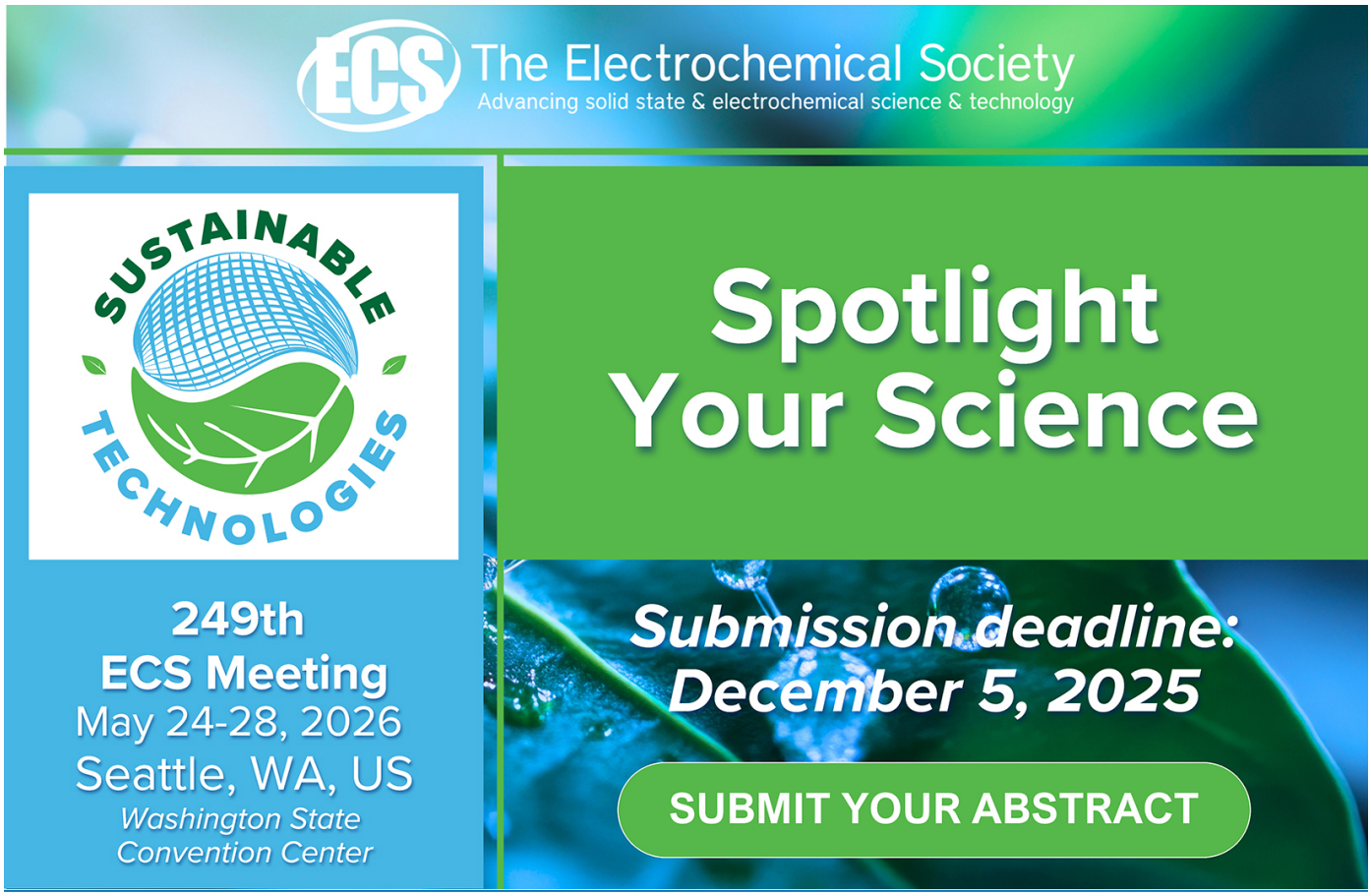
PAPER • **OPEN ACCESS**

Analysis of Data Mining in the Group of Water Pollution Areas using the K-Means Method in Indonesia

To cite this article: Hendra Jatnika *et al* 2021 *J. Phys.: Conf. Ser.* **1783** 012014



Analysis of Data Mining in the Group of Water Pollution Areas using the K-Means Method in Indonesia

Hendra Jatnika[1], Miftahul Huda[2], Ria Rizki Amelia[3], Melda Agnes Manuhutu[4], Agus Perdana Windarto[5*], Pipin Sumantrie[6], Ari Waluyo[7]

[1]Institut Teknologi PLN, Indonesia
[2]Sekolah Tinggi Ilmu Ekonomi Putra Bangsa, Kebumen, Indonesia
[3]AMIK PGRI Kebumen, Indonesia
[4]Universitas Victory Sorong, Indonesia
[5]STIKOM Tunas Bangsa, Pematangsiantar, Indonesia
[6]Akademi Keperawatan Surya Nusantara, Pematangsiantar, Indonesia
[7]Politeknik Dharma Patria, Kebumen, Indonesia

Email: *agus.perdana@amiktunasbangsa.ac.id

Abstract. Water is an essential requirement in human life. However, pollution causes the water quality to become poor, making it unsuitable for use. Pollution comes from rubbish and waste dumped into rivers, lakes and other water areas. This study aims to carry out a model for mapping areas contaminated with water pollution using artificial intelligence techniques. The data sample comes from the Indonesian Central Statistics Agency (abbreviated as BPS) which consists of 34 records. The data used are provinces in Indonesia that are polluted by water pollution in rural areas. The intelligence technique used is data mining using the k-means method. The variable used is the number of polluted villages by province. The mapping label used is the high cluster (K1) for water pollution and the low cluster (K2) for water pollution. Analysis using Rapid Miner software. The results showed that 4 provinces were included in the high cluster (K1) category, namely North Sumatra, West Java, Central Java, East Java. Testing of cluster results was carried out using Davies Bouldin (k = 2) with a value of 0.328, which means that the cluster results created were optimal. The results of the analysis are expected to be input for the government in focusing on areas polluted with water pollution.

1. Introduction

Water is one of the most important necessities of life. Without water, processes cannot take place. Although water is a natural resource that can be renewed by nature itself, the fact shows that the availability of water has not increased [1]. The need for water sources for some needs must have high purity, free from contamination of microorganisms and chemicals [2]–[4]. In Indonesia, access to clean water sources is still a problem. This is due to development accompanied by a very rapid population growth rate. Not only in cities but also in villages. The very fast population development certainly requires a source of clean water as a daily necessity [5]. Meanwhile, this development is not accompanied by sufficient education and public awareness of environmental preservation[6] which is the beginning of pollution. Pollution is a problem in the environment that can damage the sustainability of the natural environment. One of the pollution that occurs is water pollution [7]. Water pollution comes from garbage, detergent remnants, household factory waste, to large factory waste that is discharged

into sewers, rivers, lakes and the sea. The impact of pollution is decreasing water quality and the emergence of sources of disease [8]. This is a big problem because it relates to the survival of creatures. There is a lot of water pollution that occurs in Indonesia. If this problem is not resolved immediately, it will have an impact on survival in the future. The purpose of this research is to analyze areas in Indonesia by mapping the regions in Indonesia, especially rural areas.

In computer science [9]–[15], the mapping process can be done using data mining techniques [14], [16]–[19]. Data mining is a branch of artificial intelligence that can extract data to produce information and knowledge [20], [21]. Some of the well-known techniques in data mining are association, estimation, prediction, clustering and classification [17], [22]. Klastering is one technique that can be used to do leveling [23]. Popular methods are k-means and k-medoids [24], [25]. The k-means method is a simple and easy to apply method [26]. In addition, the k-means method is very flexible and adaptable [16]. Many previous studies have used k-means as a solution. Among them are Jamal (2019) [21] regarding mapping of customer loyalty. In this research, the k-means method can be applied with 5 levels of loyalty, namely Very Loyal, Fairly Loyal, Ordinary, Less Loyal, Disloyal. To test the validity used the Davies-Bouldin Index. The DBI value generated from customer clustering is 0.79074. From the DBI value, it can be concluded that the quality of the resulting clusters has a fairly good quality. It is hoped that the research results will become input for the government so that it is wiser in managing clean water sources and preventing water pollution and helping to increase public awareness of the importance of preventing water pollution, especially in rural areas so that clean water needs can be met.

2. Methodology

In the research the data source was obtained from the Central Bureau of Statistics (abbreviated as BPS) which consisted of 34 records where the data were areas that were polluted by water pollution specifically for rural areas. The variable used is the number of villages polluted by water pollution by province. Data processing uses the help of RapiMiner software. The following is the data on the number of villages polluted by water pollution as shown in the following table:

Table 1. data on the number of villages polluted by water pollution

No	Province	Polluted Village	No	Province	Polluted Village
1	Aceh	729	18	West Nusa Tenggara	282
2	North Sumatra	1205	19	East Nusa Tenggara	122
3	West Sumatra	319	20	West Kalimantan	915
4	Riau	454	21	Central Kalimantan	782
5	Jambi	614	22	South Borneo	714
6	South Sumatra	673	23	East Kalimantan	318
7	Bengkulu	286	24	North Kalimantan	139
8	Lampung	572	25	North Sulawesi	327
9	Kep. Bangka Belitung	159	26	Central Sulawesi	303
10	Kep. Riau	55	27	South Sulawesi	400
11	DKI Jakarta	126	28	Southeast Sulawesi	227
12	West Java	1890	29	Gorontalo	111
13	Central Java	1900	30	West Sulawesi	115
14	DI Yogyakarta	99	31	Maluku	105
15	East Java	1643	32	North Maluku	216
16	Banten	513	33	West Papua	155
17	Bali	130	34	Papua	249

source: BPS processed data

The following is a flowchat of the k-means method in mapping areas contaminated with water pollution by area as shown in the following figure:

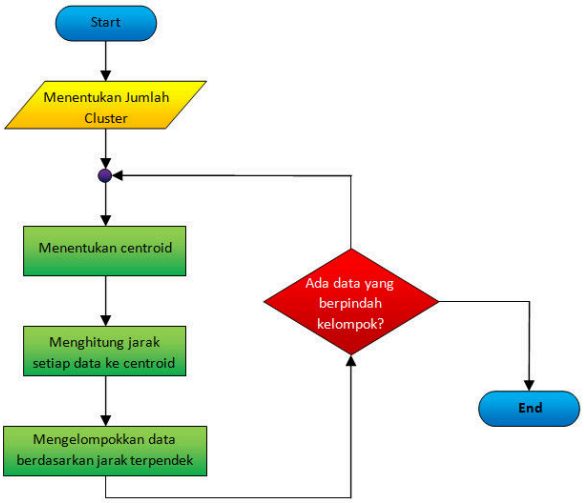


Figure 1. Flowchart of k-means method

In Figure 1, the steps for the k-means settlement process are as follows:

a) Prepare the data to be grouped.
b) Specifies the number of data clusters.
c) Determine the centroid as the center point of each cluster. In the 1st iteration, the centroid is randomly selected. Whereas in the second iteration and so on until the last iteration, the centroid is determined based on the mean of each group.
d) Calculating the distance of each data to the centroid is done with the Euclidean Distance equation as follows.

$$d_{ik} = \sqrt{\sum_j^m (x_{ij} - c_{kj})^2} \quad (1)$$

e) Determine the group based on the shortest distance to the centroid of each cluster.
f) Determine the average for each group.
g) Repeat the process from the third step until the data group in the last iteration results is the same as the data set in the previous iteration.
h) Done

3. Results and Discussion

The use of the k-means method in mapping water-polluted areas with the help of RapidMiner software, uses 2 mapping labels, namely C1: high cluster label in water-polluted areas and C2: low cluster label in water-polluted areas. The following is the k-means model and the results of mapping using the RapidMiner software:

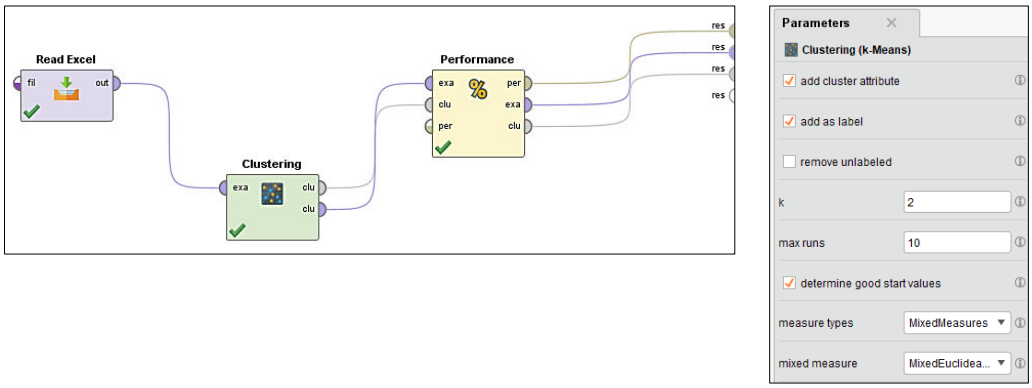


Figure 2. The k-means model in RapidMiner

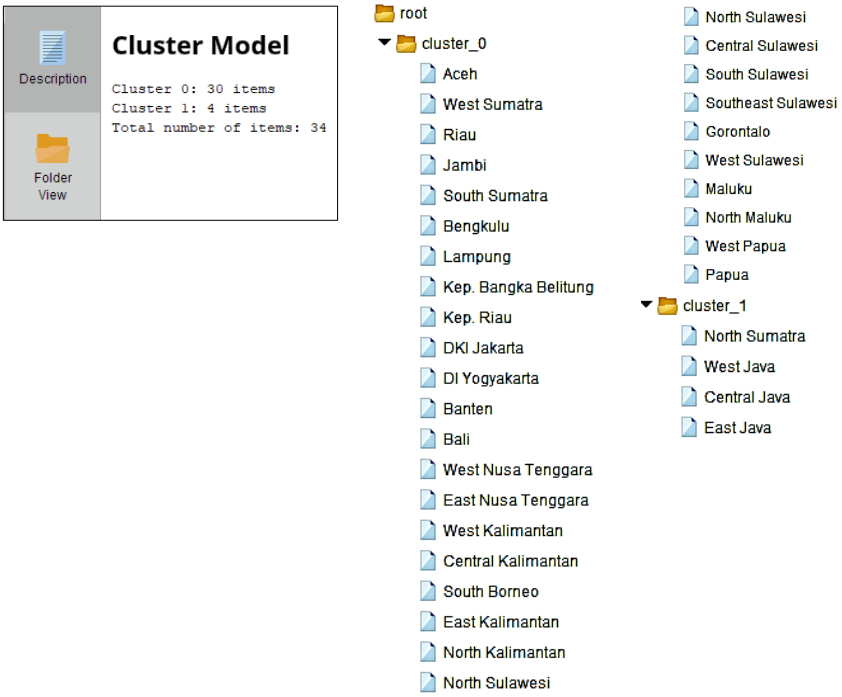


Figure 3. Cluster Results of the k-means method

In Figure 4, it can be explained that the results of the mapping of areas in Indonesia that are polluted by water pollution using the help of the RapidMiner software for the k-means analysis process. From the mapping results, there are 4 provinces in the high cluster (cluster_1) in the number of areas polluted with water pollution and 30 provinces in the low cluster (cluster_0) in the number of areas polluted with water pollution. From the mapping results, the final centroid value is obtained as shown in Table 2 below:

Table 2. Final centroid value

Attribute	C1: low cluster	C2: high cluster
Number of polluted villages	340,3	1659,5

Figure 4. Cluster results by region (province)

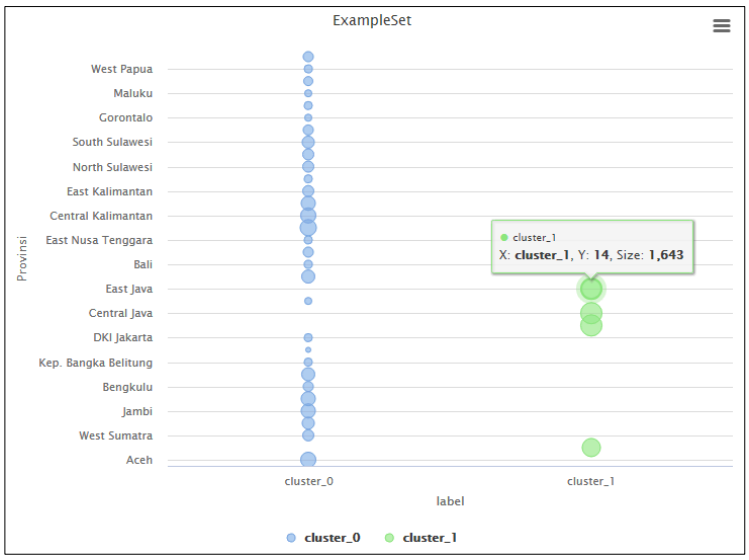


Figure 5. Cluster results by region (province)

In Figure 6, graph visualization can be done using the help of Rapid Miner software where the graph is displayed based on the results of the mapping, both in graphic visualization and scatter graphs.

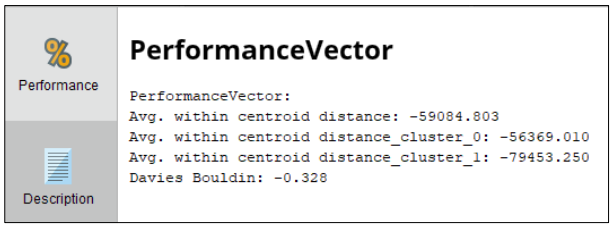


Figure 6. K-means performance test results

In Figure 7, the validity test is carried out using the Davies-Bouldin Index (DBI). DBI is a tool for measuring the value of the cluster results. Is the result of the optimal cluster formed or not. A cluster result is optimal if the DBI value is getting smaller. The DBI value generated from the water pollution cluster by province (k=2) is 0.328. From the DBI value, it can be concluded that the quality of the resulting clusters is of fairly good quality.

4. Conclusion

The research result states that the k-means method can be applied to the mapping of areas polluted with water pollution based on provinces in Indonesia. By using two mapping labels, it was found that four provinces were in the high cluster (K1), namely North Sumatra, West Java, Central Java, East Java. Meanwhile, thirty provinces are in the lower cluster (K2). To test the validity used the Davies-Bouldin Index (DBI). The resulting DBI value is 0.328. From the DBI value, it can be concluded that the resulting cluster quality is of good quality.

References

[1] C. A. Varotsos, V. F. Krapivin, and F. A. Mkrtchyan, "A Novel Approach to Monitoring the Quality of Lakes Water by Optical and Modeling Tools : Lake Sevan as a Case Study," in *Water Air Soil Pollut*, Water, Air, & Soil Pollution, 2020, pp. 1–15.

[2] A. A. Ghezelsofloo and R. V. Ardalan, "iMedPub Journals Assessment of Pollution Potential and Contaminants of Mashhad Plain The Real Case Study," *J. Water Pollut. Control*, vol. 2, no. 19, pp. 1–8, 2019.

[3] R. Karolina and Y. G. C. Sianipar, "The utilization of stone ash on cellular lightweight concrete,"

in *IOP Conference Series: Materials Science and Engineering*, 2018, vol. 309, no. 1.
[4] R. Karolina and A. L. A. Putra, "The effect of steel slag as a coarse aggregate and Sinabung volcanic ash a filler on high strength concrete," in *IOP Conference Series: Materials Science and Engineering*, 2018, vol. 309, no. 1.
[5] D. E. Puspitasari, "Dampak Pencemaran Air Terhadap Kesehatan Lingkungan dalam Perspektif Hukum Lingkungan," *Mimb. Huk.*, vol. 21, pp. 23–34, 2009.
[6] M. A. P. Handana, R. Karolina, and Steven, "Performance evaluation of existing building structure with pushover analysis," in *IOP Conference Series: Materials Science and Engineering*, 2018, vol. 309, no. 1.
[7] D. Agustiningsih and S. B. Sasongko, "Analisis Kualitas Air Dan Strategi Pengendalian Pencemaran Air Sungai Blukar Kabupaten Kendal," *Anal. Kualitas Air Dan Strateg. Pengendali. Pencemaran Air Sungai Blukar Kabupaten Kendal*, vol. 9, no. 2, pp. 64–71, 2012.
[8] A. Fransisca, "Tingkat Pencemaran Perairan Ditinjau Dari Pemanfaatan Ruang di Wilayah Pesisir Kota Cilegon," *J. Reg. City Plan.*, vol. 22, no. 2, p. 145, 2011.
[9] T. Imandasari, M. G. Sadewo, A. P. Windarto, A. Wanto, H. O. Lingga Wijaya, and R. Kurniawan, "Analysis of the Selection Factor of Online Transportation in the VIKOR Method in Pematangsiantar City," *J. Phys. Conf. Ser.*, vol. 1255, no. 012008, pp. 1–7, 2019.
[10] N. Nasution *et al.*, "Application of ELECTRE Algorithm in Skincare Product Selection," *J. Phys. Conf. Ser.*, vol. 1471, no. 1, 2020.
[11] S. R. Ningsih, R. Wulansari, D. Hartama, A. P. Windarto, and A. Wanto, "Analysis of PROMETHEE II Method on Selection of Lecturer Community Service Grant Proposals," *J. Phys. Conf. Ser.*, vol. 1255, no. 1, pp. 1–7, 2019.
[12] I. G. I. Sudipa, C. Astria, K. F. Irnanda, and A. Perdana, "Application of MCDM using PROMETHEE II Technique in the Case of Social Media Selection for Online Businesses . Application of MCDM using PROMETHEE II Technique in the Case of Social Media Selection for Online Businesses .," 2020.
[13] H. Pratiwi *et al.*, "Sigmoid Activation Function in Selecting the Best Model of Artificial Neural Networks," *J. Phys. Conf. Ser.*, vol. 1471, no. 1, 2020.
[14] F. Rahman, I. I. Ridho, M. Muflih, S. Pratama, M. R. Raharjo, and A. P. Windarto, "Application of Data Mining Technique using K-Medoids in the case of Export of Crude Petroleum Materials to the Destination Country Application of Data Mining Technique using K-Medoids in the case of Export of Crude Petroleum Materials to the Destination C," 2020.
[15] K. F. Irnanda, F. N. Arifah, M. R. Raharjo, A. Arifin, and A. P. Windarto, "The selection of Calcium Milk Products that are appropriate for advanced age using PROMETHEE II Algorithm," *J. Phys. Conf. Ser.*, vol. 1381, no. 1, 2019.
[16] A. P. Windarto *et al.*, "Analysis of the K-Means Algorithm on Clean Water Customers Based on the Province," *J. Phys. Conf. Ser.*, vol. 1255, no. 1, 2019.
[17] W. Katrina, H. J. Damanik, F. Parhusip, D. Hartama, A. P. Windarto, and A. Wanto, "C.45 Classification Rules Model for Determining Students Level of Understanding of the Subject," *J. Phys. Conf. Ser.*, vol. 1255, no. 012005, pp. 1–7, 2019.
[18] Sudirman, A. P. Windarto, and A. Wanto, "Data mining tools | rapidminer: K-means method on clustering of rice crops by province as efforts to stabilize food crops in Indonesia," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 420, p. 012089, 2018.
[19] D. Hartama, A. Perdana Windarto, and A. Wanto, "The Application of Data Mining in Determining Patterns of Interest of High School Graduates," *J. Phys. Conf. Ser.*, vol. 1339, no. 1, 2019.
[20] 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.
[21] M. B. Rahmat, A. Z. Arfianto, E. Setijadi, A. Mauludiyanto, and V. Y. P. Ardhana, "Measurement and Simulation of Microwave Absorber from Burned Rice Husk," 2019, pp. 99–102.
[22] S. Sriyanto, A. Buchori, A. Handayani, P. T. Nguyen, and H. Usman, "Implementation multi

factor evaluation process (MFEP) decision support system for choosing the best elementary school teacher," *Int. J. Control Autom.*, 2020.

[23] S. Harikumar and P. V. Surya, "K-Medoid Clustering for Heterogeneous DataSets," *Procedia Comput. Sci.*, vol. 70, pp. 226–237, 2015.

[24] I. Kamila, U. Khairunnisa, and Mustakim, "Perbandingan Algoritma K-Means dan K-Medoids untuk Pengelompokan Data Transaksi Bongkar Muat di Provinsi Riau," *J. Ilm. Rekayasa dan Manaj. Sist. Inf.*, vol. 5, no. 1, pp. 119–125, 2019.

[25] D. Marlina, N. Lina, A. Fernando, and A. Ramadhan, "Implementasi Algoritma K-Medoids dan K-Means untuk Pengelompokkan Wilayah Sebaran Cacat pada Anak," *J. CoreIT J. Has. Penelit. Ilmu Komput. dan Teknol. Inf.*, vol. 4, no. 2, p. 64, 2018.

[26] P. Arora, Deepali, and S. Varshney, "Analysis of K-Means and K-Medoids Algorithm for Big Data," *Phys. Procedia*, vol. 78, no. December 2015, pp. 507–512, 2016.