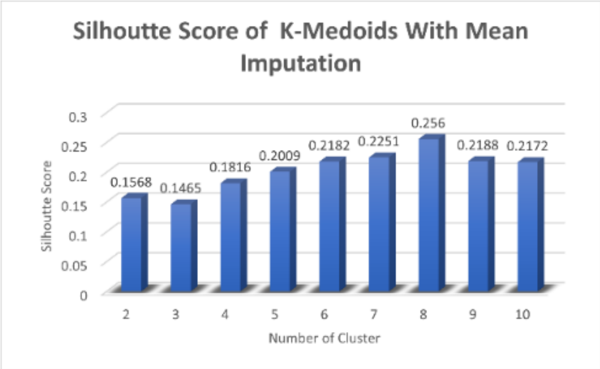


Figure 3. Correlation Visualization,
(a) Multicollinearity in K-Medoids without Treatment, (b) Multicollinearity in K-Medoids with Mean Trimming, (c) Multicollinearity in K-Medoids with Min-Max Trimming.

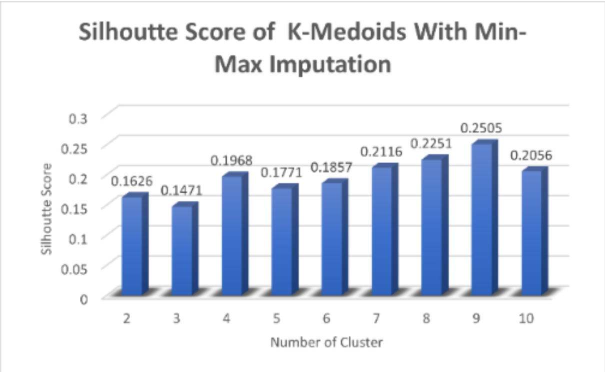
In **Figure 3**, the greener the color indicates the value is closer to 1 or contains multicollinearity between variables, then the redder the color, the less multicollinearity between variables is approaching. In the output, there are no pairs of variables that exhibit multicollinearity. Thus, further analysis in the form of clustering using K-Medoids can be carried out [35], [36], [37].



(a)



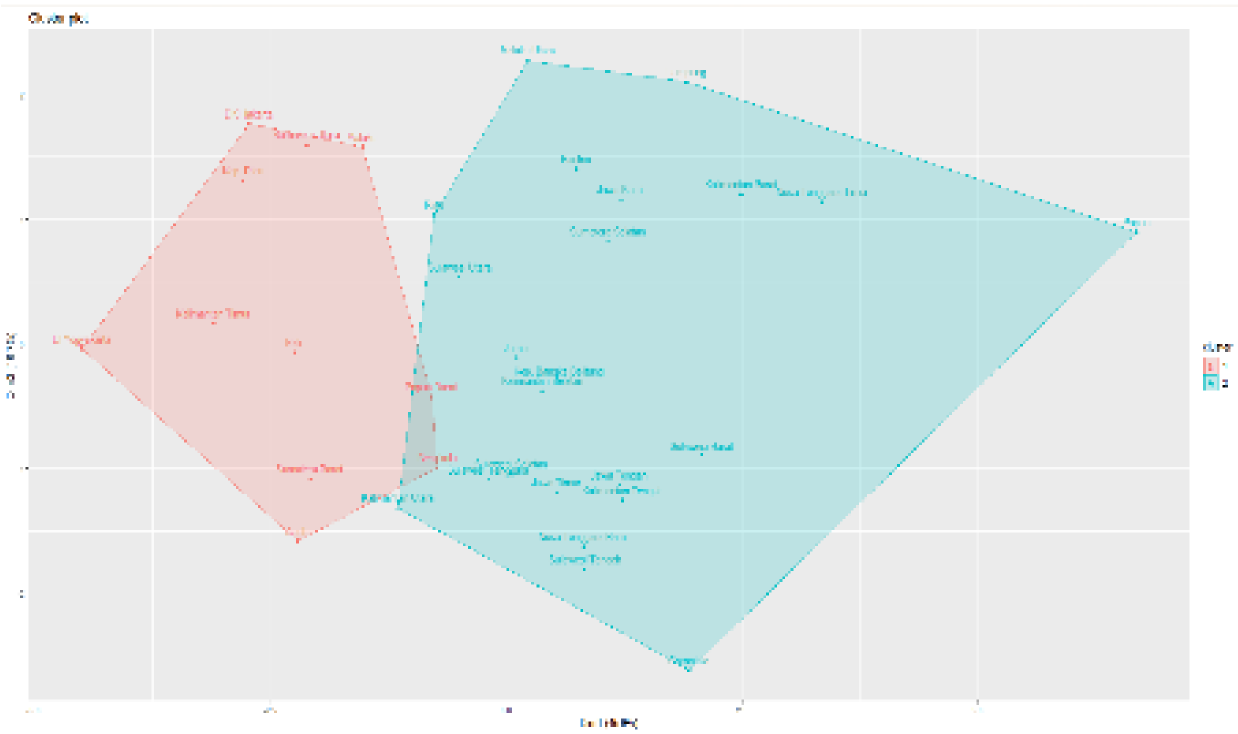
(b)



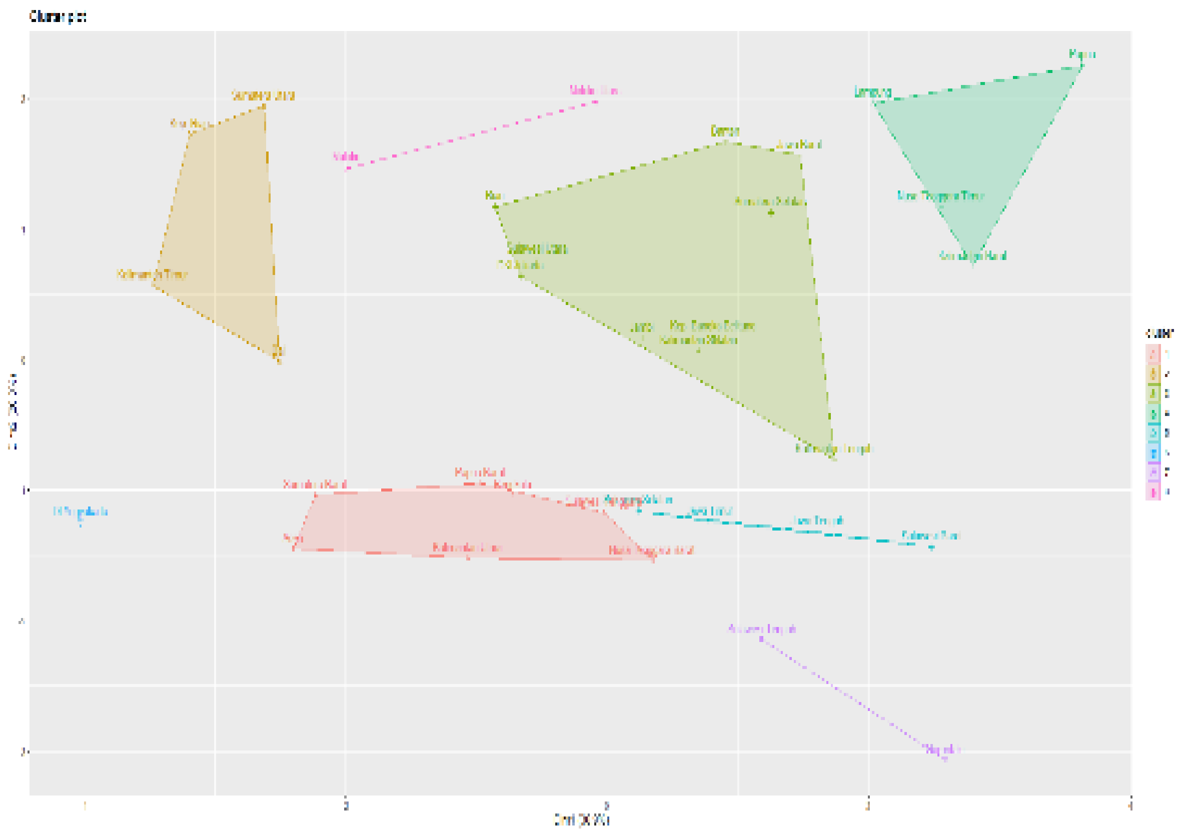
(c)

Figure 4. Visualization of the Optimal Number of Clusters (a) K-Medoids without Treatment, (b) K-Medoids without Mean Trimming, (c) K-Medoids without Min-Max Trimming

Figure 4 presents that for each optimal number of clusters using the Silhouette Coefficient, 2 clusters were obtained for K-Medoids without treatment, 8 clusters for K-Medoids with mean trimming, and 9 clusters for K-Medoids with min-max trimming.



(a)



(b)

(c)

Figure 5. Mapping Plot (a) K-Medoids without Treatment, (b) K-Medoids with Mean Trimming, (c) K-Medoids with Min-Max Trimming

In Figure 5, notice that for mapping **(a)** cluster overlap occurs between cluster 1 and cluster 2. This is because there is overlap in several dimensions in the clustering results. For mapping **(b)** and **(c)**, the cluster divisions are very numerous and detailed, but the cluster results produced in certain clusters have less than optimal inter-cluster validity values. Calculating the goodness of the K-Medoid clustering model uses the Silhouette Coefficient value.

3.3 Cluster Division

The clustering process formed groupings of provinces based on the number of clusters according to the criteria that the province has. The following is the division of clusters based on the method used.

Table 1. Cluster Division

Method	Cluster	Number of Provinces	Province
K-Medoids without Treatment	1	10	Aceh, North Sumatra, West Sumatra, Bengkulu, Riau Islands, DKI Jakarta, DI Yogyakarta, Bali, East Kalimantan, Maluku, West Papua
	2	24	Riau, Jambi, South Sumatra, Lampung, Bangka Belitung Islands, West Java, Central Java, East Java, Banten, West Nusa Tenggara, East Nusa Tenggara, West Kalimantan, Central Kalimantan, South Kalimantan, North Kalimantan, North Sulawesi, Central Sulawesi, South Sulawesi, Southeast Sulawesi, Gorontalo, West Sulawesi, North Maluku, Papua
K-Medoids with Mean Trimming	1	7	Aceh, West Sumatra, Bengkulu, West Nusa Tenggara, North Kalimantan, Southeast Sulawesi, West Papua
	2	4	North Sumatra, Riau Islands, Bali, East Kalimantan
	3	10	Riau, Jambi, South Sumatra, Bangka Belitung Islands, DKI Jakarta, West Java, Banten, Central Kalimantan, South Kalimantan, North Sulawesi
	4	4	Lampung, East Nusa Tenggara, West Kalimantan, Papua
	5	4	Central Java, East Java, South Sulawesi, West Sulawesi
	6	1	DI Yogyakarta
	7	2	Central Sulawesi, Gorontalo
	8	2	Maluku, North Maluku
K-Medoids with Min-Max Trimming	1	6	Aceh, West Sumatra, North Kalimantan, West Papua, Bengkulu, Southeast Sulawesi
	2	4	North Sumatra, Riau Islands, Bali, East Kalimantan
	3	9	Riau, North Sumatra, Jambi, Bangka Belitung Islands, South Kalimantan, Central Kalimantan, South Sumatra, Banten, West Java
	4	4	Lampung, East Nusa Tenggara, West Kalimantan, Papua
	5	1	DKI Jakarta
	6	5	Central Sulawesi, West Sulawesi, Central Java, East Java, West Nusa Tenggara
	7	1	DI Yogyakarta
	8	2	Central Sulawesi, Gorontalo
	9	2	Maluku, North Maluku

The comparison of the clustering results are as follows:

Table 2. Comparison of Clustering Methods

K-Medoids Method	Number of Clusters Formed	Silhouette Validation Index
K-Medoids without Treatment	2	0,24
K-Medoids with Mean Trimming	8	0,26
K-Medoids with Min-Max Trimming	9	0,25

By inspecting **Table 2**, the silhouette validation index or measure of the goodness of the model formed is observed to not differ significantly, but the number of clusters formed is very significant. In accordance with the findings presented in **Table 2**, and adhering to the principle of parsimony, coupled with the observation that the silhouette values do not exhibit significant variance, it is concluded that the K-Medoids algorithm without any modifications continues to be applicable for clustering methodologies in scenarios involving outliers. However, if a more diverse number of clusters is desired (for specific reasons), treating outliers in the K-Medoid method can be considered as one of the options.

3.4 Profiling

Proceeding with the best method, K-Medoids without outlier treatment, profiling is carried out to obtain the cluster **Table 3**.

Table 3. Medoid Centers

Cluster	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
1	0.3757	1.4878	0.7480	-0.321	0.4999	0.9794	-1.3417	0.4526	0.1524	0.8364
2	-0.1228	-0.446	-0.003	0.6458	-0.5806	0.0492	0.0183	-0.0899	0.6315	-0.3978

Table 4. Cluster Characteristics

Cluster	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
1	High	High	High	Low	High	High	Low	High	Low	High
2	Low	Low	Low	High	Low	Low	High	Low	High	Low

Table 4 indicates that the division of clusters is divided into two classes. Cluster 1 is labeled as a cluster with provinces that have high education quality and Cluster 2 is labeled as a cluster consisting of provinces that have low education quality based on education indicators published in 2022. To make it easier to visualize the results of clustering, the visualization results are presented in the form of a spatial map in **Figure 6.**

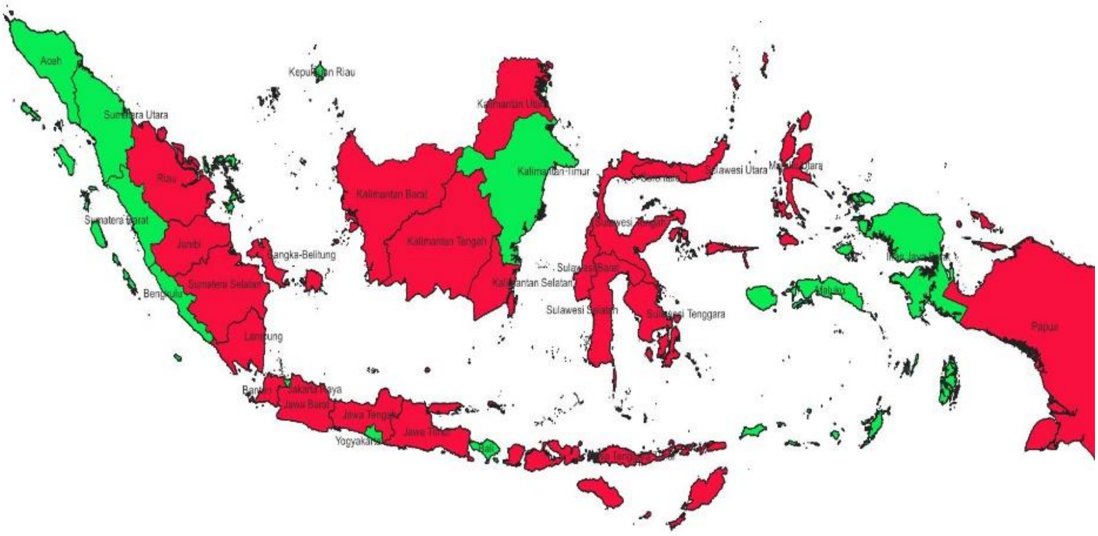


Figure 6. QGIS Visualization

The two clusters from the results of the analysis can be visualized using QGIS software, as presented above. The provinces marked in green denote cluster 1 or provinces possessing high quality education according to 10 variables, which are indicators of education in 2022. The provinces marked in red indicate low quality education based on education indicators in Indonesia in 2022 using data from the 2022 Indonesian Education Statistics Publication released by BPS RI on its website.

4. CONCLUSION

Based on the results of the analysis that has been carried out using education indicator data in Indonesia, which consists of ten variables with 34 provinces as objects, it was found that the highest average was in the qualified teacher (high school) variable achieved by the province of Gorontalo, due to the level of participation in becoming a teacher in Gorontalo still relatively high. The lowest value in the Dropout Rate (High School/Equivalent) variable is attributed to DI Yogyakarta. This is because the province of DI Yogyakarta is a prominent student city, implying that the majority of its residents are well-educated, have completed high school/equivalent education, and even pursued further education. A comparison of outlier treatments carried out on K-Medoids suggested that the best results were obtained by handling outliers in the form of trimming of the mean value. It is considered the best method because it has the highest silhouette value index, amounting to 0.26, with optimal cluster formation of 8. However, this division is too numerous and detailed, making it difficult to label each cluster. So, subjectively, the best cluster was selected, which was K-Medoids, without handling outliers which formed 2 clusters, because the resulting silhouette validation index values were not significantly different. Cluster 1 is labeled as a cluster with high quality education. Provinces in Cluster 1 include Aceh, North Sumatra, West Sumatra, Bengkulu, Riau Islands, DKI Jakarta, DI Yogyakarta, Bali, East Kalimantan, Maluku, West Papua. Meanwhile, the remaining provinces are Riau, Jambi, South Sumatra, Lampung, Bangka Belitung Islands, West Java, Central Java, East Java, Banten, East Nusa Tenggara, West Kalimantan, South Kalimantan, North Kalimantan, North Sulawesi, Central Sulawesi, Southeast Sulawesi, Gorontalo, West Sulawesi, North Maluku, Papua, are included in Cluster 2 which is labeled as a cluster with low education quality based on 2022 education indicators. Based on the conclusions above, this study advises the government and the ministry of education to grant attention towards the provinces that are included in Cluster 2. In hopes that in 2030, Indonesia can achieve one of the goals in the SDGs, namely "Quality Education" with reduced or no inequality, between one province to another so Indonesia will have quality Human Resources (HR) to accept future challenges.

REFERENCES

[1] N. Kamal Kaur, U. Kaur, and D. Singh, “K-Medoid Clustering Algorithm-A Review,” *International Journal of Computer Application and Technology (IJCAT)*, vol. 1, no. 1, 2014, [Online]. Available: www.technopublications.com

[2] P. Arora, Deepali, and S. Varshney, “Analysis of K-Means and K-Medoids Algorithm for Big Data,” *Procedia Comput Sci*, vol. 78, pp. 507–512, 2016, doi: 10.1016/j.procs.2016.02.095.

[3] G. R. Suraya and A. W. Wijayanto, “Comparison of Hierarchical Clustering, K-Means, K-Medoids, and Fuzzy C-Means Methods in Grouping Provinces in Indonesia according to the Special Index for Handling Stunting,” *Indonesian Journal of Statistics and Its Applications*, vol. 6, no. 2, pp. 180–201, Aug. 2022, doi: 10.29244/ijsa.v6i2p180-201.

[4] N. Sureja, B. Chawda, and A. Vasant, “An improved K-medoids clustering approach based on the crow search algorithm,” *Journal of Computational Mathematics and Data Science*, vol. 3, p. 100034, Jun. 2022, doi: 10.1016/j.jcmds.2022.100034.

[5] P. H. Ramsey and P. P. Ramsey, “Optimal Trimming and Outlier Elimination,” *Journal of Modern Applied Statistical Methods*, vol. 6, no. 2, pp. 355–360, Nov. 2007, doi: 10.22237/jmasm/1193889660.

[6] U. Pratap, C. Canudas-de-Wit, and F. Garin, “Outlier detection and trimmed-average estimation in network systems,” *Eur J Control*, vol. 60, pp. 36–47, Jul. 2021, doi: 10.1016/j.ejcon.2021.04.005.

[7] S. Mousavi, F. Z. Boroujeni, and S. Aryanmehr, “Improving customer clustering by optimal selection of cluster centroids in K-means and K-medoids algorithms,” *J Theor Appl Inf Technol*, vol. 98, no. 18, pp. 3807–3814, 2020.

[8] M. Adepeju, S. Langton, and J. Bannister, “Anchored k-medoids: a novel adaptation of k-medoids further refined to measure long-term instability in the exposure to crime,” *J Comput Soc Sci*, vol. 4, no. 2, pp. 655–680, Nov. 2021, doi: 10.1007/s42001-021-00103-1.

[9] Md. K. Khan, S. Mahmud Ahmed, S. Sarker, and M. H. A. Khan, “K-Cosine-Medoids Clustering Algorithm,” in *2021 5th International Conference on Electrical Information and Communication Technology (EICT)*, IEEE, Dec. 2021, pp. 1–5. doi: 10.1109/EICT54103.2021.9733540.

[10] C. Pramana, D. Chamidah, S. Suyatno, F. Renadi, and S. Syahruddin, “Strategies to Improved Education Quality in Indonesia: A Review,” *Syaharuddin Syaharuddin Turkish Online Journal of Qualitative Inquiry (TOJQI)*, vol. 12, no. 3, pp. 1977–1994, 2021.

[11] R. A. Tarigan, A. Saptono, and S. Muchtar, “Enhancing Indonesia’s Education Quality: Identifying and Addressing Key Challenges,” in *1st International Students Conference on Business*, 2023.

[12] N. W. A. Aprilia, I. G. A. M. Srinadi, and K. Sari, “Pengelompokan desa/kelurahan di kota Denpasar menurut indikator pendidikan,” *E-Jurnal Matematika*, vol. 5, no. 2, p. 38, May 2016, doi: 10.24843/MTK.2016.v05.i02.p119.

[13] N. Dwi Tsoraya, I. A. Khasanah, M. Asbari, and A. Purwanto, “Pentingnya pendidikan karakter terhadap moralitas pelajar di lingkungan masyarakat era digital,” *Literasi: Jurnal Manajemen Pendidikan*, vol. 1, no. 1, pp. 7–12, 2023.

[14] S. A. Nurfatimah, S. Hasna, and D. Rostika, “Membangun kualitas pendidikan di Indonesia dalam mewujudkan program Sustainable Development Goals (SDGs),” *Jurnal Basicedu*, vol. 6, no. 4, pp. 6145–6154, May 2022, doi: 10.31004/basicedu.v6i4.3183.

[15] New Jersey Minority Educational Development, “International Education Database,” World Best Educational Systems.

[16] P. Rosmana, S. Iskandar, N. Fadilah, N. Azhar, D. Oktavini, and A. C. Munte, “Upaya pemerataan pendidikan berkelanjutan di daerah 3T,” *Attadib: Journal of Elementary Education*, vol. 6, no. 2, pp. 405–418, 2022.

[17] Badan Pusat Statistik, “Statistik Pendidikan 2021,” Jakarta, Nov. 2021.

[18] Y. A. Priambodo and Y. J. Prasetyo, “Pemetaan penyebaran guru di provinsi Banten dengan menggunakan metode Spatial Clustering K-Means (studi kasus : wilayah provinsi Banten),” *Indonesian Journal of Computing and Modeling*, vol. 1, pp. 18–27, 2018.

[19] D. A. Alodia, A. P. Fialine, D. Endriani, and E. Widodo, “Implementasi metode K-Medoids clustering untuk pengelompokan provinsi di Indonesia berdasarkan indikator pendidikan,” *Sepren*, vol. 2, no. 2, pp. 1–13, Dec. 2021, doi: 10.36655/sepren.v2i2.606.

[20] R. A. Candra, Y. H. Chrisnanto, and P. N. Sabrina, “Segmentasi mahasiswa berdasarkan karakteristik pola belajar menggunakan metode K-Medoids clustering,” in *Prosiding Sains Nasional dan Teknologi*, Nov. 2022, p. 355. doi: 10.36499/psnst.v12i1.7047.

[21] D. M. Sinaga, A. P. Windarto, and D. Hartama, “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*, vol. 1, no. 1, 2022, [Online]. Available: https://ejurnal.pdsi.or.id/index.php/jurtidas/index

[22] N. Nurahman and D. D. Aulia, “klasterisasi pendidikan masyarakat untuk mengetahui daerah dengan pendidikan terendah menggunakan algoritma K-Means,” *Jurnal Informatika dan Rekayasa Perangkat Lunak*, vol. 5, no. 1, p. 38, Mar. 2023, doi: 10.36499/jinrpl.v5i1.7510.

[23] M. A. Putri, A. Nazir, L. Handayani, and I. Afrianty, “Penerapan algoritma K-Medoids clustering untuk mengetahui pola penerima beasiswa Bank Indonesia provinsi Riau,” *JUKI: Jurnal Komputer dan Informatika*, vol. 5, no. 1, 2023.

[24] A. Smiti, “A critical overview of outlier detection methods,” *Comput Sci Rev*, vol. 38, p. 100306, Nov. 2020, doi: 10.1016/j.cosrev.2020.100306.

[25] D. A. I. C. Dewi and D. A. K. Pramita, “Analisis perbandingan metode Elbow dan Silhouette pada algoritma clustering K-medoids dalam pengelompokan produksi kerajinan Bali,” *Matrix: Jurnal Manajemen Teknologi dan Informatika*, vol. 9, no. 3, pp. 102–109, Nov. 2019, doi: 10.31940/matrix.v9i3.1662.

[26] P. N. Safitri, R. Aristawidya, and S. B. Faradilla, “Klasterisasi faktor-faktor kemiskinan di provinsi Jawa Barat menggunakan K-Medoids clustering,” *Journal of Mathematics Education and Science*, vol. 4, no. 2, pp. 75–80, Oct. 2021, doi: 10.32665/james.v4i2.242.

[27] R. A. Wibowo, K. Nisa, H. Venelia, and W. Warsono, “Robust clustering of Covid-19 pandemic worldwide,” *BAREKENG: Jurnal Ilmu Matematika dan Terapan*, vol. 16, no. 2, pp. 687–694, Jun. 2022, doi: 10.30598/barekengvol16iss2pp687-694.

[28] E. T. Ena Tasia and M. Afdal, “Perbandingan algoritma K-Means dan K-Medoids untuk clustering daerah rawan banjir di kabupaten Rokan Hilir,” *Indonesian Journal of Informatic Research and Software Engineering (IJIRSE)*, vol. 3, no. 1, pp. 65–73, Mar. 2023, doi: 10.57152/ijirse.v3i1.523.

[29] G. KIR, A. ÜLKE KESKİN, and U. ZEYBEKOĞLU, “Clustering of precipitation in the black sea region with by Fuzzy C-Means and silhouette index analysis,” *Black Sea Journal of Engineering and Science*, vol. 6, no. 3, pp. 210–218, Jul. 2023, doi: 10.34248/bsengineering.1296734.

[30] F. N. Aini, S. Palgunadi, and R. Anggrainingsih, “Clustering business process model petri net dengan complete linkage,” *Jurnal Itsmart*, vol. 3, no. 2, 2014.

[31] LLDIKTI V, “Sebaran Perguruan Tinggi dan Program Studi LLDIKTI V,” https://lldikti5.id/evira/.

[32] S. Sugiyanto, “Yogyakarta kola pendidikan dan ekonomi alternatif,” *Cakrawala Pendidikan*, 2004.

[33] S. Haryono, “Analisis brand image Yogyakarta sebagai kota pelajar,” *Jurnal Ilmu Komunikasi*, vol. 7, no. 3, 2009, [Online]. Available: www.pdffactory.com

[34] F. R. R. Mega, “Karakter Yogyakarta sebagai kota pelajar, kota budaya, kota wisata, dan Jogja never ending Asia,” Yogyakarta, 2003.

[35] E. A. N. Fajrina, “Analysis cluster robust using k-medoids method in the data containing outlier,” Final Task, Universitas Lampung, Lampung, 2021.

[36] D. A. Rahmah, “Analisis klaster berdasarkan indikator kesejahteraan rakyat menggunakan metode Self Organizing Maps (SOM),” Final Task, Universitas Islam Negeri Maulana Malik Ibrahim, Malang, 2022.

[37] A. R. Zaidah, C. I. Septiarani, S. Nisa, A. Yusuf, and N. Wahyudi, “Komparasi algoritma k-means, k-medoid, agglomeartive clustering terhadap genre spotify,” *Jurnal Ilmiah Ilmu Komputer*, vol. 7, no. 1, 2021, [Online]. Available: https://ejournal.fikom-unasman.ac.id