KnE Social Sciences
ICASI
International Conference on Advance & Scientific Innovation
Volume 2023

Research Article

Analysis of the K-Means Algorithm for Clustering School Participation Rates in Central Java

Jati Sumarah, Ajeng Tiara Wulandari

Politeknik Dharma Patria, Kebumen, Indonesia

Abstract.

One indication of the development of educational services in Indonesia is the School Enrollment Rate (SER). Higher the rate of enrolment, the better a location offers access to training. The dataset source was collected from the Central Java Statistical Agency website. The analysis object is the percentage of SERs for ages 7-12 years, 13-15 years, and 16-18 years in the Central Java region during 2017-2019. In the Central Java province, the aim of which is the third largest province after West Java and East Java, was to analyze the level of school participation as mapped. The created research product is a mapping of locations in the District and City areas in the form of clusters. The solution is the clustering algorithm k-means. In this study, there were two groups: high (C1) and low. The clusters were separated into (C2). Cluster-mapping studies results for the years 7-12 were, that in a high cluster, 24 provinces (cluster 0) and 11 provinces (cluster 1) were in a lower cluster, whereas the 13-15-year-old cluster mapping results from 23 provinces (cluster 0) and 12 provinces (cluster 1) and the 16-18-year-old cluster mapping results from 15 provinces. Final centroid value is the basis for the determination of the clusters where the final centroid value for a cluster aged 7-12 years were high (cluster 0) {99.81, 99.87, 99.75} and low (cluster 1) {99.73, 99.43, 99.25}, whereas the final centroid value of a cluster aged 13-15 years was high (cluster 0). For all age categories, the mapping findings reveal a good proportion, that is, over 50% in the top class. In particular, 24 provinces (57%) were in the low cluster of the 16-18-year age group. Research results information can provide a macro-image of the level of SER development in recent years.

Keywords: K-Means, algorithm, clustering

Corresponding Author: Jati Sumarah; email: jatisumarahdp@gmail.com

Published 26 May 2023

Publishing services provided by Knowledge E

Selection and Peer-review under the responsibility of the ICASI Conference Committee.

1. Introduction

Poverty is one of the problems that concern the government in any country. Poverty is the inability to meet the minimum standard of living [1]. Poverty is a picture of life in many developing countries, one of which is Indonesia. Poverty alleviation has even become one of the priority programs for local governments. The percentage of the number of poor people in Indonesia, especially on the island of Java is more than 50%, Central Java Province occupies the highest absolute poverty rate in Java. The problem of poverty is influenced by several factors, one of which is School Participation Rates (SPR). School Participation Rates (SPR) is the proportion of school children at the age

of a certain level of education in the age group in accordance with the educational level. School Participation Rates (SPR) are also a measure of the success of an area's education. This shows the level of prosperity of the area. Therefore, an increase in the number of SPR will be associated with a decrease in the poverty rate. The higher the SPR score, the area is considered successful in providing access to education services. A high SPR indicates greater opportunities for access to education in general. In the age group where this opportunity occurs, it can be seen from the amount of SPR in each age group [2]. Therefore, grouping the value of School Participation Rates is very important, as information and a government barometer specifically for local governments in each district and city in Central Java Province in determining related policies in the field of education.

The specific purpose of this research is to provide input and information for the Central Java Provincial government in order to maximize efforts and concerns to increase areas that have low School Participation Rates and maintain the value of School Participation Rates to remain stable for areas with low School Participation Rates. Participation Rates are already high. The research dataset is comprised of SPR data for Central Java Province from 2017 to 2019, including 29 regencies and six cities. The data were gathered from the Central Java Statistics Agency. The clustering method used in this study is the K-Means Clustering data mining algorithm. Because data mining is an algorithm that is widely used to deal with data classification problems [3]–[7], as well as data clustering [8]–[12].

In table 1 it is explained that the School Participation Rate (SPR) is one indicator of achieving development in the field of education in a region. Among the four SPR values in the 2017 SPR table, children aged 16-18 years are the smallest, this shows that there are still many residents who have not taken college-level education. However, for the ages of 7-12 years and 13-15 years, it has reached above 99 percent, this means that the population at that age is all attending school. Numerous past research on grouping using the K-Means algorithm exist, including the following: A study was conducted to categorize disaster-prone locations in Indonesia based on provinces. The results of this study are in the form of grouping data on disaster-prone areas which are divided into 3 clusters, specifically, the high cluster comprises four provinces, the standard cluster has fourteen provinces, and the low cluster comprises sixteen provinces [14]. The following research was undertaken to categorize population density, human development index, open unemployment rate, and school enrollment rate by Indonesian provinces. The

Table 1: School Participation Rates (SPR) Central Java Region, 2017-2019 (Percent).

No	Central Java Region	7-12 Years old			13-15 Years old			16-18 Years old		
		2019	2018	2017	2019	2018	2017	2019	2018	2017
1	Cilacap District	99,73	99,70	100,00	97,27	97,38	96,25	68,23	68,12	69,84
2	Banyumas District	99,71	99,43	99,71	99,40	95,32	95,26	62,51	61,49	67,07
3	Purbalingga District	99,84	99,57	98,81	93,77	94,24	93,48	59,96	59,45	60,97
4	Banjarnegara District	99,82	100,00	99,22	89,98	89,24	90,22	64,39	64,12	62,80
5	Kebumen District	99,75	99,67	99,41	98,32	98,29	98,60	79,63	79,17	85,01
6	Purworejo District	99,82	100,00	99,69	97,21	96,91	97,83	83,84	83,76	85,24
7	Wonosobo District	99,55	99,23	99,50	94,61	94,06	93,05	59,22	57,04	55,14
8	Magelang District	99,91	99,81	98,94	97,41	96,78	96,45	68,36	68,05	70,36
9	Boyolali District	100,00	100,00	99,73	95,77	95,08	95,34	73,26	69,73	66,69
10	Klaten District	99,85	100,00	99,61	99,77	99,18	98,82	79,92	77,09	81,23
11	Sukoharjo District	99,84	100,00	99,20	99,64	100,00	98,63	81,92	82,73	82,48
12	Wonogiri District	99,69	99,67	99,04	98,88	98,69	98,24	81,02	81,82	81,61
13	Karanganyar District	99,48	99,15	99,33	96,33	96,55	96,88	83,23	83,83	79,32
14	Sragen District	99,88	99,73	99,51	96,74	96,48	95,87	82,03	79,73	78,71
15	Grobogan District	100,00	100,00	100,00	96,00	95,34	96,32	59,48	59,76	56,50
16	Blora District	99,77	99,51	99,58	98,08	98,09	97,29	73,30	73,22	67,49
17	Rembang District	99,67	99,42	100,00	97,72	97,05	97,19	68,18	68,54	68,92
18	Pati District	99,84	100,00	99,74	95,54	95,29	95,98	72,53	69,85	63,29
19	Kudus District	99,81	100,00	100,00	97,50	97,98	96,36	73,20	73,91	70,47
20	Jepara District	100,00	99,78	99,87	96,05	95,62	94,64	68,47	68,26	66,33
21	Demak District	99,84	100,00	99,65	95,71	95,30	93,78	76,31	76,27	70,89
22	Semarang District	99,81	100,00	99,82	97,35	97,20	97,18	74,69	74,39	73,34
23	Temanggung District	99,83	99,25	98,97	96,91	96,68	96,41	73,54	70,09	61,18
24	Kendal District	99,64	99,44	100,00	94,44	93,79	93,99	69,74	70,68	62,81
25	Batang District	99,64	99,74	100,00	94,54	95,12	93,96	64,20	64,62	60,90
26	Pekalongan District	99,70	99,70	99,73	90,38	90,53	90,29	66,34	66,65	60,76
27	Pemalang District	99,78	100,00	99,70	92,40	92,32	91,17	59,14	59,83	62,28
28	Tegal District	99,55	99,46	99,47	94,29	93,61	93,52	65,49	65,34	60,68
29	Brebes District	99,44	100,00	100,00	94,64	94,61	94,49	50,17	49,56	53,72
30	Magelang City	99,86	100,00	99,32	95,00	94,67	95,21	91,39	89,58	90,74
31	Surakarta City	99,88	99,75	99,15	98,85	98,59	97,83	75,80	76,92	81,28
32	Salatiga City	99,81	98,62	99,56	98,63	98,49	98,78	85,68	84,34	86,86
33	Semarang City	99,88	100,00	99,71	97,65	97,54	97,33	72,87	70,72	76,12
34	Pekalongan City	99,89	100,00	99,73	95,87	95,23	95,97	64,98	61,32	66,08
35	Tegal City	99,94	100,00	100,00	95,58	95,46	94,48	78,43	78,40	70,06

Source: Central Java Provincial Statistics Agency [13].

study's findings indicate that cluster 1 contains 12 provinces, cluster 2 contains six

provinces, cluster 3 contains one province, cluster 4 contains six provinces, and cluster 5 contains nine provinces [15]. The next research was conducted to cluster the distribution of rabies cases in the city of Palembang using K-Means data mining. Data processing in this study using RapidMiner software with the result that from 16 sub-districts in Palembang, seven sub-districts are included in the very rabies-prone area cluster (C0), while four sub-districts are included in the rabies-prone area cluster (C1), and five sub-districts are included in the regional cluster. not susceptible to rabies (C2) [16]. These studies serve as a foundation for conducting research to classify school enrollment rates in regencies and cities in Central Java Province.

2. Methods

The steps taken to solve the problem in this research are arranged in the following framework:

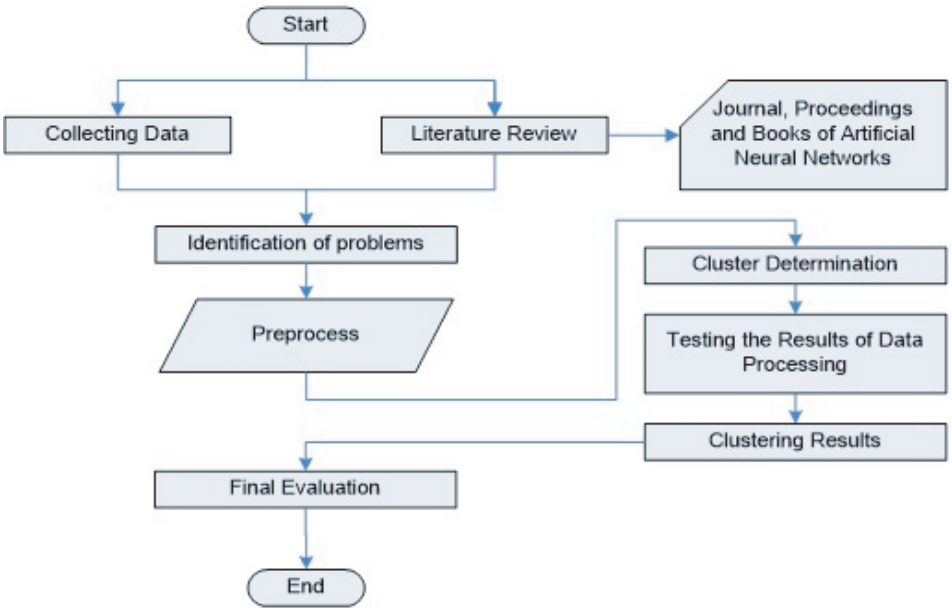

Figure 1: Research Stages.

Explanation:

Data collection in the form of School Participation Rates data in Central Java Province from 2017 to 2019 consisting of 29 Regencies and 6 Cities obtained from the Central Java Statistics Agency. Literature studies are carried out to complement the basic knowledge and theories used in research (scientific articles, books, proceedings, etc.). Problem identification is carried out after all the data is met and then the appropriate

dataset is obtained for the specified process to be carried out. Preprocessing is a stage that is carried out by making changes to several data types in the dataset attributes with the aim of facilitating understanding of the contents of the record, also making selections by paying attention to data consistency, missing values and redundant data. Cluster determination is a step to determine the number of clusters that are determined by using K-Means datamining. Testing the Results of Data Processing is a test phase of the results of data processing using RapidMiner Software after the cluster determination process is complete. Clustering results are grouping results obtained based on the number of clusters that have been determined. Final evaluation is carried out to find out whether testing the results of data processing is appropriate.

3. Results and Discussion

3.1. Cluster analysis of the SPR for 7-12 year olds

The following are the results of experiments conducted with the K-means algorithm on School Participation Rates (SPR) aged 7-12 with the number of clusters (k=2). Determine the initial center of the cluster, take the 13th data and the 15th data. Then the initial cluster:

TABLE 2: Centroid Data Iteration 1.

Atribut	2019	2018	2017
C1	99.48	99.15	99.33
C2	100	100	100

Iteration 1 : d(1,1) = $\sqrt{(99.73-99.48)^2+(99.7-99.15)^2+(100-99.33)^2}$ = 1, and so on until you get : d(35,2) = $\sqrt{(99.94-100)^2+(100-100)^2+(100-100)^2}$ = 0.06

TABLE 3: Iteration Results 1.

No	Central Java Region	C1	C2	Nearest Distance	Cluster
1	Cilacap District	1.00	0.36	0.36	C2
2	Banyumas District	0.70	0.45	0.45	C2
3	Purbalingga District	0.81	1.76	0.81	C1
...	...	...	...	...	...
33	Semarang City	1.27	0.20	0.20	C2
34	Pekalongan City	1.29	0.18	0.18	C2
35	Tegal City	1.63	0.06	0.06	C2

Table 4: Centroid Data Iteration 2.

Atribut	2019	2018	2017
C1	99.73	99.43	99.25
C2	99.81	99.87	99.75

Iteration 2: d(1,1) = $\sqrt{(99.73-99.73)^2+(99.7-99.43)^2+(100-99.25)^2}$ = 0.64 and so on until you get : d(35,2) = $\sqrt{(99.94-99.81)^2+(100-99.87)^2+(100-99.75)^2}$ = 0.21

Table 5: Iteration Results 2.

No	Central Java Region	C1	C2	Nearest Distance	Cluster
1	Cilacap District	0.64	0.17	0.17	C2
2	Banyumas District	0.30	0.23	0.23	C2
3	Purbalingga District	0.32	1.00	0.32	C1
...	...	...	...	...	...
33	Semarang City	0.69	0.09	0.09	C2
34	Pekalongan City	0.72	0.10	0.10	C2
35	Tegal City	1.10	0.21	0.21	C2

According to the study's findings, 24 provinces were classified as belonging to the high cluster (cluster 0), whereas 11 provinces were classified as belonging to the low cluster (cluster 1). The high clusters are Cilacap District, Banyumas District, Kab. Banjarnegara, Purworejo District, Boyolali District, Klaten District, Sukoharjo District, Sragen District, Grobogan District, Rembang District, Pati District, Kudus District, Jepara District, Demak District, Semarang District, Kendal District, Batang District, Pekalongan District, Pemalang District, Brebes District, Magelang City, Semarang City, Pekalongan City, Tegal City. While the low clusters are Purbalingga District, Kebumen District, Wonosobo District, Magelang District, Wonogiri District, Karanganyar District, Blora District, Temanggung District, Tegal District, Surakarta City, Salatiga City.

4. Cluster analysis of the SPR for 13-15 year olds

The following are the results of experiments conducted with the K-means algorithm on the School Enrollment Rate (SPR) aged 13-15 with the number of clusters (k=2)

Table 6: Centroid Data.

High Cluster	Low Cluster
99.42	89.81

Iteration 1: c(1,1) = $\sqrt{(96.97 - 99.42)^2}$ = 2.46, and so on until you get : c(35,2) = $\sqrt{(95.17 - 89.81)^2}$ = 5.36

TABLE 7: Iteration Results.

No	Central Java Region	Average	C1	C2	Nearest Distance	Data Clustering
1	Cilacap District	96.97	2.46	7.15	2.46	C1
2	Banyumas District	96.66	2.76	6.85	2.76	C1
3	Purbalingga District	93.83	5.59	4.02	4.02	C2
...	...	...	...	...	...	...
33	Semarang City	97.51	1.92	7.69	1.92	C1
34	Pekalongan City	95.69	3.73	5.88	3.73	C1
35	Tegal City	95.17	4.25	5.36	4.25	C1

TABLE 8: Centroid Data Iteration 3.

High Cluster	Low Cluster
97.11	93.35

Iteration 3 : c(1,1) = $\sqrt{(96.97 - 97.11)^2}$ = 0.14 and so on until you get : c(35,2) = $\sqrt{(95.17 - 93.35)^2}$ = 1.83

TABLE 9: 3^{rd} Iteration Results.

No	Central Java Region	Average	C1	C2	Nearest Distance	Data Clustering
1	Cilacap District	96.97	0.14	3.62	0.14	C1
2	Banyumas District	96.66	0.45	3.31	0.45	C1
3	Purbalingga District	93.83	3.28	0.48	0.48	C2
...	...	...	...	...	...	...
33	Semarang City	97.51	0.40	4.16	0.40	C1
34	Pekalongan City	95.69	1.42	2.34	1.42	C1
35	Tegal City	95.17	1.93	1.83	1.83	C2

TABLE 10: Centroid Data Iteration 4.

High Cluster	Low Cluster
97.19	93.50

Iteration 4 : c(1,1) = $\sqrt{(96.97 - 97.19)^2}$ = 0.22 and so on until you get : c(35,2) = $\sqrt{(95.17 - 93.50)^2}$ = 1.68

According to the study's findings, 23 provinces were classified as belonging to the high cluster (cluster 0), whereas 12 provinces were classified as belonging to the low

Table 11: 4^{th} Iteration Results.

No	Central Java Region	Average	C1	C2	Nearest Distance	Data Clustering
1	Cilacap District	96.97	0.22	3.47	0.22	C1
2	Banyumas District	96.66	0.53	3.16	0.53	C1
3	Purbalingga District	93.83	3.36	0.33	0.33	C2
...	...	...	...	...	...	...
33	Semarang City	97.51	0.32	4.01	0.32	C1
34	Pekalongan City	95.69	1.50	2.19	1.50	C1
35	Tegal City	95.17	2.02	1.68	1.68	C2

cluster (cluster 1). The high clusters are Cilacap District, Banyumas District, Kebumen District, Purworejo District, Magelang District, Boyolali District, Klaten District, Sukoharjo District, Wonogiri District, Kab. Karanganyar, Sragen District, Grobogan District, Blora District, Rembang District, Pati District, Kudus District, Jepara District, Semarang District, Temanggung District, Surakarta City, Salatiga City, Semarang City, Pekalongan City. While the low clusters are Purbalingga District, Banjarnegara District, Wonosobo District, Demak District, Kendal District, Batang District, Pekalongan District, Pemalang District, Tegal District, Brebes District, Magelang City, Tegal City.

4.1. Cluster analysis of the SPR for 16-18 year olds

The following are the results of experiments conducted with K-means on the School Participation Rate (SPR) aged 16-18 with the number of clusters (k=2). Determine the initial center of the cluster, take the 6th data and 15th data.

Table 12: Centroid Data Iteration 1.

Atribut	2019	2018	2017
C1	83.84	83.76	85.24
C2	59.48	59.76	56.5

Iteration 1: c(1,1) = $\sqrt{(68.23-83.84)^2+(68.12-83.76)^2+(69.84-85.24)^2}$ = 497.38 and so on until you get : c(35,2) = $\sqrt{(78.43-59.48)^2+(78.4-59.76)^2+(70.06-56.5)^2}$ = 550.27

Iteration 2: c(1,1) = $\sqrt{(68.23-80.00)^2+(68.12-79.51)^2+(69.84-79.56)^2}$ = 235.94 and so on until you get : c(35,2) = $\sqrt{(74.83-65.57)^2+(78.4-64.79)^2+(70.06-63.19)^2}$ = 245.39

Table 13: 1^{st} Iteration Results.

No	Central Java Region	C1	C2	Nearest Distance	Data Clustering
1	Cilacap District	497.38	256.60	256.60	C2
2	Banyumas District	847.43	117.75	117.75	C2
3	Purbalingga District	1203.89	20.56	20.56	C2
...	...	...	...	...	...
33	Semarang City	264.19	518.46	264.19	C1
34	Pekalongan City	889.52	99.71	99.71	C2
35	Tegal City	264.57	550.27	264.57	C1

Table 14: Centroid Data Iteration 2.

Atribut	2019	2018	2017
C1	80.00	79.51	79.56
C2	65.57	64.79	63.19

Table 15: 2^{nd} Iteration Results.

No	Central Java Region	C1	C2	Nearest Distance	Data Clustering
1	Cilacap District	235.94	57.99	57.99	C2
2	Banyumas District	498.17	28.98	28.98	C2
3	Purbalingga District	767.96	39.02	39.02	C2
...	...	...	...	...	...
33	Semarang City	96.22	209.68	96.22	C1
34	Pekalongan City	527.56	20.96	20.96	C2
35	Tegal City	93.00	245.39	93.00	C1

The results indicated that 15 provinces were classified as belonging to the high cluster (cluster 0), whereas 20 provinces were classified as belonging to the low cluster (cluster 1). The high clusters are Kebumen District, Purworejo District, Klaten District, Sukoharjo District, Wonogiri District, Kab. Karanganyar, Sragen District, Kudus District, Demak District, Semarang District, Magelang City, Surakarta City, Salatiga City, Semarang City, Tegal City. While the low clusters are Cilacap District, Banyumas District, Purbalingga District, Banjarnegara District, Wonosobo District, Magelang District, Boyolali District, Grobogan District, Blora District, Rembang District, Pati District, Jepara District, Temanggung District, Kendal District, Batang District, Pekalongan District, Pemalang District, Tegal District, Brebes District, Pekalongan City.

5. Conclusions

The results of mapping in the form of clusters on the use of the K-Means algorithm on the School Participation Rate in Central Java can be implemented. Overall mapping results show a good percentage for all age groups, namely above 50 percent in the high cluster. While in detail in the 16-18 year age group, there are 24 provinces (57%) which are in the low cluster. Information from the results of the study can provide a macro picture of the level of development of the School Enrollment Rate (SPR) over the last few years.

Acknowledgements

Our thanks go to Direktorat Riset dan Pengabdian Masyarakat, Deputi Bidang Penguatan Riset and Pengembangan, Kementerian Riset and Teknologi/ Badan Riset and Inovasi Nasional (Ristek/BRIN) who has provided novice lecturer research grants (PDP) in 2020 for 2021 funding. Furthermore, thank you to the chairman and staff of the Dharma Patria Polytechnic LPPM who have facilitated PDP activities starting from the preparation of proposals to research reports

References

[1] Bednar DJ, Reames TG. Recognition of and response to energy poverty in the United States. *Nat Energy*. 2020;5(6):432–439.

[2] Sandra H, Majid SA, Dawood TC, Hamid A. What causes children to work in Indonesia? J Asian Fin Econ Bus. 2020;7(11):585–593.

[3] Damanik IS, Windarto AP, Wanto A, Poningsih SR, Andani SR, Andani SW. "Decision tree optimization in C4.5 algorithm using genetic algorithm." *J Phys Conf Ser*. 2019 Aug;1255(1):1–6.

[4] Katrina W, Damanik HJ, Parhusip F, Hartama D, Windarto AP, Wanto A. C.45 classification rules model for determining students level of understanding of the subject. *J Phys Conf Ser*. 2019;1255(1):1–7.

[5] Siahaan H, Mawengkang H, Efendi S, Wanto A, Perdana Windarto A. Application of classification method C4.5 on selection of exemplary teachers. *J Phys Conf Ser*. 2019;1235(1):012005.

[6] Parlina I, Yusuf Arnol M, Febriati NA, Dewi R, Wanto A, Lubis MR, et al. Naive Bayes algorithm analysis to determine the percentage level of visitors the most dominant zoo visit by age category. *J Phys Conf Ser*. 2019;1255(1):1–5.

[7] Hartama D, Perdana Windarto A, Wanto A. The application of data mining in determining patterns of interest of high school graduates. *J Phys Conf Ser*. 2019;1339(1):1–6.

[8] Hanafiah MA, Wanto A, Indonesia PB. Implementation of data mining algorithms for grouping poverty lines by district/city in North Sumatra. Int J Inf Sys Technol. 2020;3(2):315–322.

[9] Febriyati NA, GS AD, Wanto A. "GRDP growth rate clustering in Surabaya City uses the K- Means Algorithm." Int J Inf Sys Technol. 2020;3(2):276–283.

[10] Sudirman S, Windarto AP, Wanto A. Data mining tools | RapidMiner : K-Means method on clustering of rice crops by province as efforts to stabilize food crops in Indonesia. IOP Conf Series Mater Sci Eng. 2018;420(012089):1–8.

[11] Abbas SA, Aslam A, Rehman AU, Abbasi WA, Arif S, Kazmi SZ. K-Means and K-Medoids: Cluster analysis on birth data collected in City Muzaffarabad, Kashmir. IEEE Access. 2020;8:151847–151855.

[12] Hutagalung J, Ginantra NL, Bhawika GW, Parwita WG, Wanto A, Panjaitan PD. COVID-19 cases and deaths in Southeast Asia clustering using K-Means algorithm. *J Phys Conf Ser*. 2021;1783(1):012027.

[13] BPS. "Angka Partisipasi Sekolah (APS) (Persen), 2017-2019," Badan Pusat Statistik Provinsi Jawa Tengah, 2020. [Online]. Available: https://jateng.bps.go.id/indicator/28/71/1/angka-partisipasi-sekolah-aps-.html. [Accessed: 25-Oct-2020].

[14] Supriyadi B, Windarto AP, Soemartono T, Mungad. Classification of natural disaster prone areas in Indonesia using K-Means. *Int J Grid Distrib Comput*. 2018;11(8):87–98.

[15] Ahmar AS, Napitupulu D, Rahim R, Hidayat R, Sonatha Y, Azmi M. Using K-Means clustering to cluster provinces in Indonesia. *J Phys Conf Ser*. 2018;1028(1):1–6.

[16] Rahayu K, Novianti L, Kusnandar M. Implementation data mining with K-Means algorithm for clustering distribution rabies case area in Palembang City. *J Phys Conf Ser*. 2020;1500(1):1–9.