Digital Zone
Jurnal Teknologi Informasi dan Komunikasi
Vol: 15 No 02 2024 E-ISSN: 2477-3255
Accepted by the Editor: 24-10-2024| Revision: 10-11-2024 | Published: 30-11-2024

Comparison of K-Means and K-Medoids Algorithms for Clustering Poverty Data in South Sumatra Using DBI Evaluation

M. Dandi Akhda[1], Ken Ditha Tania[2*],
[1,2]Program Studi Sistem Informasi Bilingual Fakultas Ilmu Komputer Universitas Sriwijaya
[1,2]Jl. Srijaya Negara, Bukit Lama, Kec. Ilir Barat I, Kota Palembang, Sumatera Selatan
e-mail: [1]mdandiakhda123@gmail.com, [2*]kenya.tania@gmail.com

Abstract

This research focuses on the implementation and comparison of the K-Means and K-Medoids algorithms that function as poverty data clustering in South Sumatra Province, the poverty data is taken from the Central Statistics Agency of Indonesia (BPS Indonesia). This research also aims to analyze the poverty level in South Sumatra Province by including additional variables such as average years of schooling and per capita expenditure in the community in each regency or city in South Sumatra Province. Data clustering is done by both algorithms and then the performance value is Evaluated using Davies Bouldin Index DBI shows that K-Means gives better results, with a lower DBI value (0.204 at K=5) while K-Medoids has a DBI value of 0.239 at K=5, which indicates more compact and separated clusters. The superiority of K-Means is due to the homogeneous and minimal outlier characteristics of the dataset, which makes the centroid approach more optimal than medoids in K-Medoids. With these results, K-Means was chosen as the best algorithm for clustering poverty data in the region. The use of the K-Means algorithm produces a pattern in clusters related to education, economic inequality, and poverty distribution in various regions in South Sumatra. This implementation provides insight into how data clustering techniques can be applied to socio-economic data to provide policy makers in a region with information about the region, especially information about poverty-stricken areas.

Keywords: *K-Means, K-Medoids, Poverty Data, Clustering, Davies Bouldin index (dbi)*

1. Introduction

Poverty is one of the problems often faced in cities or villages in Indonesia [1], including cities and villages in South Sumatra Province, although according to the Indonesian Central Bureau of Statistics the average percentage of Poverty in South Sumatra Province decreased by 1.29% in 2023 but social inequality between regions is still quite visible. Poverty in a region can also be caused by the development of education in each region in the province of South Sumatra which is uneven so that many people in South Sumatra Province do not have proper education, people who have inadequate education can make productivity decrease and result in low average per capita community expenditure [2][3]. Some of the factors mentioned above are a problem for social inequality in South Sumatra Province.

Knowing this, the author wants to group the regions in South Sumatra Province based on poverty levels using data mining techniques with clustering methods. Clustering algorithms such as K-Means and K-Medoids have generally been widely applied in various socioeconomic contexts in Indonesia, but few studies focus on combining several socioeconomic indicators to cluster regional poverty data in South Sumatra Province using DBI as a comparison. This research bridges the gap by analyzing the effectiveness of the algorithm in clustering regional poverty data.

Data mining itself is a tool used to analyze large data that was previously unknown, implicit, and considered meaningless into information, knowledge, and patterns [4]. Clustering is a method that will be used by the author, this method is one of the methods that can be used to group areas based on the same poverty level. The K-Means algorithm that will be used is one of the algorithms in the clustering method that is often used to analyze data by grouping data into several clusters, K-Means itself is suitable for numerical data that is homogeneous or diversely distributed and minimal Outliers [5], [6]. The K-Medoids algorithm is also used in this study as a comparison of which is the best algorithm in clustering, the K-Medoids algorithm itself has similarities with the k-means algorithm but K-Medoids is considered more resistant to data that deviates far from the patterns or values in a dataset or is often called Outliers, and also K-Medoids is considered more effective for datasets that have nominal features or non-numeric data which is often called Heterogeneous datasets [7], the difference between these two algorithms lies in the cluster center where K-Means uses Centroid, which is the average of all data in the cluster [8], while K-Medoids uses Medoid, which is the actual data that has the smallest total distance to other data in the cluster. In terms of computational speed, K-Means is much faster than K-Medoids, this is because the centroid update process is simpler and more efficient than K-Medoids which must select actual data points (medoids) that minimize the total distance in the cluster, which requires more calculations and is more complicated. K-Medoids tend to be slower, especially on large datasets due to the complexity in finding the optimal medoid [9], [10].

The author utilizes software such as Rapidminer to help the data processing and writing process, RapidMiner itself is a Machine Learning software by utilizing predictive and descriptive data analysis to provide knowledge to its users, RapidMiner can be used for Data Mining, data analysis, and statistical techniques both descriptive and inferential in processing data [11]. Clustering carried out by analyzing socio-economic indicators such as the average length of school time per population, per capita expenditure, and the number of poor people can provide knowledge and information to policy makers in South Sumatra Province to formulate data-based policies that are effective and in accordance with the poverty characteristics of each region.

Researchers cited several previous journals in helping solve problems in this study. The first research is a discussion of the method of finding the optimal number of clusters in the K-Means and K-Medoids algorithms, the use of this method can help and facilitate the use of both algorithms in the problem at hand [12].
The second research is the use of DBI as an evaluation technique for the K-Means and K-Medoids algorithms, the result is that the K-Medoids algorithm is superior to K-Means with a smaller DBI score, but the research conducted by [13] uses a complex dataset so that there are many outliers that make K-Means affected. K-Medoids as previously mentioned has a resistance to outliers.

The third research is the use of clustering algorithms, especially K-Means, in grouping households based on socio-economic conditions. This study obtained 3 clusters using the K-Means algorithm. The three clusters include stable, critical, and at-risk clusters. The approach taken by this study can help researchers in finding patterns obtained in the clusters formed [14].

https://doi.org/10.31849/digitalzone.v15i2.23624

2. Research methods

This research uses quantitative methods, which are systematic and structured studies that use numerical data collection and analysis to describe social phenomena. To draw empirical findings, the data is examined using statistical or mathematical techniques, therefore in this study the flow of studies to be carried out will be carried out as follows.

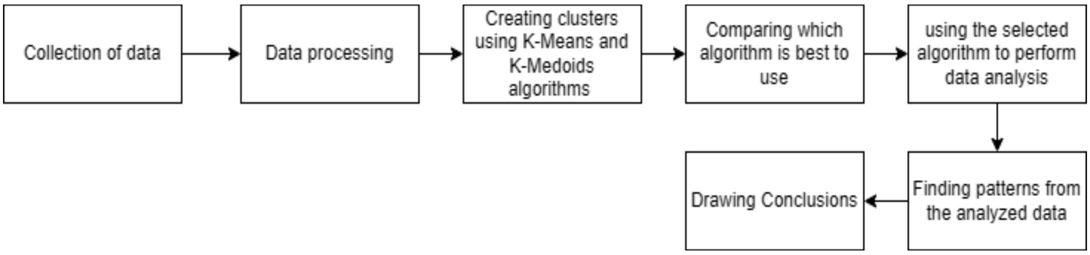

Collection of data
Data processing
Creating clusters using K-Means and K-Medoids algorithms
Comparing which algorithm is best to use
using the selected algorithm to perform data analysis
Finding patterns from the analyzed data
Drawing Conclusions

Figure 1: Research framework

2.1. Data Collection Stage

At this stage the author collects datasets from the Indonesian Central Bureau of Statistics specifically in the South Sumatra region, the datasets taken are in the form of the average length of schooling of the population per Regency / City in South Sumatra, the average number of poor people per Regency / City in South Sumatra, and the average Expenditure per Capita per Regency / City from 2019 to 2023. This dataset is in the form of raw data taken directly from BPS, the following data display in tabular form.

Table 1. BPS Data on the Average Number of Poor People

District/City	Average Number of Poor Population per District/City in South Sumatra from 2019 to 2023 (thousand people)				
	2019	2020	2021	2022	2023
Ogan Komering Ulu	46.840	47.3	47.5	44.2	44.11
Ogan Komering Ilir	124.14	123.34	124.78	113.79	114.48
Muara Enim	78.75	79.27	80.4	73.53	73.24
Lahat	65.03	65.75	68.4	65.39	63.36
Musi Rawas	53.82	54.95	57.46	55.8	59.75
Musi Banyuasin	105.83	105.38	105.23	102.24	101.63
Banyu Asin	96.55	96.27	94.08	88.55	85.88
Ogan Komering Ulu Selatan	37.92	39.5	41.23	39.61	39.3
Ogan Komering Ulu Timur	70.4	71.1	72.89	69.69	69.91
Ogan Ilir	57.06	57.97	60.5	54.55	59.33
Empat Lawang	30.68	31.89	34.11	31.06	30.78
Penukal Abab Lematang Ilir	25.47	24.17	25.1	23.14	21.72
Musi Rawas Utara	36.63	37.75	39.5	36.65	36.67
Kota Palembang	180.67	182.61	194.12	181.65	179.45
Kota Prabumulih	21.62	21.83	23.6	22.12	22.33
Kota Pagar Alam	12.37	12.71	13.27	12.05	12.73
Kota Lubuklinggau	29.98	29.8	31.61	30.68	31.02

Table 2: BPS Data on Average Years of Schooling of Communities

District/City	Average Years of Schooling per District/City in South Sumatra from 2019 to 2023 (years)				
	2019	2020	2021	2022	2023
Ogan Komering Ulu	8,69	8,69	8,85	8,7	8,71
Ogan Komering Ilir	7,03	7,03	7,08	7,04	7,05

District/City	Average Years of Schooling per District/City in South Sumatra from 2019 to 2023 (years)				
	2019	2020	2021	2022	2023
Muara Enim	7,78	7,78	8,14	7,79	7,8
Lahat	8,45	8,45	8,56	8,46	8,52
Musi Rawas	7,51	7,51	7,56	7,52	7,53
Musi Banyuasin	7,61	7,61	7,68	7,62	7,63
Banyu Asin	7,19	7,19	7,46	7,2	7,44
Ogan Komering Ulu Selatan	7,83	7,83	8,05	7,84	7,85
Ogan Komering Ulu Timur	7,54	7,54	8,07	7,55	7,56
Ogan Ilir	7,85	7,85	8,08	7,86	7,87
Empat Lawang	7,39	7,39	7,66	7,6	7,64
Penukal Abab Lematang Ilir	6,75	6,75	7,08	7,04	7,05
Musi Rawas Utara	6,5	6,5	7,5	6,84	7,09
Kota Palembang	10,52	10,52	10,92	10,53	10,75
Kota Prabumulih	9,72	9,72	10,35	9,96	9,97
Kota Pagar Alam	9,14	9,14	9,43	9,39	9,4
Kota Lubuklinggau	9,81	9,81	9,93	9,89	9,9

Table 3. BPS Data on Average Community Per Capita Income

District/City	Average Expenditure per Capita per District/City (Ribu Rupiah/month)				
	2019	2020	2021	2022	2023
Ogan Komering Ulu	910.096	940.790,6	1.073.099	1.126.201	1.154.192
Ogan Komering Ilir	835.028	941.429,6	1.127.907	1.130.922	1.160.304
Muara Enim	945.635	1.015.267	1.043.238	1.008.113	1.072.846
Lahat	971.705	9.902.52,8	915.708,4	1.118.652	1.374.192
Musi Rawas	849.314	898.831,7	865.759,7	995.141,1	993.094
Musi Banyuasin	855.378	1.078.593	1.130.116	1.127.575	1.249.650
Banyu Asin	890.509	958.325,7	1.178.180	1.237.838	1.244.567
Ogan Komering Ulu Selatan	767.767	688.848,6	800.794,2	777.368,6	984.573
Ogan Komering Ulu Timur	888.182	850.854,2	1.020.198	1.033.977	960.288
Ogan Ilir	904.663	910.976,2	873.021,7	938.379,9	897.070
Empat Lawang	653.127	785.246,9	770.937,9	908.774,6	1.000.059
Penukal Abab Lematang Ilir	785.736	844.245,4	816.787,8	869.729,4	904.973
Musi Rawas Utara	884.222	883.104,1	918.546,3	969.819,4	1.071.605
Kota Palembang	1.273.229	1.361.933	1.424.768	1.507.689	1.607.054
Kota Prabumulih	1.007.104	1.014.842	952.847,4	1.107.191	1.201.533
Kota Pagar Alam	887.292	894.535,9	943.736,4	959.267,5	978.948
Kota Lubuklinggau	1.110.135	1.190.145	1.148.281	1.176.459	1.218.088

2.2. Data processing stage

Furthermore, the data that has been taken from BPS will be processed before it can be processed into clusters. the data that has been obtained will go through a normalization process by standardizing the value of each data variable, the process is by changing the per capita expenditure data and the number of poor people into units of thousands, while the average length of schooling remains in units of years, this normalization aims to reduce the scale imbalance between variables so that the clustering results by the algorithm itself are not biased [12][13]. Furthermore, the calculation of the average value of each data variable is carried out in order to simplify the data value of the variables used so that there is no inconsistency in the centroid value in the algorithm used. The following dataset has passed the data processing stage

Table 4. Average of all data

District/City	Average Years of Schooling per District/City in South Sumatra from 2019 to 2023 (years)	Average Number of Poor Population per District/City in South Sumatra from 2019 to 2023 (thousand people)	Average Expenditure per Capita per District/City (Ribu Rupiah/month)
Ogan Komering Ulu	8,736	45.990	1.048.776
Ogan Komering Ilir	7,054	120.106	1.039.118
Muara Enim	7,882	77.038	1.070.120
Lahat	8,504	65.586	1.071.042
Musi Rawas	7,534	56.356	920.428
Musi Banyuasin	7,638	104.062	1.088.263
Banyu Asin	7,348	92.266	1.101.884
Ogan Komering Ulu Selatan	7,886	39.512	803.870
Ogan Komering Ulu Timur	7,704	70.798	950.960
Ogan Ilir	7,914	57.882	904.822
Empat Lawang	7,588	31.704	832.609
Penukal Abab Lematang Ilir	6,996	23.920	844.294
Musi Rawas Utara	7,038	37.440	945.496
Kota Palembang	10,726	183.700	1.434.938
Kota Prabumulih	10,04	22.300	1.056.704
Kota Pagar Alam	9,354	12.626	932.756
Kota Lubuklinggau	9,888	30.618	1.168.622

2.3. Data Clustering Stage

The data that has been processed earlier will be analyzed and the clustering process will be carried out with RapidMiner tools, two K-Means and K-Medoids algorithms will be applied to see which algorithm will be used for clustering poverty data in South Sumatra Province. Several clustering experiments will be conducted with the K-Means and K-Medoids algorithms in order to obtain the appropriate cluster value.

2.3.1 K-Means

Clustering experiments were conducted using 2 to 8 clusters to get the right K value. The following experimental results using RapidMiner tools can be seen in table 5 below.

Table 5. Clustering Results with K-Means

Experiment with K Value	Cluster								
	0	1	2	3	4	5	6	7	8
K=2	9	8	-	-	-	-	-	-	-
K=3	8	8	1	-	-	-	-	-	-
K=4	8	5	1	3	-	-	-	-	-

Experiment with K Value	Cluster								
	0	1	2	3	4	5	6	7	8
K=5	1	5	1	3	7	-	-	-	-
K=6	4	1	5	3	1	3	-	-	-
K=7	2	2	1	3	5	3	1	-	-
K=8	1	2	4	1	3	1	2	3	-

Next, the cluster will be tested for validity using the Davies Bouldin Index (DBI) technique with the aim of getting the best cluster.

Table 6. DBI Value of K-Means

Cluster	DBI Value
2	0.568
3	0.306
4	0.239
5	0.204
6	0.302
7	0.326
8	0.319

Based on the DBI calculation, the best K-Means cluster is in trial K=5 or cluster 5 with a DBI value of 0.204 which divides the data into 5 clusters with 1 member from cluster 0, 5 members from cluster 1, 1 member from cluster 2, 3 members from cluster 3, and 7 members from cluster 4.

2.3.2 K-Medoids

This research also uses the K-Medoids algorithm as a comparison with the same clustering experiment as K-Means. The experimental results can be seen in Table 7 below.

Table 7: Clustering Results with K-Medoids

Experiment with K Value	Klaster								
	0	1	2	3	4	5	6	7	8
K=2	7	10	-	-	-	-	-	-	-
K=3	6	10	1	-	-	-	-	-	-
K=4	7	1	8	1	-	-	-	-	-
K=5	7	1	5	3	1	-	-	-	-
K=6	7	1	2	1	1	5	-	-	-
K=7	1	2	5	2	3	1	3	-	-
K=8	1	5	2	1	2	2	3	1	-

Furthermore, the same as the previous algorithm, the validity test will be carried out using the Davies Bouldin Index (DBI).

Table 8: DBI Value of K-Medoids

Cluster	DBI Value
2	0.696
3	0.482
4	0.347
5	0.239
6	0.537
7	0.348

Cluster	DBI Value
8	0.475

Based on the DBI calculation for K-Medoids, the best cluster for the K-Medoids algorithm is at K=5 or cluster 5 with a DBI value of 0.239 which divides the data into 5 clusters with 7 members from cluster 0, 1 member from cluster 1, 5 members from cluster 2, 3 members from cluster 3, and 1 member from cluster 4.

2.3.3 Determining the Best Algorithm

Furthermore, after clustering using the K-Means algorithm, K-Medoids and cluster validation using DBI in RapidMiner, the author will compare the DBI scores of the two algorithms to find out which algorithm is better used in clustering poverty data in South Sumatra based on the dataset that has been used.

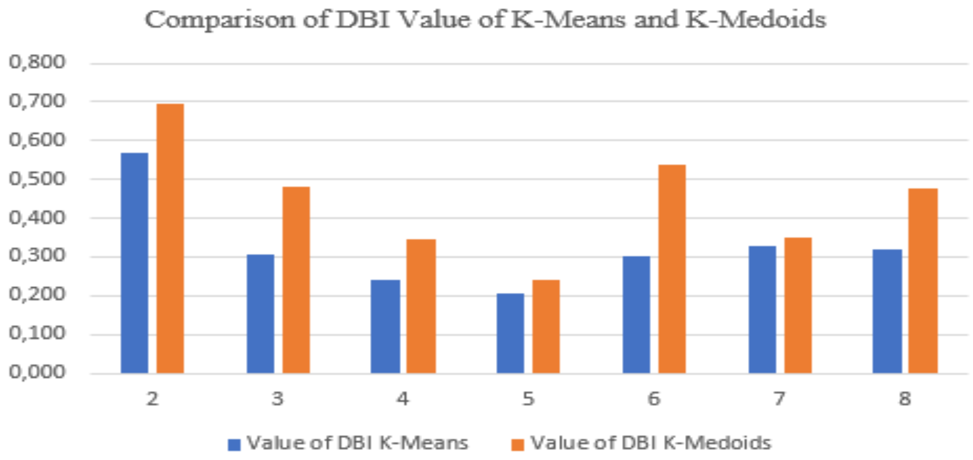


Figure 2: Comparison of DBI Values of both Algorithms

Table 9. Comparison of DBI values of both Algorithms

Cluster	DBI Value	
	K-Means	K-Medoids
2	0.568	0.696
3	0.306	0.482
4	0.239	0.347
5	0.204	0.239
6	0.302	0.537
7	0.326	0.348
8	0.319	0.475

For the DBI value, these two algorithms both have the smallest DBI value in Cluster 5 or K = 5 but the value of the DBI itself is different, with the K-Means algorithm having a smaller DBI value than the K-Medoids algorithm. The DBI index involves calculating the cluster diameter which measures the spread of points in the cluster and the distance between cluster centroids. The lower the DBI value, the better the cluster configuration, as this indicates that the clusters are more separated and better organized [17], [18].

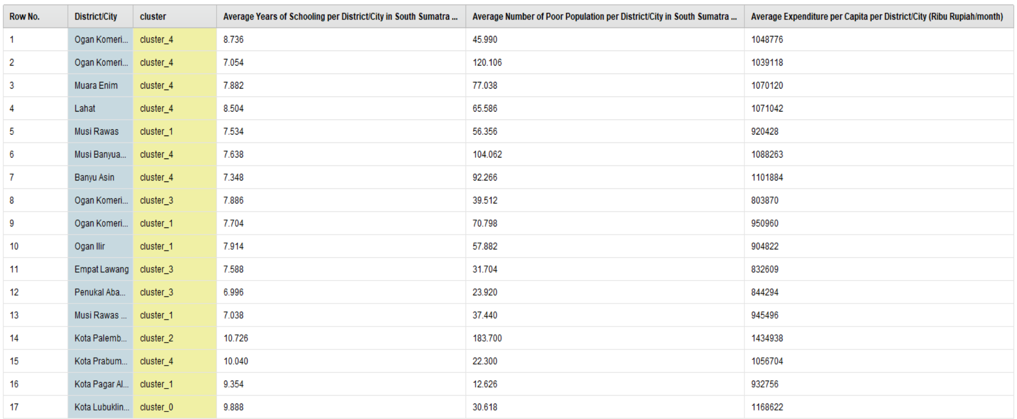
In this dataset K-means has an advantage over K-medoids, this is because the dataset used tends not to have many outliers and is relatively homogeneous, K-Means with its centroid calculation has better accuracy than K-Medoids which uses Medoid or original data points as its calculation method.

The lower DBI value of K-Means which is 0.204 at the value of K = 5 compared to K-Medoids which is 0.239 at the same value of K = 5 shows that K-Means has data points in the cluster closer to each other so that the cluster is more compact and the clusters made by K-Means are more separated so that the patterns made between clusters are more visible and clear differences [19].

3. Results and Discussion

3.1 Cluster Division based on Data Variables

With the determination of the algorithm to be used, namely K-Means with a value of K = 5, at this stage it will be explained what the contents of the cluster that has been created based on the existing data variables, this is useful for knowing which district or city in the province of South Sumatra has the largest number of poor people and is useful for drawing conclusions in the hope that it can help policy makers with the support of other data such as data on the average length of schooling of each community, and data on the average per capita expenditure of each community. The display of cluster division by the RapidMiner tool can be seen in Figure 5 below.



Row No.	District/City	cluster	Average Years of Schooling per District/City in South Sumatra ...	Average Number of Poor Population per District/City in South Sumatra ...	Average Expenditure per Capita per District/City (Ribu Rupiah/month)
1	Ogan Komeri...	cluster_4	8.736	45.990	1048776
2	Ogan Komeri...	cluster_4	7.054	120.106	1039118
3	Muara Enim	cluster_4	7.882	77.038	1070120
4	Lahat	cluster_4	8.504	65.586	1071042
5	Musi Rawas	cluster_1	7.534	56.356	920428
6	Musi Banyua...	cluster_4	7.638	104.062	1088263
7	Banyu Asin	cluster_4	7.348	92.266	1101884
8	Ogan Komeri...	cluster_3	7.886	39.512	803870
9	Ogan Komeri...	cluster_1	7.704	70.798	950960
10	Ogan Ilir	cluster_1	7.914	57.882	904822
11	Empat Lawang	cluster_3	7.588	31.704	832609
12	Penukal Aba...	cluster_3	6.996	23.920	844294
13	Musi Rawas ...	cluster_1	7.038	37.440	945496
14	Kota Palemb...	cluster_2	10.726	183.700	1434938
15	Kota Prabum...	cluster_4	10.040	22.300	1056704
16	Kota Pagar Al...	cluster_1	9.354	12.626	932756
17	Kota Lubuklin...	cluster_0	9.888	30.618	1168622

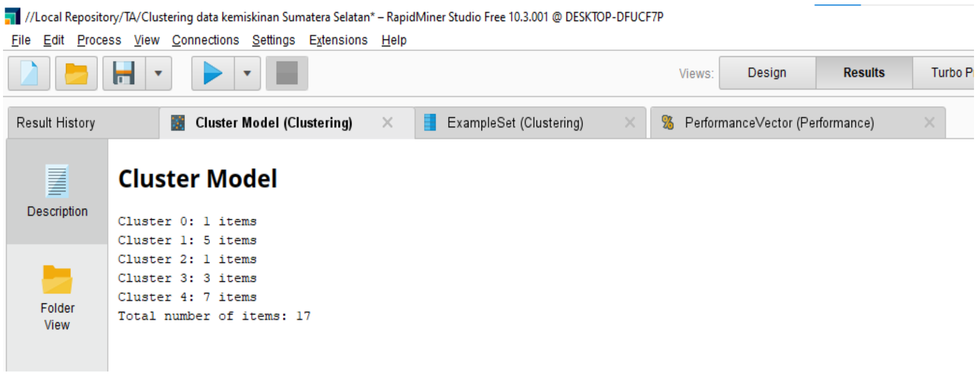
Figure 3: Cluster division view

3.1.1 Explanation of Members of Each Cluster Created

The clusters that have been formed have different members in each cluster, each member is grouped into 5 clusters, among others:

- 1 member is in cluster 0,
- 5 members are in cluster 1,
- 1 member is in cluster 2,
- 3 members are in cluster 3,
- 7 members are in cluster 4.

https://doi.org/10.31849/digitalzone.v15i2.23624


//Local Repository/TA/Clustering data kemiskinan Sumatera Selatan* – RapidMiner Studio Free 10.3.001 @ DESKTOP-DFUCF7P
File Edit Process View Connections Settings Extensions Help
Views: Design Results Turbo P
Result History
Cluster Model (Clustering)
ExampleSet (Clustering)
PerformanceVector (Performance)
Description
Folder View

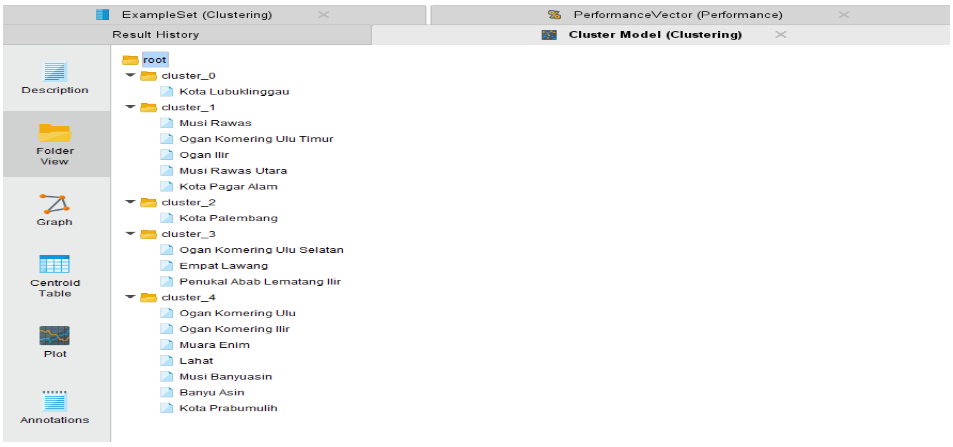
Cluster Model

```
Cluster 0: 1 items
Cluster 1: 5 items
Cluster 2: 1 items
Cluster 3: 3 items
Cluster 4: 7 items
Total number of items: 17
```

Figure 4: Visualization of cluster data sharing in RapidMiner.

From the initial dataset, the division of cluster members consists of districts or cities in South Sumatra with the distribution distributed by RapidMiner to the 5 clusters with calculations that have been made, the division includes:

- Cluster 0: Consists of Lubuklinggau City
- Cluster 1: Consists of Musi Rawas District, East Ogan Komering Ulu District, Ogan Ilir District, North Musi Rawas District, and Pagar Alam City.
- Cluster 2: Consists of Palembang City
- Cluster 3: Consists of South Ogan Komering Ulu District, Empat Lawang District, and Penukal Abab Lematang Ilir District.
- Cluster 4 : Consists of Ogan Komering Ulu District, Ogan Komering Ilir District, Muara Enim District, Lahat District, Musi Banyuasin District, Banyu Asin District, Prabumulih City.



```
root
  cluster_0
    Kota Lubuklinggau
  cluster_1
    Musi Rawas
    Ogan Komering Ulu Timur
    Ogan Ilir
    Musi Rawas Utara
    Kota Pagar Alam
  cluster_2
    Kota Palembang
  cluster_3
    Ogan Komering Ulu Selatan
    Empat Lawang
    Penukal Abab Lematang Ilir
  cluster_4
    Ogan Komering Ulu
    Ogan Komering Ilir
    Muara Enim
    Lahat
    Musi Banyuasin
    Banyu Asin
    Kota Prabumulih
```

Figure 5. Member View of Each Cluster in RapidMiner

The division of members of each cluster above is measured based on the distance of the data to the data center or *Centroid*, the display of the *Centroid* itself can be seen in RapidMiner. The following table *Centroid* data that has been determined by RapidMiner based on available data using the *Euclidean Distance* Theory distance calculation.

Table 10. Cluster Data *Centroid* Table

Atrribute	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Average Years of Schooling per District/City in South Sumatra from 2019 to 2023 (years)	9.888	7.909	10.726	7.490	8.172

Average Number of Poor Population per District/City in South Sumatra from 2019 to 2023 (thousand people)	30.618	47.0204	183.7	31.712	75.335
Average Expenditure per Capita per District/City (Ribu Rupiah/month)	1168622	930892	1434938	826924	1067986.714

3.2 Analysis of each Cluster to Get the Final Result

Based on the clustering results conducted by RapidMiner on poverty data in South Sumatra, a pattern emerges in each cluster with the following characteristics:

1. Cluster 0:
   - Member: Lubuk Linggau City
   - Characteristics:
     - The average years of schooling is quite high (9.8 years).
     - Relatively low number of poor people (30,618 thousand people).
     - High per capita expenditure (1,168,622 thousand rupiah/month).
   - Conclusion: This cluster can be categorized as a region with a high level of education, few poor people, and relatively high per capita expenditure.

2. Cluster 1

   - Members: from Musi Rawas Regency, East Ogan Komering Ulu Regency, Ogan Ilir Regency, North Musi Rawas Regency, and Pagar Alam City.
   - Characteristics:
     - Average years of schooling between 7-9 years, indicating a secondary level of education.
     - The number of poor people varies, but tends to be moderate (12,626 to 70,798 thousand people).
     - Per capita expenditure is moderate, ranging from 904,822 to 950,960 thousand rupiah/month.
   - Conclusion: The regions in this cluster show a fairly good level of education with a medium poverty rate and not too high expenditure.

3. Cluster 2:

   - Member: Palembang City
   - Characteristics:
     - Highest average years of schooling (10.7 years).
     - The largest number of poor people (183,700 thousand people).
     - Highest per capita expenditure (1,434,938 thousand rupiah/month).
   - Conclusion: Palembang City is in a unique category, with very high education, a large number of poor people, but also very high per capita expenditure. This indicates significant inequality in the city, with potentially large economic disparities.

4. Cluster 3:

- Members: South Ogan Komering Ulu Regency, Empat Lawang Regency, and Penukal Abab Lematang Ilir Regency.
- Characteristics:
  - Relatively low average years of schooling (around 7 years).
  - Low to moderate poverty (23,920 to 39,512 thousand people).
  - Low per capita expenditure (803,870 to 844,294 thousand rupiah/month).
- Conclusion: This cluster represents regions with low education, low poverty rates, and low per capita expenditure as well, signaling the potential for economically underdeveloped regions.

5. Cluster 4:

- Members: from Ogan Komering Ulu Regency, Ogan Komering Ilir Regency, Muara Enim Regency, Lahat Regency, Musi Banyuasi Regency, Banyu Asin Regency, Prabumulih City.
- Characteristics:
  - Average years of schooling vary from moderate to high (7 to 10 years).
  - The number of poor people is high, especially in Ogan Komering Ilir (120,106 thousand people), but there are also lower ones such as in Prabumulih City (22,300 thousand people).
  - Per capita expenditure is at a moderate to high level (1,038,118 to 1,108,263thousand rupiah/month).

Conclusion: This cluster includes regions with fairly good education, varying poverty rates, and fairly high per capita expenditure. This indicates areas with more advanced economic potential, although there are inequalities in terms of poverty.

In a previous study titled Applying Clustering Algorithm on Poverty Analysis in a Community in the Philippines [14]. Several clusters were obtained with different patterns for each cluster, Different from that study which only used 3 clusters, here the researcher used 5 clusters due to the discovery that the value of K = 5 is the optimal value in this dataset and added several factors such as average years of schooling and per capita population expenditure. The analysis shows that Cluster 0 and Cluster 2 represent regions that are more advanced in terms of education and economy. However, there is a significant disparity between the two, particularly in the number of poor residents. In Cluster 2, Palembang City, despite having high education levels and per capita spending, also has a substantial number of poor residents. Cluster 1 represents developing areas with relatively stable social and economic conditions. Meanwhile, Cluster 3 indicates regions that are still lagging behind in both educational and economic aspects, but with potential for further development. Cluster 4 reflects areas with a wide variation in education and economy, where some regions have advanced, but still face significant social inequality challenges. Understanding the characteristics of each cluster is crucial to formulating more targeted and directed development policies for policymakers.

4. Conclusion

Based on the results of the DBI comparison on the K-Means and K-Medoids algorithms, K-Means proved to be more effective in clustering this dataset and also obtained a value of K = 5 which is the best number of clusters in both algorithms, but because the DBI value of K-Means is superior, the authors only use the K-Means algorithm in clustering. This is because the dataset used in this study is relatively homogeneous without many outliers, which supports the performance of K-Means in producing more compact and separate clusters. Then the results of clustering using the selected algorithm, K-Means, show the existence of 5 clusters that reflect the characteristics of the regions in South Sumatra, based on the variables of average years of schooling, number of poor people, and per capita expenditure. Cluster 0 and Cluster 2 represent regions that are more advanced in terms of education and economy, but with significant disparities in terms of poverty, especially in Palembang City (Cluster 2). Cluster 1 represents developing regions, with relatively stable socioeconomic conditions. Cluster 3 reflects regions that are still underdeveloped, with lower levels of education and economy, but have the potential to be developed further. Cluster 4 shows regions with wide variations in education and economy, with some more advanced regions but still facing challenges in social inequality.

References

[1] N. Nursini, "Micro, small, and medium enterprises (MSMEs) and poverty reduction: empirical evidence from Indonesia," *Development Studies Research*, vol. 7, no. 1, pp. 153–166, Jan. 2020, https://doi.org/10.1080/21665095.2020.1823238.

[2] L. Sugiharti, M. A. Esquivias, M. S. Shaari, A. D. Jayanti, and A. R. Ridzuan, "Indonesia's poverty puzzle: Chronic vs. transient poverty dynamics," *Cogent Economics and Finance*, vol. 11, no. 2, 2023, https://doi.org/10.1080/23322039.2023.2267927.

[3] W. Liu, J. Li, and R. Zhao, "The effects of rural education on poverty in China: a spatial econometric perspective," *J Asia Pac Econ*, vol. 28, no. 1, pp. 176–198, Jan. 2023, https://doi.org/10.1080/13547860.2021.1877240.

[4] H. Aggarwal, V. Kumar, and H. D. Arora, "Data mining algorithm based on Renyi fuzzy association rule: an application for the selection of suitable course," *Research in Statistics*, vol. 1, no. 1, Oct. 2023, https://doi.org/10.1080/27684520.2023.2271902.

[5] H. S. Kim, S. K. Kim, and L. S. Kang, "BIM performance assessment system using a K-means clustering algorithm," *Journal of Asian Architecture and Building Engineering*, vol. 20, no. 1, pp. 78–87, 2021, https://doi.org/10.1080/13467581.2020.1800471.

[6] P. Wang, H. Shi, X. Yang, and J. Mi, "Three-way k-means: integrating k-means and three-way decision," *International Journal of Machine Learning and Cybernetics*, vol. 10, no. 10, pp. 2767–2777, 2019 https://doi.org/10.1007/s13042-018-0901-y.

[7] A. Sobrinho Campolina Martins, L. Ramos de Araujo, and D. Rosana Ribeiro Penido, "K-Medoids clustering applications for high-dimensionality multiphase probabilistic power flow," *International Journal of Electrical Power and Energy Systems*, vol. 157, Jun. 2024, https://doi.org/10.1016/j.ijepes.2024.109861.

[8] N. H. M. M. Shrifan, M. F. Akbar, and N. A. M. Isa, "An adaptive outlier removal aided k-means clustering algorithm," *Journal of King Saud University - Computer and Information Sciences*, vol. 34, no. 8, pp. 6365–6376, Sep. 2022, https://doi.org/10.1016/j.jksuci.2021.07.003.

[9] S. Mousavi, F. Zamani 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, 2020, [Online]. Available: www.jatit.org

[10] N. H. M. M. Shrifan, M. F. Akbar, and N. A. M. Isa, "An adaptive outlier removal aided k-means clustering algorithm," *Journal of King Saud University - Computer and Information Sciences*, vol. 34, no. 8, pp. 6365–6376, Sep. 2022, https://doi.org/10.1016/j.jksuci.2021.07.003.

[11] E. D. Madyatmadja, D. J. M. Sembiring, S. M. Br Perangin Angin, D. Ferdy, and J. F. Andry, "Big Data in Educational Institutions using RapidMiner to Predict Learning Effectiveness," *Journal of Computer Science*, vol. 17, no. 4, pp. 403–413, 2021, https://doi.org/ 10.3844/jcssp.2021.403.413.

https://doi.org/10.31849/digitalzone.v15i2.23624

[12] J. Heidari, N. Daneshpour, and A. Zangeneh, "A novel K-means and K-medoids algorithms for clustering non-spherical-shape clusters non-sensitive to outliers," *Pattern Recognit*, vol. 155, Nov. 2024, https://doi.org/10.1016/j.patcog.2024.110639.

[13] S. Ghaida Muthmainah and A. Id Hadiana, "Comparative Analysis of K-Means and K-Medoids Clustering in Retail Store Product Grouping," *International Journal of Quantitative Research and Modeling*, vol. 5, no. 3, pp. 280–294, 2024.

[14] M. P. Repollo, R. Aurelius, and C. Robielos, "Applying Clustering Algorithm on Poverty Analysis in a Community in the Philippines."

[15] P. I. Dalatu and Midi, "MALAYSIAN JOURNAL OF MATHEMATICAL SCIENCES New Approaches to Normalization Techniques to Enhance K-Means Clustering Algorithm," 2020.

[16] S. Sinsomboonthong, "Performance Comparison of New Adjusted Min-Max with Decimal Scaling and Statistical Column Normalization Methods for Artificial Neural Network Classification," *Int J Math Math Sci*, vol. 2022, 2022, https://doi.org/10.1155/2022/3584406.

[17] Y. A. Wijaya, D. A. Kurniady, E. Setyanto, W. S. Tarihoran, D. Rusmana, and R. Rahim, "Davies Bouldin Index Algorithm for Optimizing Clustering Case Studies Mapping School Facilities," *TEM Journal*, vol. 10, no. 3, pp. 1099–1103, Aug. 2021, https://doi.org/10.18421/TEM103-13.

[18] M. Li, D. Xu, D. Zhang, and J. Zou, "The seeding algorithms for spherical k-means clustering," *Journal of Global Optimization*, vol. 76, no. 4, pp. 695–708, 2020, https:doi.org/10.1007/s10898-019-00779-w.

[19] Z. Feng, W. Niu, R. Zhang, S. Wang, and C. Cheng, "Operation rule derivation of hydropower reservoir by k-means clustering method and extreme learning machine based on particle swarm optimization," *J Hydrol (Amst)*, vol. 576, pp. 229–238, 2019, https://doi.org/10.1016/j.jhydrol.2019.06.045.