Parameter: Journal of Statistics
Available at https://bestjournal.untad.ac.id/index.php/parameter
Vol. 3 No. 2 2023, 61 - 67
DOI: *https://doi.org/10.22487/27765660.2023.v3.i2.16906*
eISSN: 2776-5660

IMPLEMENTATION OF THE K-MEDOIDS METHOD IN CLUSTERING HUMAN DEVELOPMENT INDEXES IN INDONESIA

Gustriza Erda[1*], Radhiatul Khaira Usdika[2], Rizka Pitri[3], Zulya Erda[4]

[1,2]Statistic Study Program, Riau University
[3]Islamic Library and Information Science Study Program, Raden Intan State Islamic University
[4]Poltekkes Kemenkes Tanjung Pinang

e-mail: gustrizaerda@lecturer.unri.ac.id

ABSTRACT

The Human Development Index (HDI), which takes into account three fundamental aspects of human existence, a long and healthy life, knowledge, and a reasonable level of living, is one tool used to assess the effectiveness of human progress. Clustering provinces based on the human development index is important so that development disparities can be identified and help identify provinces with high, medium or low levels of development. The purpose of this study was to use the k-medoids approach to perform a cluster analysis of HDI in Indonesia based on life expectancy, average years of schooling, expected years of schooling, and expenditure per capita adjusted for 2022. The analysis indicate that two clusters were created: cluster 1 had a high human development index, while cluster 2 had a low human development index. More provinces belonged to cluster 1 than cluster 2 suggesting that human development index in Indonesia in 2022 was largely in the high category.

Keywords: *Clustering, Human Development, Human Development Index, K-Medoids*

Cite: Erda, G., Usdika K. R., Pitri, R., Erda, Z., (2023). *Implementation of the K-Medoids Method in Clustering Human Development Indexes in Indonesia*. *Parameter: Journal of Statistics, 3*(2), 61-67, https://doi.org/10.22487/27765660.2023.v3.i2.16906.

*Received: Dec 2023; Reviewed: Dec 2023; Published: Dec 2023

INTRODUCTION

The Human Development Index (HDI) is a tool for measuring the success of human development which includes three basic dimensions of human life, namely long and healthy life (measured by the indicator Life Expectancy in years), knowledge (measured by the indicator Expected Years of Schooling in years and Average -average length of schooling in years), and decent living standards (measured by the Per Capita Expenditure indicator in thousands of rupiah). One of the main benefits of HDI is that there is comparability between countries in terms of progress in human development. By comparing HDI between countries, it is possible to see and better understand differences in human development progress around the world. In addition, IPM also helps identify specific issues and challenges faced by certain countries to achieve sustainable human development goals. By observing these differences and issues, policy makers can develop more effective policies and programs to improve human development progress in their country (Badan Pusat Statistik, 2022).

After experiencing difficulties in 2020 due to the COVID-19 pandemic, Indonesia's HDI began to increase in 2021 and 2022. HDI grew by 0.49% in 2021 and 0.86% in 2022, higher than the growth in 2020 during the COVID pandemic -19 started to hit Indonesia, which only grew by 0.03%. HDI growth in 2022 is even greater than growth before the COVID-19 pandemic in 2019, which grew by 0.74%. Increasing the dimensions of a decent standard of living, represented by the adjusted real per capita expenditure indicator, is the main factor driving Indonesia's HDI improvement in 2022. This indicator grew 2.90 percent in 2022, after growing 1.30 percent in the previous year. DKI Jakarta has the highest HDI (81.65) and Papua has the lowest HDI (61.39).

One of the statistical analysis techniques used to group data based on the characteristics of natural variables is clustering analysis. The word "cluster analysis" refers to a group of statistical techniques created especially to identify patterns in large, intricate data sets (Gao et al., 2021; Gore, 2000). The concept of clustering is usually used to group a set of objects into several groups without understanding more about the groups. The main goal of clustering is to group a data set into groups that have almost the same characteristics and each group has unique characteristics that differentiate them from each other. One of the clustering techniques used is the k-medoids clustering technique (Paramartha et al., 2017).

The k-medoids clustering algorithm, which is also often called partitioning around medoids (PAM), is a non-hierarchical clustering method which is a variation of the k-means method. K-medoids can overcome the weakness of k-means which tends to be sensitive to outliers that may deviate from the data distribution (Wicaksono & Yolanda, 2021). Furthermore, research on Human Development Index (HDI) in Riau Province classifying the factors that influence HDI in Riau Province in 2021 was discussed by Erda et al., (2023). It was found that there were three categories, namely service quality, health facilities and economic conditions. This research will then be developed to classify HDI in Riau Province in 2022 based on life expectancy, average years of schooling, expected years of schooling, and expenditure per capita using the k-medoids method.

MATERIALS AND METHODS

This research uses Human Development Index data in the form of Life Expectancy (UHH), Average Years of Schooling (RLS), Expected Years of Schooling (HLS), and Expenditure per Capita adjusted for 2022 which was obtained from the Central Statistics Agency website. The analysis method used in this research is clustering using the k-medoids algorithm. K-Medoids is an algorithm used to find medoids in a group (cluster) which is the central point of a group (cluster). The K-Medoids algorithm is better than K-Means because K-Medoids can find k as representative objects to minimize the number of dissimilarities in data objects, whereas K-Means uses the number of Euclidean distances for data objects (Sindi et al., 2020).

The group analysis used in this research is the non-hierarchical cluster K-Medoid method using Euclidean distance (Supranto, 2004). Euclidean distance can be used when each variable used is orthogonal or uncorrelated, and has the same units and measurement scale (Asroni & Adrian, 2015). Euclidean distance of an object $x' = [x_1, x_2, \dots, x_p]$ and $y' = [y_1, y_2, \dots, y_p]$ has dimension p, written in equation (1) below.

$$d(x, y) = \sqrt{(\mathrm{x1} - \mathrm{y1})^2 + (\mathrm{x2} - \mathrm{y2})^2 + \cdots + (\mathrm{xp} - \mathrm{yp})^2} \quad (1)$$

$$= \sqrt{\sum_{i=1}^{p} (x_i - y_i)^2}$$

With:

$d(x, y)$ = Distance of data from x to cluster center y

$x1, y1$ = First point coordinates

$x2, y2$ = Second point coordinates

p = Lots of observations

The K-Medoids algorithm is called partitioning around medoids. In the partition method, data consisting of n objects is divided into k groups with the number k ≤ n. Medoids are objects that are considered to represent groups and group centers. The k-medoids algorithm uses objects in a collection of objects that represent groups. This group is formed by calculating the distance between the center point and non-center objects. This analysis uses the absolute error (E) value to minimize inequality for each object in the cluster Han & Kamber (2006).

$$E = \sum_{c-1}^{k} \sum_{i=1}^{n_c} |P_{ic} - O_c|$$

with:

E = sum of absolute errors for all objects c

n_c = Number of objects in the c-th cluster

P_{ic} = Non-medoids object i in the cth cluster

O_c = The value of medoids in the c-th cluster

In this k-medoid method, the initial medoids are k objects chosen at random, then non-medoid objects that are similar to the medoids are grouped so that they become one. If the hope is to get good cluster quality, then the process of substituting medoid objects with non-medoids is carried out through an iterative process. This quality can be calculated using the absolute error (E) value both before and after the replacement process until Oc does not change (Mustajab et al., 2021).

RESULTS AND DISCUSSION

Descriptive statistics

Table 1 shows that life expectancy in Indonesia in 2022 were in the range of 65 to 75 years with an average of 69.94 years and a standard deviation of 2.470. For the average length of schooling, in general the length of schooling in Indonesia is 12,74 years or the equivalent of a high school/vocational school education level, with a minimum of 11 years or the equivalent of junior high school and a maximum of 15 years or the equivalent of high school. The expected length of schooling, the average is 8,38 years with a standard deviation of 0.992 and is in the range of 7 to 11 years. Meanwhile, from an economic perspective, the average per capita expenditure has a minimum value of 7.00 and a maximum value of 18.00 with an average value of 10.56.

Table 1. Description of Human Development Index Indicator Data in Indonesia

Variabel	Minimum	Maximum	Mean	Standar Deviaton
Life Expectancy (UHH)	65,00	75,00	69,94	2,470
Average Years of Schooling (HLS)	11,00	15,00	12,74	0,742
Long School Expectations (RLS)	7,00	11,00	8,38	0,992
Expenditure per Capita (PP)	7,00	18,00	10,56	2,246

Detecting Outliers

Before continuing to group data, outlier detection will first be carried out on the data used. Data that contains outliers can be identified using boxplot visualization. The results of the boxplot can be seen in Figure 1 below:

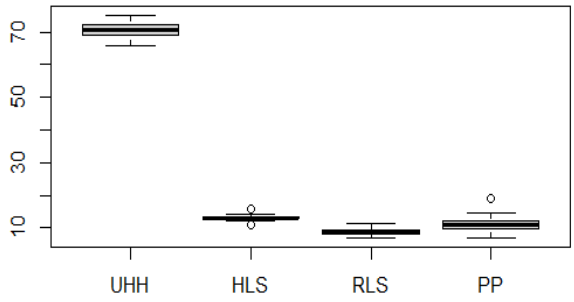

Figure 1. Boxplot of Human Development Index Data in Indonesia in 2022

Figure 1 presents the results of outlier testing using a boxplot. Outlier data is data that is too far from the data center as indicated by data that is out of the boxplot range. Figure 1 shows that there are no outlier data on the Life Expectancy (UHH) variable and the Average Years of Schooling (RLS) variable and there are outlier data on the Average Years of Schooling (HLS) in East Java province with a value of 3.288976 and Expenditure per capita (PP) in West Java province with a value of 3.492831 which is at the top of the boxplot. In this study, outlier data is maintained so that each observation (province) remains represented.

Determination of K-Optimal Silhouette Coefficient for Clustering

One way to determine the optimal cluster is by using the silhouette coefficient method. This method measures how similar an object is to its own group compared to other groups. Silhouette Coefficient values range from -1 to 1, where 1 indicates that the groups are well separated, 0 indicates that the groups are not different, and -1 indicates that the groups are attributed incorrectly (Hartama & Anjelita, 2022). A higher Silhouette score indicates better clustering. The graphic results of the silhouette coefficient method can be seen in Figure 2 below:

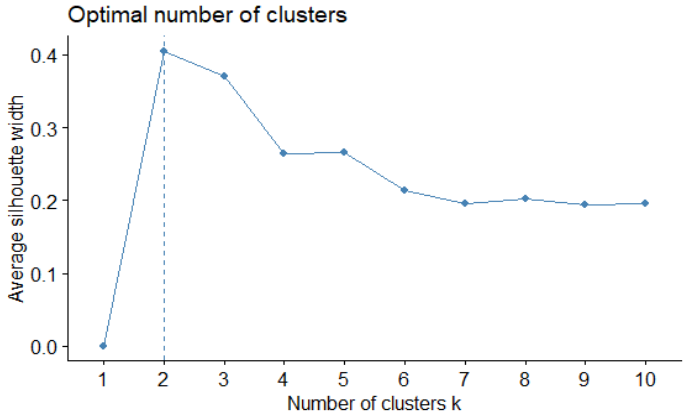

Figure 2. Silhouette Coefficient Graph

Figure 2 shows that the cluster with the highest coefficient value was in cluster 2 (k=2), that was 0.4. Meanwhile, other clusters had coefficient values below that value.This is indicated by the appearance of a dotted vertical line on the x axis at number 2 meaning k = 2 was the best cluster for grouping HDI in Inodesia in 2022.After the number k is determined, the next step is to carry out the cluster process.

K-Medoids Clustering Results

After determining the number of clusters (k), the next step is to group them using the k-medoids clustering method. In Table 2, the central point or medoid values of each cluster are presented, which are obtained from the cluster results.

Table 2. Cluster Center Point

K	*Medoids*	UHH	RLS	HLS	PP
1	27	70,97	13,53	8,63	11,43
2	33	66,46	13,21	7,84	8,10

Based on the results displayed in Table 2, it was found that the medoid observation or central point of the 1st cluster was the 27th observation, that was province of South Sulawesi and the center point of the 2nd cluster was the 33rd observation that was province of West Papua. This region was formed naturally based on the similarity measured from the distance of each non-medoid point to the nearest medoid point. The non-medoid objects closest to each medoid will merge into a single cluster.

The cluster results obtained using the k-medoids method obtained from all data, there were 2 clusters with cluster 1 totaling 27 provinces and cluster 2 totaling 7 provinces. The k-medoids results can be seen in Table 3

Table 3. Results of K-Medoids Clustering Provinces in Indonesia

Cluster	Hasil *Cluster*
Cluster 1	Aceh, Bali, Banten, Bengkulu, DI Yogyakarta, DKI Jakarta, Jambi, West Java, Central Java, East Java, Bangka Belitung Islands, West Kalimantan, Central Kalimantan, South Kalimantan, East Kalimantan, North Kalimantan, Riau Islands, Lampung, Riau , North Sumatra, West Sumatra, South Sumatra, North Sulawesi, South Sulawesi, Southeast Sulawesi, Central Sulawesi.
Cluster 2	Maluku, North Maluku, West Nusa Tenggara, East Nusa Tenggara, West Sulawesi, West Papua, Papua.

Based on Table 3, it can be seen that the values for life expectancy, average length of schooling, expected length of schooling, and high per capita income for the provinces included in cluster 1 are Sumatra, Java, Kalimantan, and parts of Sulawesi. Cluster 1 is a cluster with a high development index, this is characterized by the presence of infrastructure, good health and high education. The provinces included in this cluster are among the provinces that are developing rapidly.

In cluster 2, the value of life expectancy, average years of schooling, expected years of schooling, and adjusted per capita income are lower when compared to cluster 1. The provinces included in this cluster are NTT, NTB, West Sulawesi, North Maluku, Maluku, Papua and West Papua. Low life expectancy in an area should be followed by health development programs and other social programs including environmental health, nutritional and calorie adequacy, including education programs. In terms of characteristics and demographics, the Sabang area and its surroundings will certainly be superior to the areas that are part of Merauke and its surroundings. This is because the location of Merauke is far from the capital so that the infrastructure is still minimal, transportation is inadequate and other things.

From the cluster results, it can be seen that the Human Development Index data is mostly found in cluster 1, namely 27 out of 34 provinces in Indonesia. Next, using R 4.3.1 software, a visualization of the formation of 2 clusters is obtained which is presented in Figure 3.

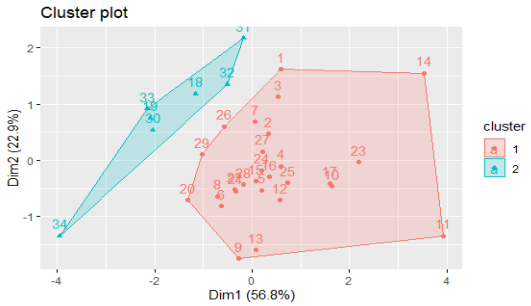

Figure 3. Visualization of Provincial Cluster Results in Indonesia

The visualization results in Figure 3 show that the red part of the plot is a member of cluster 1 and the blue part of the plot is a member of cluster 2. Based on the cluster results obtained, it can be seen that most of the data is gathered in cluster 1 with the diversity value produced by dimension 1 amounting to 56.8% and dimension 2 amounting to 22.9%, so that overall the diversity that can be explained by these two dimensions is 79.7% of the total diversity of the actual data.

Interpretation of Cluster Results

The characteristics of each cluster formed can be seen in the interpretation of the results of each cluster. This can be seen from the results of calculating the average value of the variables in each cluster, which are presented in Table 4.

Table 4. Average of variables in each cluster

K	UHH	RLS	HLS	Expenditure	N
1	71,4	13,2	8,99	11,7	27
2	66,9	13,2	8,24	8,63	7

Information:
k : Cluster
AHH : Life Expectancy (Years)
RLS : Average Years of Schooling (Years)
HLS : Expected Years of Schooling (Years)
PP : Adjusted Per Capita Expenditure (Million Rupiah)
n : The number of regions in the cluster

Interpretation of cluster results is the final process of cluster analysis. Based on the total average of the four variables shown in Table 4, it shows that the total average in cluster 1 wass greater than cluster 2, so that cluster 1 was categorized as a group that has a high human development index and cluster 2 was categorized as a group that had a low level of human development index. The interpretation of the two clusters is as follows:

1st Cluster

The results of cluster 1 were a grouping of human development index indicators based on high levels of HDI. Based on the average value of cluster 1 in Table 4, the life expectancy variable had an average value of 71.4 years, average length of schooling was 13.2 years, expected length of schooling was 8.99 years and per capita expenditure was 11 ,7 million. Therefore, cluster 1 was categorized as a cluster with a high level of human development index.

2nd Cluster

The results of cluster 2 were a grouping of human development index indicators based on low HDI levels. Based on the average value of cluster 2 in Table 4, the life expectancy variable had an average value of 66.9 years, average length of schooling was 13.2 years, expected length of schooling was 8.24 years, and per capita expenditure was 8.63 million rupiah. Therefore, cluster 2 wass categorized as a cluster with a low level of human development index.

CONCLUSION

The results of the human development index indicator data cluster based on life expectancy, expected length of schooling, average length of schooling, and per capita expenditure using the k-medoids method produced 2 clusters. The results in cluster 1 were a grouping of human development index indicators based on a high level of human development index which consists of the provinces of Aceh, North Sumatra, West Sumatra, Riau, Jambi, South Sumatra, Bengkulu, Lampung, Bangka Belitung Islands, Riau Islands, DKI Jakarta, West Java, Central Java, DI Yogyakarta, East Java, Banten, Bali, West Kalimantan, Central Kalimantan, South Kalimantan, East Kalimantan, North Kalimantan, Gorontalo, North Sulawesi, South Sulawesi, Southeast Sulawesi, Central Sulawesi. The results in cluster 2 were a grouping of human development index indicators based on a low level of human development index which consists of the provinces of West Nusa Tenggara, East Nusa Tenggara, West Sulawesi, Maluku, North Maluku, West Papua, Papua. This shows that there is a need for more attention from the government to the 7 provinces so that the HDI in the provinces in cluster 2 increases further.

REFERENCES

Badan Pusat Statistik. (2022). Indeks Pembangunan Manusia. Badan Pusat Statistik, https://news.ge/anakliis-porti-aris-qveynis-momava.

Erda, G., Mega Aulia, S., & Erda, Z. (2023). Classifiying The Factors Influencing The Human Development Index in Riau Province using Principal Component Analysis. Parameter: Journal of Statistics, 2(3), 17–23. https://doi.org/10.22487/27765660.2022.v2.i3.16203

Gao, S., Meng, F., Gu, Z., Liu, Z., & Farrukh, M. (2021). Mapping and Clustering Analysis on Environmental, Social and Governance Field a Bibliometric Analysis Using Scopus. Sustainability, 13(13), 7304. https://doi.org/10.3390/su13137304

Gore, P. A. (2000). Cluster Analysis. In Handbook of Applied Multivariate Statistics and Mathematical Modeling (pp. 297–321). Elsevier. https://doi.org/10.1016/B978-012691360-6/50012-4

Hartama, D., & Anjelita, M. (2022). Analysis of Silhouette Coefficient Evaluation with Euclidean Distance in the Clustering Method (Case Study: Number of Public Schools in Indonesia). Journal Mantik, 6(3), 667–3677.

Mustajab, R., Aristawidya, R., Puspita, L., & Widodo, E. (2021). Aplikasi Metode K-Medoid pada Pengelompokan Kabupaten/Kota di Jawa Barat berdasarkan Indikator Indeks Pembangunan Manusia Tahun 2020. Jurnal Statistika Dan Aplikasinya, 5(2), 221–229.

Paramartha, G. N. W., Ratnawati, D. E., & Widodo, A. W. (2017). Analisis Perbandingan Metode K-Means Dengan Improved Semi-Supervised Analisis Perbandingan Metode K-Means Dengan Improved Semi- Supervised K-Means Pada Data Indeks Pembangunan Manusia (IPM). Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer, Vol. 1(9), 813–824.

Sindi, S., Ningse, W. R. O., Sihombing, I. A., R.H.Zer, F. I., & Hartama, D. (2020). Analisis Algoritma K-Medoids Clustering Dalam Pengelompokan Penyebaran Covid-19 Di Indonesia. Jurnal Teknologi Informasi, 4(1), 166–173. https://doi.org/10.36294/jurti.v4i1.1296

Wicaksono, A. S., & Yolanda, A. M. (2021). Pengelompokkan Kabupaten / Kota di Provinsi Nusa Tenggara Timur Berdasarkan Indikator Indeks Pembangunan Manusia Menggunakan K-Medoids Clustering Penyedia Data Statistik Berkualitas untuk Indonesia Maju Pengelompokkan Kabupaten / Kota di Provinsi Nusa Ten. Statistika Terapan, 1(1), 79–90.