Journal of Artificial Intelligence and Engineering Applications

Website: https://ioinformatic.org/

15th February 2022. Vol. 1. No. 2

Implementation Of The K-Means Method In Grouping Districts And Cities In North Sumatra On Social Welfare Problems

Edisman Rahul Gonjales Siahaan[1]*, Poningsih[2], Yuegilion Pranayama Purba[3], Solikhun[4], Wendi Robiansyah[5]

[1,3,5]STIKOM Tunas Bangsa Pematangsiantar, North Sumatra, Indonesia
[2,4]AMIK Tunas Bangsa Pematangsiantar, North Sumatra, Indonesia
**edismansiahaan@gmail.com*

Abstract

Social welfare problems are obstacles, difficulties or disturbances experienced by a person, difficulty or disturbance. or groups that cannot carry out their social functions and cannot establish harmonious and harmonious relationships with the surrounding environment. In this case, the problem of social welfare that the author does is in the province of North Sumatra which has 33 districts and cities. The source of this research data comes from the Central Bureau of Statistics of North Sumatra. The aim of this is to apply the k-means method in grouping districts and cities on social welfare issues that can help the government and social services in making decisions which areas should be dominantly assisted in solving social welfare problems in order to save costs. The k-means method is one of the methods in data mining to group data sets that are similar to others. The data are grouped into 3 clusters, namely high, medium and low clusters, the results of the high cluster are 2 regencies/cities, the medium cluster is 6 regencies/cities and the low cluster is 25 regencies/cities, these results can be a record for the local government and agencies in dealing with social welfare problems in regencies and cities in North Sumatra.

Keywords: *Social Welfare Problems, Data Mining, K-Means, North Sumatra*

1. Introduction

North Sumatra is a province located in the northern part of the island of Sumatra which is part of the State of Indonesia. North Sumatra province has 25 regencies and 8 cities with Medan as its capital[1]. In 2020 North Sumatra has a population of 15,136,552 people. A small part of the population of North Sumatra experience social welfare problems that are not good, causing anxiety to the environment. This social welfare problem is fostered by the local government and the social service of North Sumatra. North Sumatra has 33 regencies/cities which are one of the most populous provinces in Indonesia. In North Sumatra, there are many people in every district and city who experience social welfare problems, be it neglected children, the poor, victims of violence and so on. The condition of the large number of people experiencing social welfare problems can make it difficult for the government or social services to carry out guidance or monitoring in each district and city. If the development process is carried out on social welfare issues in each district/city simultaneously, it will require a lot of time and costs to reduce or overcome social welfare problems.

In the description above, the regencies and cities are grouped on social welfare issues using themethod K-Means clustering[2],[3],[4],[5]. K-Means clustering is done to group data sets that are similar to other or dissimilar data in other groups[6],[7]. The method has high accuracy in object size so that it is more valuable and efficient in processing large objects[8]. By using themethod K-Means, it is possible to group districts and cities on social welfare issues in order to find out which areas in North Sumatra are priority areas to be fostered by the local government in order to achieve a reduction in social welfare problems.

2. Research methodology

North Sumatra is a province located in the northern part of the island of Sumatra which is part of the State of Indonesia. North Sumatra province has 25 regencies and 8 cities with Medan as its capital. In 2020 North Sumatra has a population of 15,136,552 people. A small part of the population of North Sumatra experience social welfare problems that are not good, causing anxiety to the environment. This social welfare problem is fostered by the local government and the social service of North Sumatra. North Sumatra has 33 regencies/cities which are one of the most populous provinces in Indonesia. In North Sumatra, there are many people in every district and city who experience social welfare problems, be it neglected children, the poor, victims of violence and so on. The condition of the large number of people experiencing social welfare problems can make it difficult for the government or social services to carry out guidance or monitoring in each district and city. If the development process is carried out on social welfare issues in each district/city simultaneously, it will require a lot of time and costs to reduce or overcome social welfare problems.

In the description above, the regencies and cities are grouped on social welfare issues using themethod k-means clustering. K-means clustering is done to group data sets that are similar to other or dissimilar data in other groups[9]. K-means has high accuracy in object size so that it is more valuable and efficient in processing large objects. By using themethod k-means, it is possible to group districts and cities on social welfare issues in order to find out which areas in North Sumatra are priority areas to be fostered by the local government in order to achieve a reduction in social welfare problems.

The research design can be seen in the image below to be able to find out the description and problem solving process in this study.

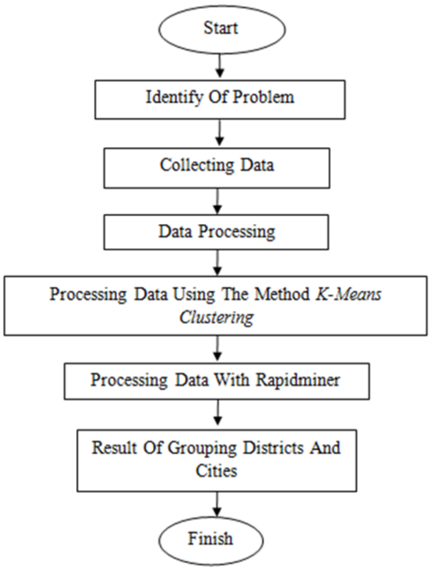

Start
Identify Of Problem
Collecting Data
Data Processing
Processing Data Using The Method *K-Means Clustering*
Processing Data With Rapidminer
Result Of Grouping Districts And Cities
Finish

Figure 1. Research Scheme

In Figure 1 The explain of design process research done to determine clustering counties and cities on social welfare issues with method k-means following is an explanation: Identification of Problems is to analyze the problems relating to classify the counties and cities on social welfare issues. In this study, the criteria were data on social welfare problems in 2019. Collecting Data for this research were obtained from the Central Statistics Agency of North Sumatra with the website https://sumut.bps.go.id. Data Processing with excel to obtain results whose data can be processed to the next process, so that the resulting data is factual and accurate. Processing Data Using the K-Means Clustering, Processing the data first manually using the method k-means, then grouping the existing data into three groups, namely high, medium and low by using the closest distance in the method k-means clustering. Processing Data With RapidMiner and testing of data is carried out using an application, namely Rapid miner. Then the results obtained from the comparison in manual data processing with the results of data processing with software, if the results are the same then the processing is successful. Result of Grouping Districts and Cities of this study are the grouping of districts and cities on social welfare problems based on the calculation of the method k-means. clustering use fault the North Sumatra Social Service.

3. Results And Discussion

The results of this study present the data processing process which is divided into two parts, namely the manual section using themethod k-means clustering and equating the results from manual calculations with testing in the application using RapidMiner 5.3. The data used is data on social welfare problems in the province of North Sumatra 2019. Following steps are data processing using the k-means method clustering:

3.1. Manual Calculation of the K-Means Method

The following is the process carried out insocial welfare issues using k-means clustering. Determining the amount of data to be clustered. Grouping districts and cities in North Sumatra onIn grouping districts and cities in North Sumatra on social welfare issues by using 33 districts/cities. The following list of tables of social welfare problems in districts and cities in North Sumatra in the use of data can be seen in table 1:

Table 1. Data on Social Welfare Problems

NO	District - City	TYPE OF SOCIAL WELFARE PROBLEMS																								
		A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
1	Nias	90	315	10	0	1	1	0	298	1	7	311	6	1	12	17	929	0	121	0	6	28	1	2	5	1	1369
2	Mandailing Natal	2	1	12	0	4	0	4	0	0	48	67	0	10	4	1785	46255	0	0	0	45	0	0	0	0	0	365
3	Tapanuli Selatan	2	0	0	0	19	0	30	150	1	0	676	0	0	0	5043	17972	0	0	0	0	0	0	0	0	0	135
4	Tapanuli Tengah	22	101	31	0	14	38	9	1098	0	192	448	20	3	31	209	15700	0	209	0	51	56	28	0	0	40	2292
5	Tapanuli Utara	4	10	0	0	30	0	0	1260	0	45	1508	3	0	0	1	28688	0	3	0	1	18	5	0	0	32	675
..	..	..	..	..	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
33	Gunung Sitoli	1	94	30	0	19	1	5	245	8	2	1317	6	4	15	18	17094	1	0	0	0	35	1	0	0	2	816

Information: A: Abandoned Toddler, B: Abandoned Child, C: Street Child, D: Jermal Child, E: Naughty Child, F: Victim of Violence, G: Narcotics Victims, H: Socio-Economic Vulnerable Women, I: Women Victims of Violence, J: Tuna Susila, K: Persons with Disabilities, L: Homeless, M: Beggars, N: Former Convicts, O: Victims of Natural Disasters, P: Poor Fakir, Q: Trafficking, R: Remote Indigenous Communities, S: Migrant Workers with Problems, T: Victims of Social Disasters, U: Families with Social Psychology Problems, V: Scavengers, W: Minority Groups, X: People with HIV/AIDS, Y: Children Need Special Protection, Z: Elderly.

Determine the value of k for the number of clusters. The number of clusters is 3 clusters, namely low (CO), medium (C1), and high (C2) clusters. Determining theValue Centroid (Center Cluster), it is cluster initial determined randomly from the data, the value for the low(clustercluster 0) is taken from the lowest value, the medium(cluster 1) is taken from the average value, and the cluster high(cluster 2) is taken from the highest score on each social welfare problem. The following is a list of centroids initial:

Table 2. Centroid Initial Data

	Cluster 0	*Cluster* 1	*Cluster* 2
A	0	30.30	360
B	0	190.18	2294
C	0	9.27	106
D	0	0	0
E	0	10.30	60
F	0	6.79	112
G	0	53.55	923
H	0	809	4413
I	0	3.91	60
J	0	22.88	192
..	…	……	……

After the centroid is determined, the next thing to do is to calculate the distance of each data to the center of the cluster. The process of searching for the shortest data in iterations can be seen in the calculation below:

$$E_{Nias,C0}= \sqrt{\begin{array}{c}(90-0)^2+(315-0)^2+(10-0)^2+(0-0)^2+(1-0)^2+(1-0)^2\\ +(0-0)^2+(298-0)^2+(1-0)^2+(7-0)^2+(311-32)^2+(6-0)^2\\ +(1-0)^2+(12-0)^2+(17-0)^2+(929-0)^2+(0-0)^2+\\ (121-0)^2+(0-0)^2+(6-0)^2+(28-0)^2+(1-0)^2+(2-0)^2+\\ (5-0)^2+(1-0)^2+(1369-0)^2\end{array}}$$

$$=1739{,}91$$

The results of the entire calculation can be seen in table 3:

Table 3. Results of Calculation of Data Distances on Centroid Iteration 1

No	Regency – City	C0	C1	C2	Distance Shortest	Result
1	Nias	1739,91	29128,07	220068,50	1739,91	C0
2	Mandailing Natal	46290,93	16478,03	174998,68	16478,03	C1
3	Tapanuli Selatan	18678,37	13195,41	203128,54	13195,41	C1
4	Tapanuli Tengah	15914,43	14334,36	205262,51	14334,36	C1
5	Tapanuli Utara	28761,56	2514,92	192452,63	2514,92	C1
…	………….	………	…………	……..	…………	……
33	Gunung Sitoli	17163,74	13077,51	204002,78	13077,51	C0

Grouping Data Based on Distance to the centroid nearest. If the lowest value is in cluster 0 then it will enter k into cluster 0, if the lowest value is in cluster 1 then it will enter into cluster 1, and if the lowest value is in cluster 2 then it will enter cluster 2.

Table 4. Results Cluster Iteration 1

Cluster	Value
Low (C0)	15
Medium (C1)	16
High (C2)	2

Repeat Step 3, perform iterations so that the data grouping is the same as the previous iteration data grouping. If they are not the same, then repeat steps 3-5 until the results of the cluster have iteration the same value. Calculate the centroid using the results of the iteration of each member in each cluster, the following is the calculation:

$$D_{C0,A=}\frac{\begin{array}{c}90+0+1+360+2+0+9+107\\+0+52+12+0+0+7+0\end{array}}{15}=42{,}67$$

$$D_{C1,A=}\frac{\begin{array}{c}2+2+22+4+21+25+6+0\\+238+0+1+2+0+4+0+1\end{array}}{16}=20{,}5$$

$$D_{C2,A=}\frac{1+31}{2}=16$$

The overall results of the calculations of C0,AZ, C1,A-Z, C2,AZ, are in the following table:

Table 5. Result Centroid 2

	C0	C1	C2
A	42,67	20,50	16
B	251,53	134,69	174
C	9,33	9,63	6
D	0	0	0
E	6,33	12,63	21,50
F	9	5,25	2,5
G	76,33	25,88	104
H	495,67	1033,25	1365
I	1,8	5,25	9
J	12	30,75	41,5
….	…..	…..	……
Z	1828,27	2367,31	9378,5

The calculation is carried out until iteration 7 with the final result as follows:

Table 6. Result of Calculation of Data Distance on Centroid Iteration 7

No	Regency – city	C0	C1	C2	Distance Shortest	Result
1	Nias	10190,98	56219,37	184966,65	10190,98	C0
2	Mandailing Natal	35287,76	11110,83	139777,93	11110,83	C1
3	Tapanuli Selatan	8683,91	39467,01	168079,68	8683,91	C0
4	Tapanuli Tengah	4669,78	41450,02	170166,97	4669,78	C0
5	Tapanuli Utara	17736,15	28467,39	157271,80	17736,15	C0
6	Toba	8497,72	49692,20	177467,53	8497,72	C0
7	Labuhan Batu	7803,91	39667,53	168037,95	7803,91	C0
8	Asahan	30751,69	15752,30	144230,14	15752,30	C1
9	Simalungun	66516,59	20455,76	108493,49	20455,76	C1
10	Dairi	20809,88	25446,36	154168,69	20809,88	C0
…..	……….	………	………..	………….	………….	……
33	Gunung Sitoli	6271,11	40063,04	168844,62	6271,11	C0

Grouping Data Based on Distance to the centroid nearest:

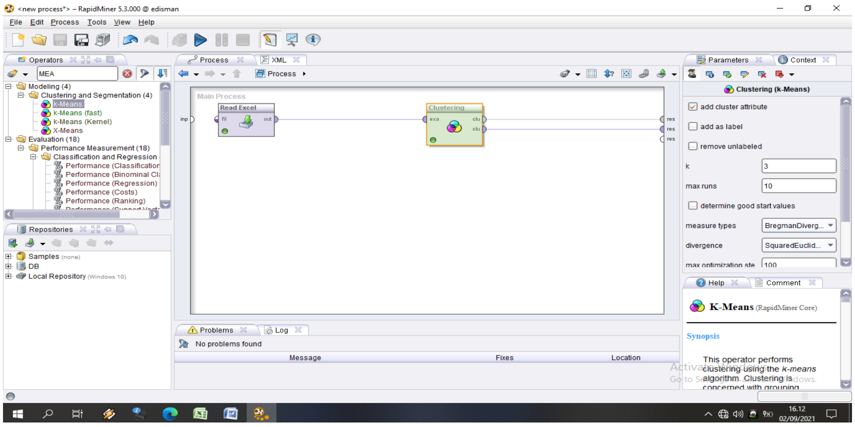
Table 7. Results of Cluster Iteration 7

Cluster	Value
Low (C0)	25
Medium (C1)	6
High (C2)	2

Based on manual calculations carried out by the author on data on social welfare problems that have obtained the final results in iteration 6 and iteration 7 have the same value, at C0 is 25 , C1 is 6 while C2 is 2 where there is no change in position then the process is stopped and concluded: Cluster High (C0) with a high number of social welfare problems, namely Nias, South Tapanuli, Central Tapanuli, North Tapanuli, Toba, Labuhan Batu, Dairi , Langkat, South Nias, Humbang Hasundutan, Pakpak Bharat, Samosir, Serdang Bedagai, North Padang Lawas, Padang Lawas, Labuhan Batu Selatan, Labuhan Batu Utara, North Nias, West Nias, Sibolga, Tanjung Balai, Pematangsiantar, Binjai, Padang Sidempuan, Mount Sitoli. Cluster Medium (C1) the number of existing social welfare problems in the areas of Mandailing Natal, Asahan, Simalungun, Deli Serdang, Tebing Tinggi and Medan. Low Cluster (C2) the number of social welfare problems is low in the districts of Karo and Batubara.

3.2. Testing using RapidMiner Software

After that process the data with the k-means method into a worksheet then read excel relationship with the k-means operator and determine the number of clusters.



☑ add cluster attribute

□ add as label

□ remove unlabeled

k: 3

max runs: 10

□ determine good start values

measure types: BregmanDiverg...

divergence: SquaredEuclid...

max optimization ste: 100

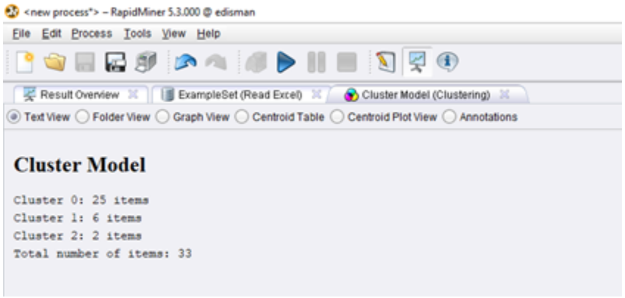
K-Means (RapidMiner Core)

Synopsis

This operator performs clustering using the *k-means* algorithm. Clustering is concerned with grouping

Figure 2. Data Processing

After determining the number of clusters and then clicking Run (blue triangle image) then the results will appear.


<new process*> – RapidMiner 5.3.000 @ edisman
File Edit Process Tools View Help
Result Overview
ExampleSet (Read Excel)
Cluster Model (Clustering)
● Text View ○ Folder View ○ Graph View ○ Centroid Table ○ Centroid Plot View ○ Annotations

Cluster Model

```
Cluster 0: 25 items
Cluster 1: 6 items
Cluster 2: 2 items
Total number of items: 33
```

Figure 3. Result

Description: Cluster 0 (High) as many as 25 regencies/cities, Cluster 1 (Medium) as many as 6 Regencies/cities, Cluster 2 (Low) as many as 2 Regencies and Cities.

4. Conclusion

Based on the results of data calculations carried out by researchers using the method k-means, the authors can draw the conclusion that data mining techniques using themethod k-means can be applied in classifying district and city data on social welfare problems. The data processed in the calculation of the k-means results are 2 high regencies/cities, 6 medium regencies/cities and 25 low regencies/cities. The results of the high cluster can help the social service and local government in improving the work system to reduce social welfare problems in high areas. The results of the trial using software as a tool to prove that the manual data calculation results are the same as the tests carried out using Rapidminer 5.3.

Acknowledgement

Acknowledgments to the supervisors and examiners who are lecturers at AMIK and STIKOM Tunas Bangsa so that this research can be arranged as one of the requirements for completing Bachelor's education (S1) at STIKOM Tunas Bangsa. I hope this research can be a reference for other research related to the methods and algorithms used. I hope for constructive suggestions for the readers for the perfection of this research in the future.

References

[1] L. A. N. Gsa, L. A. N. Gkat, K. O. Ta, U. T. A. R. A. Helatoba-tarutung, and U. T. A. Ra, “North sumatra,” pp. 1–2, 2013.
[2] S. Dini and A. Fauzan, “Clustering Provinces in Indonesia based on Community Welfare Indicators,” *EKSAKTA J. Ilmu-ilmu MIPA*, vol. 20, pp. 56–63, Feb. 2020, doi: 10.20885/EKSAKTA.vol1.iss1.art9.
[3] P. M. Hasugian, H. D. Hutahaean, B. Sinaga, Sriadhi, and S. Silaban, “Villages Status Classification Analysis Involving K-Means Algorithm to Support Kementerian Desa Pembangunan Daerah Tertinggal dan Transmigrasi Work Programs,” *J. Phys.*

Conf. Ser., vol. 1641, no. 1, 2020, doi: 10.1088/1742-6596/1641/1/012058.

[4] N. A. Khairani and E. Sutoyo, "Application of K-Means Clustering Algorithm for Determination of Fire-Prone Areas Utilizing Hotspots in West Kalimantan Province," *Int. J. Adv. Data Inf. Syst.*, vol. 1, no. 1, pp. 9–16, 2020, doi: 10.25008/ijadis.v1i1.13.

[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] I. R. Munthe, B. H. Rambe, R. Pane, D. Irmayani, and M. Nasution, "Jurnal Mantik," *J. Mantik*, vol. 3, no. January, pp. 31–38, 2019.

[7] C. S. Li, "Cluster center initialization method for K-means algorithm over data sets with two clusters," *Procedia Eng.*, vol. 24, pp. 324–328, 2011, doi: 10.1016/j.proeng.2011.11.2650.

[8] R. Dash, D. Mishra, A. Rath, and M. Acharya, "A hybridized K-means clustering approach for high dimensional dataset," *Int. J. Eng. Sci. Technol.*, vol. 2, Sep. 2010, doi: 10.4314/ijest.v2i2.59139.

[9] M. Hossain, M. N. Akhtar, R. B. Ahmad, and M. Rahman, "A dynamic K-means clustering for data mining," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 13, p. 521, Feb. 2019, doi: 10.11591/ijeecs.v13.i2.pp521-526.