JURNAL SISFOTENIKA
Vol. 12 No. 2, Juli 2022
p-ISSN: 2087-7897 ; e-ISSN : 2460-5344
DOI : 10.30700/jst.v12i2.1241

Pengelompokkan Kebutuhan Air Bersih Di Indonesia Periode 2012-2017 Menggunakan Algoritma K-Means

Clustering of Clean Water Needs in Indonesia for the 2012-2017 Period Using the K-Means Algorithm

Daniati Uki Eka Saputri*[1], Taopik Hidayat[2], Siti Masturoh[3]
[1,2,3]Universitas Nusa Mandiri; Jl. Raya Jatiwaringin No.2, Cipinang Melayu, Makasar, Jakarta Timur, Jakarta 13620
[1,3]Information System, Faculty of Information Technology Universitas Nusa Mandiri
[2]Data Science, Faculty of Information Technology Universitas Nusa Mandiri
e-mail: *1daniati.due@nusamandiri.ac.id, 2taopik.toi@nusamandiri.ac.id, 3siti.uro@nusamandiri.ac.id

Abstrak

Kebutuhan air bersih penting untuk menunjang segala aktivitas kelangsungan hidup manusia. Data Badan Pusat Statistik (BPS) pada Tahun 2017 menunjukkan angka tertinggi penyaluran air bersih disetiap provinsi hanya sebesar 72,04%. Data tersebut menunjukkan bahwa akses air bersih untuk memenuhi kebutuhan harian masih jauh dari cukup. Penelitian ini bertujuan untuk mengelompokkan kebutuhan air bersih periode Tahun 2012-2017 menggunakan algoritma K-Means. Sumber data diperoleh dari situs resmi BPS yaitu data volume air bersih yang disalurkan kesetiap provinsi di Indonesia Tahun 2012-2017. Dilakukan proses replace missing value terhadap data yang hilang kemudian data dikelompokkan menjadi tiga cluster yaitu rendah (C0) sebanyak 25 provinsi, tinggi (C1) sebanyak 4 provinsi, dan sedang (C2) sebanyak 5 provinsi menggunakan algoritma K-Means. Nilai centroid untuk cluster C0 sebesar 150588.24, data centroid untuk cluster C1 sebesar 1939461, data centroid untuk cluster C2 sebesar 857876,6. Hasil clustering K-Means diuji menggunakan Davies Bouldin Index (DBI) Validation sebanyak 3 cluster dengan nilai 0,534, hasil cluster tersebut optimal karena nilai DBI sudah mendekati angka 0.

Kata kunci*— Pengelompokan, Air Bersih, Metode K-Means, Rapidminer, Davies Bouldin Index*

Abstract

The need for clean water is important to support all activities of human survival. Data from the Central Statistics Agency (BPS) in 2017 showed the highest number of clean water distribution in each province was only 72.04%. These data indicate that access to clean water to meet daily needs is still far from sufficient. This study aims to classify the need for clean water for the period 2012-2017 using the K-Means algorithm. The data source was obtained from the official BPS website, namely data on the volume of clean water distributed to each province in Indonesia in 2012-2017. The process of replacing missing values was carried out on the missing data, then the data were grouped into three clusters, namely low (C0) in 25 provinces, high (C1) in 4 provinces, and moderate (C2) in 5 provinces using the K-Means algorithm. The centroid value for the C0 cluster is 150588.24, the centroid data for the C1 cluster is 1939461, the centroid data for the C2 cluster is 857876.6. The results of the K-Means *clustering were tested using the Davies Bouldin Index (DBI) Validation as many as 3 clusters with a value of 0.534, the cluster results were optimal because the DBI value was close to 0.*

Keywords— *Clustering, Clean Water, K-Means Method, Rapidminer, Davies Bouldin Index*

1. PENDAHULUAN

The need for clean water is very important need for humans in supporting survival, because all human activities in various aspects of life require clean water such as toilets (bathing, washing, and defecating) even for consumption. According to BPS data that the need for proper clean water is one of the basic problems in Indonesia and at the district level, average percentage of the population with access to clean water has only reached 49% with a distribution range of 1% to 100% [14]. Other data explains that there is an increase in households that have access to clean water in Indonesia, there is no province that has access to clean water sources up to 100%. In 2017 it was only 72.04% of data from the National Socio-Economic Survey [4]. This shows that access to clean water to meet daily needs is still far from sufficient.

The fulfillment of the need for clean water allows the community to live healthily. Most of the people of Indonesia so far meet their needs for clean water from well water or from water provided by the Regional Drinking Water Company (PDAM). PDAM is a clean water provider company that is spread in every province, district, and municipality in Indonesia.

Clean water statistics for the period 2012-2017 in Indonesia show that the volume of clean water distributed by clean water companies managed by the government has increased by an average of 7.42% per year, from 2710 million m3 increasing to 3609 million m3 in 2017 [9]. Based on these data, it is a challenge for the central and regional governments as well as private companies to ensure that the amount of distribution of clean water needs in each province is balanced, there is even no shortage of water distribution from the company every year, besides that it is a consideration for the central and regional governments as well as the private sector. in the implementation and determination of policies.

Research by [17] made the grouping of water company customers against BPS data for 2010-2015 using the K-Means algorithm, the study resulted in 3 clusters, namely low, medium and high, the study was carried out manually in excel and rapidminer had validation values the same one. A similar study by Pratama [13] regarding the grouping of provinces in Indonesia that need to obtain access to clean drinking water and proper drinking water using the IPA and K-Means algorithms, obtained as many as 4 clusters, namely cluster 1, low levels of drinking water use and drinking water. high clean, cluster 2 the level of use of drinking water is high and clean drinking water is high, cluster 3 the level of use of drinking water is low and clean drinking water is low, cluster 4 the level of use of drinking water is high and clean drinking water is low.

The data clustering process can be used with several methods, namely the partition-based method, the hierarchy-based method, the density-based method, the lattice-based method, and the partition method. For the partition method itself, there are several yahoos such ask-means, k-modes, k-medoids, fuzzy c-means, and others. The K-Means algorithm is a non-hierarchical data clustering algorithm that tries to partition existing data into one or more clusters/groups. The K-Means algorithm is easier to implement and the process is fast [15].

Based on this case, the authors conducted a study to create data cluster of clean water needs for the period 2012-2017 in Indonesia which requires clean water from the highest level, medium level and low level based on the attributes of the volume of clean water distributed by clean water companies in each province within the period 5 years time using the K-Means algorithm. Clean water demand data from BPS has several missing values, so it is necessary to carry out a cleaning process so that the data is ready to be tested.

The results of clustering are then evaluated using the Davies Bouldin Index method to measure cluster validity to maximize the inter-cluster distance and at the same time try to minimize the distance between points in a cluster. This research is important to do to find out regional clusters in Indonesia based on the volume of clean water distributed to customers, so that each cluster contains data that is as similar as possible and different from data in other clusters. From the data seen, it can be seen which areas most need clean water distribution**.**

2. RESEARCH METHODS

2.1 Research stage

This research was conducted with several steps or stages of research. The steps of the research carried out are described in Figure 1 below:

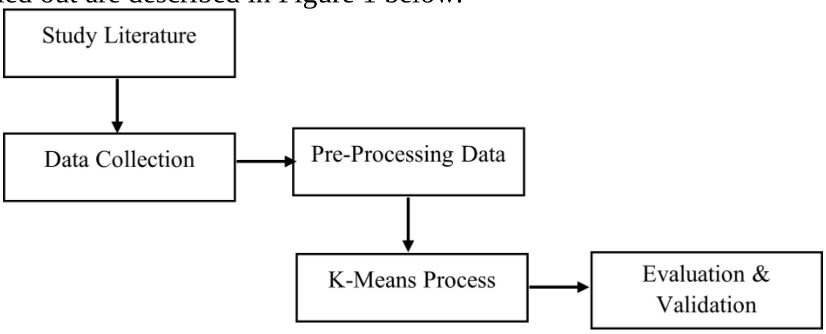

Figure 1. Research Methodology

1. Study Literature

Study Literature is a way of obtaining information and collecting data by reading and studying various literature such as books, journals, modules, internet references, and others related to the problem taken [18]. The library method (Literature) is done by reading sources from books that support writing, scientific journals, playing sites on the internet.

2. Data Collection

Data collection can be done in two ways, namely primary data collection and secondary data collection where primary data can be obtained by means of observations, interviews, questionnaires, etc., which are collected directly from data sources. Primary data collection requires more time and cost than secondary data. While secondary data collection can be obtained directly from other people, both published and unpublished, such as books, journals, documentation, etc. related to the problem to be studied. The data were previously used for different purposes. Secondary data is relatively fast and low cost [3] [1].

The author uses the method of observation or direct observation of the object of research on the official website of the Central Statistics Agency (BPS). The object obtained is data on the volume of clean water distributed by clean water companies by province, 2012-2017. However, the data does not contain 2016 data because the data is presented integrated with the SE2016-Advanced publication. The data contains 5 attributes from 2012 to 2017 and 34 provincial records that are distributed clean water, the data type of all attributes is numeric. Each attribute has blank data and has the total volume of clean water that has been distributed for one year.

3. Pre-Processing Data

Some machine learning cannot work on dirty data such as missing data and outliers that can affect the performance results of the model built. Data cleaning techniques are able to overcome these problems and make the data ready to be used as input data from the built model. Data cleaning techniques can be done with Missing Values Handling and Outlier Detection [19].

At this stage, the missing data will be replaced by the process of replacing missing values using the Rapidminer application version 9.0.3. Missing data will be filled in automatically by the system so that there is no more empty data so that the data is ready to be processed with the algorithm that will be used.

4. K-Means Process

The K-Means algorithm is an iterative clustering algorithm to assign a random cluster value (K) as the cluster center value/centroid/mean/means. From the existing data, the closest distance to each centroid is sought using the Euclidian formula [16]. Meanwhile, according to Sadewo et al., explained that K-Means is an algorithm used in partitioning grouping which separates data into different groups [16].

At this stage, the K-Means clustering algorithm is applied to the existing research objects. The first step is to determine the number of clusters in the data, then determine the value of the center/centroid, then calculate the closest distance to the centroid with Euclidean Distance to get a group of objects based on the closest optimal distance.

The steps taken in using the K-Means algorithm according to Novianti & Haviluddin [7] are as follows:

1. Determine k as the number of clusters to be formed.
2. Determine the initial k centroids randomly or randomly, using the following formula:

$$v = \sum_{i=1}^{n} \frac{x_i}{n} \qquad (1)$$

   Description :
   i = 1,2,3,…n
   v = centroid on cluster xi
   xi = objek to-i
   n = the number of objects / number of objects that are members of the cluster
3. Calculate the distance of each object to each centroid of each cluster using the Euclidian Distance.

$$d(x,y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \qquad (2)$$

4. Allocate each object to the nearest centroid
5. Perform iterations, then determine the position of the new centroid using equation (2).
6. Go back to step 3 if the position of the new centroid with the old centroid is not the same

Table 1. Initial Data Centroid

Attribute	**C0**	**C1**	**C2**
Distributed clean water	22387	465056,8824	2551955

Table 1 is the initial center point (centroid) to determine the initial cluster C0, the average value (average) for cluster C1, the largest (maximum) value for C2 then the value of the initial center point is used to take the closest distance from each data to each -each centroid to generate a new center point value.

Table 2. New Data Centroid Iteration 1

Attribut	*C0*	*C1*	C2
The volume of clean water distributed	92386,7	620635,5	1939461,25

Table 2 is the value of the center point (centroid) resulting from the second iteration process using the K-Means method. Each cluster has a new value from the previous value.

Table 3. Amount of Data On Centroid

Iteration	C0	C1	C2	Total
1	20	10	4	34
2	22	7	5	34
3	22	7	5	34

Table 3 is the value of the center point (centroid) resulting from three iterations, the iteration process stops because in the second and third iterations the position of the new centroid and the old centroid has the same value.

5. Evaluation & Validation

The next stage is to evaluate whether the resulting cluster is good based on the quantity and proximity of the cluster data. while validation is to ensure whether the results are the same as those carried out independently [12]. The validation test process in this study used the Davies Bouldin Index (DBI) as many as 3 clusters. The Davies-Bouldin Index is one of the methods used to measure the validity or the most optimal number of clusters in a grouping method where cohesion is defined as the sum of the proximity of the data to the cluster center point of the cluster being followed [8].

Evaluation with the Davies-Bouldin index (DBI) has an internal cluster evaluation scheme, namely the good or bad results seen from the quantity and proximity of the clustered data, because this method is used to measure cluster validity in the grouping method [2]. Davies-Bouldin index (DBI) is one of the internal evaluation methods that regulates cluster evaluation in a grouping method based on cohesion and separation values [6].

The steps for calculating DBI are as follows [6]:

a. Calculating Sum of squares within cluster (SSW) to find out the cohesion matrix in an i-th cluster.

$$SSW_i = \frac{1}{m_i}\sum_{j=i}^{m_i} d(x_j, c_i) \quad (3)$$

b. Calculating the Sum of square between clusters (SSB) to determine the separation between clusters

$$SSB_{i,j} = d(c_i, c_j) \quad (4)$$

c. Then after knowing the value of cohesion and separation, the next step is to calculate the ratio (Rij) to determine the comparison value between the i-th cluster and the j-th cluster.

$$R_{i,j} = \frac{SSW_i + SSW_j}{SSB_{i,j}} \quad (5)$$

d. The last step is to calculate the value of the Davies-Bouldin index (DBI) using the ratio value obtained

$$\text{DBI} = \frac{1}{K}\sum_{i=1}^{k} max_{i \neq j}(R_{i,j}) \quad (6)$$

3. RESULT AND DISCUSSION

Implementation of the K-Means algorithm to classify the data volume of water distributed by the water company to every province in Indonesia the period 2012 to 2017 using the application RapidMiner produce 3 clusters. Based on the data that has been generated, each cluster has a maximum value, of the maximum data that can be compared to determine whether these clusters fall into the category of low-level needs, medium, and high. Cluster 0 486 199 has a maximum value, for Cluster 1 has a maximum value of 2,551,955, while Cluster 2 has a maximum value of 1366336. From these data, it can be said that the cluster 0 included in the category of low needs for maximum value is less than the maximum belonging to cluster 1, and cluster 1 includes a high need category for the highest maximum value compared to other clusters, whereas cluster 2 including the categories of needs being for the maximum value between the *two other clusters.*

The cluster model resulting from grouping the volume of clean water distributed by the company using the K-Means clustering algorithm is shown in table 4. which contains the number of clusters generated from the rapidminer application. The centroid value in cluster 0 is 150588.24, the centroid in cluster 1 is 1939461.25 and the centroid value in cluster 2 is 857876.6 as shown in table 5. Meanwhile for Cluster 0 consists of 25 data, cluster 1 consists of 4 data, cluster 2 consists of 5 data.

Based on the data that has been grouped, the value range of the total volume of water distributed by the company for cluster 0 is 22387 to 486199, then the value range for cluster 1 is 1598634 to 2551955, while the minimum value range for cluster 2 is 613827 to 1366366.

Table 4. Cluster Model Results

Cluster 0	**Cluster 1**	**Cluster 2**	**Total**
25	4	5	34

Table 5 Centroid on Rapidminer

Attribut	**Cluster 0**	**Cluster 1**	***Cluster 2***
The volume of clean water distributed	150588.24	1939461.25	857876.6

From the total data of 34 provinces, it can be seen that 4 provinces are included in cluster C1 with a high level of need, namely DKI Jakarta, West Java, Central Java, and East Java, then 5 provinces are included in the C2 cluster with a moderate level of need, namely North Sumatra, South Sumatra, Banten, Bali and East Kalimantan, and 25 provinces in the C0 cluster with low needs, namely Aceh, West Sumatra, Riau, Jambi, Bengkulu, Lampung, Bangka Belitung Islands, Riau Islands, DI Yogyakarta, West Nusa Tenggara, East Nusa Tenggara, Kalimantan West, Central Kalimantan, South Kalimantan, North Kalimantan, North Sulawesi, Central Sulawesi, South Sulawesi, Southeast Sulawesi, Gorontalo, West Sulawesi, Maluku, North Maluku, West Papua, and Papua, the data is presented in figure 2, figure 3, and figure 4 as follows:

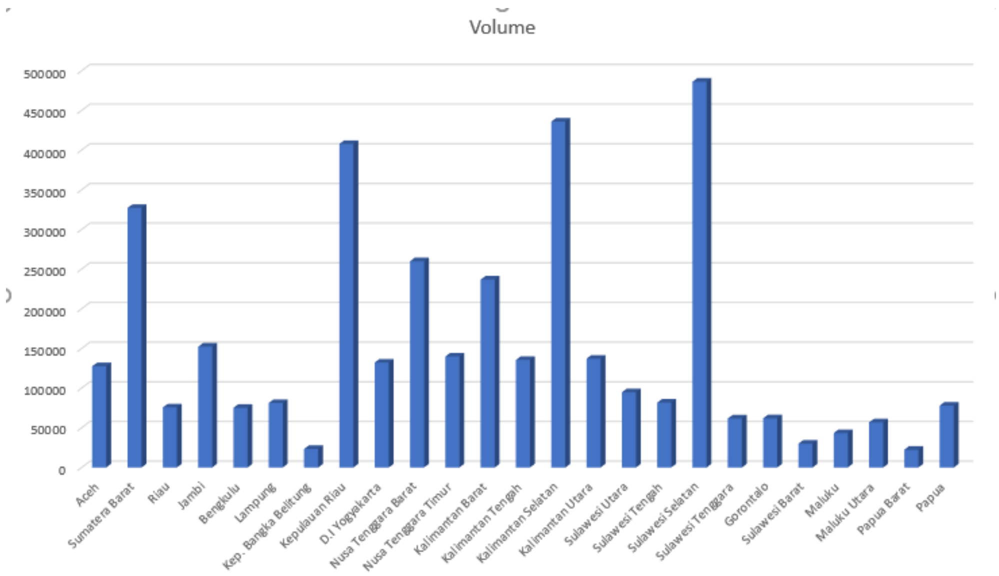

Figure 2. Results of Grouping Cluster_0

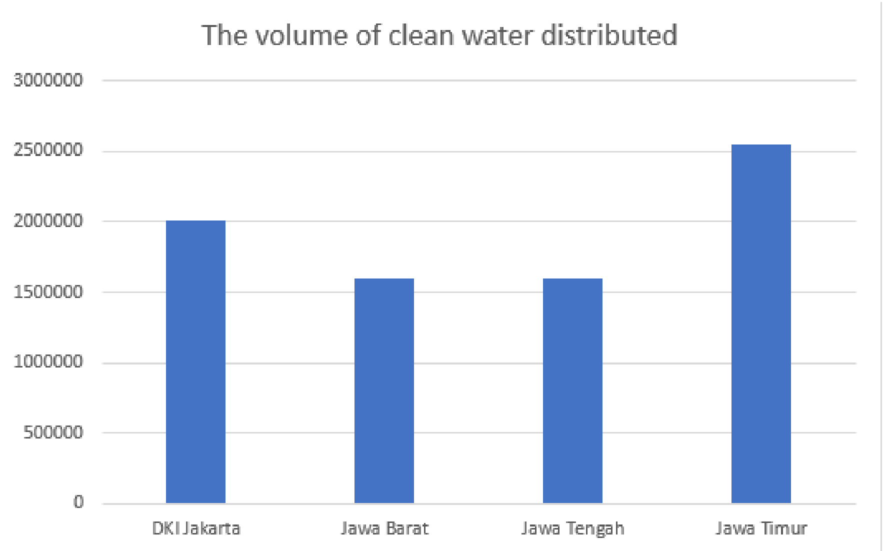

Figure 3. Results of Grouping Cluster_1

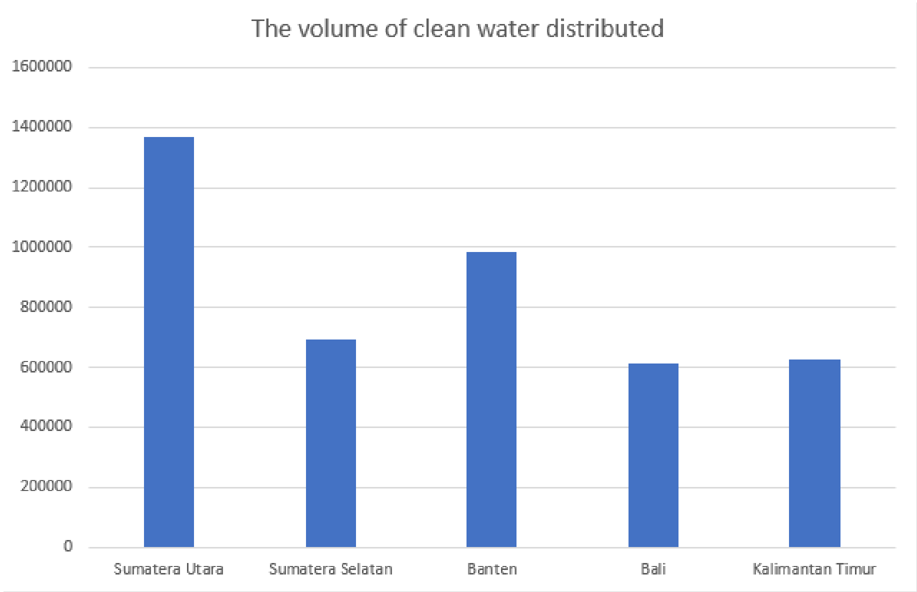

Figure 4. Results of Grouping Cluster_2

A. Evaluation

At this evaluation stage, using a rapidminer application with operator performance to evaluate the performance of centroid-based clustering from the results of K-Means clustering using the Davies Bouldin Index (DBI) method, this DBI approach aims to maximize the distance between one cluster and another and try to minimize the distance between objects in a cluster. The results of the validation of the DBI method obtained a value of 0.534, this value is a good cluster result because the DBI value is close to 0. Then for the results of the average centroid distance, which is 42653057248.250, the average centroid distance of cluster 0 is 16982492031.782, the average centroid distance of cluster 1 is 152895574047.688, the average distance of the centroid of cluster 2 is 82811869891.040.

4. CONCLUSION

The conclusions that can be drawn based on the discussion in the previous chapter regarding the grouping of clean water needs in Indonesia for the period 2012-2017 using the K-Means clustering algorithm are:

a. The application of the K-Means clustering method for grouping clean water needs by the province in Indonesia based on low, medium, and high needs using the rapidminer application has been successful and easy to implement.
b. The application of the K-Means clustering method for grouping has been effective because the results of grouping are 3 groups, namely C0 low level with a value range of 22387 to 486199 and totaling 25 data, C1 is high level with a range of 1598634 to 2551955 and totaling 4 data, C2 is medium level with a range of 5 members and 3 clusters tested using the Davies Bouldin Index (DBI) validation with a value of 0.534. as a measuring value for the validation of the processed data

5. RECOMMENDATION

Based on the results of the research that has been done, for further research can use data by district or sub-district so that research on the need for clean water in Indonesia is more detailed and clear areas that need clean water distribution. In addition, research can use other algorithms for the process of grouping clean water needs in Indonesia.

.

REFERENCES

[1] A. Saifudin. 2018. Metode Data Mining Untuk Seleksi Calon Mahasiswa Pada Penerimaan Mahasiswa Baru Di Universitas Pamulang, *J. Teknol.*, vol. 10, no. 1, pp. 25–36, [Online]. Available: https://www.academia.edu/35836119/Metode_Data_Mining_untuk_Seleksi_Calon_Mahasiswa_pada_Penerimaan_Mahasiswa_Baru_di_Universitas_Pamulang.

[2] Rustam, Suhardi. 2018. Analisa *Clustering Phising* Dengan *K-Means* Dalam Meningkatkan Keamanan Komputer. ISSN: 2087-1716, *Jurnal Ilmiah ILKOM,* Kursor Vol. 10 2 agustus 2018: 175–81. Diambil dari: https://media.neliti.com/media/publications/269188-analisa-clustering-phising-dengan-k-mean-c3e07d80.pdf

[3] Sugiyono. 2016. Metode Penelitian Kuantitatif, Kualitatif dan Kombinasi (*Mixed Methods*). Bandung: CV. Alfabeta.

[4] Gerintya, Scholastica. 2018. Bagaimana Mutu Dan Akses Air Bersih Di Indonesia?, Diambil dari: https://tirto.id/bagaimana-mutu-dan-akses-air-bersih-di-indonesia-cGrk.

[5] Gusril, Henny. 2016. Studi Kualitas Air Minum PDAM di Kota Duri Riau. ISSN: 2085-8167, Jurnal Geografi, Kursor Vol. 8 No. 2 2016: 190–96. Diambil dari: https://jurnal.unimed.ac.id/2012/index.php/geo/article/view/5783/5179

[6] Waworuntu, M. Nanda Variestha, and Muhammad Faisal Amin. 2018. Penerapan Metode *K-Means* Untuk Pemetaan Calon Penerima Jamkesda. ISSN: 2406-7857, *Klik* - Kumpulan *Jurnal Ilmu Komputer,* Kursor Vol. 5 No. 2, September 2018: 190–200. Diambil dari: http://klik.ulm.ac.id/index.php/klik/article/download/157/pdf7

[7] Novianti, Puspitasari, and Haviluddin Haviluddin. 2016. Penerapan Metode *K-Means* Dalam Pengelompokkan Curah Hujan di Kalimantan Timur. ISSN: 2443-048X, Proceedings Snrik. Vol. 1, 14 Desember 2016: 20–26. Diambil dari: https://www.academia.edu/30456634/Penerapan_Metode_K-Means_Dalam_Pengelompokkan_Curah_Hujan_Di_Kalimantan_Timur

[8] Muningsih, Elly, Ina Maryani, Vembria Rose Handayani. 2021. Penerapan Metode K-Means dan Optimasi Jumlah Cluster dengan Index Davies Bouldin untuk Clustering Propinsi Berdasarkan Potensi Desa. Jurnal Sains dan Manajemen. ISSN: 2338-8161 E-ISSN: 2657-0793. Vol 9 No. 1.

[9] Pradipta, Ika Wahyu. 2018. Statistik Air Bersih 2012-2017. ISSN: 0853-6449. November 2018: 1-74 Diambil dari: https://www.bps.go.id/menu/1/informasi-umum.html#masterMenuTab1

[10] Sadewo, Mhd Gading, Agus Perdana Windarto, and Dedy Hartama. 2018. Penerapan Algoritma *Clustering* Dalam Mengelompokkan Banyaknya Desa/Kelurahan Menurut Upaya Antisipasi/ Mitigasi Bencana Alam Menurut Provinsi Dengan *K-Means*. ISSN: 2597-4645, *KOMIK (Konferensi Nasional Teknologi Informasi Dan Komputer,* Kursor Vol. 2 No. 1, Oktober 2018: 311–19. Diambil dari: https://doi.org/10.30865/komik.v2i1.943

[11] Sadewo, Mhd Gading, Agus Perdana Windarto, and Anjar Wanto. 2017. Penerapan *Datamining* Pada Populasi Daging Ayam Ras Pedaging di Indonesia Berdasarkan Provinsi Menggunakan *K-Means Clustering*. ISSN: 2540-7600, *InfoTekJar (Jurnal*

Nasional Informatika Dan Teknologi Jaringan), Kursor Vol. 2 No.1, September 2018: 60–67. Diambil dari: https://jurnal.uisu.ac.id/index.php/infotekjar/article/view/164

[12] Saifudin, Aries. 2018. Metode Data Mining Untuk Seleksi Calon Mahasiswa pada Penerimaan Mahasiswa Baru di Universitas Pamulang ISSN: 2085-16669, Jurnal Teknologi, Kursor Vol. 10 No. 1, Januari 2015. Diambil dari: https://www.academia.edu/35836119/Metode_Data_Mining_untuk_Seleksi_Calon_Mahasiswa_pada_Penerimaan_Mahasiswa_Baru_di_Universitas_Pamulang.

[13] Pratama, Bayu. dkk. 2018. CImplementasi Importance Performance Analysis dan Algoritme K-means untuk Wilayah Indonesi. Jurnal Matematika, Statistika & Komputasi (JMSK). Vol. 15, No. 1. e-ISSN: 2614-8811.

[14] Sukartini, Ni Made. 2016. Akses Air Bersih di Indonesia. ISSN: 2301-8968, *Jurnal Ekonomi Kuantitatif Terapan,* Kursor Vol. 9 No. 2, Agustus 2016: 89-98, Diambil dari *https://ojs.unud.ac.id/index.php/jekt/article/view/27421/17372*

[15] Suyanto. 2019. Data Mining Untuk Klasifikasi dan Klasterisasi Data. Bandung: Informatika.

[16] Vulandari, Retno Tri. 2017.Data Mining Teori dan Aplikasi *Rapidminer*. Yogyakarta: Gava Media.

[17] Siregar, A.R, dkk. 2019. Clustering Pada Pelanggan Perusahaan Air Bersih Dengan Algoritma K-Means. Jurnal Penerapan Kecerdasan Buatan. Vol. 1, No. 1. ISSN: 2715-9906.

[18] Wijianto, R., & Anggoro, A. 2018. Sistem informasi pendaftaran seleksi kerja berbasis web pada BKK (Bursa Kerja Khusus) Tunas Insan Karya SMK Negeri 2 Banyumas. Jurnal Evolusi, 76-84.Diambil dari: https://ejournal.bsi.ac.id/ejurnal/index.php/evolusi/article/view/3584/2282.

[19] C. Li,. 2019. Preprocessing Methods and Pipelines of Data Mining: An Overview. *Semin. DATA Min.*, no. June, pp. 1–7, [Online]. Available: http://arxiv.org/abs/1906.08510