Grouping of General Hospitals Based on Specialist Doctors in the Special Region of Yogyakarta with the K-Means Clustering and Visualization on the Dashboard Tableau

Erma Shofi Utami[1], Mercynanda Yuliany Alang[1], Rokhana Dwi Bekti[1*], Edhy Sutanta[2]

[1]Department of Statistics, Institut Sains & Teknologi AKPRIND Yogyakarta, Indonesia
[2]Department of Informatics, Institut Sains & Teknologi AKPRIND Yogyakarta, Indonesia
**Corresponding author: rokhana@akprind.ac.id*

ABSTRACT

One of the most needed health workers in hospitals is a specialist doctor. The number of general hospitals in Special Region of Yogyakarta (DIY) is 55, with different characteristics of specialist doctors. **Purpose:** Information on the grouping of general hospitals based on the characteristics of specialist doctors will help provide information about the distribution of health workers. **Methods:** This study uses K-Means Clustering as a non-hierarchical clustering method that can group large amounts of data with fast and efficient computation time and is easy to adapt. **Results:** This study used the number of clusters (k) 2, 3, 4, and 5 and then compared them. Based on the largest silhouette index coefficient, the best number of clusters is k = 2. A total of 35 hospitals in cluster 2 represent hospitals with an adequate number of specialists. A total of 2 hospitals in cluster 2 represent hospitals with a sufficient number of specialists. Visualization using the Tableau dashboard has also provided benefits, namely providing information on the number of specialist doctors in each general hospital, cluster results, and cluster profiling. **Implications:** The K-Means method can provide a reference for the distribution of specialist doctors in the Special Region of Yogyakarta.

Keywords: *hospitals, K-Means, specialist doctor, Tableau*

1. INTRODUCTION

Hospital is a health service institution that provides complete individual health services that provide inpatient, outpatient, and emergency services. The level of community welfare can be seen from the population health condition, such as health facilities, health workers, and the number of sick people in the community. A hospital is one of the health facilities with adequate health equipment, such as medical equipment, health workers, and medicines, so people often prefer to come directly to the hospital for an examination. In addition, hospitals generally provide 24 hours health services to meet the unexpected and unexpected needs of the community.

Health workers are an essential factor that can support the quality of service in a hospital. A health worker is every person who devotes and has knowledge and expertise in the health sector. Health workers are divided into several clumps. According to Law No. 36 of 2014, health workers consist of medical staff, clinical psychology staff, nursing staff, midwifery workers, pharmaceutical workers, public health workers, environmental health workers, nutritionists, physical therapy personnel, medical technicians, biomedical engineering personnel, traditional health workers, and other health workers. Specialist doctors are health workers widely available in public hospitals and specialize in a particular field of disease or body part. Meanwhile, there are 45 types of specialist doctors in Indonesia. In addition, a pandemic that has not ended like this will certainly require more health workers to improve health services.

Special Region of Yogyakarta (DIY) is a province that has quite several general hospitals with various types of specialist doctors. These types of specialists include internal medicine specialists, surgeons, obstetricians and gynecologists, anesthesiologists, dental specialists, pathology specialists, and many others. According to Bappeda data, in 2021, DIY has 5,602 posyandu, 808 puskesmas, 9 regional general hospitals (RSUD), 45 general hospitals (RSU), and 23 special hospitals. There are many general hospitals in DIY, each of which has different service characteristics and the number/type of specialist doctors. Hospitals that have complete specialist doctors will certainly make it easier for people to get health services. With so many types of specialist doctors in certain general hospitals, it is necessary to group general hospitals based on specialist doctors. The clustering method that can be used is K-Means clustering.

K-Means algorithm is used to group data into several groups with several clusters [1]. It is often used to group data based on certain characteristics because it is easy to implement. The time required is relatively fast, flexible, and uses simple principles so that it can be explained in non-statistical terms. This method is included in unsupervised machine learning [2], namely the process of grouping data that does not have a label. K-Means is a non-hierarchical data clustering method that seeks to partition existing data into one or more clusters or groups so that data with the same characteristics are grouped into the same group. Initially, it takes some of the population components to be used as the initial cluster center. At this stage, the cluster is randomly selected from a population data set. Then each element in the population is tested and marked to one of the defined cluster centers, depending the minimum distance between components and each cluster. The cluster center position will be recalculated until all data components are classified into each cluster center, and finally, a new cluster center position will be formed [3].

Visualization of descriptive and clustering results is very important to facilitate the interpreting the results. Therefore, the researcher also created a dashboard for visualization. According to [4], visualization is defined as a method for presenting data or problems in a graphic format or image form that is easy to understand. This data visualization will make it easier for readers to understand the information quickly and effectively conveyed by researchers using various interactive graphics or images that are interesting to readers. In this study, the researcher visualized the data using various graphs that were put together in a dashboard. A dashboard is a user interface between the data and the design that displays some of the results of the analysis of the data. According to [5], the dashboard is the result of a representative data visualization. A dashboard is a visual display of the most important information needed to achieve one or more goals, combined and arranged on a screen so that the information needed can be seen at a glance, so it doesn't take long to understand the information displayed.

This study grouped general hospitals in DIY based on the number of specialist doctors. The aim is to obtain information about the types of hospitals that have specialist doctor facilities, from incomplete to the most complete. These results can also provide information about the distribution of specialist doctors in DIY. The method used is K-Means which has an easy-to-implement algorithm. Furthermore, the grouping results will be visualized on the Tableau dashboard to provide more interesting and easy-to-understand illustrations

2. METHODS

The research design used in this study is a descriptive quantitative approach. The approach is presented in numerical or numerical form, and the interpretation of the results is carried out in the form of a description of the results of grouping hospitals. The object of this research is 37 general hospitals in DIY. The data used is secondary data obtained from the DIY Health Office in 2021. The variables are a number of doctor specialists of 7 types (surgeons, ob-gyn specialists, anesthesiologists, dental specialists, clinical pathologists, anatomical pathologists, and internal medicine specialists).

This study uses the K-Means clustering method, which aims to group general hospitals in DIY based on the number of specialist doctors. The data processing is carried out in the Rstudio and Tableau software. Rstudio is used to perform clustering analysis with K-Means, and Tableau is used to visualize the clustering results. The following are data analysis steps: 1) Test the grouping assumptions, namely the multicollinearity test and the outlier test. The multicollinearity test was carried out by looking at the output results of the correlation value between variables where the correlation value was not more than 0.95. Outlier detection is done by Boxplot [6]; 2) Determine the optimal number of clusters using the Elbow method; 3) Formation of clusters with K-Means algorithm, with k = 2,3,4, and 5; 4) Comparing clustering results based on silhouette index values [7]; 5) Perform profiling in each cluster that is formed; 5) Arrange visualization of clustering results on Tableau [8].

Clustering is the grouping of a number of data or objects into clusters so that each data in the cluster will contain data that is as similar as possible and different from objects in other clusters [9]. Clustering is the process of dividing data in a set into several groups whose data similarity in one group is greater than the similarity of the data with data in other groups.

Clustering is a data mining method that is unsupervised or without direction. It means that this method is carried out without training and output. Clustering is divided into two types, namely hierarchical clustering and non-hierarchical clustering. Hierarchical clustering starts by grouping two or more objects that have the closest similarity. Then the process is passed on to another object that has a second proximity, and so on until the cluster will form a kind of tree where there is a clear hierarchy between objects from the most similar to the least similar. Dendrograms are usually used to help clarify the hierarchical process. The methods included in hierarchical clustering are the Single Linkage Method, Complete Linkage Method, Average Linkage Method, Ward's Method, Centroid Method, and Median Method. Meanwhile, non-hierarchical begins by first determining the number of clusters and then doing the clustering process. The K-Means K-Median method and the fuzzy method are non-hierarchical methods. Another method is clustering by spatial methods, such as Spatial 'Kluster analysis by tree edge removal or SKATER [10] and Local Getis Ord-G Statistics [11].

K-Means is an algorithm partitioning grouping and separating data into different groups. K-Means is included in the cluster analysis where k is the number of clusters. This algorithm is simple to implement and run. It can group large amounts of data with relatively fast and efficient computation time and is easy to adapt [12]. The concept is to get the minimum variation value where each cluster with the minimum distance between the data and the cluster midpoint. If a cluster still has large variations, then the cluster can still be divided into two different clusters. Before performing the analysis with K-Means, it starts with standardizing the data if the variables used have different units. Standardization aims to homogenize data values that are inputted in an inconsistent format among variables.

The steps in the K-Means analysis are testing assumptions, forming clusters, and validating and profiling clusters. The following are the steps in forming a cluster:

a. Determine the number of clusters to be formed.

In determining the optimal number of clusters, researchers can use various methods, such as the Elbow method, the Silhouette method, or the number of clusters that the researcher has determined. This study use Elbow method to determine the optimal number of clusters. The Elbow determines the best number of clusters by looking at the percentage of comparison results between the number of clusters that will form an elbow at a point [13]. To get a comparison, calculate each cluster value's Sum of Square Error (SSE). The larger the number of cluster k values, the smaller the SSE value will be small. Here's the SSE formula:

$$SSE = \sum_{k=1}^{k} \sum_{x_i} |x_i - c_k|^2 \quad (1)$$

with k is cluster, x_i is object data distance I, and c_k is the center of the *i-th* cluster.

b. Determine the value of the centroid or the center point of the cluster.

The determination of the centroid is done randomly. Then it is done using an iteration stage. The formulas that can be used are:

$$\overline{v_{\iota J}} = \frac{1}{N} \sum_{k=0}^{Ni} x_{kj} \quad (2)$$

where v_{ij} is the centroid or average of the i-th *cluster* and j-th variable, N_i is the amount of data that is a member of the i-th cluster, x_{kj} is the value of the k-th data in the cluster for the j-th variable.

c. Calculate the distance of each centroid point with the point of each object.
This steps use the Euclidean distance (d_{ij}).

$$d_{ij} = \sqrt{\sum_{k=1}^{p}\left(x_{ik} - x_{jk}\right)^2} \quad (3)$$

d. Grouping data into clusters with the closest distance.

$$min \sum_{k}^{i} = d_{ij} = \sqrt{\sum_{k=1}^{p}\left(x_{ik} - x_{jk}\right)^2} \quad (4)$$

e. Calculate the center of the new cluster by finding the average value of the data that is a member of the cluster.

$$C_{kj} = \frac{\sum_{k}^{i} x_{ij}}{p} \quad (5)$$

where p is the number of k-th cluster members.

f. Repeat steps two through four until no more data is moved to another cluster.

3. RESULTS AND DISCUSSION

Characteristics of Specialist Doctors in each General Hospital

Descriptive analysis can be used to describe the characteristics of the number of specialist doctors in 37 general hospitals in DIY. The highest number of specialists are internal medicine specialists; that is, on average there are 4.7 or 5 internal medicine specialists in each general hospital. The hospital that has the most specialists is RSUP Dr. Sardjito, with 39 doctors. Meanwhile, out of 37 general hospitals, there is still 1 hospital that does not yet have an internal medicine specialist. The maximum number of surgeons is 13 doctors at Panti Rapih Hospital. The average surgeon specialist in 37 hospitals is 2.81 or 3 doctors in each general hospital. Meanwhile, there are still 4 hospitals that do not yet have surgeons. Panti Rapih Hospital also has the highest number of dentists, which is 10.

Dr. Hospital Sardjito is still a general hospital with the highest number of specialist doctors, namely, 25 obstetricians, 13 anesthesiologists, 15 clinical pathology specialists, and 14 anatomy specialists. Specialist doctors who are still limited are anatomical pathology specialists because the average number of these doctors is less than 1 or there are still many public hospitals that do not yet have.

Assumption Test

Assumption tests performed before using the K-Means algorithm were sample adequacy test, multicollinearity test, and outlier detection. Good data for K-Means is data that has tested the sample's adequacy, no multicollinearity, and no outlier data. The data of this study have met the adequacy of the sample because the data used is population data. The results of the multicollinearity test are based on the correlation value between the variables. The results are the correlation value between variables are small. It means that there is no multicollinearity between variables or the multicollinearity assumption has been met. Detection of outlier data in this study was carried out using a boxplot. The results show that there are outlier data. However, this study still maintains outlier data because all data are important. As an alternative, this study uses data that has been standardized on the K-Means process.

K-Means Clustering

The first step in K-Means clustering is to determine the number of clusters. This study uses the Elbow method, one of the methods commonly used to determine the optimum number of clusters by looking at the comparison percentage between the number of clusters that will form an elbow at a point. The results of the Elbow method are presented in Figure 1. Based on the figure, it can be seen that the line that undergoes a fracture that forms an elbow is at k = 2. Thus, the best number of clusters is 2. However, this study also performs the formation of clusters with k = 3, k = 4, and k = 5 to get a comparison of results.

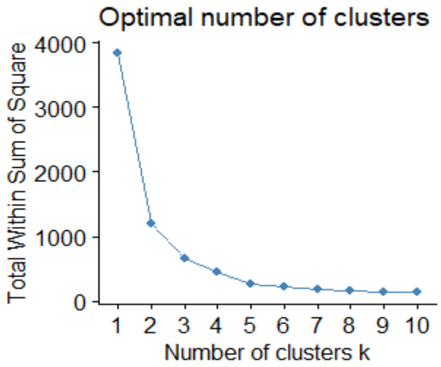


Figure 1. Elbow graph for get optimal number clusters

The results of grouping with the K-means algorithm are visualized on a 2-dimensional plot as shown in Figure 2. In the results of 2 clusters, members of cluster 1 are red and cluster 2 are blue. Based on the comparison, it can also be seen that grouping with k=2 is better than k=3, 4, and 5. With k=2, objects in clusters have higher homogeneity than between clusters. In addition, there is high heterogeneity between clusters, which is indicated by objects in cluster 1 having a long-distance from objects in cluster 2.

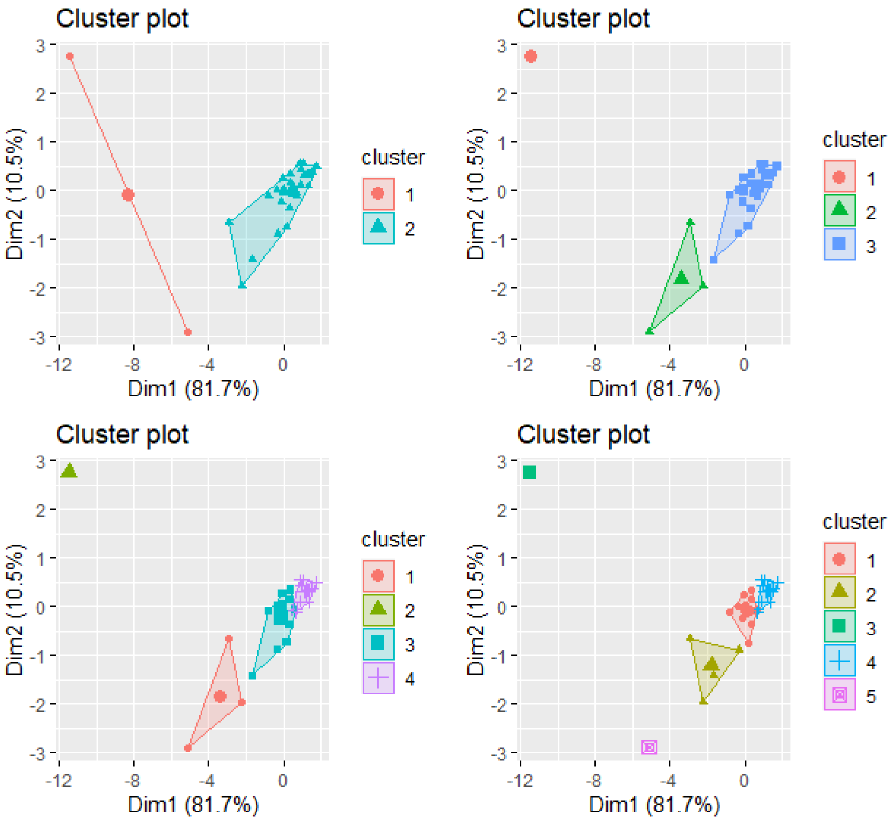


Figure 2. 2-dimensional plot for K-Means grouping results

After knowing the results of clustering, the next step is to validate to find out whether the cluster results obtained are valid or not. This is done by looking at the silhouette index coefficient value. Good and valid groupings are those that have a silhouette index value close to the value 1. The largest silhouette index value is 0.81 with the number of clusters k = 2.

Once it is known that k-2 gives the best results, the next step is to identify hospitals in each cluster and develop a profile. The names of public hospitals in clusters 1 and 2 are as follows:

Cluster 1 : RSU Panti Rapih, RSUP Dr.Sardjito

Cluster 2 : RSU PKU Muhammadiyah Bantul, RSU Nur Hidayah, RSU Rachma, Husada, RSU Permata Husada, RSU Griya Mahardika Yogyakarta, RS Rajawali Citra, RSPAU Dr. Suhardi Harjolukito, RSUD Wonosari, RSU Nur Rohmah, RSU Pelita Husada, RSU Panti Rahayu, RS Bethesda Wonosari, RS Happy Land Medical Centre, RSU PKU Muhammadiyah Wonosari, RS Islam Hidayatullah Yogyakarta, RS Tk. III 04.06.03 Dr. Soetarto, RS PKU Muhammadiyah Yogyakarta, RS Bethesda Yogyakarta, RSUD Kota Yogyakarta, RS Ludira Husada Tama, RS Bethesda Lempuyangan, UPT RS Pratama Kota Yogyakarta, RS Siloam Yogyakarta, RSUD Wates, RSU St Yusup Boro, RSU Rizki Amalia Medika, RSU Kharisma Paramedika, RSU PKU Muhammadiyah Wates, RSU PKU Muhammadiyah Nanggulan, RSUD Nyi Ageng Serang, RSU Pura Raharja Medika, RSUD Sleman, RSU Panti Baktiningsih, RSU Queen Latifa, RS JIH.

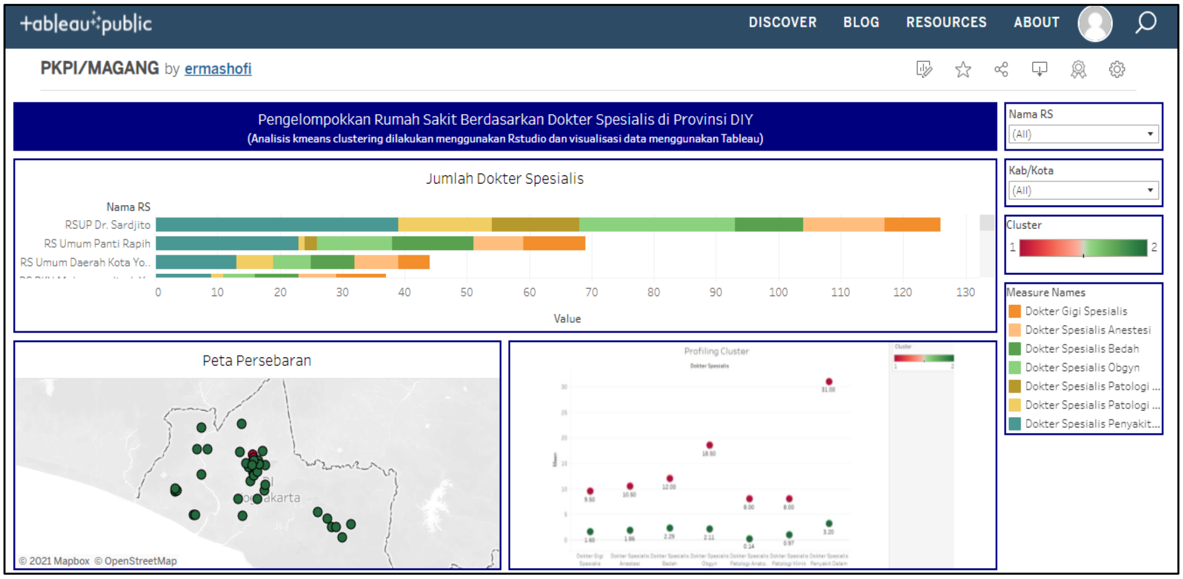
Cluster profiling is compiled using the average value of each variable in each cluster which is presented in Table 1. The average value of the number of specialist doctors in hospitals in cluster 1 is more than cluster 2. Therefore, it can be said that cluster 1 represents a hospital with a very sufficient number of specialists and cluster 2 represents a hospital with a sufficient number of specialists.

Table 1. Profiling cluster with k = 2

Cluster	Surgeon Specialist Doctor	Surgeon Specialist Doctor	Anaesthesiologist	Dentist	Clinical Pathology Specialist	Anatomical Pathology Specialist	Internal Medicine Specialist
1	12.00	18.50	10.50	9.50	8.00	8.00	31.00
2	2.29	2.37	2.06	1.77	1.03	0.17	3.17

Visualization on the Tableau Dashboard

Data visualization is used to make it easier for readers to understand information. Hospital data based on specialist doctors in DIY obtained two clusters, namely cluster 1 and cluster 2. The visualization that has been built is presented in Figure 3 or can be accessed at https://tabsoft.co/39vMKkn . This visualization is done using Tableau software with a dashboard based on the best grouping, namely two clusters. This dashboard displays a map of the distribution of hospitals based on their clusters and a stacked bar chart of the number of specialist doctors by hospital.



PKPI/MAGANG by ermashofi

Pengelompokkan Rumah Sakit Berdasarkan Dokter Spesialis di Provinsi DIY
(Analisis kmeans clustering dilakukan menggunakan Rstudio dan visualisasi data menggunakan Tableau)

Nama RS: (All)

Kab/Kota: (All)

Figure 3. Visualization results on the Dashboard Tableau

4. CONCLUSION

This study used the number of clusters (k) 2, 3, 4, and 5 and then compared them. Based on the largest silhouette index coefficient, the best number of clusters is k = 2. A total of 2 hospitals in cluster 1 represent hospitals with high a number of specialists. A total of 35 hospitals in cluster 2 represent hospitals with a high number of specialists. Visualization using the Tableau dashboard has also provided benefits: providing information on the number of specialist doctors in each general hospital, cluster results, and cluster profiling.

ACKNOWLEDGMENT

The authors would like to thank the Institut Sains & Teknologi AKPRIND Yogyakarta, Indonesia, that support this research through grant funding and facilities to carry out this research.

REFERENCES

[1] N. Dwitri, J. A. Tampubolon, S. Prayoga, F. I. R. H Zer, and D. Hartama, “Penerapan algoritma K-means dalam menentukan tingkat penyebaran pandemi Covid-19 di Indonesia,” *J. Teknol. Inf.*, vol. 4, no. 1, 2020, doi: 10.36294/jurti.v4i1.1266.

[2] K. P. Sinaga and M. S. Yang, “Unsupervised K-means clustering algorithm,” *IEEE Access*, vol. 8, 2020, doi: 10.1109/ACCESS.2020.2988796.

[3] Y. D. Darmi and A. Setiawan, “Penerapan metode clustering K-means dalam pengelompokan penjualan produk,” *J. Medi Infotama*, vol. 12, no. 2, 2017, doi: 10.37676/jmi.v12i2.418.

[4] K. Kurniawan and D. Antoni, “Visualisasi data penduduk dalam membangun e-government berbasis Sistem Informasi Geografis (GIS),” *J. Sisfokom (Sistem Inf. dan Komputer)*, vol. 9, no. 3, 2020, doi: 10.32736/sisfokom.v9i3.828.

[5] M. Silvana, R. Akbar, and R. Tifani, “Penerapan dashboard system di Perpustakaan Universitas Andalas menggunakan Tableau Public,” *Semin. Nas. Sains dan Teknol. 2017*, November, 2017.

[6] I. Hussain, “Outlier detection using graphical and nongraphical functional methods in hydrology,” *Int. J. Adv. Comput. Sci. Appl.*, vol. 10, no. 12, 2019, doi: 10.14569/ijacsa.2019.0101259.

[7] I. L. Kirilyuk and O. V. Senko, “Assessing the validity of clustering of panel data by Monte Carlo methods (using as example the data of the Russian regional economy),” *Comput. Res. Model.*, vol. 12, no. 6, 2020, doi: 10.20537/2076-7633-2020-12-6-1501-1513.

[8] J. Kokina, D. Pachamanova, and A. Corbett, “The role of data visualization and analytics in performance

management: Guiding entrepreneurial growth decisions," *J. Account. Educ.*, vol. 38, 2017, doi: 10.1016/j.jaccedu.2016.12.005.

[9] N. Damanik and M Sigiro, "Penerapan data mining menggunakan algoritma K-Means clustering pada penerimaan mahasiswa baru sebagai metode promosi," *Jutisal J. Tek. Inform. Univers.*, 2021.

[10] Y. Setyawan, R. D. Bekti, and F. Isarlin, "Application of SKATER and Ward's methods in grouping Indonesian provinces based on monthly expenditure per capita of food commodity groups," in *IOP Conference Series: Materials Science and Engineering*, 2020, vol. 807, no. 1. doi: 10.1088/1757-899X/807/1/012017.

[11] R. D. Bekti, G. E. Dirgantara, and E. Sutanta, "Distance and AMOEBA weights matrices in local getis Ord-G statistics to identify spatial cluster of gini ratio," 2021. doi: 10.1109/ICERA53111.2021.9538666.

[12] H. Sulastri and A. I. Gufroni, "Penerapan data mining dalam pengelompokan penderita thalassaemia," *J. Nas. Teknol. dan Sist. Inf.*, vol. 3, no. 2, 2017, doi: 10.25077/teknosi.v3i2.2017.299-305.

[13] R. Sammouda and A. El-Zaart, "An optimized approach for prostate image segmentation using K-Means clustering algorithm with Elbow method," *Comput. Intell. Neurosci.*, vol. 2021, doi: 10.1155/2021/4553832.