Journal of Artificial Intelligence and Engineering Applications

Website: https://ioinformatic.org/

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

Implementation of K-Means Clustering on High School Students Management

Anggriani Dwi Kartina[1*], M. Safii[2]

[1] STIKOM Tunas Bangsa Pematangsiantar, Sumatera Utara, Indonesia*
[2] AMIK Tunas Bangsa Pematangsiantar, Sumatera Utara, Indonesia
[1,2]Jln. Sudirman Blok A No. 1-3 Pematangsiantar, Sumatera Utara
[1] anggrianadwi25@gmail.com; [2] m.safii@amiktunasbangsa.ac.id*

Abstract

The quality of national education and teaching needs to be monitored continuously in every stage and step of educational activities. The monitoring is intended as an effort to control the quality of education and furthermore as a guarantee of the quality of education. Therefore, a method is needed to facilitate the grouping of high school student data. With the k-means clustering approach, the division of student groups can be done based on the national final exam scores. In this study, students were clustered using the K-Means algorithm. By using K-Means, it aims to facilitate the grouping of the highest and lowest Pemtangssiantar High School students. The result is a picture that shows the grouping of students based on national final exam scores.

***Keywords**: Grouping, Data Mining, Clustering, K-Means, High School Students*

1. Introduction

One way that can be used to measure the quality of education is to group the UAN scores obtained by each school[1],[2]. Therefore, an analysis is needed to obtain more detailed results in the grouping of schools so that the information obtained is a description of the quality of the school based on the results of the National Examination and other value components that influence it. The relatively different national final exam scores in each school can be used as a reference by the government in order to improve and equalize the quality of education in Indonesia in general and in Pematangsiantar in particular. Therefore, the central government and local governments must also pay attention to the acquisition of UAN scores obtained by each school. Indeed, the national final exam cannot be used as the sole measure of the quality of education in schools, but the national final exam is the first and most visible indicator in society to measure the quality of education.

A fairly popular method to answer this problem is cluster analysis[3],[4],[5]. Cluster analysis is a name for groups in a multivariate technique which essentially aims to group objects based on the characteristics of the object. The results of grouping objects must be able to show high internal homogeneity (within clusters) and high external heterogeneity as well (between clusters).

In the hierarchical method, the determination or selection of the number of clusters is carried out by the clustering process[6],[7], in other words the number of clusters cannot be known beforehand, the results of which are entirely left to the researcher by prioritizing subjectivity in accordance with the research objectives. This resulted in the clusters formed could be 4, 5, 3 or 2 related to the subjectivity of the researcher. Whereas in the non-hierarchical method, the determination or selection of the number of clusters must be determined at the beginning before the clustering process runs, so that the end result will form the same number of clusters[8].

2. Research Methodology

Grouping Pematangsiantar high school students using the K-Means Clustering method. In this section the author will explain how the data collection procedures will be used in this study. In research, several procedures for data collection such as library research are used, namely using libraries[9], books, proceedings or journals as a medium for reference material in determining the parameters used in research and data sources taken from the branch office of the Pematangsiantar City Education office.

2.1. Data Analisys

The data analysis process is carried out after data collection and processed into Microsoft Excel whose results will be applied to RapidMiner. The author will analyze descriptive statistical data, namely methods related to collecting data, presenting a data set so as to provide useful information. The type of data used in this study is secondary data, namely data obtained not from the source directly but has been collected by other parties and has been processed and has a relationship with the problems studied.

Data on the achievement of high school national exam scores in Pematangsiantar used in this study can be seen in the following table:

TABLE 1
Data on the achievement of high school national in Pematangsiantar

No	Name Of Education Unit	Average Value On Test Eyes							
		Indonesian	English	Mathematic	Physics	Chemical	Biology	Sosiology	Geographic
1	SMA Negeri 1	71.23	58	41.93	47.26	45.96	53.7	52.77	59
2	SMA Negeri 2	69.13	58.07	42.03	43.44	55.44	51.79	55.46	63.5
3	SMA Negeri 3	71.51	57.71	44.16	55.13	49.17	56.75	50.41	45.44
4	SMA Negeri 4	78.76	68.39	60.62	45.95	64.9	60.33	60.15	47
5	SMA Swasta Rk Bintang Timur	75.37	62.36	42.47	47.76	52.64	50.63	63.68	67.5
6	SMA Swasta YP HKBP 1	58.67	43.53	34.25	40.83	45.83	45.6	43.71	47.5
7	SMA Swasta Harapan	50.12	46.94	32.21	42.92	45.78	44.77	42	56
8	SMA Swasta Seminari	81	72	47.5	80	55.88	53.21	75.73	45
9	SMA Swasta Pelita	67.59	63.59	57.78	68	66.87	58.75	45	56
10	SMA Swasta YPI	62.5	48.75	34.69	51.25	56.66	42.5	43	57
11	SMA Swasta Melati	50.16	47.47	30.41	55	56	42.09	46	37.45
12	SMA Swasta Taman Siswa	63.21	44.8	33.11	33.89	45	46.26	55	44.43
13	SMA Swasta PGRI 10	57.25	40	34.06	54	64	39.38	56	67
14	SMA Swasta Kartika I-4	61.11	46.73	37.58	43.89	41.67	48.72	45.88	41.2
15	SMA Swasta Perguruan Keluarga	67.94	53.77	34.92	36.88	44.69	46.18	50.64	42.26
16	SMA Swasta Teladan	67.32	47.14	36.82	38.75	48.33	47.38	55.35	57.75
17	SMA Swasta Kristen Kalam Kudus	76.82	71.64	45.61	56.19	55.57	65.23	64.49	63
18	SMA Swasta Methodist	76.86	72.08	55.78	63.36	74.55	62.5	51.25	57.89
19	SMA Swasta Surya	48	39.67	26.25	53	65	39.17	28	37.67
20	SMA Swasta Budi Mulia	81.76	82.01	65.05	63.46	64	81.89	68.79	71.44
21	SMA Swasta Erlangga	58.67	40	31	56	62	38.83	44.45	57
22	SMA Swasta Kampus Hkbp Nomensen	64.92	48.75	36.25	35.25	45.77	43.33	56.2	46.4
23	SMA Swasta Advent	70.62	56.95	38.03	43.25	48.75	53.65	52	45.13
24	SMA Swasta Sultan Agung	73.54	61.1	41.98	44.38	48.44	53.86	52.37	48.18
25	SMA Swasta Mars	66.31	43.92	31.54	36.5	63	38.93	47.13	60
26	SMA Swasta Tri Sakti	45.43	37.71	27.86	48	45	35.36	67	58
27	SMA Negeri 5	66	47.26	35.84	38.13	36.96	45.84	45.98	40.73
28	SMA Negeri 6	65.9	54.1	38.6	38.61	49.42	50.14	50	47.28

2.2. Research Contribution

It is hoped that the results of this study can contribute to the Pematangsiantar local government so that students who enter the low cluster get more attention and improve welfare and should be given more guidance by the Pematangsiantar city government in the future.

3. Result and Discussion

In this study, the data were grouped into 2 clusters, namely the highest and lowest national examination scores. The following is a description of the manual calculation process for the K-Means clustering algorithm[10],[11],[12].

3.1. Data processing

The following are the steps in data processing using the K-Means Algorithm:

a. Determining the Data to be Clustered
The data on the achievement of high school national exam scores in Pematangsiantar used in this study consisted of 28 schools with the scores shown in Table 1.

b. Determining the Value of k Number of Clusters
The number of clusters is 2 clusters. The clusters formed are high clusters (C1) and low clusters (C2).

c. Determining the Centroid Value (Cluster Center)
Determination of the initial cluster center is determined randomly which is taken from the data in the range. The value for the high cluster (cluster 1) is taken from the highest value in table 1 and the value for the lowest cluster (cluster 2) is taken from the lowest value in table 1. The following is the data centroid table in table 2.

TABLE 2
Initial Data Centroid

	Ind	Eng	Math	Phis	Chem	Bio	Sosio	Geo
C I	81.76	82.01	65.05	80	74.55	81.89	75.73	71.44
C II	45.43	37.71	26.25	33.89	36.96	35.36	28	37.45

d. Calculating the Distance of Each Data to the Centroid (Cluster Center)
After the initial cluster center value data is determined, the next step is to calculate the distance of each data to the cluster center[13]. The process of finding the shortest distance in iteration 1 can be seen in the calculations and tables below:

$$D_{BI,c1} = \sqrt{\begin{array}{c}(71{,}23-81{,}76)^2+(58-82{,}01)^2+\\(41{,}93-65{,}05)^2+(47{,}26-80)^2+(45{,}96-74{,}55)^2+(53{,}7-81{,}89)^2+\ (52{,}77-75{,}73)^2+(59-71{,}44)^2\end{array}}$$

$= 67.73317$

$$D_{BI,c2} = \sqrt{\begin{array}{c}(71{,}23-45{,}43)^2+(58-37{,}71)^2+\\(41{,}93-26{,}25)^2+(47{,}26-33{,}89)^2+(45{,}96-36{,}96)^2+(53{,}7-35{,}36)^2+\ (52{,}77-28)^2+(59-37{,}45)^2\end{array}}$$

$= 54.74719$

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

TABLE 3
Calculation Result of Center Cluster Distance Iteration 1

NO	NAME OF EDUCATION UNIT	C I	C II	Nearest distance
1	SMA Negeri 1	67.73317	54.74719	54.74718623
2	SMA Negeri 2	65.9114	57.97016	57.97016388
3	SMA Negeri 3	65.90696	54.9976	54.99760449
4	SMA Negeri 4	52.6454	76.86261	52.64539771
5	SMA Swasta Rk Bintang Timur	59.95157	67.93079	59.95156712
6	SMA Swasta YP HKBP 1	90.75789	29.20059	29.20059417
7	SMA Swasta Harapan	90.9181	30.50729	30.50728601
8	SMA Swasta Seminari Menengah	47.73851	89.61288	47.73850752
9	SMA Swasta Pelita	50.14087	73.43892	50.14086856
10	SMA Swasta YPI	79.85523	42.80353	42.80352556
11	SMA Swasta Melati	89.35232	36.22682	36.22682294

NO	NAME OF EDUCATION UNIT	C I	C II	Nearest distance
12	SMA Swasta Taman Siswa	90.42652	37.07598	37.0759774
13	SMA Swasta Pgri 10	79.55721	54.90801	54.9080067
14	SMA Swasta Kartika I-4	87.84337	33.01549	33.01548576
15	SMA Swasta Perguruan Keluarga	86.04328	39.51283	39.51282956
…	……….	……….	……….	……….

Cluster results can be seen in the following Table 4

TABLE 4
Cluster results Iteration 1

NO	NAMA SATUAN PENDIDIKAN	C1	C2
1	SMA Negeri 1		1
2	SMA Negeri 2		1
3	SMA Negeri 3		1
4	SMA Negeri 4	1	
5	SMA Swasta Rk Bintang Timur	1	
6	SMA Swasta YP HKBP 1		1
7	SMA Swasta Harapan		1
8	SMA Swasta Seminari	1	
9	SMA Swasta Pelita	1	
10	SMA Swasta YPI		1
11	SMA Swasta Melati		1
12	SMA Swasta Taman Siswa		1
13	SMA Swasta Pgri 10		1
14	SMA Swasta Kartika I-4		1
15	SMA Swasta Perguruan Keluarga		1
….	……….	…..	…..

e. Determining the Position of the Cluster or Grouping
The K-Means process will continue to iterate until the data grouping is the same as the previous iteration data grouping[14]. The process will continue to iterate until the data in the last iteration is the same as the previous iteration.

f. Calculates the new centroid using the results in each cluster. After getting the results of the distance from each object in the 1st iteration, then proceed to the 2nd iteration.

TABLE 5
Cluster result iteration 2

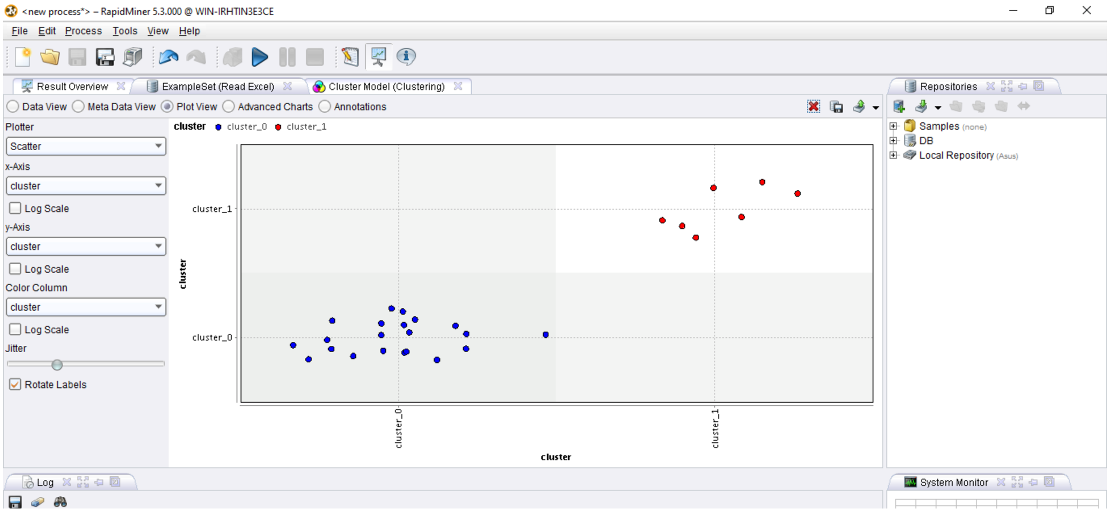
NO	NAME OF EDUCATION UNIT	C 1	C 2
1	SMA Negeri 1		1
2	SMA Negeri 2		1
3	SMA Negeri 3		1
4	SMA Negeri 4	1	
5	SMA Swasta Rk Bintang Timur	1	
6	SMA Swasta YP HKBP 1		1
7	SMA Swasta Harapan		1
8	SMA Swasta Seminari Menengah	1	
9	SMA Swasta Pelita	1	
10	SMA Swasta YPI		1

NO	NAME OF EDUCATION UNIT	C 1	C 2
11	SMA Swasta Melati		1
12	SMA Swasta Taman Siswa		1
13	SMA Swasta Pgri 10		1
14	SMA Swasta Kartika I-4		1
15	SMA Swasta Perguruan Keluarga		1
16	SMA Swasta Teladan		1
17	SMA Swasta Kristen Kalam Kudus	1	
18	SMA Swasta Methodist	1	
19	SMA Swasta Surya		1
20	SMA Swasta Budi Mulia	1	
21	SMA Swasta Erlangga		1
22	SMA Swasta Kampus Hkbp Nomensen		1
23	SMA Swasta Advent		1
24	SMA Swasta Sultan Agung		1
25	SMA Swasta Mars		1
26	SMA Swasta Tri Sakti		1
27	SMA Negeri 5		1
28	SMA Negeri 6		1

The data above obtained the final result where in iteration 1 and iteration 2 the data grouping carried out on 2 clusters obtained the same results. The results of the second iteration are C1 = 7 and C2 = 21 in the data position of each cluster. So that the position of the cluster in the data does not change again, the iteration process stops until iteration 2.

3.2. Implementation of the K-Means Algorithm with RapidMiner

The implementation of school grouping is also carried out using RapidMiner tools. The processes and results that can be seen are as follows in Figure 2:



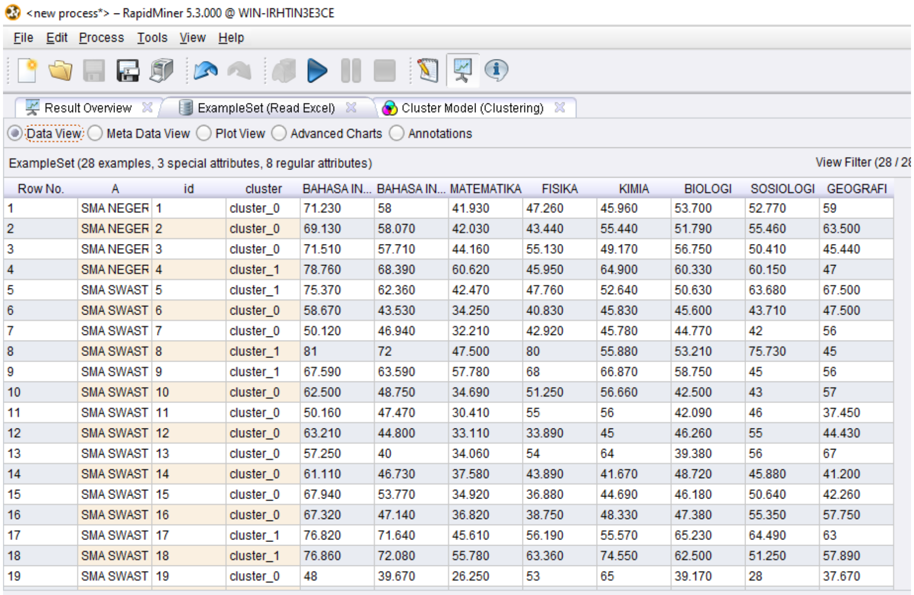
<new process*> – RapidMiner 5.3.000 @ WIN-IRHTIN3E3CE

Plotter
Scatter
x-Axis
cluster
□ Log Scale
y-Axis
cluster
□ Log Scale
Color Column
cluster
□ Log Scale
Jitter
☑ Rotate Labels

File Edit Process Tools View Help
Result Overview
ExampleSet (Read Excel)
Cluster Model (Clustering)
○ Data View ○ Meta Data View ● Plot View ○ Advanced Charts ○ Annotations
cluster ● cluster_0 ● cluster_1
cluster_1
cluster_0
cluster
cluster_0
cluster_1
cluster
Repositories
Samples (none)
DB
Local Repository (Asus)
Log
System Monitor

Figure 2: Results of The Grouping

Based on Figure 2 above, it can be seen that the low group has 7 nodes in red, while the high group has 21 nodes in blue. The results obtained from processing the K-Means Algorithm on RapidMiner are as follows:



<new process*> – RapidMiner 5.3.000 @ WIN-IRHTIN3E3CE
File Edit Process Tools View Help
Result Overview
ExampleSet (Read Excel)
Cluster Model (Clustering)
● Data View ○ Meta Data View ○ Plot View ○ Advanced Charts ○ Annotations
View Filter (28 / 28

ExampleSet (28 examples, 3 special attributes, 8 regular attributes)

Row No.	A	id	cluster	BAHASA IN...	BAHASA IN...	MATEMATIKA	FISIKA	KIMIA	BIOLOGI	SOSIOLOGI	GEOGRAFI
1	SMA NEGER	1	cluster_0	71.230	58	41.930	47.260	45.960	53.700	52.770	59
2	SMA NEGER	2	cluster_0	69.130	58.070	42.030	43.440	55.440	51.790	55.460	63.500
3	SMA NEGER	3	cluster_0	71.510	57.710	44.160	55.130	49.170	56.750	50.410	45.440
4	SMA NEGER	4	cluster_1	78.760	68.390	60.620	45.950	64.900	60.330	60.150	47
5	SMA SWAST	5	cluster_1	75.370	62.360	42.470	47.760	52.640	50.630	63.680	67.500
6	SMA SWAST	6	cluster_0	58.670	43.530	34.250	40.830	45.830	45.600	43.710	47.500
7	SMA SWAST	7	cluster_0	50.120	46.940	32.210	42.920	45.780	44.770	42	56
8	SMA SWAST	8	cluster_1	81	72	47.500	80	55.880	53.210	75.730	45
9	SMA SWAST	9	cluster_1	67.590	63.590	57.780	68	66.870	58.750	45	56
10	SMA SWAST	10	cluster_0	62.500	48.750	34.690	51.250	56.660	42.500	43	57
11	SMA SWAST	11	cluster_0	50.160	47.470	30.410	55	56	42.090	46	37.450
12	SMA SWAST	12	cluster_0	63.210	44.800	33.110	33.890	45	46.260	55	44.430
13	SMA SWAST	13	cluster_0	57.250	40	34.060	54	64	39.380	56	67
14	SMA SWAST	14	cluster_0	61.110	46.730	37.580	43.890	41.670	48.720	45.880	41.200
15	SMA SWAST	15	cluster_0	67.940	53.770	34.920	36.880	44.690	46.180	50.640	42.260
16	SMA SWAST	16	cluster_0	67.320	47.140	36.820	38.750	48.330	47.380	55.350	57.750
17	SMA SWAST	17	cluster_1	76.820	71.640	45.610	56.190	55.570	65.230	64.490	63
18	SMA SWAST	18	cluster_1	76.860	72.080	55.780	63.360	74.550	62.500	51.250	57.890
19	SMA SWAST	19	cluster_0	48	39.670	26.250	53	65	39.170	28	37.670

Figure 3: RapidMiner Processing Results

Explained that the results of the manual calculation of the k-means algorithm and Microsoft excel data have the same value, namely between several clusters, namely high 21 and low 7 clusters, and entering Microsoft excel calculations into rapidminer has the same value as well.

4. Conclusion

The data processed to obtain the results of the National High School Exam Score Achievement in Pematangsiantar applying the K-Means Clustering method can determine the centroid value in 2 clusters, namely the highest and lowest clusters. The highest cluster produces 21 schools and the low cluster produces 7 schools. These results are expected to be input for the Education Party to pay more attention to schools that have the lowest National Examination Scores.

References

[1] M. Brown, G. McNamara, and J. O'Hara, "Quality and the rise of value-added in education: The case of Ireland," *Policy Futur. Educ.*, vol. 14, no. 6, pp. 810–829, 2016, doi: 10.1177/1478210316656506.

[2] O. Little, L. Goe, and C. Bell, "A practical guide to evaluating teacher effectiveness. Washington, DC: National Comprehensive Center for Teacher Quality," *Natl. Compr. Cent. Teach. Qual.*, no. April, pp. 1–32, 2009, [Online]. Available: https://eric.ed.gov/?id=ED543776%0Ahttps://www.wested.org/wp-content/uploads/teacher-effectiveness-guide.pdf.

[3] G. Punj and D. W. Stewart, "Cluster Analysis in Marketing Research: Review and Suggestions for Application," *J. Mark. Res.*, vol. 20, no. 2, p. 134, 1983, doi: 10.2307/3151680.

[4] D. T. Utari and D. S. Hanun, "Hierarchical Clustering Approach for Region Analysis of Contraceptive Users," vol. 2, no. 2, pp. 99–108, 2021, doi: 10.20885/EKSAKTA.vol2.iss1.art.

[5] M. Z. Rodriguez *et al.*, *Clustering algorithms: A comparative approach*, vol. 14, no. 1. 2019.

[6] 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.

[7] Nurmalasari *et al.*, "Implementation of Clustering Algorithm Method for Customer Segmentation," *J. Comput. Theor. Nanosci.*, vol. 17, no. 2, pp. 1388–1395, 2020, doi: 10.1166/jctn.2020.8815.

[8] I. Sukmadewanti, R. Arifudin, and E. Sugiharti, "Use of K-Means Clustering and Analytical Methods Hierarchy Process in Determining the Type of MSME Financing in Semarang City," *Sci. J. Informatics*, vol. 5, no. 2, pp. 148–158, 2018, doi: 10.15294/sji.v5i2.16221.

[9] I. N. Rachmawati, "Pengumpulan Data Dalam Penelitian Kualitatif: Wawancara," *J. Keperawatan Indones.*, vol. 11, no. 1, pp. 35–40, 2007, doi: 10.7454/jki.v11i1.184.

[10] C. Devi, O. Soleman, N. Pramaita, and M. Sudarma, "Classification Of Loyality Customer Using K-Means Clustering, Studi Case : PT. Sucofindo (Persero) Denpasar Branch," *Int. J. Eng. Emerg. Technol.*, vol. 5, no. 2, 2020.

[11] A. Chaerudin, D. T. Murdiansyah, and M. Imrona, "Implementation of K-Means++ Algorithm for Store Customers Segmentation Using Neo4J," *Indones. J. Comput.*, vol. 6, no. 1, pp. 53–60, 2021, doi: 10.34818/indojc.2021.6.1.547.

[12] M. Algorithm, “JurnalMantik,” vol. 4, no. 3, pp. 1855–1867, 2020.

[13] B. Haviluddin, A. Fanany, and O. Gafar, “Proceedings of the Eleventh International Conference on Management Science and Engineering Management,” *Proc. Elev. Int. Conf. Manag. Sci. Eng. Manag.*, vol. 2, 2018, doi: 10.1007/978-3-319-59280-0.

[14] B. Supriyadi, A. P. Windarto, T. Soemartono, and Mungad, “Classification of natural disaster prone areas in Indonesia using K-means,” *Int. J. Grid Distrib. Comput.*, vol. 11, no. 8, pp. 87–98, 2018, doi: 10.14257/ijgdc.2018.11.8.08.