Journal of Artificial Intelligence and Engineering Applications
Websites:https://ioinformatic.org/

15th October 2021. Vol. 1. No. 1

Grouping of Toddlers with Malnutrition Based on Provinces in Indonesia Using K-Medoids Algorithm

Sri Anita Siallagan[1]*, M.Safii[2]

[1]*STIKOM Tunas Bangsa Pematangsiantar, North Sumatra, Indonesia*
[2]*AMIK Tunas Bangsa Pematangsiantar, North Sumatra, Indonesia*
[3]*Jln. Sudirman Block A No. 1-3 Pematangsiantar, North Sumatra*
** anitasallagan03@gmail.com, * m.safii@ amiktunasbangsa.ac.id;*

Abstract

Malnutrition is a poor health condition in infants and toddlers caused by a lack of nutritional intake. Babies and toddlers who suffer from malnutrition will experience conditions of slowness in development, slowness in thinking, underweight and so on. Malnutrition can be prevented by complete immunization from birth, providing good nutrition for their development, and so on. The purpose of this study was to determine the results of the grouping of provinces with the highest malnutrition sufferers using the K-Medoids method which is part of Data Mining. The K-Medoids method is a clustering method that can break the dataset into several groups. In this study, the data used were sourced from the Central Statistics Agency in 2016 – 2018. The results of this clustering will later show the province which is the toddler with the highest malnutrition. This research is expected to provide information for the government regarding the grouping of children under five with malnutrition in Indonesia.

Keywords: *Data Mining, K-Medoids, Clustering, Malnutrtion*

1. Introduction

Data Mining is a process or activity to collect large data and then extract the data into data mining information that can be used later [1]. Data mining is a method that can be used for large-scale data processing, therefore data mining has an important role in several fields, namely, industry, finance, weather, science and technology. In general, studies in data mining discuss methods such as clustering, classification, regression, variable selection, and market basket analysis[2].

Data mining is a discipline that studies methods to extract knowledge or find patterns from data. The results of data processing using data mining methods can be used to make decisions in the future. Data mining is also often known as pattern recognition[3].
There are several clustering methods contained in data mining, one of which is the K-Medoids Algorithm. The K-Medoids algorithm is a clustering algorithm that is related to the K-Means algorithm and the "medoidshift" algorithm. The K-Medoids algorithm was developed by Leonard Kaufman and Peter J. Rousseeuw in 1987. The K-Medoids algorithm is also often called the Partitoning AroundMedoids (PAM) algorithm. The K-Medoids algorithm has similarities with the K-Means algorithm, which is both a partitioning algorithm. Partitioning algorithm is an algorithm for grouping data into a number of clusters without any hierarchical structure being carried out between one another. The K-Medoids algorithm has advantages over the K-Means algorithm. K-Medoids has a more optimal performance and is better if the amount of data used is only small. This K-Means algorithm uses objects in a collection of objects,[4],[5].

Clustering is one of the unsupervised learning techniques where we do not need to train the method or in other words, there is no learning phase. The purpose of the clustering method is to group a number of data or objects into clusters so that each cluster will be filled with data that is as similar as possible.[6]Malnutrition and malnutrition are the status of a person's condition who is undernourished, or whose nutrition is below average. Malnutrition is a lack of nutrients in the body such as protein, carbohydrates, fats, and vitamins needed by the body. One way to assess nutritional status can be done with anthropometric, clinical, biochemical, and biophysical measurements. This anthropometric measurement can be done with several measurements, namely measurements of weight, height, upper arm circumference, and so on[7].

In Indonesia, the prevalence of under-fives suffering from malnutrition is still relatively high, so it is important for the government to group together in order to minimize the number of infants with malnutrition in Indonesia. From the data that has been obtained after adding up since 2016 - 2018 the lowest malnutrition rate was found in Bali Province, which was 8.82 percent and the highest was in East Nusa Tenggara Province, which was 41.79 percent. After calculating the average of the total number of provinces, the number of children under five experiencing malnutrition is809.91 percent in Indonesia.

The quality of balanced nutrition and nutrition is not yet fully evenly distributed in Indonesia, thus causing so many toddlers to experience malnutrition and not only the balance of nutritional quality and nutritional balance that causes malnutrition for toddlers, environmental factors, parental awareness and the poverty line also contribute to the causes of malnutrition. bad for toddlers, so it is important for the government to emphasize more on solutions for parents and pay more attention to the problem of malnutrition for toddlers. In this study, the data source used was obtained from the Central Statistics Agency (BPS). The algorithm used in classifying children under five with malnutrition in Indonesia with the highest and lowest clusters is K-Medoids. K-Medoids or Partitioning Around Medoids (PAM) is a clustering algorithm which is almost the same as the K-Means algorithm. The difference between these two algorithms is that the K-Medoids or PAM algorithm uses an object as a representative (medoid) as the center of the cluster for each cluster, while K-Means uses the average value (mean) as the center of the cluster.[8].

2. Research methods

The data analysis process can be carried out after data collection. In conducting the research, the writer carried out the process of analyzing descriptive statistical data. The type of data used in this study is secondary data. Secondary data is data that is not obtained directly but data that has been collected by other parties and has been processed and has relevance to the problems studied. This study aims to classify children under five with malnutrition using the RapidMiner application and using the K-Medoids algorithm from Data Mining. This study was conducted to determine the lowest and highest clusters to minimize the number of children under five with malnutrition. The research conducted by the author uses data from the official website, namely the Indonesian Central Statistics Agency (BPS).

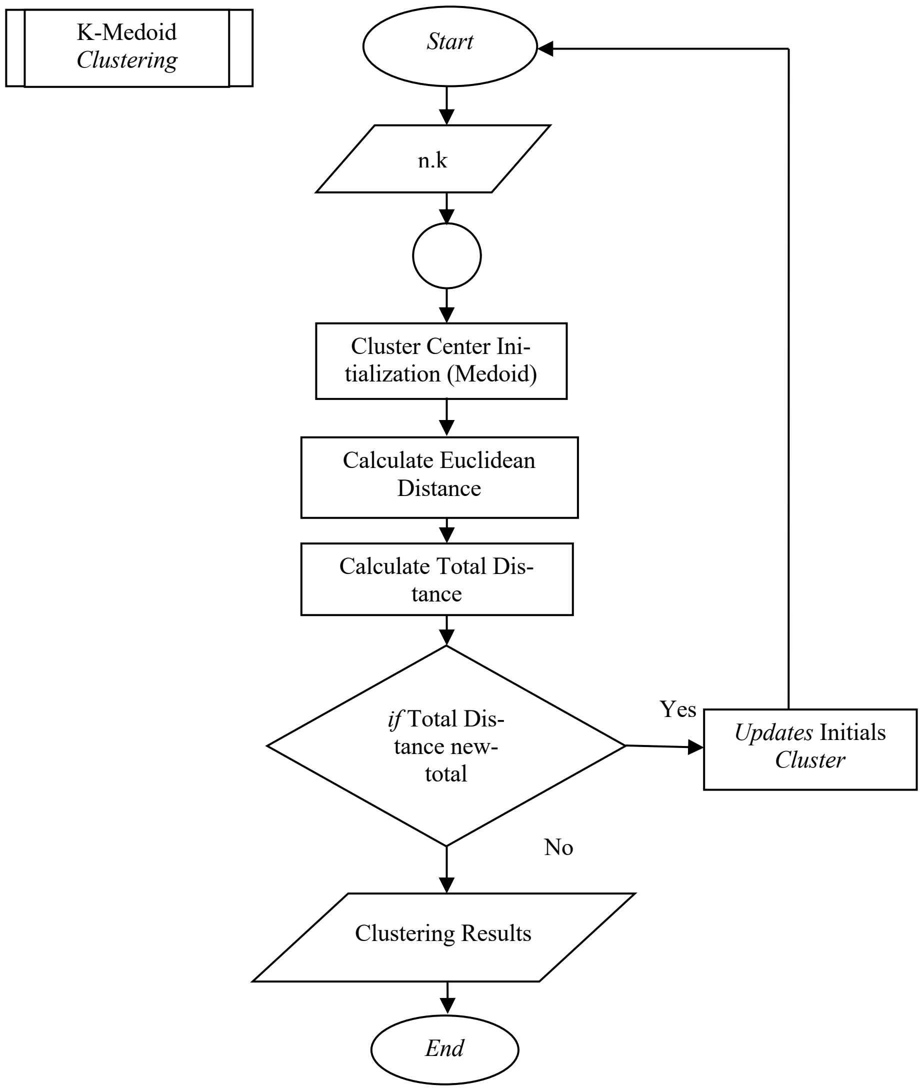

Figure 1. Flowchart of the K-Medoids method

The following is an explanation of Figure 1 above [9] :

1. Initialization of k cluster centers (number of clusters)
2. Allocate each data (object) that has been obtained to the nearest cluster using the Euclidian Distance measure equation with the equation:

$$dij = ia -)2 = i -)' - \sqrt{\textstyle\sum_{a=1}^{p}(x\, x_{ja}\sqrt{(xx_j(x_i x_j)}}$$

where i = 1,.....,n; j = 1,.....,n and p are the number of variables, and V is the covariance variance matrix.

3. Randomly select an object in each cluster as a new medoid.
4. Calculate the distance for each object that is in each cluster with the new medoid.
5. Calculate the total deviation (S) by calculating the new total cost - the old total cost. If S < 0, then swap objects with cluster data to form a new set of objects as medoids.

6. Repeat steps 3 to 5 until there is no medoid change, so that clusters and their respective cluster members are obtained.

The following is the data that will be used in conducting the research. The data used is from 2016 to 2018 in 34 provinces in Indonesia.

Table 1. Research data
(Source: bps.go.id)

Province	Prevalence of malnourished children under five by province in Indonesia (PSG)		
	Toddler Malnutrition		
	2016	2017	2018
ACEH	4.7	10.7	13
NORTH SUMATRA	5.91	9.9	11.2
WEST SUMATRA	3.76	6.7	7.4
RIAU	3.85	8.5	8.7
JAMBI	4.75	5.6	8.7
SOUTH SUMATRA	3.54	4	8.8
BENGKULU	2.18	4.8	5.2
LAMPUNG	3.48	6.4	5.8
KEEP. BANGKA BELITUNG	4.02	8.4	8
KEEP. RIAU	7.84	6.4	5.9
DKI JAKARTA	5.92	7.4	3.8
WEST JAVA	4.7	5.8	5.1
CENTRAL JAVA	5.68	5.7	6.6
IN YOGYAKARTA	4.42	5.2	3.9
BALI	1.52	3.7	3.6
WEST NUSA TENGGARA	5.49	7.7	10.3
EAST NUSA TENGGARA	13.39	14.2	14.2
WEST KALIMANTAN	13.26	12.7	10.4
CENTRAL KALIMANTAN	10.24	11.3	10.2
SOUTH KALIMANTAN	7.73	8.2	9.4
EAST KALIMANTAN	7.21	8.9	6.6
NORTH KALIMANTAN	8.64	9.5	4.1
NORTH SULAWESI	2.82	6.3	10.5
CENTRAL SULAWESI	9.78	10.9	8.6
SOUTH SULAWESI	9.42	9.4	9.4
SOUTHEAST SULAWESI	3.66	11	11.9
GORONTALO	8.65	11.5	14.9
WEST SULAWESI	9.16	9.8	12
SHAME	10.91	10.5	17.7
NORTH MALUKU	4.08	8.3	12.1
WEST PAPUA	11.23	11.7	9.2
PAPUA	6.85	13.3	9.6

3. Results and Discussion

The following are the steps of the K-Medoids algorithm for solving the problems contained in this study:

1. Initialization of k cluster centers (number of clusters) [9]

Table 2. Initial Medoid

province	Information	X	Y	Z
ACEH	1st data as cluster center	4.7	10.7	13
NORTH SUMA-TRA	2nd data as cluster center	5.91	9.9	11.2

2. Allocate each data (object) that has been obtained to the nearest cluster using the Euclidian Distance measure equation with the equation:

$$\text{dij} = \text{ia -)2} = \text{i -)' -} \sqrt{\sum_{a=1}^{p}(x\, x_{ja}\sqrt{(xx_j}(x_i x_j)}$$

where i = 1,.....,n; j = 1,.....,n and p are the number of variables, and V is the covariance variance matrix.

$D_{Aceh(C1)} = \sqrt{((4{,}7-4{,}7)^2+(10{,}7-10{,}7)^2+(13-13)^2} = 0$

$D_{Aceh(C2)} = \sqrt{((4{,}7-5{,}91)^2+(10{,}7-9{,}9)^2+(13-11{,}2)^2} = 2{,}311730953$

$D_{North\ Sumatra(C1)} = \sqrt{((5{,}91-4{,}7)^2+(9{,}9-10{,}7)^2+(11{,}2-13)^2} = 6{,}945761297$

$D_{North\ Sumatra(C2)} = = 0\sqrt{((5{,}91-5{,}91)^2+(9{,}9-9{,}9)^2+(11{,}2-11{,}2)^2}$

$D_{Bengkulu(C1)} = \sqrt{((2{,}18-4{,}7)^2+(4{,}8-10{,}7)^2+(5{,}2-13)^2} = 7{,}992221218$

$D_{Bengkulu(C2)} = \sqrt{((2{,}18 - 5{,}91)^2 + (4{,}8 - 9{,}9)^2 + (5{,}2 - 11{,}2)^2} = 6{,}796094467$

Table 3. Calculation Results of the 1st Iteration

No	province	Distance to Medoid		Nearby	*Cluster* The Approaching
		C1	C2		
1	Aceh	0	2.311730953	0	1
2	North Sumatra	2.311730953	0	0	2
3	West Sumatra	6.945761297	5,413178364	5,413178364	2
4	Riau	4.904334817	3.528965854	3.528965854	2
5	Jambi	6.671019412	5,107406387	5,107406387	2
6	South Sumatra	7.992221218	6.796094467	6.796094467	2
7	Bengkulu	10.09952474	8,713374777	8,713374777	2
8	Lampung	8.474573736	6.878582703	6.878582703	2

3. Randomly select an object in each cluster as a new medoid.

Table 4. New Medoid (Non-Medoid) 2nd Iteration

province	Information	X	Y	Z
KEEP. BANGKA BELITUNG	9th data as cluster center	4.02	8.4	8
KEEP. RIAU	10th data as cluster center	7.84	6.4	5.9

$D_{Aceh(C1)} = 5.54548465\sqrt{(4{,}7 - 4{,}02)^2 + (10{,}7 - 8{,}4)^2 + (13 - 8)^2} =$

$D_{Aceh(C2)} = 8.874660557\sqrt{((4{,}7 - 7{,}84)^2 + (10{,}7 - 6{,}4)^2 + (13 - 5{,}9)^2} =$

$D_{Riau\ Islands(C1)} = 4.796081734\sqrt{((7{,}84 - 4{,}02)^2 + (6{,}4 - 8{,}4)^2 + (5{,}9 - 8)^2} =$

$D_{Riau\ Islands(C2)} = 0\sqrt{((7{,}84 - 7{,}84)^2 + (6{,}4 - 6{,}4)^2 + (5{,}9 - 5{,}9)^2} =$

$D_{Bali(C2)} = \sqrt{((1{,}52 - 7{,}84)^2 + (3{,}7 - 6{,}4)^2 + (3{,}6 - 5{,}9)^2} = 7{,}247233955$

$D_{Bali((C1)} = \sqrt{((1{,}52 - 4{,}02)^2 + (3{,}7 - 8{,}4)^2 + (3{,}6 - 8)^2} = 6.90651866$

Table 5. 2nd Iteration Calculation Results

No	province	Distance to Medoid		Nearby	*Cluster* The Approaching
		C1	C2		
1	Aceh	5.54548465	8.874660557	5.54548465	0
2	North Sumatra	4.007754983	6.6381398	4.007754983	0
3	West Sumatra	1.821428011	4.357338637	1.821428011	0
4	Riau	0.727255113	5.307551224	0.727255113	0
5	Jambi	2,97706231	4.245951012	2,97706231	0
6	South Sumatra	4.497821695	5,714892825	4.497821695	0
7	Bengkulu	4.917885725	5.923309886	4.917885725	0
8	Lampung	3.021853736	4.361146638	3.021853736	0

4. Calculate the total deviation (S) by calculating the new total cost - the old total cost. If $S < 0$, then swap objects with cluster data to form a new set of objects as medoids.
   S = New Total Cost – Old Total Cost
   = 373,9708076 - 417,3550176
   = -43,38421
5. Repeat steps 3 to 5 until there is no medoid change, so that clusters and their respective cluster members are obtained.

Table 6. New Medoid (Non-Medoid 2) Iteration k-3

province	Information	X	Y	Z
PAPUA	34th data as cluster center	6.85	13.3	9.6
NORTH MALUKU	The 32nd data as the center of the cluster	4.08	8.3	12.1

$Daceh(C1) = \sqrt{((4{,}7 - 6{,}85)^2 + (10{,}7 - 13{,}3)^2 + (13 - 9{,}6)^2} = 4.789832982$
$DAceh(C2) = \sqrt{((4{,}7 - 4{,}08)^2 + (10{,}7 - 8{,}3)^2 + (13 - 12{,}1)^2} = 2{,}637119641$
$DBanten(C1) = = \sqrt{((7{,}43 - 6{,}85)^2 + (7{,}6 - 13{,}3)^2 + (6{,}6 - 9{,}6)^2} 6{,}467333299$
$DBanten(C2) = = \sqrt{((7{,}43 - 4{,}08)^2 + (7{,}6 - 8{,}3)^2 + (6{,}6 - 12{,}1)^2} 6{,}477846865$
$DJambi(C1) = = \sqrt{((4{,}75 - 6{,}85)^2 + (5{,}6 - 13{,}3)^2 + (8{,}7 - 9{,}6)^2} 8.031811751$
$DJambi(C2) = = \sqrt{((4{,}75 - 4{,}08)^2 + (5{,}6 - 8{,}3)^2 + (8{,}7 - 12{,}1)^2} 4.393051331$

Table 7. Calculation Results of the 3rd Iteration

No	province	Distance to Medoid		Nearby	*Cluster* The Approaching
		C1	C2		
1	Aceh	4.789832982	2,637119641	2,637119641	1
2	North Sumatra	3.873448076	2.592084104	2.592084104	1
3	West Sumatra	7.612364941	4.975178389	4.975178389	1
4	Riau	5.731491952	3,413634427	3,413634427	1
5	Jambi	8.031811751	4.393051331	4.393051331	1
6	South Sumatra	9.903842689	5,4471644	5,4471644	1
7	Bengkulu	10.64983098	7.966806135	7.966806135	1
8	Lampung	8.567782677	6.607571415	6.607571415	1

After obtaining a positive total cost result, the iteration search is stopped. The results of the total cost of the third iteration deviation are as follows.

S = New Total Cost – Old Total Cost
= 439,0338205 - 373.9708076
= 65.0630129

3.1. Results of Data Processing With RapidMiner

*RapidMiner*is software / software for data processing. Using data mining principles and algorithms, RapidMiner extracts patterns from large datasets by combining statistical methods, artificial intelligence and databases.[10]. RapidMiner is software that was created to make it easier for users to use this software. The results shown by RapidMiner can also be displayed visually with graphs, making RapidMiner one of the software of choice for extracting data using data mining methods.[11].

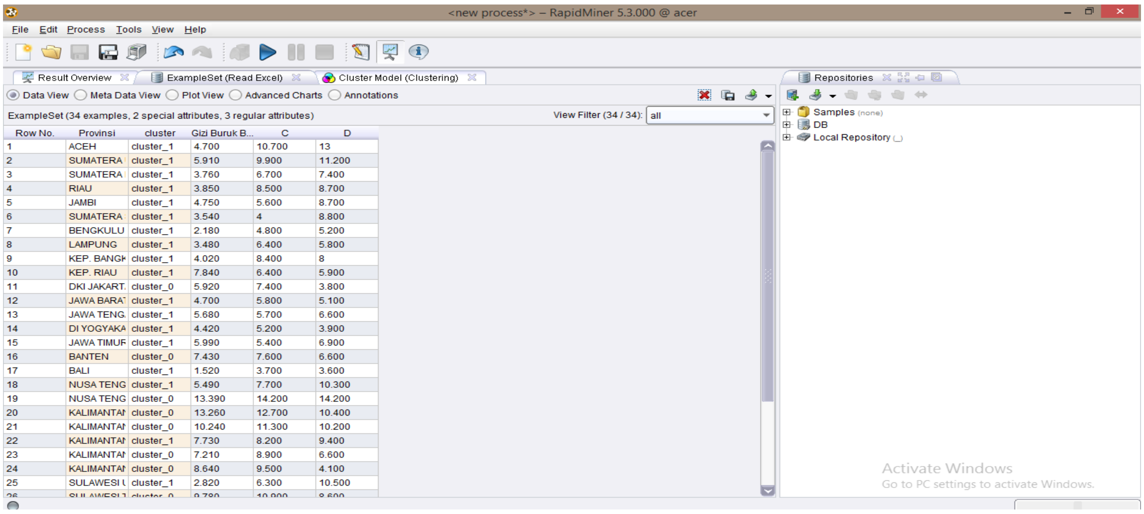

Figure 2. Cluster Data Results

In Figure 2 we can see the results of cluster_0 and cluster_1. Each existing data has been grouped based on its respective clusters.

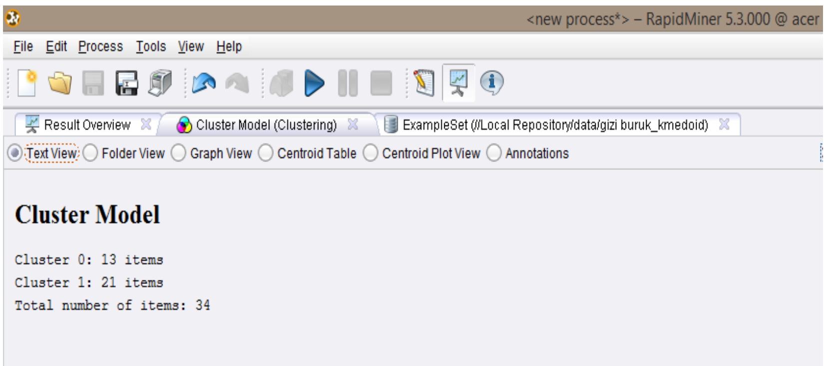

Figure 3. Cluster Model Results

In Figure 3 we can see the cluster results that have been obtained. The acquisition of cluster_0 is 13 items and cluster_1 is 21 items.

3.2. Final Results of K-Medoids and RapidMiner Tools

The following are grouping results from the manual calculations of K-Medoids and RapidMiner 5.3.

Table 8. Calculation Results and Data Testing

province	Calculation of K-Medoids	Testing with RapidMiner 5.3
ACEH	*cluster_1*	*cluster_1*
NORTH SUMATRA	*cluster_1*	*cluster_1*
WEST SUMATRA	*cluster_1*	*cluster_1*
RIAU	*cluster_1*	*cluster_1*
JAMBI	*cluster_1*	*cluster_1*
SOUTH SUMATRA	*cluster_1*	*cluster_1*
BENGKULU	*cluster_1*	*cluster_1*
LAMPUNG	*cluster_1*	*cluster_1*
KEEP. BANGKA BELITUNG	*cluster_1*	*cluster_1*
KEEP. RIAU	*cluster_1*	*cluster_1*
DKI JAKARTA	*cluster_0*	*cluster_0*
WEST JAVA	*cluster_1*	*cluster_1*
CENTRAL JAVA	*cluster_1*	*cluster_1*
IN YOGYAKARTA	*cluster_1*	*cluster_1*
EAST JAVA	*cluster_1*	*cluster_1*
BANTEN	*cluster_0*	*cluster_0*
BALI	*cluster_1*	*cluster_1*
WEST NUSA TENGGARA	*cluster_1*	*cluster_1*
EAST NUSA TENGGARA	*cluster_0*	*cluster_0*
WEST KALIMANTAN	*cluster_0*	*cluster_0*
CENTRAL KALIMANTAN	*cluster_0*	*cluster_0*
SOUTH KALIMANTAN	*cluster_1*	*cluster_1*
EAST KALIMANTAN	*cluster_0*	*cluster_0*
NORTH KALIMANTAN	*cluster_0*	*cluster_0*
NORTH SULAWESI	*cluster_1*	*cluster_1*
CENTRAL SULAWESI	*cluster_0*	*cluster_0*
SOUTH SULAWESI	*cluster_0*	*cluster_0*
SOUTHEAST SULAWESI	*cluster_1*	*cluster_1*
GORONTALO	*cluster_0*	*cluster_0*
WEST SULAWESI	*cluster_0*	*cluster_0*
SHAME	*cluster_1*	*cluster_1*
NORTH MALUKU	*cluster_1*	*cluster_1*
WEST PAPUA	*cluster_0*	*cluster_0*
PAPUA	*cluster_0*	*cluster_0*
Results	*Cluster_0* = 13	*Cluster_0* = 13
	Cluster_1= 21	*Cluster_1*= 21

4. Conclusion

Based on the results of the discussion on the grouping of under-fives with malnutrition described above, this thesis using the K-Medoids algorithm draws the conclusion that the research that has been carried out using the K-Medoids algorithm can be used in classifying toddlers with malnutrition. This study used 2 clusters, namely low clusters and high clusters, with 21 provinces obtaining low clusters and 13 provinces with high clusters. Data obtained from the Central Statistics Agency, namely the prevalence of under-fives with malnutrition from 2016-2018 in 34 provinces in Indonesia can be used to obtain information on provinces with the highest malnutrition-to-be under fives by implementing data mining methods, one of which is the K-Medoids algorithm. The clustering results obtained were tested again using Rapid Miner 5.3 software for more accurate results. The test results on Rapidminer 5.3 with manual calculations from the K-Medoids algorithm obtained the same results, namely high clusters of 13 provinces and low clusters of 21 provinces.

References

[1] C. Fadlan, S. Ningsih, and A. P. Windarto, "Penerapan Metode Naïve Bayes Dalam Klasifikasi Kelayakan Keluarga Penerima Beras Rastra," *J. Tek. Inform. Musirawas*, vol. 3, no. 1, p. 1, 2018, doi: 10.32767/jutim.v3i1.286.

[2] N. Rofiqo, A. P. Windarto, and D. Hartama, "Penerapan Clustering Pada Penduduk Yang Mempunyai Keluhan Kesehatan

Dengan Datamining K-Means," *KOMIK (Konferensi Nas. Teknol. Inf. dan Komputer)*, vol. 2, no. 1, pp. 216–223, 2018, doi: 10.30865/komik.v2i1.929.

[3] H. Sulastri and A. I. Gufroni, "Penerapan Data Mining Dalam Pengelompokan Penderita Thalassaemia," *J. Nas. Teknol. dan Sist. Inf.*, vol. 3, no. 2, pp. 299–305, 2017, doi: 10.25077/teknosi.v3i2.2017.299-305.

[4] S. R. Ningsih, I. S. Damanik, A. P. Windarto, H. S. Tambunan, J. Jalaluddin, and A. Wanto, "Analisis K-Medoids Dalam Pengelompokkan Penduduk Buta Huruf Menurut Provinsi," *Pros. Semin. Nas. Ris. Inf. Sci.*, vol. 1, no. September, p. 721, 2019, doi: 10.30645/senaris.v1i0.78.

[5] A. M. H. Pardede *et al.*, "Implementation of Data Mining to Classify the Consumer's Complaints of Electricity Usage Based on Consumer's Locations Using Clustering Method," in *Journal of Physics: Conference Series*, 2019, vol. 1363, no. 1, doi: 10.1088/1742-6596/1363/1/012079.

[6] M. Safii, "Implementasi Data Mining Dengan Metode Pohon Keputusan Algoritma Id3 Untuk Menentukan Status Mahasiswa," *J. Mantik Penusa*, vol. 2, no. 1, pp. 82–86, 2019.

[7] D. Alamsyah, M. Mexitalia, A. Margawati, S. Hadisaputro, and H. Setyawan, "Beberapa Faktor Risiko Gizi Kurang dan Gizi Buruk pada Balita 12-59 Bulan (Studi Kasus di Kota Pontianak)," *J. Epidemiol. Kesehat. Komunitas*, vol. 2, no. 1, p. 46, 2017, doi: 10.14710/jekk.v2i1.3994.

[8] D. F. Pramesti, Lahan, M. Tanzil Furqon, and C. Dewi, "Implementasi Metode K-Medoids Clustering Untuk Pengelompokan Data," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 1, no. 9, pp. 723–732, 2017, doi: 10.1109/EUMC.2008.4751704.

[9] S. Sundari, I. S. Damanik, A. P. Windarto, H. S. Tambunan, J. Jalaluddin, and A. Wanto, "Analisis K-Medoids Clustering Dalam Pengelompokkan Data Imunisasi Campak Balita di Indonesia," *Pros. Semin. Nas. Ris. Inf. Sci.*, vol. 1, no. September, p. 687, 2019, doi: 10.30645/senaris.v1i0.75.

[10] B. Rahmat C.T.I. *et al.*, "Implemetasi k-means clustering pada rapidminer untuk analisis daerah rawan kecelakaan," *Semin. Nas. Ris. Kuantitatif Terap. 2017*, no. April, pp. 58–60, 2017.

[11] M. D. Chandra, E. Irawan, I. S. Saragih, A. P. Windarto, and D. Suhendro, "Penerapan Algoritma K-Means dalam Mengelompokkan Balita yang Mengalami Gizi Buruk Menurut Provinsi," *BIOS J. Teknol. Inf. dan Rekayasa Komput.*, vol. 2, no. 1, pp. 30–38, 2021, doi: 10.37148/bios.v2i1.19.