Application of Data Mining Using the K-Medoids Algorithm for Poverty Index Clustering

Muhammad Faisal[*1], Wiranti Sri Utami[2]
[1,2]Informatics Technology Study Program Faculty of Science and Technology
Raharja University
E-mail: [*1]**muhammad.faisal@raharja.info**, [2]wiranti.utami@raharja.info

Abstract

Poverty index is a term for measuring poverty, this is done by a government agency or commonly referred to as the Central Statistics Agency (BPS). The poverty index or poverty rate is the percentage of the population in a province who is below the poverty line, which is the minimum in obtaining an adequate standard of living. In the government's efforts to reduce the level of poverty in a province, the government often provides special assistance programs for people belonging to the poverty line. Based on the explanations that have been discussed, a conclusion can be drawn. This research can be done using the Data Mining technique to group the total Poverty Index by Province in Indonesia using the K-Medoids Algorithm, then by determining the Clusters randomly. The results of this study are expected to assist the government in providing assistance to the affected population below the poverty line.

Keywords *— Data Mining, K-Medoids, Poverty Index*

1. INTRODUCTION

Poverty is a condition when you are unable to meet basic needs such as food, clothing, education, housing, and health. Poverty can be caused by difficulty in obtaining basic needs and not having the opportunity to get a proper education or job. Poverty is a global problem, This does not only happen in Indonesia but also in several other countries. In the government's efforts to reduce poverty, the government is aggressively providing assistance including direct cash assistance, education assistance, health assistance, and others.

Each province in the Republic of Indonesia has various categories of Poverty Index, this is stated in the form of an information table on the annual poverty index compiled by the government agency, namely the Central Statistics Agency (BPS). To obtain information about the classification of the poverty index into low and high categories, a study is needed. The research that will be conducted by the author is related to the Poverty Index using data mining techniques with the K-Medoids method. The K-Medoids method can group each data obtained based on the cluster to be tested.

2. RESEARCH METHOD

Based on the results of the explanation of the problems behind this research, this research will be carried out through several stages, including the following:

2.1. Identification of problems

In this study, the first step is to determine the K-Medoids algorithm in defining the problem and the final goal to be achieved, by determining the number of clusters to be selected on the Poverty Index problem. In this study, the author uses data obtained from the Central Statistics Agency (BPS). The data is a poverty index in 2020 to 2021. The data obtained will then be selected by determining the attributes, normalization and transformation of the data that will be selected for clustering.

2.2. Literatur Review

1) Research conducted by Muh. Arifandi, et al (2021) entitled "Implementation of the K-Medoids Algorithm for Clustering Areas Infected with Covid19 Cases in DKI Jakarta." This study discusses the grouping or clustering of areas in DKI Jakarta that are infected with the COVID19 Virus. The results of this study are expected to assist the DKI Jakarta Government in provide handling of COVID19 cases in accordance with the clusters obtained using the K-Medoids algorithm [1].
2) Research conducted by Fitri Hardiyanti, et al (2019) with the title "Application of the K-Medoids Clustering Method in Handling Diarrhea Cases in Indonesia". , the results obtained from this study are for grouping cases of handling diarrheal diseases according to clusters in 31 provinces in Indonesia [2].
3) Further research conducted by Nurliana Pulungan, et al (2019) entitled "Application of the K-Medoids Algorithm for Grouping the Population 15 Years Old and Over by Main Job." This study discusses the main occupations of the age group 15 years and over. By applying the data mining method using the K-Medoids algorithm, it is expected to be able to find out the type of main job grouping aged 15 years and over [3].
4) Research conducted by Siti Nurlaela, et al (2020) entitled "K-Medoids Algorithm for Ulcer Clustering in Karawang Regency" This study discusses the percentage of ulcer disease in 2017 to 2019, Application of the K-Medoids algorithm in this study to knowing and giving priority to areas with the highest ulcer disease in Karawang district with data mining methods and data grouping or clusters [4].
5) Research conducted by Siti Asmiatun, et al (2020) with the title "Application of the K-Medoids Method for Grouping Road Conditions in Semarang City." This study discusses the condition of road damage in Semarang City when the rainy season arrives. This problem can be overcome by providing updated information about damaged road conditions by grouping road condition data using the K-Medoids method [5].
6) Research conducted by Daffa Rafif Agustian, et al (2022) with the title "Dengue Hemorrhagic Fever Clustering Analysis with the K-Medoids Algorithm (Case Study of Karawang Regency)". This study discusses dengue hemorrhagic fever caused by Aedes aegyti and Aedes albopictus mosquitoes. The purpose of this study was to classify areas with the highest and lowest dengue cases in Karawang Regency using Data Mining with the K-Medoids method [6].
7) Research conducted by Sukma Sindi, et al (2020) with the title "K-Medoids Clustering Algorithm Analysis in Grouping the Spread of Covid-19 in Indonesia". This study

discusses the spread of the Covid-19 virus in Indonesia, due to the large amount of data that is accommodated, a method is needed to classify data using K-Medoids [7].

2.3. Method Data Mining

Data mining is a technique for data analysis that aims to obtain definite relationships and provide useful results for data owners because they can obtain previously unknown results. Data Mining is also called a method that can be used to extract previously unknown predictive information data on a database set.

2.4. K-Medoids

The K-Medoids algorithm or Partitioning Around Method (PAM) is a type of non-hierarchical cluster method and is a method of the K-Means variant. The workings or formula of the K-Medoids algorithm is as follows:

$$d_{ij} = \sqrt{\sum_{a}^{p} = 1 \left(x_{ia} - x_{ja}\right)^2} = \sqrt{(x_i} - x_j)' (x_i - x_j)' (1)$$

The K-Medoids algorithm computationally has a weakness compared to the K-Means algorithm, due to the calculation of the medoids value based on the frequency that has occurred.

3. RESEARCH RESULTS AND DISCUSSION

This study uses data obtained from the Central Statistics Agency (BPS). This data is the Poverty Index in 2020 and 2021.

Table 1. Poverty Index of the Central Statistics Agency (BPS)

Provinsi	Indeks Kedalaman Kemiskinan (P1) Menurut Provinsi dan Daerah (Persen)																	
	Perkotaan						Perdesaan						Perkotaan & Perdesaan					
	2020			2021			2020			2021			2020			2021		
	Semester 1 (Maret)	Semester 2 (September)	Tahunan	Semester 1 (Maret)	Semester 2 (September)	Tahunan	Semester 1 (Maret)	Semester 2 (September)	Tahunan	Semester 1 (Maret)	Semester 2 (September)	Tahunan	Semester 1 (Maret)	Semester 2 (September)	Tahunan	Semester 1 (Maret)	Semester 2 (September)	Tahunan
Aceh	1.55	1.61	-	1.79	1.70	-	3.28	3.46	-	3.40	3.59	-	2.72	2.85	-	2.86	2.95	-
Sumatera Utara	1.48	1.54	-	1.50	1.40	-	1.55	1.67	-	1.54	1.51	-	1.51	1.60	-	1.52	1.45	-
Sumatera Barat	0.74	0.80	-	0.87	0.74	-	1.07	1.17	-	1.21	1.18	-	0.92	0.99	-	1.04	0.96	-
Riau	0.97	1.02	-	1.01	1.12	-	1.25	1.53	-	1.09	1.08	-	1.14	1.32	-	1.06	1.09	-
Jambi	1.71	1.77	-	2.20	1.81	-	0.80	0.89	-	0.85	0.74	-	1.10	1.18	-	1.29	1.09	-

Provinsi	Indeks Kedalaman Kemiskinan (P1) Menurut Provinsi dan Daerah (Persen)																	
	Perkotaan						Perdesaan						Perkotaan & Perdesaan					
	2020			2021			2020			2021			2020			2021		
	Semester 1 (Maret)	Semester 2 (September)	Tahunan	Semester 1 (Maret)	Semester 2 (September)	Tahunan	Semester 1 (Maret)	Semester 2 (September)	Tahunan	Semester 1 (Maret)	Semester 2 (September)	Tahunan	Semester 1 (Maret)	Semester 2 (September)	Tahunan	Semester 1 (Maret)	Semester 2 (September)	Tahunan
Maluku Utara	0.38	0.65	-	0.97	0.96	-	1.16	1.26	-	0.97	0.93	-	0.94	1.09	-	0.97	0.94	-
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
Papua	0.62	0.82	-	0.85	0.86	-	8.37	9.34	-	7.52	8.53	-	6.16	6.90	-	5.60	6.31	-

3.1. Data Processing

The data that has been obtained is then carried out in a data cleaning process to avoid noise or inconsistent data. The process of data cleaning that has been done can be seen in the table below.

Table 2. Data Cleaning

Provinsi	Indeks Kedalaman Kemiskinan (P1) Menurut Provinsi dan Daerah (Persen)					
	Perkotaan		Perdesaan		Perkotaan & Perdesaan	
	2020		2020		2020	
	Semester 1 (Maret)	Semester 2 (September)	Semester 1 (Maret)	Semester 2 (September)	Semester 1 (Maret)	Semester 2 (September)
Aceh	1.55	1.61	3.28	3.46	2.72	2.85
Sumatera Utara	1.48	1.54	1.55	1.67	1.51	1.60
Sumatera Barat	0.74	0.80	1.07	1.17	0.92	0.99
Riau	0.97	1.02	1.25	1.53	1.14	1.32
Jambi	1.71	1.77	0.80	0.89	1.10	1.18
Maluku Utara	0.38	0.65	1.16	1.26	0.94	1.09
...	...	...	...	...	...	...
Papua	0.62	0.82	8.37	9.34	6.16	6.90

3.2. Define Attributes

Poverty Index data obtained from BPS has many attributes. In order to get the appropriate cluster results, attribute determination is carried out. These attributes can be seen in the table below.

Table 3. Attributes

ATRIBUT	INISIAL
Perkotaan 2020 Semester 1 (Maret)	X1
Perkotaan 2020 Semester 2 (September)	X2
Perdesaan 2020 Semester 1 (Maret)	X3
Perdesaan 2020 Semester 2 (September)	X4
Perkotaan &Perdesaan2020 Semester 1 (Maret)	X5
Perkotaan & Perdesaan2020 Semester 2 (September)	X6

3.3. Data Normalization

The next process is to normalize the data, normalization is a transformation process to change the data value. Normalization is used as a process to equalize the attribute scale into a smaller specific range. The calculation process of data normalization can be seen in the formula below.

$$V^1 = \frac{V - min_a}{max_a - min_a} = (new\_max_A = new\_min_A)$$

Information:
V^1 = Normalization result
V = What will be normalized
min_a = Lowest value (Minimum)
max_a = Highest score (Maximum)
new_min_A = New minimum value, 0
new_max_A = New maximum value, 1

Before carrying out the process of calculating the normalization value, first determine the Minimum and Maximum values based on Poverty Index data obtained from BPS. Minimum and Maximum values can be seen in the table below.

Table 4. Minimum and Maximum Value

Minimal	0,38	0,43	0	0	0,52	0,61
Maksimal	2,54	2,85	9,29	9,74	6,16	6,9

After getting the Minimum and Maximum values then enter the calculation process to find the Normalization value. The following is a way to find the normalized value.

$$X1 = \frac{1{,}55 - 0{,}38}{2{,}54 - 0{,}38} = 0{,}542$$

$$X2 = \frac{1{,}61 - 0{,}43}{2{,}85 - 0{,}43} = 0{,}488$$

$$X3 = \frac{3{,}28 - 0}{9{,}29 - 0} = 0{,}353$$

$$X4 = \frac{3{,}46 - 0}{9{,}74 - 0} = 0{,}355$$

$$X5 = \frac{2{,}72 - 0{,}52}{6{,}16 - 0{,}52} = 0{,}390$$

$$X6 = \frac{2{,}85 - 0{,}61}{6{,}9 - 0{,}61} = 0{,}356$$

Then the normalization value is calculated for each province in each attribute. The results of the overall normalization value can be seen in the table below.

Table 5. Normalization Results

Provinsi	X1	X2	X3	X4	X5	X6
Aceh	0,542	0,488	0,353	0,355	0,390	0,356
Sumatera Utara	0,509	0,459	0,167	0,171	0,176	0,157
Sumatera Barat	0,167	0,153	0,115	0,120	0,071	0,060
Riau	0,273	0,244	0,135	0,157	0,110	0,113
Jambi	0,616	0,554	0,086	0,091	0,103	0,091
Maluku Utara	0	0,091	0,125	0,129	0,074	0,076
…	…	…	…	…	…	…
Papua	0,111	0,161	0,901	0,959	1	1

3.4. Application of the K-Medoids Algorithm

This process is a data processing stage using the K-Medoids algorithm, based on the Poverty Index data obtained from BPS. The stages in performing manual data mining calculations using the K-Medoids cluster include:

Determine the initial cluster center by taking 3 data from a number of data used. The data is taken randomly, the cluster data taken can be seen in the table below.

Table 6. Initial Centeroid Value

Cluster	X1	X2	X3	X4	X5	X6
Kep. Bangka Belitung (C1)	0,44	0,64	0,79	0,94	0,6	0,77
Bali (C2)	0,47	0,55	0,65	0,75	0,52	0,61
Kalimantan Selatan (C3)	0,66	0,76	0,74	0,78	0,7	0,77

A. The next step is to determine the closest distance using the Euclidiean Distance equation. The initial calculation starts from the Aceh Province and continues to the Papua Province, the Euclidiean Distance calculation process can be considered below:

$$C1 = \sqrt{\begin{matrix}(0{,}44-0{,}542)^2+(0{,}64-0{,}488)^2\\+(0{,}79-0{,}353)^2+(0{,}94-0{,}355)^2\\+(0{,}6-0{,}390)^2+(0{,}77-0{,}356)^2\end{matrix}} = 0{,}884$$

$$C2 = \sqrt{\begin{matrix}(0{,}47-0{,}542)^2+(0{,}55-0{,}488)^2\\+(0{,}65-0{,}353)^2+(0{,}75-0{,}355)^2\\+(0{,}52-0{,}390)^2+(0{,}61-0{,}356)^2\end{matrix}} = 0{,}578$$

$$C3 = \sqrt{\begin{matrix}(0{,}66-0{,}542)^2+(0{,}76-0{,}488)^2\\+(0{,}74-0{,}353)^2+(0{,}78-0{,}355)^2\\+(0{,}7-0{,}390)^2+(0{,}77-0{,}356)^2\end{matrix}} = 0{,}828$$

From the results of the equation calculation using Euclidiean Distance, then the distance obtained in each cluster, cluster Kep. Bangka Belitung, Bali, and South

Kalimantan are in cluster 2. The results of the process of calculating the literacy distance of 2 (two) can be seen in the table below:

Table 7. Results of 1 (One) Literacy Calculation

ID	Jarak Medoid			Terdekat	Cluster
	C1	C2	C3		
1	0,884186	0,57825	0,828067	0,57824985	2
2	1,253839	0,949495	1,209422	0,94949532	2
3	1,490974	1,198416	1,528741	1,19841635	2
4	1,378173	1,079366	1,396674	1,07936621	2
5	1,400955	1,103081	1,328323	1,10308114	2
6	1,148391	0,884829	1,0095	0,88482944	2
7	1,160893	0,922398	0,976622	0,92239793	2
8	1,125088	0,820259	1,119354	0,82025858	2
9	1,606856	1,324361	1,665469	1,32436093	2
10	1,32003	1,021056	1,337531	1,02105605	2
11	1,685601	1,394779	1,720607	1,39477864	2
12	1,343548	1,044089	1,355224	1,0440891	2
...	...	...	...	...	...
34	0,750354	0,877005	0,92775	0,75035391	2
Number of Proximity				**32,9801377**	

B. After finding the first literacy value has been done, the next step is to find the value or perform calculations on new medoids or non-medoids. The data that will be used as new medoids is data taken at random, the data are DKI Jakarta, Central Kalimantan, and South Sumatra Provinces. The data can be seen in the table below:

Table 8. Centeroid 2 . Value

Cluster	X1	X2	X3	X4	X5	X6
DKI Jakarta (C1)	0,59	0,67	0	0	0,59	0,67
Kalimantan Tengah (C2)	0,82	0,91	0,79	0,86	0,8	0,88
Sumatera Selatan (C3)	1,48	1,54	1,55	1,67	1,51	1,6

Then proceed with the process of calculating the value of the second literacy, which is to find the value of new medoids using the equations that have been used previously in the first literacy. The process is considered as follows:

$$C1 = \sqrt{\begin{matrix}(0{,}59-0{,}542)^2+(0{,}67-0)^2 \\ +(0-0{,}353)^2+(0-0{,}355)^2 \\ +(0{,}59-0{,}390)^2+(0{,}67-0{,}356)^2\end{matrix}} = 0{,}651$$

$$C2 = \sqrt{\begin{matrix}(0{,}82-0{,}542)^2+(0{,}91-0{,}488)^2 \\ +(0{,}79-0{,}353)^2+(0{,}86-0{,}355)^2 \\ +(0{,}8-0{,}390)^2+(0{,}88-0{,}356)^2\end{matrix}} = 1{,}069$$

$$C3 = \sqrt{\begin{matrix}(1{,}48-0{,}542)^2+(1{,}54-0{,}488)^2 \\ +(1{,}55-0{,}353)^2+(1{,}67-0{,}355)^2 \\ +(1{,}51-0{,}390)^2+(1{,}6-0{,}356)^2\end{matrix}} = 2{,}819$$

Based on the results of the equation calculation using the Euclidiean Distance above, the distance in each cluster is obtained, the DKI Jakarta, Central Kalimantan, and South Sumatra clusters are in cluster 2. The results of the 1 (one) literacy distance calculation process can be seen in the table below:

Table 9. Results of Literacy Calculations 2 (Two)

ID	Jarak Medoid			Terdekat	Cluster
	C1	C2	C3		
1	0,651879	1,069615	2,819705	0,651879	1
2	0,736859	1,440411	3,183121	0,736859	1
3	1,056099	1,790569	3,542742	1,056099	1
4	0,930416	1,652647	3,405988	0,930416	1
5	0,776504	1,537548	3,257767	0,776504	1
6	0,643539	1,178587	2,864407	0,643539	1
7	0,706366	1,104354	2,73609	0,706366	1
8	0,750484	1,371619	3,125162	0,750484	1
9	1,191559	1,932094	3,682324	1,191559	1
10	0,929841	1,592928	3,343622	0,929841	1
11	1,156931	1,980289	3,73521	1,156931	1
12	0,905172	1,608559	3,361382	0,905172	1
...	...	...	...	...	...
34	1,580048	1,06761	2,306951	1,06761	2
Number of Proximity				**31,0905**	

C. Then after getting the results of the second literacy value, the next step is to calculate the total distance between the first literacy and the second literacy. After getting the results of the calculation of the distance between the first literacy and the second literacy, the total deviation is summed by finding the difference between the new total distance and the old value distance. If S<0, it will be replaced by the object value with the new medoids. The calculation process can be seen below:

S = New Total – Old Total

= 32,9801377 - 31,0905

= 1,88963565

D. Because the value of S>0 in the second literacy, there is no need to continue in the next literacy process. Then the clustering process in the second literacy is stopped. In the end, the minimum total distance was 1.88963565 in the second literacy.

3.5. Testing

To get the best results in the clustering process, this research refers to testing data that has an impact on the K value. This test uses data from 34 data from the Province of the Republic of Indonesia. To find out the highest and lowest number of clusters, you can pay attention to the Silhouette Coefficient value. The cluster values that will be tested are the values of K 6 to K 12. The results of these tests can be seen in the table below.

Table 10. Test Results on the Number of Clusters

Test Data	Value K	Average Silhouette Coefficient
34	6	1,374943
	7	1,305835
	8	1,092973
	9	1,111452
	10	1,098097
	11	1,205585
	12	1,15347

The results of the cluster above show a test of cluster quality with the best category being in cluster 6 with an average Silhoutte Coefficient value of 1.374943.

4. CONCLUSION

Based on the research that has been done, it can be concluded that by using data mining techniques for the process of grouping Provinces with a Poverty Index using the K-Medoids method and manual calculations using Microsoft Excel, the results are 2 clusters with low and high poverty index categories. From the results of testing the quality of clusters using the Silhoutte Coefficient showing the value of cluster 6, the best results are obtained with the results of the Silhoutte Coefficient 1.374943.

5. SUGGESTED

After conducting the research process with data mining techniques using the K-Medoids method on the Poverty Index, the data can still be done using other methods, for example, K-Means, Naive Bayes, and others. In addition, it can also be done using other data processing techniques such as Decision Support Systems.

6. REFERENCES

[1] Arifandi, Muh. Hermawan, Arief. Avianto, Donny. 2021. "Implementasi Algoritma K-Medoids Untuk Clustering Wilayah Terinfeksi Kasus Covid19 Di Dki Jakarta". Jurnal JTT Vol. 7 No. 2 September 2021. ISSN 2549-1938

[2] Hardiyanti, Fitri. Tambunan, Satria, Heru. Saragih Syahputra, Ilham. 2019. "Penerapan Metode K-Medoids Clustering Pada Penanganan Kasus Diare Di Indonesia". Jurnal KOMIK Vol. 3 No. 1 Oktober 2019. ISSN 2597-4610

[3] Pulungan, Nurliana. Suhada. Suhendro, Dedi. 2019. "Penerapan Algoritma K-Medoids Untuk Mengelompokkan Penduduk 15 Tahun Keatas Menurut Lapangan Pekerjaan Utama". Jurnal KOMIK Vol. 3 No. 1 Oktober 2019. ISSN 2597-4610

[4] Nurlaela, Siti. Primajaya, Aji. Padilah, Nur, Tesa. 2020. "Algoritma K-Medoids Untuk Clustering Penyakit Maag Di Kabupaten Karawang". Jurnal Informatika Vol. 12 No. 2 Desember 2020. ISSN 1979-0694

[5] Asmiatun, Siti. Wakhidah, Nur. Putri, Novita, Astrid. 2020. "Penerapan Metode K-Medoids Untuk Pengelompokkan Kondisi Jalan Di Kota Semarang". Jurnal Informatika dan Sistem Informasi Vol. 6 No. 2 Maret 2020. ISSN 2407-4322

[6] Agustian, Rafif, Daffa. Dermawan, Arif, Budi. 2022. "Analisis Clustering Demam Berdarah Dengue Dengan Algoritma K-Medoids (Studi Kasus Kabupaten Karawang)". Jurnal Informatika dan Komputer (JIKO) Vol. 6 No. 1 Februari 2022. ISSN 2447-3964

[7] Sukma, Sindi. Ratnasari,Weni. Agustika, Irma. Ilmi, Fikrul. Hartama, Dedy 2020. "Analisis Algoritma K-Medoids Clustering Dalam Pengelompokan Penyebaran Covid-19 Di Indonesia ". Jurnal Teknologi Informasi (JTI) Vol. 4 No. 1 2020. ISSN 2580-7927