Journal homepage: https://pcijournal.org/index.php/jmscowa

Journal of Mathematics and Scientific Computing with Applications

Published: Pena Cendekia Insani

IMPLEMENTATION OF K-MEANS METHOD IN POPULATION GROUPING BASED ON PUBLIC HEALTH INDICATORS OF NORTH SUMATRA

Nur Lela[1] , Nurleli[2], Yolanda Novita Zebua[3], Mukminah Mardiah[4] , Putri Rahma Novia[5]
[1,3]Department of Mathematics, Faculty of Science and Technology, Universitas Islam Negeri Malang

Article Info

Article history:

Received April 19, 2020
Revised May 22, 2020
Accepted June 21, 2020

Keywords:

Implementation Of K-Means Method In Population Grouping Based On Public Health Indicators Of North Sumatra

ABSTRACT

Basically, health comes from the word healthy which means well-being from all diseases, both psychological and physical. Health is a source of strength for life because of the state of the body that is prosperous and free from various diseases. Health indicators can be used to determine the condition of public health in an area. Where there are several opinions about the variables that make up the Public Health Indicators. The condition of public health in North Sumatra which is not homogeneous will make it difficult to develop a healthy city when conducting monitoring. To eradicate this problem, it is very necessary to group public health in each city and district. Based on the grouping requires several methods, the grouping of public health in North Sumatra in each city and district will be carried out using the K-Means method where this method is very popular and the method is often used.The K-Means (Cluster Analysis) method is one of the non-hierarchical cluster analysis methods that can be used to partition objects into groups based on the proximity of characteristics, so that objects that have the same characteristics are grouped in the same cluster and objects that have similar characteristics. different groups are grouped into other clusters. The purpose of clustering is to minimize the objective function set in the clustering process, which basically seeks to minimize variation within a cluster and maximize variation between clusters.

Corresponding Author:

Nur Lela
Department of Mathematics,
Universitas Islam Negeri Malang, Indonesia
Email: nurlela.nl225@gmail.com

1. INTRODUCTION

North Sumatra Province consists of 33 district/city governments, which are divided into 8 cities and 25 districts, with a total of 440 sub-districts and a total of 6,129 villages/kelurahan. North Sumatra Province is also the largest province after West Java, East Java, and Central Java, which have the largest population. Based on data from the BPS, North Sumatra province in 2018, it was recorded that it had a population of 14,441,391 consisting of 7,193,200 men and 7,222,191 women. Which has the characteristics of different degrees of health. According to Law Number 36 of 2009 concerning Health, it is stated that a health service facility is a tool or place used to carry out health service efforts, whether promotive, preventive, curative, or rehabilitation carried out by the government, be it local government or the community.

In grouping the population based on health indicators, there are several factors consisting of morbidity figures, mortality rates, health service indicators, environmental conditions indicators, health management indicators, nutritional status indicators, indicators of access and quality of health services, indicators of health resources , indicators of community life behavior, as well as indicators of the contribution of related sectors (Dinkes RI, 2010). Gerring, et al (2013) conducted a study to measure the performance of the health system using health indicators by considering regional factors. The results of his research stated that not all variables that make up health indicators can be used to measure the level of health in an area because each region has different characteristics, and the study also said that the variables that affect health indicators include economic variables, education variables. , and epidemiological variables. In this study, the grouping of districts/cities in North Sumatra will be based on the variables that make up the health indicators, namely, the economic index, the education index, and the health index, using the K-means method.

2. LITERATURE REVIEW

2.1 Clustering

Clustering is the grouping of records, observations or cases into classes that have similarities to their objects. Clustering is a collection of records that are similar to, or not similar to, records from other clusters. The clustering algorithm is used to determine the overall segment of the data set into relatively similar subgroups or clustering with the similarity of records in the clustering is maximized and the similarity of records outside the clustering is minimized. The purpose of clustering is to group objects on the basis of their characteristics. The result of clustering an object must have high internal homogeneity and high external heterogeneity. If the clustering is successful, then objects in one cluster will be close to each other if plotted geometrically and different clusters will move away from each other. Cluster validation is a procedure that evaluates the results of cluster analysis quantitatively and objectively. There are three approaches to explore cluster validity, namely external, internal and relative validation. In this study, internal validation will be used, namely by using the Connectivity index, Silhouette, and Dunn index .

2.2 k-means method

K-Means Cluster Analysis is a non-hierarchical cluster analysis method that can be used to partition objects into groups based on the proximity of characteristics, so that objects that have the same characteristics are grouped into the same cluster and objects that have different characteristics are grouped into groups. in another cluster. The purpose of clustering is to minimize the objective function set in the clustering process, which basically seeks to minimize variation within a cluster and maximize variation between clusters.

Basically, the use of algorithms in the clustering process depends on the existing data and the conclusions to be reached. For this reason, the K-Means algorithm is used which makes the following rules:

1. The number of clusters must be entered
2. It only has numeric typed attributes .

Basically, the K-Means Algorithm is a non-hierarchical method that initially takes some of the population components to be used as the initial cluster center. At this stage the cluster center is chosen randomly from a set of population data. Next K-Means tests each component in the data population and marks the component to one of the cluster centers that have been defined depending on the minimum distance between components and each cluster. The position of the cluster center will be recalculated until all data components are classified into each cluster. -each cluster center and finally a new cluster center position will be formed. The K-Means algorithm basically performs two processes, namely the process of detecting the central location of each cluster and the process of searching for members of each cluster.

2.3 Public Health Index

Health indicators can be used to determine the condition of public health in an area. Where there are several opinions about the variables that make up the Public Health Indicators. According to the Human Health Development Index (IPKM), there are 24 health indicators used as a measure of the level of human health development in an area. The health indicators are the prevalence of malnourished children under five, the prevalence of very short and short toddlers, the prevalence of very thin and thin children under five, the prevalence of obese children under five, the prevalence of diarrhea, the prevalence of pneumonia, the prevalence of hypertension, the prevalence of mental disorders, the prevalence of asthma, the prevalence of dental and oral diseases, the prevalence of disability, prevalence of injury, prevalence of joint disease, prevalence of ARI, proportion of hand washing behavior, proportion of smoking daily, access to clean water, access to sanitation, coverage of deliveries by health workers, coverage of neonatal 1 examination, coverage of complete immunization, coverage of weighing under five, ratio of doctors /Puskesmas, and the ratio of midwives/village.

3. RESEARCH METHOD

1) Data Source

In this research, the data used is secondary data about public health indicator variables in 33 regencies/cities taken from the results of Riskesdas of North Sumatra Province in 2018.

2) Research Variables

The research variables used are public health indicator variables, with the following details:

Variable	Variable name
	Puskesmas in providing services according to standards
	Nutritional Status Prevalence in toddlers
	Measles / MR immunization in infants

The following is an explanation of each research variable used as follows:

1. Variables that stated percentage in all health centers that exist in North Sumatra to provide services appropriate standard .
2. Variables that prevalence of nutritional status in infants stating the number of people in the population that the status of nutrition more , nutrition good , nutritional less and nutrient poor .
3. The variable is Measles Immunization Coverage in infants which states the number of people in the measles immunization population .

3) Analysis Step

The initial analysis step is to describe the characteristics of each public health indicator variable based on the Regency/City in North Sumatra by using the average value, variance, maximum value and minimum value. The next step is to group districts/cities in North Sumatra using the K-Means method.

4) Thought Framework

1. Start

2. Describing variables
3. Grouping data or variables using the K-means . method
4. Done

4. ANALYSIS AND DISCUSSION

Based on the grouping of public health indicators in each district and city in North Sumatra, there are several stages that need to be done, namely creating districts and cities in North Sumatra into 3 clusters for the K-Means method .

4.1 Descriptive Statistics

Before analyzing, the data description of the variables that make up public health indicators in North Sumatra is carried out in order to determine the optimal number of groups in each method, both the K-Means and Kohonen methods. Several measures of data centering were used to describe the data consisting of the average value, the minimum value variance and the maximum value.

4.2 Application of K-Means Clusterring Method Metode

At this stage, the main process will be carried out, namely segmentation or grouping of North Sumatran public health indicator data, using the K-Means *clustering* method . Experiments were carried out using the following parameters:

1. Number of clusters = 3
2. Number of variables = 3
3. Total data = 33 cities/districts

No	Name of Regency/City	Variable 1	Variable 2	Variable 3
1	Padang Sidempuan	100.00	1.7	86.7
2	Mount Sitoli	50.00	0.00	62.59
3	Field	95.12	0.46	93.06
4	Binjai	100.00	1.86	86.76
5	High cliff	100.00	1.70	92.60
6	Pematang Siantar	68.42	3.51	88.88
7	Tanjung Balai	100.00	0.40	105.05
8	Sibolga	100.00	3.54	93.88
9	West Nias	100.00	0.97	48.33
10	North Nias	45.45	2.17	63.72
11	North Batu Labuhan	82.35	6.40	93.91
12	South Batu Labuhan	82.35	1.03	92.96
13	Old Field	75.00	19.76	37.47
14	North Lawas	64.71	1.68	65.96
15	Coal	40.00	0.48	103.45
16	Serdang Bedagai	80.00	2.04	83.32
17	Sammosir	83.33	2.26	75.89
18	Mr. Bharat	100.00	4.34	72.91
19	Humbang Hasundutan	91.67	3.64	53.18

20	South Nias	28.57	23.56	89.40
21	Langkat	100.00	0.54	65.72
22	Deli Serdang	100.00	1.08	99.24
23	Karo	89.47	0.69	88.14
24	Dairi	27.78	5.30	86.23
25	Simalungun	91.18	1.71	102.59
26	sharpen	24.00	0.00	85.37
27	Labuhan Batu	92.32	4.62	96.58
28	Toba Samosir	21.05	5.87	91.79
29	North Tapanuli	100.00	0.77	61.58
30	Middle Tapanuli	73.91	10.97	106.17
31	South Tapanuli	50.00	4.00	95.34
32	Christmas Mandailing	69.23	0.47	74.91
33	Nias	10.00	1.92	90.38

Table 4.1 District/city data in variable 1, variable 2 and variable 3.

4.2.1 1st Iteration

1. Determination of the initial center of the *cluster*

The initial center of the *cluster* or centroid is obtained randomly, for the initial determination of the cluster are:

C 1 = (95.12 ; 0.46 ; 93.06)

C2 = (75.00 ; 19.76 ; 37.47)

C3 = (27.78 ; 5.30 ; 86.23)

2. *Cluster* center distance calculation

To measure the distance between the data and the center of the cluster, *Euclidien distance is* used , then the distance matrix will be obtained as follows:

d = distance

= criteria data

= centroid in the j-th cluster

For the results of the centroid distance in iteration 1, we can see in the following table:

literacy 1				
i-th data	c1	c2	c3	Cluster
1	8.11	58.09	72.31	1
2	54.45	40.58	32.87	3
3	0.00	62.19	67.86	1
4	8.09	58.09	72.30	1

5	5.06	63.17	72.59	1
6	27.20	54.32	40.77	1
7	12.95	74.61	74.79	1
8	5.83	63.80	72.65	1
9	45.00	33.11	81.68	2
10	57.71	43.26	28.79	3
11	14.11	58.46	55.12	1
12	12.78	59.03	55.15	1
13	62.19	0.00	69.40	2
14	40.75	35.28	42.28	2
15	56.09	77.14	21.66	3
16	18.05	49.41	52.40	1
17	20.91	43.03	56.59	1
18	21.09	46.03	73.44	1
19	40.16	28.01	71.95	2
20	70.54	69.76	18.55	3
21	27.77	42.34	75.23	1
22	7.90	69.21	73.50	1
23	7.50	56.04	61.89	1
24	67.86	69.40	0.00	3
25	10.39	69.49	65.58	1
26	71.54	72.70	6.57	3
27	6.13	63.43	65.37	1
28	74.28	77.81	8.75	3
29	31.86	39.58	76.45	1
30	27.06	69.27	50.57	1
31	45.32	64.98	24.05	3
32	31.62	42.51	43.24	1
33	85.17	85.69	18.57	3

Table 4.2 Results of iteration 1

4.2.2. 2nd iteration

1. Determination of new cluster center

Then we determine the position of the new centroid by calculating the average value of the existing data on the same centroid.

So that the new center point or centroid is obtained, namely:

C1 : (90.38 ; 2.50 ; 88.04)

C2 : (82.85 ; 6.51 ; 51.24)

C3 : (32.98 ; 6.19 ; 85.36)

2. Cluster center distance calculation

Calculate the Euclidean distance from all data to the new center point (C1, C2, C3) as was done in step 1 after we get the calculation results, then we compare the results. If the result of the cluster position in the 2nd iteration is the same as the position of the first iteration, then the process is stopped, and if not, the process is continued to the 3rd iteration.

literacy 2				
i-th data	c1	c2	c3	Cluster
1	9.74	39.69	67.18	1
2	47.80	35.36	29.09	3
3	7.20	44.01	62.87	1
4	9.72	39.72	67.17	1
5	10.67	45.04	67.56	1
6	22.00	40.43	35.71	1
7	19.65	56.81	70.09	1
8	11.30	46.06	67.61	1
9	40.89	18.26	76.75	2
10	51.10	39.66	25.30	3
11	10.68	42.68	50.10	1
12	9.53	42.09	50.21	1
13	55.61	20.65	65.14	2
14	33.87	23.85	37.46	2
15	52.73	67.81	20.22	3
16	11.42	32.52	47.24	1
17	14.05	25.02	51.38	1
18	18.02	27.73	68.19	1
19	34.90	9.48	66.98	2
20	65.32	68.51	18.37	3
21	24.38	23.23	70.06	2
22	14.83	51.27	68.63	1
23	2.03	37.94	56.82	1
24	62.69	65.26	5.35	3
25	14.59	52.25	60.86	1
26	66.48	68.34	10.91	3
27	9.01	46.36	60.41	1
28	69.52	73.92	13.56	3
29	28.21	20.84	71.32	2

30	25.92	55.84	46.16	1
31	41.07	55.05	19.85	3
32	24.98	27.97	38.15	1
33	80.42	82.82	23.91	3

Table 4.3 Results of the 2nd iteration

Because the results of the cluster position in the 2nd iteration are not the same as the 1st iteration position, the process will continue to the 3rd iteration.

4.2.3 3rd iteration

In this 3rd iteration we have to determine the new centroid. So we can produce:

C1: (89.32 ; 2.71 ; 90.75)

C2 : (88.56 ; 4.56 ; 55.37)

C3 : (32.98 ; 6.19 ; 85.36)

Then we do it again to find or calculate the Euclidean distance from all data to the new center point (C1, C2, C3) as we did in the previous step. So we can generate the data as follows:

literacy 3				
i-th data	c1	c2	c3	Cluster
1	11.47	33.47	67.18	1
2	48.44	39.50	29.09	3
3	6.64	38.47	62.87	1
4	11.44	33.51	67.17	1
5	10.89	39.05	67.56	1
6	20.99	39.11	35.71	1
7	18.00	51.15	70.09	1
8	11.16	40.18	67.61	1
9	43.78	13.90	76.75	2
10	51.53	43.98	25.30	3
11	8.49	39.08	50.10	1
12	7.50	38.26	50.21	1
13	57.75	27.12	65.14	2
14	34.95	26.26	37.46	2
15	50.97	68.46	20.22	3
16	11.94	29.34	47.24	1
17	16.03	21.30	51.38	1
18	20.86	20.94	68.19	1
19	37.66	3.91	66.98	2

20	64.24	71.54	18.37	3
21	27.30	15.94	70.06	2
22	13.74	45.47	68.63	1
23	3.31	33.01	56.82	1
24	61.76	68.17	5.35	3
25	12.02	47.37	60.86	1
26	65.59	71.34	10.91	3
27	6.83	41.38	60.41	1
28	68.35	76.72	13.56	3
29	31.13	13.55	71.32	2
30	23.31	53.26	46.16	1
31	39.60	55.54	19.85	3
32	25.68	27.79	38.15	1
33	79.32	86.05	23.91	3

Table 4.4 Results of the 3rd iteration

Because the results of the position of the 2nd iteration and the 3rd iteration are the same, the process is stopped. Then we can generate the data grouping as follows:

C1 = { Padang Sidempuan, Medan, Binjai, High Cliffs, Pematang Siantar, Tajung Balai, Sibolga, Labuhan Batu Utara, Labuhan Batu Selatan, Serdang Bedagai, Samosir, Pakpak Bharat, Deli Serdang, Karo, Simalungun, Labuhan Batu, Tapanuli Tengah, Christmas mandailing }.

C2 = { West Nias, Old Padang, North Old Padang, Humbang Hasundutan, Langkat, North Tapanuli }.

C3 = { Gunung Sitoli, North Nias, Coal, South Nias, Dairi, Asahan, Toba Samosir, South Tapanuli, Nias }

5. CONCLUSION

By using the application of the k-means *clustering* method in processing data from public health indicator variables, we can combine into 1 type of data that produces a different value or result. From the output that has been obtained from the process of applying the k-means *clustering* method , the results of the analysis and evaluation obtained are in the form of a data group that has been clustered from each group, where we can group the population based on health indicators which are divided into 3 groups that produce cluster 1 being the highest group, cluster 2 being the medium group and cluster 3 being the lowest group in the factors of public health indicators.

REFERENCES

[1] Zulaikah, Ika. (2017). "Thesis Grouping Using *Self Organizing Maps Clustering* (Case Study: Informatics Engineering Study Program, Universitas Nusantara PGRI Kediri)". simki.unpkediri.ac.id. 4-5.

[2] Marsudi Putri, Marina. "Clustering of Districts/Cities in East Java Province Based on Public Health Indicators Using SOM and K-Means Kohonen Methods". Journal of Science and ITS Vol.4, No.1, (2015) 2337-3520, 12-16

[3] Darmi, Yulia and Agus Setiawan. 2016. "Application of K-Means Clustering Method in Product Sales Grouping". Infotama Media Journal. Vol.12. No. 2.

[4] Wardhani, Anindya Krishna. 2016. "Implementation of the K-Means Algorithm for Grouping Patient Diseases at the Kajen Pekalongan Health Center". Journal of Transformation. Vol.14. No. 1.

[5] Bastian, Ade. 2018. "Application of the K-Means *Clustering* Analysis Algorithm in Human Infectious Diseases (Case Study of Majalengka Regency". Journal of Information Systems, Vol. 14. No. 1.

[6] Halim, Ninda Nurul and Edy Widodo. 2017. " *Clustering* of Earthquake Impacts in Indonesia Using Kohonen *Self Organizing Maps* ". Vol. 1. No. 1.

[7] The Central Bureau of Statistics for the Mandailing Natal Regency, "The health condition of the residents of the Mandailing Natal Regency in 2019", 2020.

[8] Central Bureau of Statistics for Asahan Regency , " Health condition of the population of Asahan Regency in 2019", 2020.

[9] The Central Bureau of Statistics of the Deli Serdang Regency, "The health condition of the population of the Deli Serdang Regency in 2019", 2020.

[10] The Central Bureau of Statistics for the Karo Regency "Karo Regency People's Welfare Indicator 2019" 2020.