BAREKENG: Jurnal Ilmu Matematika dan Terapan
June 2022 Volume 16 Issue 2 Page 743–750
P-ISSN: 1978-7227 E-ISSN: 2615-3017
https://doi.org/10.30598/barekengvol16iss2pp743–750

APPLICATION OF CLUSTERING ANALYSIS TO DATA DISTRIBUTION OF COVID-19 IN BENGKULU PROVINCE

Nurul Hidayati*

Statistics Study Program, Faculty of Mathematics and Natural Sciences, Bengkulu University
W. R. Supratman St., Kandang Limun, Bengkulu, 38371, Indonesia

*Corresponding author's e-mail: * nurulhidayati@unib.ac.id*

Abstract. *Bengkulu Province is one of the provinces in Indonesia. Based on the results of the Population Census (SP) in September 2020, carried out by BPS, there were 2,010,670 inhabitants in Bengkulu Province. The area of Bengkulu Province is 19,813 km2, consisting of 10 regencies/cities. The large area and population encourage an effort to anticipate the transmission of COVID-19 that is soaring high in Bengkulu Province. One is by grouping regencies/cities in Bengkulu Province based on several variables that characterize objects using the Clustering method. This study aimed to group districts/cities in Bengkulu Province based on several variables that characterize objects related to the spread of COVID-19 in Bengkulu Province. The method used was the clustering method. The data used in this study was secondary data about the variable of the spread of COVID-19 in Bengkulu Province from January 1, 2021, to May 31, 2021. It is accessed through the official website of the Bengkulu Province government to convey information to the public regarding the increase of COVID-19 Cases in Bengkulu Province. The grouping using the Hierarchical Clustering method obtained the best model as complete linkage, with the number of clusters K = 2 and the K-Means method with K = 2. The results obtained are good because it has relatively tiny variability within the cluster, and the value of variability in both clusters is relatively large.*

Keywords: *Bengkulu, COVID-19, clustering, hiraerchical clustering, K-means*

Article info:

Submitted: 6th February 2022 *Accepted: 21st April 2022*

How to cite this article:
N. Hidayati, "APPLICATION OF CLUSTERING ANALYSIS TO DATA DISTRIBUTION OF COVID-19 IN BENGKULU PROVINCE", *BAREKENG: J. Il. Mat. & Ter.*, vol. 16, iss. 2, pp. 743–750, June, 2022.

1. INTRODUCTION

A virus that spreads through droplets of liquid from the mouth and nose that come from an infected person when coughing and sneezing and is found in humans and animals whose symptoms are similar to flu is called coronavirus-19 [1]. This virus is spreading quite rapidly. More than 240 countries have been infected with this virus [2]. In December 2019, COVID-19 was discovered in China, Wuhan Province [3].

For the first time, two positive cases of COVID-19 were reported in Indonesia on March 2, 2020. In March 2020, there were 1,528 confirmed cases and 136 deaths. The number of deaths in Indonesia is the highest in Southeast Asia, with a COVID-19 death rate of 8.9% [4].

Indonesia has 34 provinces, one of which is Bengkulu province, located on the southern part of Sumatra Island. Bengkulu is the 32nd province affected by the coronavirus, with 101 confirmed patient cases in June 2020 [5].

The population of Bengkulu Province is 2,010,670 people, the 27th largest population in Indonesia. The 9th in Sumatra, based on the Population Census (SP) results in September 2020 conducted by the Central Statistics Agency (BPS). The population has increased by 295,152 people from 2020 to 2021, or an average of 24,596 people yearly. The area of Bengkulu Province is 19,813 km2. It is the 24th largest city in Indonesia and the 8th largest in Sumatra, with ten districts or cities [6].

The large area and population encourage an effort to handle the COVID-19 virus to prevent a high spike in transmission in Bengkulu Province. One is by grouping district/city areas in Bengkulu Province based on the variables that characterize the object by using clustering. [7].

K-Means is a simple and commonly used clustering method that can group large amounts of data with relatively fast and efficient computational data processing. However, the weakness of the K-Means method caused by the initial determination of the given cluster can cause the clustering result to be a locally optimal solution. On the other hand, the hierarchical clustering method can be used to overcome the problem of determining the cluster center in K-Means [8]. Therefore, this study combined these two methods, namely the K-Means method and Hierarchical Clustering. The methods included in the hierarchical clustering group were compared to see which cluster gave better clustering results. The purpose of grouping was to determine the similarity or closeness between data.

Many studies on data clustering using hierarchical and K-Means methods have been carried out, including "Comparative Analysis of Hierarchical Clustering Methods K-Means and the Combination of the Two in Cluster Data (Case Study: Practical Work Problem, Department of Industrial Engineering ITS)". The study results explain that the best cluster results are obtained with the cluster variance test parameter and the Silhouette coefficient method by combining the Single Linkage Clustering method with K-Mean [8]. The results of the study entitled "Clustering Analysis Using the K-Means and Hierarchical Clustering Method (Case Study: Thesis Document of Chemistry Department, FMIPA, Sebelas Maret University)" show that the variety of research themes carried out by students and the number of studies in a theme are to the interest students and lecturer projects in the Department of Chemistry are influenced by the expertise of the lecturers [9].

2. RESEARCH METHOD

2.1 Data source

This study used secondary data regarding Corona Virus (COVID-19) cases spread in 10 regencies/cities of Bengkulu Province from January 1, 2021, to May 31, 2021. It was obtained from the official website of the Task Force for the Acceleration of Handling COVID-19 for Bengkulu Province (https://covid19.bengkuluprov.go.id).

2.2 Research variable

The variables used in this study were the number of confirmed people(X_1), number of people or patients who died (X_2), the number of people or patients who recovered (X_3) and the number of people who were suspected (X_4). The research variables are briefly presented in Table 1, below:

Table 1. Research Variables

Variable	Interpretation
X_1	Confirmed
X_2	Died
X_3	Recovered
X_4	Suspected

2.3 Analysis Steps

This study used cluster analysis with the K-means clustering method. The steps of analysis for this research were:

1. Collecting secondary data from the Task Force's official website for the acceleration of handling COVID-19 for Bengkulu Province.
2. Performing cluster analysis using the hierarchical method.
   a. The hierarchical method starts the grouping stage with two or more objects with the closest similarity. Furthermore, the grouping process continues to the next object with second proximity, so the cluster will form a structure similar to the tree diagram structure, which is presented as a dendrogram [10]. The advantages of hierarchical clusters are that they speed up data processing and save time because the inputted data forms a hierarchy (tiers). Meanwhile, the weakness of the hierarchical cluster method is the difference in the size of the distance used, and there are irrelevant variables [11].

      Hierarchical clustering is divided into two, namely: the agglomerative method and the divisive method. The agglomerative method consists of the linkage, centroid, and variance methods. The agglomerative method begins the grouping stage by placing objects in different clusters. Then the objects are grouped gradually into larger clusters. 12]. The agglomerative clustering method assumes that every piece of data that exists as a cluster at the beginning of the process, the object of the study is considered a cluster. If in the first stage there are *n* objects, there are *n* clusters. The algorithm of the agglomerative clustering method is [9]:

      a. Calculating the distance between clusters using the Euclidean distance. The equations used are:

$$d = \frac{\sum_{i=1}^{n}(x_i - y_i)^2}{n} \quad (1)$$

      Where:

      y_i = cluster center

      x_i = object of observation to $-\ i$

      n = the number of objects that are members of the cluster

      b. Choosing two clusters that have a minimum distance based on the results of the distance calculation using equation (1).

      c. Merging two clusters that are minimally spaced.

      d. Repeating step one to step three, so that all objects are merged into one cluster.

3. Determining the optimum number of clusters based on the Elbow and Silhouette methods.

   a. Elbow Method

      The Elbow method can determine the optimal number of clusters by comparing the percentage results with the number of clusters that will form an elbow at a point. The comparison between the first and second cluster values was obtained by calculating each cluster value's Sum of Error (SSE). If the SSE value has decreased at the maximum, then the value of the cluster is the right one. If the SSE value is getting smaller, the number of cluster K values will get bigger. The SSE equation is as follows:

$$SSE = \sum_{k=1}^{k}\sum_{x_i \epsilon S_k}\|x_i - c_k\|_2^2 \quad (2)$$

With *k* is the number of clusters, *K-Means* x_i is the number of objects i and c_k is the number of the *k-th* cluster.

b. Silhouette Method

This method combines the cohesion method with separation, which aims to see the quality and strength of the cluster, namely how well an object is placed in a cluster is called the Silhouette Coefficient method. The Silhouette Coefficient method can also be used to measure the close relationship between objects in a cluster.[13]. The value obtained from the silhouette coefficient method lies in the range of values from -1 to 1. If the silhouette coefficient value is close to the value 1, the grouping of objects in one cluster becomes better. On the other hand, if the Silhouette Coefficient approaches the value of -1, the grouping of objects in one cluster becomes worse [13].

4. Performing cluster analysis with non-hierarchical method using K-Means.

The K-Means method is a non-hierarchical method that groups data in the form of one or more clusters or groups. The K-means method is a partition method used to analyze data and treat data observations as objects based on the location and distance between each data [14]. K-Means aimed to group objects into clusters according to their similar characteristics so that objects with similar/same attributes are grouped into one cluster and objects with different attributes are grouped into different clusters [14]. This method is used for data with large sample sizes because it has a higher speed than the hierarchical method.

The grouping process using the K-means method was as follows [15]:

i) Determine the number of *K* clusters.
ii) Determine the average value of objects in the cluster or the centroid value of the cluster.
iii) Use the Euclidean distance measure to calculate the closest distance of each object to each centroid or center.
iv) Recalculate the object mean for the newly formed cluster.
v) If the average value of the object does not change, then the process is terminated. However, if the average value of the object changes, then repeat steps two to four. The process is stopped if there is no more object transfer between clusters.

5. Conclusions and recommendations.

3. RESULTS AND DISCUSSION

3.1 Analysis *Hierarchical Clustering*

a. Calculating the Euclidean distance between two district/city grouping objects in Bengkulu Province.

```
> jarak
           1          2          3          4          5          6          7
2   34640.371
3   79801.326  45939.076
4   35088.631  69565.105 114764.881
5   22527.888  12316.170  57397.884  57585.494
6   60656.566  26680.009  19267.394  95658.456  38192.751
7   68883.320  35258.247  11061.110 103789.963  46546.605   8774.577
8   60400.865  26925.681  19444.436  95329.375  38068.406   2719.518   8490.812
9  411110.524 445385.934 490907.732 376222.896 433565.383 471753.024 479975.557
10  57901.778  24835.067  22095.836  92766.044  35682.237   4972.459  11075.764
           8          9
2
3
4
5
6
7
8
9  471498.902
10   2915.605 468956.002
```

Figure 2. Euclidean distance matrix between two objects of Regency/City grouping in Bengkulu Province

The Muko-Muko Regency (data 1) and North Bengkulu (data 2) have the closest distance between other districts with a distance of 34640.37, which means that these two districts have similar characteristics. Meanwhile, those with the furthest distance between other regencies are Muko-Muko Regency and Lebong Regency, 79801, 326. It shows no similarity in characteristics between Muko-Muko and Lebong Regencies. Therefore, it can be concluded that if the distance between the two objects has a minimum value. Then, the two objects have similar characteristics and vice versa. The results of this distance calculation use equation 1.

b. Determination of the Best Cluster Method

A cluster validity test was conducted by looking at the value of the cophenetic correlation coefficient to get the best cluster method. The correlation coefficient between the original elements of the dissimilarity matrix (Euclidian distance matrix) and the elements generated by the cophenetic matrix is called the cophenetic correlation coefficient [16].
The grouping of districts or cities can be done by looking at the results of the comparison of the single linkage, complete linkage, and ward's methods by looking at the cophenetic correlation values as follows:

Table 3. Cophenetic correlation results

Single Method	Complete method	Ward method
0,9965	0,9966	0,9927

Based on the cophenetic correlation value, it can be seen that the best method used is the complete linkage method, because it has the highest correlation value, which is 0.99666.

c. Cluster Analysis Process using Complete Linkage Method

The complete linkage method is one of the methods in the agglomerative process or the merging method. The process of grouping or clustering can be presented with a dendrogram. The dendrogram of the results of the grouping of districts/cities in the case of the spread of COVID-19 using the complete linkage method is as follows:

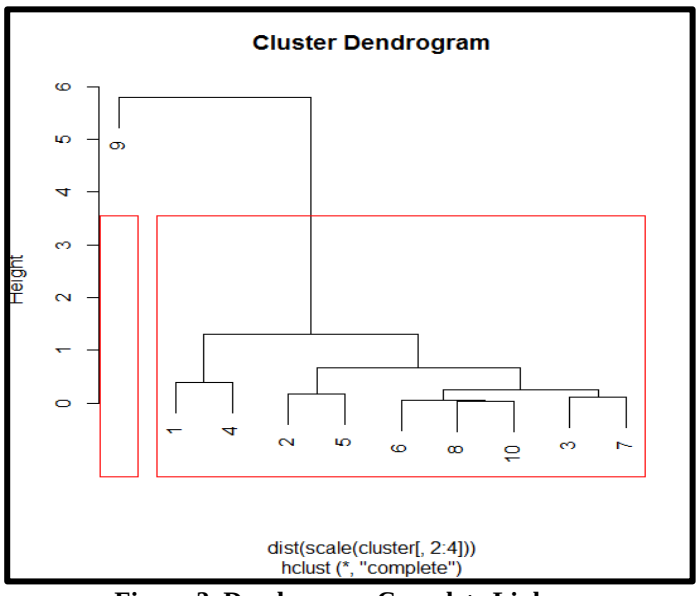

Figure 3. Dendrogram Complete Linkage

Therefore, the conclusion of the number of members of each cluster can be seen in Table 4, below:

Table 4. Members of each cluster

Cluster	Members	Country/city
1	9	Muko-muko, North Bengkulu, Lebong, Rejang Lebong, Kepahyang, South Bengkulu, Kaur, Seluma, Central Bengkulu
2	1	Bengkulu city

From Table 4, it can be seen that the regencies or cities in Bengkulu Province are grouped into 2 clusters (groups). Each cluster has its number or composition. The first cluster consists of 9 regencies, namely Muko-muko Regency, North Bengkulu, Lebong, Rejang Lebong, Kepahyang, South Bengkulu, Kaur, Seluma, Central Bengkulu while the second cluster consists of 1 city, namely Bengkulu City.

3.2 Determine the optimum number of clusters based on the Elbow and Silhouette method.

a. Elbow Method

Based on the results of the analysis using the Elbow method, the optimum cluster, as many as 2 clusters, can be seen in the image below:

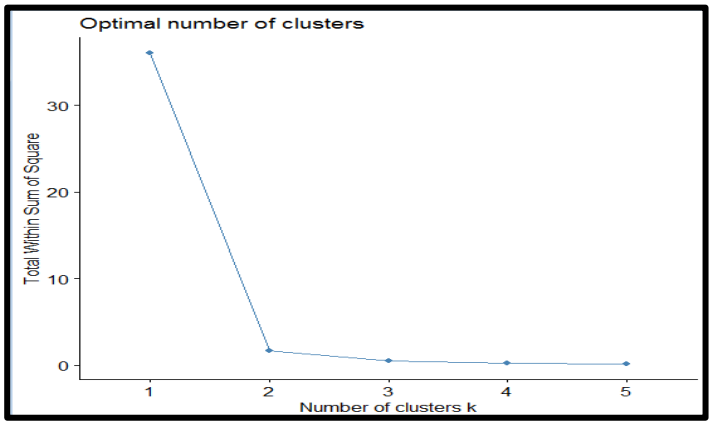

Figure 4. Graph of the optimal number of clusters using the Elbow method

In Figure 4, it can be seen that the value of K drops drastically and forms an elbow, which is at K=2.

b. Silhouette Method

The Silhouette method uses an average value approach to estimate the quality of the clusters formed. If the silhouette coefficient value is close to 1, the better the grouping of objects in one cluster. On the other hand, if the Silhouette Coefficient approaches the value of -1, the grouping of objects in one cluster becomes worse [13]. The silhouette coefficient value is presented in Figure 5, below:

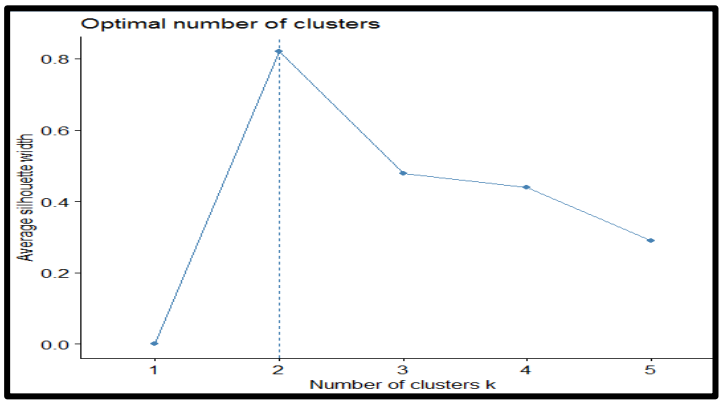

Figure 5. Graph of the optimal number of clusters in the Silhouette Method

Based on Figure 5, it is obtained that many optimum clusters are formed at k=2, because the average value is the highest and the dotted line shows K=2.

From the analysis of these two methods, it can be conluded that the optimum number of clusters obtained is K=2.

3.3 K-Means Clustering Analysis

With K-Means clustering, it is known that K=2, the classification of districts/cities is as follows:

Table 5. Within Cluster Sum of Squares

Cluster	Members	Country/city
1	9	Muko-muko, North Bengkulu, Lebong, Rejang Lebong, Kepahyang, South Bengkulu, Kaur, Seluma, Central Bengkulu
2	1	Bengkulu City

Based on the K-Means Clustering analysis, the variability value in each cluster is obtained or called Within Cluster Sum of Square (WCSS) which is summarized as follows:

Table 6. Within Cluster Sum of Square

Cluster	
1	**2**
1.635	0.0000

Table 6 shows that the value of variability within the cluster is relatively small, which means it is pretty good. Then, the variability value of the two clusters is obtained or called *between Cluster Sum of Squares* (BCSS) of 95.5%, which means that the variability value of the two clusters is relatively large, which means it is pretty good, so the results of the analysis can be said to be a good fit. The average value of each variable in the two clusters is as follows:

Table 7. Cluster Means

Cluster	Variable			
	Confirmed	Died	Recovered	Suspected
1	-0.307	-0.308	-0.307	-0.314
2	2.762	2.775	2.762	2.822

Based on Table 7, it can be concluded that cluster two has an average value of each variable that is higher than cluster 1. It means that cluster 2 is included in the category of higher confirmed, died, and suspected levels. It indicates that there is still high mobility of the population or community discipline related to social and physical distancing that may not be complied with. It is because Bengkulu City has the highest population compared to other regencies in Bengkulu Province, which is 373.59 thousand [6].

Cluster 2 shows a higher cure rate than cluster 1. If confirmed with additional data from BPS, Bengkulu City has many health services for private hospitals and hospitals, as many as 6, 22 health centers, 45 sub-health centers, 18 polyclinics, 285 doctors, 1375 nurses, 586 midwives, 296 pharmacists, and 137 nutritionists. [6].

4. CONCLUSION

Based on the results of the analysis, several conclusions were obtained from this study, namely:

1. The cophenetic correlation value shows that the best method for grouping the characteristics of active COVID-19 cases is complete linkage with similarity measurements using Euclidean Distance.
2. Districts/cities in Bengkulu Province were grouped using the complete linkage method into 2 clusters (groups). Each cluster has its number or composition. The first cluster consists of 9 regencies, namely Muko-muko Regency, North Bengkulu, Lebong, Rejang Lebong, Kepahyang, South Bengkulu, Kaur, Seluma, and Central Bengkulu while the second cluster consists of 1 city, namely Bengkulu City.
3. The method used to find the optimal number of clusters was the Elbow method and the Silhouette method. In this study, both methods resulted in the value of K=2.
4. Based on the K-Means clustering analysis results, it can be seen that cluster 2 has an average value of each variable that is higher than cluster 1, which means that cluster 2 is included in the category of higher confirmed, died, and suspected. It indicates that there is still high population mobility or

community discipline regarding social and physical distancing, which may not be fully complied with. It is because Bengkulu City has the highest population compared to other regencies in Bengkulu Province, which is 373.59 thousand [6]. However, cluster 2 also shows a higher recovery rate than cluster 1. It is because health facilities and health workers in Bengkulu City are pretty adequate. It can be seen from the data from BPS that Bengkulu City has some health services for private hospitals and public hospitals, as many as 6 22 health centers, 45 sub-health centers, 18 polyclinics, 285 doctors, 1375 nurses, 586 midwives, 296 pharmacists, and 137 nutritionists [6]

REFERENCES

[1] Y. A. E. d. S. Rizkiana Prima. R., "Research Gate," may 2020. [Online]. Available: https://www.researchgate.net/publication/342697385_Analisis_Cluster_Virus_Corona_COVID-19_di_Indonesia_pada_2_Maret_2020_-_12_April_2020_dengan_Metode_K-Means_Clustering. [Accessed August 2021].

[2] A. T. R. D. Raditya Novidianto, "Analisis KLaster Kasus Aktif Covid-19 Menurut Provinsi di Indonesia Berdasarkan Data Deret Waktu," ["Cluster Analysis of Active Covid-19 Cases by Province in Indonesia Based on Time Series Data,"] *Jurnal Aplikasi Statistika dan Komputasi Statistika*, vol. 12, no. 2, pp. 15-24, 2020.

[3] K. K. Achmad Solichin, "Klaterisasi Persebaran Virus Corona (Covid-19) di DKI Jakarta Menggunakan Metode K-Means," ["Clustering the Distribution of Corona Virus (Covid-19) in DKI Jakarta Using the K-Means Method,"] *Fountain of Informatics Journal*, vol. 5, no. 3, pp. 52-59, November 2020.

[4] J. A. T. ,. S. P. P. F. I. R. Z. Nayuni Dwitri, "Penerapan Algoritma K-Means Dalam Menentukan Tingkat Penyebaran Pandemi Covid-19 di Indonesia," ["The Application of the K-Means Algorithm in Determining the Spread of the Covid-19 Pandemic in Indonesia,"] *Jurnal Teknologi Informasi*, vol. 4, no. 4, pp. 128-132, Juni 2020.

[5] A. E. L. H. E. L. A. U. Mochammad Yusa, "Sistem Pakar: Implementasi Metode Bayes ProbabiltiesuntukPenentuan Kriteria PasienCOVID-19 Berdasarkan Fitur Gejala," ["Expert System: Implementation of the Bayes Probabilties Method for Determining the Criteria for COVID-19 Patients Based on Symptom Features," *Jurnal Teknologi Informasi dan Terapan (J-TIT)*, vol. 8, no. 5, pp. 13-20, Juni 2021.

[6] B. P. S. P. Bengkulu, PROVINSI BENGKULU DALAM ANGKA [BENGKULU PROVINCE IN NUMBERS], B. P. S. P. Bengkulu, Ed., Bengkulu: Badan Pusat Statistik Provinsi Bengkulu, 2021.

[7] M. H. Arief Rachman, "Klaterisasi Sumber Penyebaran Virus Covid-19 dengan Menggunakan Metode K-Means di Daerah Kota Cimahi dan Kab.Bandung Barat," ["Clustering the Sources of the Spread of the Covid-19 Virus using the K-Means Method in the Cimahi City and West Bandung Districts,"] *Jurnal Teknik : Media Pengembangan Ilmu dan Aplikasi Teknik*, vol. 19, no. 6, pp. 62-72, November 2020.

[8] B. S. A. R. B. Tahta Alfina, "Analisa Perbandingan Metode Hierarchical Clustering, K-Means dan Gabungan Keduanya dalam Cluster Data (Studi Kasus: Problem Kerja Praktek Jurusan Teknik Industri ITS)," ["Comparative Analysis of Hierarchical Clustering Methods, K-Means and their Combination in Cluster Data (Case Study: Practical Work Problems, ITS Industrial Engineering Department)," *Jurnal Teknik ITS*, vol. 1, no. 8, pp. A-521-A-525, September 2012.

[9] S. W. S. E. S. Lynda Rahmawati, "ANALISA CLUSTERING MENGGUNAKAN METODE K-MEANS DAN HIERARCHICAL CLUSTERING (STUDI KASUS : DOKUMEN SKRIPSI JURUSAN KIMIA, FMIPA, UNIVERSITAS SEBELAS MARET)," ["CLUSTERING ANALYSIS USING K-MEANS AND HIERARCHICAL CLUSTERING METHODS (CASE STUDY: THESIS DOCUMENT DEPARTMENT OF CHEMISTRY, FMIPA, UNIVERSITY ELEMENTA MARCH)," *Jurnal Universitas Sebelas Maret*, vol. 2, pp. 66-73, 2014.

[10] R. AWALIAH, "ANALISIS CLUSTERING UNTUK MENGELOMPOKKAN TINGKAT," Makassar, 2018.

[11] R. WIndasari, "Analisis Cluster Hirarki Metode Average Linkage Berdasarkan Jumlah Kriminalitas di INdonesia Tahun 2019," ["Analysis of Hierarchical Clusters of Average Linkage Methods Based on the Number of Crimes in Indonesia in 2019,"] Malang, 2020.

[12] Y. E. A. Sunarso, "Analisis Cluster dan Aplikasinya," ["Cluster Analysis and Its Applications,"] Yogyakarta, 2008.

[13] G. R. Prima, "Analisa Perbandingan Nilai K Terbaik Untuk Clustering K-Means Menggunakan Pendekatan Elbow dan Silhouette Pada Citra Aksara Jawa," ["Comparative Analysis of the Best K Values for K-Means Clustering Using Elbow and Silhouette Approaches on Javanese Script Imagery,"] Yogjakarta, 2021.

[14] N. R. Y. Y. Wiyli Yustanti, "Klastering Wilayah Kota/Kabupaten Berdasarkan Data Persebaran Covid-19 di Provinsi Jawa Timur dengan Metode K-Means," ["Clustering of City/Regency Areas Based on Data on the Distribution of Covid-19 in East Java Province with the K-Means Method,"] *Journal Information Engineering and Educational Technology*, pp. 1-8, 2020.

[15] S. JUMAROH, "Jumaroh, S. 2015. Analisis Klaster K-Means Dari Data Luas Grup Sunspot dan Data Grup Sunspot Klasifikasi Mc. Intosh Yang Membangkitkan Flasre X-Ray dan H_α," [K-Means Cluster Analysis From Sunspot Group Area Data and Sunspot Group Data Mc Classification. The Intosh That Generated the X-Ray and H_α Flashes,"] Malang, 2015.

[16] "ANALISIS CLUSTER PENDERITA DISABILITAS MENTAL DI PROVINSI DAERAH ISTIMEWA YOGJAKARTA TAHUN 2016," ["CLUSTER ANALYSIS OF PEOPLE WITH MENTAL DISABILITIES IN THE PROVINCE OF THE SPECIAL REGION OF YOGYAKARTA IN 2016,"] YOGJAKARTA, 2018.