Available online at http://iorajournal.org/indx.php/orics /index

Operations Research: International Conference Series

e-ISSN: 2722-0974
p-ISSN: 2723-1739

Vol. 4, No. 4, pp. 141-148, 2023

Comparison of K-Medoids and Clara Algorithm in Poverty Clustering Analysis in Indonesia

Ananda Rizki Dwi Ardini[1*], Haposan Sirait[2]

[1,2]*Faculty of Mathematics and Natural Sciences, Riau University, Bina Widya Campus, Pekanbaru 28293*

**Corresponding author email: ananda.rizki3910@student.unri.ac.id*

Abstract

The Covid-19 pandemic entered Indonesia in March 2020, so the government imposed restrictions on people's movement in various regencies. The imposition of restrictions on people's movement will have an impact on the economy to the point of poverty. Poverty is influenced by several factors such as population, health, education, employment and economic factors. The poverty of a district/city in Indonesia is grouped to assist the government in alleviating poverty more efficiently. The process of grouping data in data mining is to group districts/cities in Indonesia based on factors that affect poverty with the K-Medoids and CLARA algorithms, then compare the two methods based on the average value of the ratio of the standard deviations. The variables used in this study consist of 4 variables, namely human development index (HDI), gross regional domestic product (GRDP), unemployment rate, and population density. The results of this study indicate that using the K-Medoids obtained 2 clusters, while using the CLARA algorithm obtained 3 clusters. Based on the results of grouping the two algorithms, the best algorithm was obtained using cluster validation, namely the CLARA algorithm because it has the average value of the ratio of the smallest standard deviation of 0.106.

Keywords: Data mining, clustering, K-Medoids algorithm, CLARA algorithm, cluster validation.

1. Introduction

Data mining is a collection of methods and techniques for exploring and analyzing large volumes of data sets in an automatic or semi-automatic way to analyze the data into useful information. Data mining is divided into several groups based on its tasks, namely description, estimation, prediction, classification, clustering and association (Tufféry, 2011) . Cluster analysis or clustering is a type of multiple variable analysis (multivariate analysis) which is used to group objects in such a way that objects in one group are very similar and objects in various groups are quite different. Clustering techniques in data mining have two methods, namely hierarchical and non-hierarchical (level) clustering (Shmueli et al, 2018) . Non-hierarchical clustering methods include the K-Medoids algorithm and Clustering Large Application (CLARA).

In March 2020, the Covid-19 pandemic entered Indonesia, causing the government to impose restrictions on people's movement in various districts/cities. The implementation of restrictions on people's movement has an impact on various sectors, such as the business sector, which then affects the economy and causes poverty (Sari & Ediwijojo, 2023). Poverty is a complex fundamental problem that has become the center of attention of governments in various countries, one of which is Indonesia. Based on data from the Central Statistics Agency (BPS), during the Covid-19 pandemic, the percentage of poor people in Indonesia increased by 0.97%. In September 2019 the percentage of poor people was 9.22% and in September 2020 it increased to 10.19% (Central Statistics Agency, 2021) . There are several factors that influence poverty, based on the explanation from Ipmawan et al. (2022) showing that the rate of Gross Regional Domestic Product (GRDP), the Human Development Index (HDI), and the unemployment rate significantly influence poverty in Indonesia. According to Safri (2021) explained that economic growth and population density have a significant effect on poverty levels in Jambi Province. Regional grouping/clustering can explain the poverty conditions in each district/city in Indonesia, making it easier for the government to implement targeted policies for each district/city.

Previous research was conducted by Safitri et al (2021) regarding the clustering of poverty factors in West Java Province using the K-Medoids method . The results of this research showed that the number of clusters was 3 clusters and cluster 2 was a cluster with poor poverty. Research conducted by Hanafiah & Wanto (2020) on the grouping of

districts/cities in Indonesia based on poverty information using the K-Means clustering method resulted in a grouping of 4 clusters. The formation of 4 clusters can be seen that the fourth cluster needs more attention in alleviating poverty because it has the smallest average variable value of 57.09. Another research conducted by Marsita et al (2021) regarding clustering of earthquake-prone areas in West Sumatra using the CLARA method and the DBSCAN method (Density Based Spatial Clustering Of Applications With Noise) obtained the same number of clusters as 5 clusters and a better algorithm is the algorithm CLARA with an Average Silhouette Width value of 0.57.

Based on the description above, researchers are interested in grouping poverty using a grouping method by comparing the K-Medoids and CLARA algorithms. The aim of this research is to group districts/cities in Indonesia based on factors that influence poverty in 2020 data. The research results can be implemented to help the government in alleviating poverty.

2. K-Medoids Algorithm , ClARA Algorithm, Silhouette Coefficient , and Cluster Validation

Cluster analysis aims to group objects based on similarities or similarities in their characteristics. Similarity between objects can be measured using distance measurements. Several distance measures that are often used to measure the degree of similarity of objects include Euclidean , Minkowski , and Manhattan /city block distances. (Johnson & Wichern, 2007) . In this study, Euclidean distance was used.

K-Medoids is a clustering algorithm that is used to find medoids which are the center point of a cluster. The K-Medoids algorithm is less sensitive to outliers when compared to other algorithms. K -Medoids can find k as representative objects to minimize the number of dissimilarities in data objects (Arora et al., 2016) . The steps that can be used to apply the K- Medoids algorithm are as follows:

a) Determines the number of clusters.
b) Determines initial medoids randomly.
c) Calculate the distance of each data to the initial medoids using Euclidean distance with formula (1).

$$d(x, y_m) = \sqrt{\sum_{i=1}^{a} (x_i - y_{im})^2}\,, \tag{1}$$

with $d(xy_m)$ represents the distance between objects x and y, a represents the number of attributes, m represents medoids, $x_i$represents the i- th data ,y_i represents the ith medoids data .

d) Calculate the total distance value by adding up all the calculation results in step 3.
e) Next, the data in each cluster is randomly selected as a candidate for new medoids.
f) Use the new medoids to calculate the distance of each data in each cluster and calculate the total distance value.
g) Calculate the total deviation (S) with the following formula.

$$S = \sum d(x, y_{mb}) - \sum d(x, y_{ma}). \tag{2}$$

h) If S < 0 is obtained then repeat steps 5 to 7, but if S > 0 then stop selecting medoids randomly.
i) Distribute each data into clusters with the smallest distance.

CLARA (Clustering Large Application) was introduced by Kauffman and Rousseeuw to handle large data sets. The CLARA algorithm is one of the methods in the K-Medoids algorithm where the CLARA algorithm is almost the same as the Partition Around Medoids (PAM) algorithm. The CLARA algorithm is a sampling- based algorithm , where samples are taken randomly (Chinchmalatpure & Dhore, 2020) . The stages of the CLARA algorithm are as follows:

a) Determines the number of clusters.
b) Select a sample of 40+2 k samples or it could be called m samples.
c) Choose k initial cluster centers (medoid s) from m samples.
d) Calculate the distance of each data to each medoid s in each cluster using Euclidean distance with formula (1).
e) Calculate the total distance by adding up all the calculation results in step 4.
f) Randomly select a new candidate medoid s then calculate the distance of each data to the new candidate medoid s using Euclidean distance and calculate the total distance.
g) Calculate the total deviation (S) by calculating the total distance in the new medoids minus the total distance in the initial medoids using the formula in equation (2).
h) If you get S < 0 then repeat steps 6 to 7, but if S > 0 then stop selecting medoids randomly.
i) Distribute each data into clusters with the smallest distance.

In data analysis, it is important to identify outliers that are far from outside the overall pattern of data. An outlier can occur due to various things, for example errors in data collection, errors in the sampling process, and so on. One test tool that can be used to detect outliers is to detect observations/values that are above the upper limit and below the

lower limit on the boxplot. In making a boxplot, it is first necessary to determine the upper and lower limits of a set of data using quartile values and IQR (Interquartile Range). The following is the formula for calculating the upper and lower limits (Weiss, 2011) .

$$lower\ limit = Q_1 - 1.5(IQR), \tag{3}$$

$$upper\ limit = Q_3 + 1.5(IQR), \tag{4}$$

$$IQR = Q_3 - Q_1, \tag{5}$$

with Q_1 is the first quartile of observational data andQ_3 is the third quartile of observational data.

Clustering analysis must meet the assumption that there will be no symptoms of multicollinearity between the independent variables. The tool used to test multicollinearity in this research is Pearson correlation. The Pearson correlation formula between independent variables X_i and $X_j$with n being the number of observations is as follows (Weiss, 2011) ,

$$r_{x_i x_j} = \frac{n\sum_{i=1}^{n} X_i X_j - \sum_{i=1}^{n} X_i \sum_{i=1}^{n} X_j}{\sqrt{\left(n\sum_{i=1}^{n}(X_i)^2 - \left(\sum_{i=1}^{n} X_i\right)^2\right)\left(n\sum_{i=1}^{n}\left(X_j\right)^2 - \left(\sum_{i=1}^{n} X_j\right)^2\right)}} \tag{6}$$

Testing with Pearson correlation with the criterion that if the correlation value $(r_{x_i x_j})$is > 0.8 then it states that there is multicollinearity between variables.

The silhouette coefficient is useful in determining the number of clusters or k and how good or bad the placement of n objects in a cluster is. This method is a combination of separation and cohesion values. The formula for calculating the silhouette coefficient is as follows (Rousseeuw, 1987) ,

$$s(i) = \frac{b(i) - a(i)}{max(a(i), b(i))}\ , i = 1,2, \dots, n, \tag{7}$$

with, $s(i)$ states the silhouette coefficient value of the ith data, $a(i)$ states the average distance of the i- th data, $b(i)$ states the average distance of the ith data from all observations.

According to Bunkers & Miller Jr (1994), to see the performance between one method and another or to see the quality of a method or the best clustering algorithm, you can use criteria based on standard deviation values. The standard deviations seen are the standard deviation within clusters (S_w) and the standard deviation between clusters (S_b). Calculating the value $S_w$can use formula (8) as follows:

$$S_w = \frac{1}{K}\sum_{k=1}^{K} S_k\ , \tag{8}$$

with Kstates the number of clusters formed, $S_k$expresses the standard deviation of the kth cluster . If a cluster is given c_k, where $k = 1, \dots, K$, and each cluster has members x_{ki}, where $i = 1, \dots, n_k$and $n_k$are the number of members of each cluster, then to find the standard deviation value of the kth group , namely,

$$S_k = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n_k}(x_{ki} - \bar{x}_k)^2}\ , \tag{9}$$

with notes $\bar{x}_k$is

$$\bar{x}_k = \frac{\sum_{1}^{n_k} x_{ki}}{n_k}. \tag{10}$$

If there is an average variable in each cluster k which is denoted by $\bar{x}_k$then the components of each cluster are different, and the standard deviation between groups (S_b) can be formulated as follows,

$$S_b = \sqrt{\frac{1}{(K-1)}\sum_{k=1}^{K}(\bar{x}_k - \bar{X})^2}\ , \tag{11}$$

with notes $\bar{X}$is

$$\bar{X} = \frac{\sum_1^K \bar{x}_k}{K}. \tag{12}$$

The smaller the value $S_w$and the greater the value S_b, the method has good performance, in the sense that the method has high homogeneity. The criterion for the best algorithm is if the ratio of standard deviation within clusters (S_w) to standard deviation between clusters (S_b) is the smallest among other algorithms.

RESEARCH METHODOLOGY

This research uses secondary data found on the BPS website for each province in Indonesia. The data used in this research is data on factors that influence poverty in Indonesia in 2020 there were 514 districts/cities in Indonesia. The variables used are the Human Development Index (HDI), Gross Regional Domestic Product (GRDP), Open Unemployment Rate (TPT), Population Density.

The steps in the research are:

Carrying out collection data which is a variable in research.

Perform data preprocessing which includes checking missing values and outlier detection.

Perform multicollinearity testing.

Applying the K-Medoids and CLARA algorithms to the data on factors that influence poverty that have been obtained.

Perform cluster validation to determine the best algorithm.

Draw conclusions from the results of the cluster analysis that has been carried out and see which algorithm is better between the K-Medoids and CLARA algorithms.

3. K-Medoids and CLARA Clustering Analysis of Data on Factors Affecting Poverty in Indonesia

Before grouping using the K-Medoids and CLARA methods, data preprocessing can be carried out , namely by checking for missing values and outliers in the data. The following are the results of checking missing values and outlier data.

Table 1: Results of missing value checks

HDI	GRDP	TPT	KP
Mode: Logical	Mode: Logical	Mode: Logical	Mode: Logical
False: 514	False: 514	False: 514	False: 514

Based on Table 1, which is the processed result of the Rstudio software output, it can be seen that for each variable the result is False: 514, meaning there are no missing values in the data on factors that influence poverty in 2020, meaning they total 514. Next is the detection of outliers which can be seen in Figure 1.

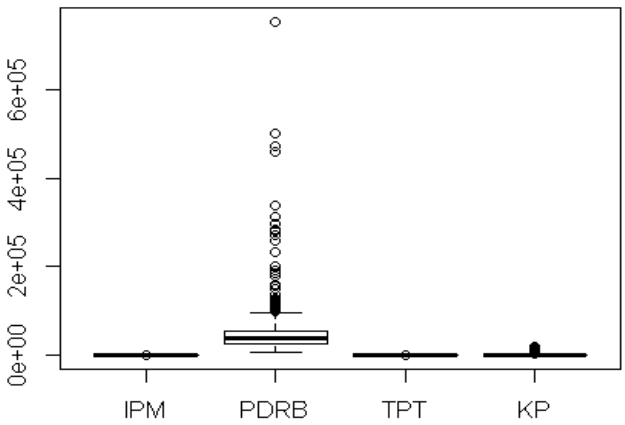

Figure 1: Outlier detection results

From Figure 1 above, it can be seen that there are outliers in the GRDP variable, so the data must be standardized first before analysis is carried out without eliminating the outlier data.

Next, testing the symptoms of multicollinearity can be seen from the correlation coefficient between independent variables which can be calculated based on equation (6). The results of the multicollinearity test can be seen in Table 2.

Table 2: Multicollinearity test results

Variable	HDI	GRDP	TPT	KP
HDI	1.000	0.307	0.585	0.505
GRDP	0.307	1.000	0.261	0.344
TPT	0.585	0.261	1,000	0.525
KP	0.505	0.344	0.525	1.000

From Table 2 it can be seen that there are no correlation coefficient values above 0.8, so the assumption that multicollinearity does not occur is met. In connection with the absence of multicollinearity, the next step is to determine the number of clusters.

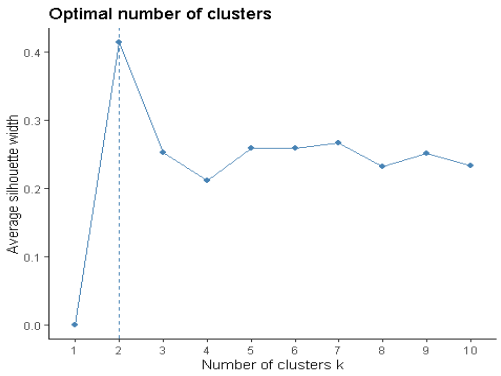

Figure 2: Value of the number of clusters for the K-Medoids algorithm

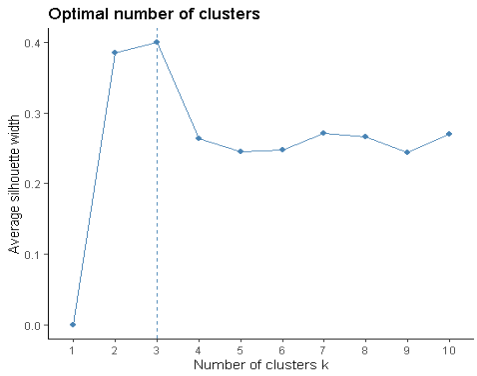

Figure 3: Value of the number of algorithm clusters CLARA

Based on Figure 2, it can be seen that the highest point is for the $k = 2$K-medoids algorithm. This indicates that the optimal number of clusters is 2 clusters. Meanwhile, in Figure 3, the highest point is $k = 3$for the CLARA algorithm. This indicates that the optimal number of clusters is 3 clusters. Next, carry out the analysis stage using the K-Medoids algorithm.

At this stage the data used is standardized data. The first step in the K-Medoids (PAM) algorithm is determining the value of k. In the previous discussion, the value of k for the K-Medoids algorithm was obtained, namely 2 clusters and then selecting k medoids. As for how to choose the initial medoids, namely randomly, the medoids chosen are as follows.

Table 3: Initial Medoids in the K-Medoids algorithm

District/City	HDI	GRDP	TPT	KP
South Tapanuli	0.075	-0.042	-0.413	-0.399
North Aceh	-0.046	-0.367	1,099	-0.348

Based on Table 3 it can be seen that South Tapanuli Regency $(y_1)$and North Aceh Regency (y_2)selected to be the initial medoids. Next, calculate the closest distance value for each research object $(x)$with the initial medoids that have been determined using the Euclidean distance in equation (1). After getting the results, continue with the second iteration in the same way , namely selecting the second candidate medoids , as in Table 4.

Table 4: Second Medoids in the K-Medoids algorithm

District/City	HDI	GRDP	TPT	KP
North Buton	-0.270	-0.314	-0.527	-0.405
Sibolga City	0.614	0.154	0.894	0.397

Based on Table 4 , it can be seen that North Buton Regency $(y_1{}^*)$and Sibolga City $(y_2{}^*)$were chosen as the second medoids , so we can recalculate the distance of each research object $(x)$to the second medoids using the same distance, namely Euclidean distance .

After obtaining the results of the first and second iterations, then look for the total deviation S. The condition is if $S < 0$then exchange the objects with cluster data r to create a new set of k objects as medoid s . The S deviation is calculated by finding the difference between the total distance in the second iteration and the total distance in the first iteration and produces an S deviation value of 27.03. It can be seen that the deviation value S > 0, so the iteration process is stopped. After going through an iteration process, the members of each cluster are obtained. The results obtained were that there were cluster 1 with 342 districts/cities and cluster 2 with 172 districts/cities.

The next data processing analysis of factors that influence poverty in Indonesia uses the CLARA algorithm. The following are the calculation steps with a total of 3 clusters. Next, randomly select samples $40 + 2k$where k is known, namely 3, then 46 samples are taken from the research data. The next step is to select the initial medoids. The initial medoids of k were selected from the 46 sample data above, so the medoids selected are as shown in Table 5 below.

Table 5: Initial medoids in the CLARA algorithm

District/City	HDI	GRDP	TPT	KP
East Waringin City	0.258	0.038	-0.110	-0.408
Barru	0.210	-0.189	0.306	-0.359
Central Jakarta City	1.805	11.233	1.979	7.169

Based on Table 5, it can be seen that those selected as initial medoids were East $(y_1{}^{\#})$Waringin City, Barru Regency $(y_2{}^{\#})$, and Central Jakarta City $(y_3{}^{\#})$. Next, calculate the closest distance value for each research object $(x)$with the initial medoids using Euclidean distance calculations. After getting the results, continue with the second iteration in the same way, namely selecting the second candidate medoids, as in Table 6 below.

Table 6: Second Medoids in the CLARA algorithm

District/City	HDI	GRDP	TPT	KP
Sambas	-0.399	-0.255	-0.673	-0.383
Sibolga City	0.614	0.154	0.894	0.397
Central Jakarta City	1.805	11.233	1.979	7.169

Based on Table 6, it can be seen that the second medoids selected were Sambas $(y_1{}^{\blacksquare})$Regency, Sibolga City $(y_2{}^{\blacksquare})$, and Central Jakarta City $(y_3{}^{\blacksquare})$. Next, the distance between each research object and the second medoid can be recalculated (x)using Euclidean distance.

second iterations, then calculate the total deviation S. The condition is if $S < 0$then exchange the objects with the cluster data r to create a new set of k objects as medoid s . The S deviation is calculated by finding the difference value between the total distance in the second iteration and the total distance in the first iteration and produces an S deviation value of 1 6.91. It can be seen that the deviation value S > 0, so the iteration process is stopped. The results obtained were that there were cluster 1 with 334 districts/cities, cluster 2 with 179 districts/cities, and cluster 3 with only 1 city.

Clustering visualization using the K-Medoids and CLARA algorithms on data on factors influencing poverty in Indonesia is presented in the form of a map shown in Figure 4 and Figure 5.

Figure 4: Map of clustering results with K-Medoids

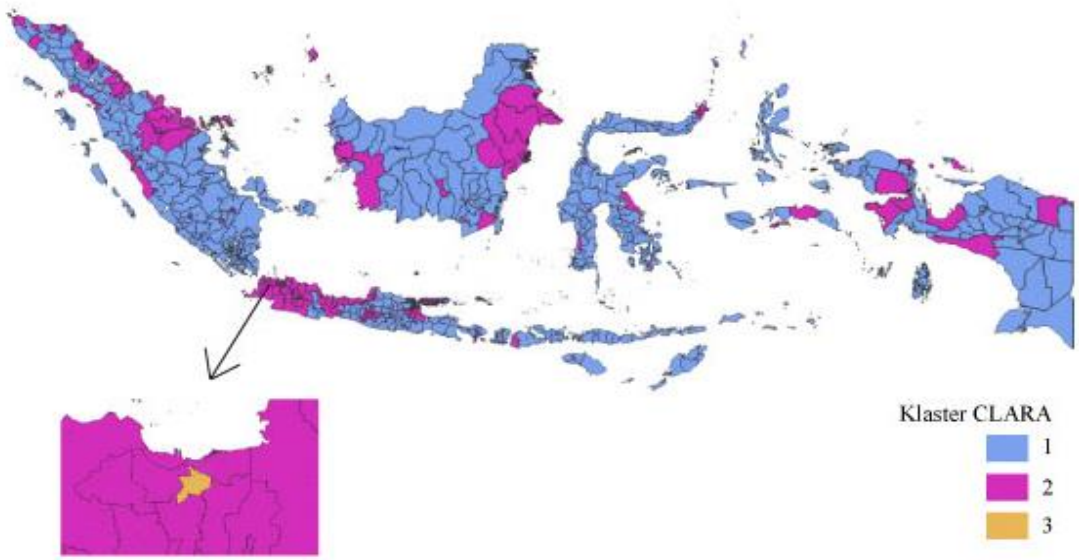

Figure 5: Map of clustering results with CLARA

Furthermore, from the clustering results in the K-Medoids algorithm , the average cluster value is shown in Table 7.

Table 7: Average cluster value in the K-Medoids algorithm

Cluster	HDI	GRDP	TPT	KP
1	-0.421	-0.243	-0.545	-0.341
2	0.838	0.484	1,080	0.678

Based on Table 7, it can be seen that cluster 1 has a smaller average value than cluster 2, so that cluster 1 is a collection of districts/cities that are included in the high poverty category. Next, using the CLARA algorithm, the average cluster value is obtained which is shown in Table 8 below.

Table 8: Average cluster value in the CLARA algorithm

Cluster	HDI	GRDP	TPT	KP
1	-0.432	-0.234	-0.571	-0.342
2	0.797	0.373	1.050	0.598
3	1.800	11.200	1.980	7.170

Based on Table 8, it can be seen that the smallest average value of the 3 clusters is in cluster 1, so that in the CLARA algorithm cluster 1 is a collection of districts/cities that are included in the high poverty category and require primary attention in poverty alleviation.

Next, cluster validation was carried out to see which algorithm was better used for data on factors influencing poverty in Indonesia in 2020. The cluster validation results calculated using equation (8) and equation (11) can be seen in Table 9.

Table 9: Cluster validation results

Method	S_w	S_b	S
K-Medoids	0.546	0.819	0.667
CLARA	0.336	3,161	0.106

Based on Table 9, it shows that the cluster validation value for the K-Medoids algorithm is 0.667, while the cluster validation value for the CLARA algorithm is 0.106. This means that the CLARA algorithm is better than the K-Medoids algorithm because the CLARA algorithm has a smaller cluster validation value

4. Conclusion

Based on the results and discussion, it can be concluded that in the K-Medoids algorithm there were 2 clusters, while in the CLARA algorithm there were 3 clusters. The best algorithm is the CLARA algorithm with the smallest average ratio and standard deviation value, namely 0.106.

References

Arora, P., Deepali, & Varshney, S. (2016). Analysis of k-means and k-medoids algorithm for big data. *Procedia Computer Science*, *78*, 507–512.

Badan Pusat Statistik. (2021). Poverty Profile Statistics in Indonesia. In Central Bureau of Statistics: Jakarta (Ed.), *Poverty Profile in Indonesia September 2020*.

Bunkers, M. J., & Miller Jr, J. R. (1994). Definition of climate regions in northern plains using an objective cluster modification technique. *Journal of Climate*, *9*(1), 130–146.

Chinchmalatpure, M., & Dhore, M. P. (2020). Quality healthcare prediction using k-means and CLARA partition based clustering algorithm for big data analytics. *International Journal of Engineering and Advanced Technology (IJEAT)*, *9*(3), 1140–1144.

Hanafiah, M. A., & Wanto, A. (2020). Implementation of Data Mining Algorithms for Grouping Poverty Lines by District/City in North Sumatra. *IJISTECH (International Journal of Information System and Technology)*, *3*(2), 315-322.

Ipmawan, H., Kristanto, D., Hendrawan, K., & Kuncoro, A. W. (2022). The Influence of The Human Development Index, Unemployment Rate, and Illiteracy Population on Poverty Level in Indonesia for the Period 2015-2020. *MUHARRIK: Jurnal Dakwah dan Sosial*, *5*(1), 89-103.

Johnson, R. A., & Wichern, D. W. (2007). *Applied multivariate statistical analysis (6th ed)*. New Jersey: Pearson education, Inc**.**

Marsita, D., Utami, T. W., & Al Haris, M. (2021). *Clustering daerah rawan gempa di Sumatra Barat menggunakan metode Clustering Large Application dan metode Density-Based Spatial Clustering of Application With Noise*. Universitas Muhammadiyah Semarang

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, *20*(1), 53–65.

Safitri, P. N., Aristawidya, R., & Faradilla, S. B. (2021). Klasterisasi faktor-faktor kemiskinan di Provinsi Jawa Barat menggunakan k-medoids clustering. *Journal of Mathematics Education and Science*, *4*(2), 75–80.

Safri, M. (2021). The Analysis Related To The Factors Which Affect The Poverty Levels Of Districts/Cities In Jambi Province During 2014-2018. *Dinasti International Journal of Education Management and Social Science*, *2*(3), 451-462.

Sari, F. D. R., & Ediwijojo, S. P. (2023). Clustering Analysis Using K-Medoids on Poverty Level Problems in Central Java by District/City. *KnE Social Sciences*, 78-87.

Shmueli, G., Bruce, P. C., Yanhav, I., Petel, N. R., & Lichtendahl Jr, K. C. (2018). *Data mining for bussiness analytics*. United States: John Wiley & Sons.

Tufféry, S. (2011). *Data mining and statistics for decision making* (1st ed.). United Kingdom: John Wiley & Sons.

Weiss, N. A. (2011). *Elementary statistics* (8th ed., Vol. 4, Issue 1). Boston: Pearson.