OPEN ACCESS
ISSN 2460-9056
http://socj.telkomuniversity.ac.id/ijoict/

Intl. Journal on ICT
Vol. 8, No. 1, June 2022. pp. 1-9
doi: 10.21108/ijoict.v8i1.606

Java Island Health Profile Clustering using K-Means Data Mining

Muhammad Andryan Wahyu Saputra[1*], Sri harini[2]

[1,2] *Magister of Informatics, Faculty of Science and Technology, Maulana Malik Ibrahim State Islamic University Malang*
Jl. Gajayana No. 50, Malang, Jawa Timur, Indonesia

[*] 210605210010@student.uin-malang.ac.id , sriharini@mat.uin-malang.ac.id

Abstract

This study aims to use the Clustering Algorithm in grouping each region (regency and city) on the island of Java to determine the level of health so that counseling, services, and providing assistance can be more accurate and on target. Sources of data for this study came from the Central Bureau of Statistics and the Ministry of Health. The data used in this study is data consisting of 85 regencies and 34 cities on the island of Java. The method used in this study is the K-means algorithm by determining the optimum number of clusters using the Elbow method. The data will be processed by clustering into 4 clusters, namely clusters of high health quality levels, clusters of fairly high levels of health quality, clusters of fairly low and low levels of health quality. This results in an assessment of 1 city with a high level of health quality, namely Central Jakarta, 53 districts/cities with a fairly high level of health quality, 52 districts/cities with a fairly low level of health quality, and 13 other provinces including a low level of health quality. This can be an input for the government to pay more attention to each area that has low health quality through improving public health services so that the population of each area becomes healthier.

Keywords: Clustering, Datamining, K-Means, Java Island Health Profile

I. INTRODUCTION

Indicators of the success of a country's development can be seen from the level of achievement of the country in providing health insurance. In Indonesia, the government has set several indicators to measure the progress of health development in all provinces, districts, and sub-districts.

Every year, the Ministry of Health collects, and processes population health data to produce provincial and district/city health rankings. The results of this analysis and ranking become an important reference for the government to formulate a work plan to improve health status in its region. In addition, the results of this analysis serve as the basis for the central government in formulating health problem areas and as the basis for determining the allocation of central government health funding to regions to support the Ministry of State for Vulnerable Regions in regional/urban development [1].

Regional clustering is carried out to determine the level of health in the region, especially Java Island. Quoted from the Central Bureau of Statistics, Java Island consists of 85 regencies and 34 cities. So far, the grouping of environmental health indicator data is still based on manual computing techniques, where the calculations still have several problems, especially in the field of data consistency. In terms of improving the health services of the Health Service, it is necessary to have a system that groups healthy areas based on environmental health data, so that counseling, services, and helping can be more accurate and on target.

K-Means is a non-hierarchical clustering technique that tries to partition existing data into one or more clusters. This method partitions data into clusters, so that data with the same characteristics are grouped into

Received on 27 Dec 2021. Revised on 17 April 2022. Accepted and Published on 21 June 2022.

the same cluster and data with different characteristics are grouped into different clusters. K-Means Clustering is a data analysis technique or data mining technique that performs an unsupervised modeling process (unsupervised) and is a data grouping technique using a partition system. There are two types of data clustering that are often used in the data grouping process, namely hierarchical and non-hierarchical, and K-Means is a method of clustering non-hierarchical data or segregated clustering [2]. K-Means clustering tries to group existing data into groups. Here, the data in one group have the same characteristics with each other and different characteristics from the data in other groups. The k-means method can group a large number of observational objects with the proximity of their distance so that they form one or more groups with similar characteristics in each group [3].

The main purpose of this study is to divide each region (district and city) into groups, and the results of the grouping can be used as a reference to describe the distribution of cities/groups based on the health profile of the district.

II. LITERATURE REVIEW

The application of data mining in the health sector has been put forward by many researchers, because it has the ability to extract large amounts of data to obtain useful information [4]. Data mining techniques can be widely used in hospitals, clinics and pharmacies by health care providers to provide better and more affordable healthcare to patients [5]. Clustering is a technique of grouping records in a database based on certain conditions. The basic concept of clustering is to group several objects into clusters, where a good cluster is a cluster that has a high degree of similarity between objects in the cluster and a high degree of dissimilarity with other cluster objects. The reason for using the K-means clustering algorithm is because this algorithm has a fairly high accuracy and is efficient in processing large numbers of objects. In addition, the advantage of the K-Means method is that it is flexible because the user can determine the number of clusters to be created [6].

Based on a previous study entitled "K-Means Algorithm for Grouping Patients' Diseases at the Cigugur Tengah Health Center" in 2020, in this study, patients were grouped based on acute and non-acute diseases using data mining techniques with clustering methods with the k-means algorithm and k-medoids algorithm as a comparison. Based on the test results from the k-means algorithm and the k-medoids algorithm, the Davies Bouldin value for the k-means algorithm is -0.453 and the k-medoids algorithm is -1.276. From these results it can be said that the algorithm that produces the smallest Davies Bouldin value is considered a better algorithm, it can be concluded that the kmeans algorithm is better than the k-medoids algorithm [7].

Another study entitled "Implementation of the K-Means Clustering Algorithm on the Spread of Acute Respiratory Tract Infections (ARI) in Riau Province" in April 2021 found that the K-Means algorithm was able to classify the causes of the spread of ARI into 2 clusters, of which cluster 1 provides high recommendations totaling 10 districts, cluster 2 providing low recommendations totaling 2 districts. In this test, using two applications, the first using a Microsoft Excel number processing application and using the Rapidminer 5.3 application using K-Means to display 2 clusters from the classification results [8].

There is also a study entitled "Application of the K-Means Clustering Analysis Algorithm in Human Infectious Diseases (Case Study of Majalengka Regency)" which states that the K-means algorithm is not affected by the order of objects used, this is proven when the author tries to randomly determine the central starting point. cluster of any of the objects at the start of the computation. The number of cluster memberships generated is the same when using another object as the starting point for the cluster center. However, this only affects the number of iterations performed. Grouping objects (object clustering) is one of the processes of object mining that aims to partition existing objects into one or more object clusters based on their characteristics. This study examines how the K-means Cluster Analysis Algorithm is used in case studies of human infectious diseases on an object. This study examines the K-means Cluster Analysis method in infectious diseases in humans based on the set of variables formed per sub-district of each health center, there are 32 Puskesmas offices in Majalengka Regency [9].

Based on the results of previous studies, the k-means algorithm has better results than other algorithms, and the k-means algorithm is also able to group health data. The difference between this study and previous research is that there is no research for grouping districts/cities in Java based on the level of health quality with 8 parameters using a clustering algorithm that is often used in previous studies, namely k-means, which is useful for relevant agencies in giving more attention to each areas with low health quality.

III. RESEARCH METHOD

A. Level of Data Collection

Data collection for this study cites publications from the Central Statistics Agency and the Ministry of Health. This study uses data from 2017 [10], [11].

The data in this study were sourced from the publications of the Central Statistics Agency and the Ministry of Health. This study uses data from 2017 with datasets from 85 districts and 34 cities on the island of Java. The variables used in this study are :

- Life expectancy
- Health Center Ratio
- Hospital Ratio
- Percentage of PHBS RT
- Percentage of Eligible Sanitation RT
- Percentage of Babies with LBW
- Percentage of Babies Who Get Exclusive Breastfeeding
- Diarrhea Pain Rate

B. Data Processing Stage

Processed data is mainly processed for grouping. The previous step summarizes the data for each city for each aspect, so that in this step the units and data coverage used are different, so the data obtained is first standardized and then processed, in the grouping or clustering step [12].

C. Clustering Stage

1. K - Means Algorithm

K-Means clustering is a distance-based clustering technique that tries to partition data into multiple clusters [8]. K-Means Clustering is a cluster analysis method that determines the number of k clusters that the user wants to create. Object grouping is based on the shortest distance between objects and the average of their centroid/centroid/cluster. The working principle or clustering algorithm of K-Means is as follows: [13].

1) Determine k as the number of clusters to be formed
2) Start k with randomly generated cendroid
3) Calculate the distance from each data point to each centroid using the Euclidean distance equation as shown in equation (1) below.

$$D(P,Q) = \sqrt{\sum_{j=1}^{p} (x_j(P) - x_j(Q))^2} \quad (1)$$

where : D = distance between P and Q
P = data size

4) Grouping data based on the shortest distance between the data and the centroid
5) Determine the location of the new centroid (k) If the position of the new centroid and the old centroid do not match, repeat step 3.

2. Elbow Method

Elbow method is one method to determine the optimal number of cluster. The algorithm to find the optimal cluster value using the elbow method is as follows [14].

1. Initialize cluster value
2. Increase the cluster value
3. Calculate the Sum of Squared Error value from each cluster
4. See a drastic reduction in the number of squared errors (SSE).
5. Set the cluster value in the form of a square.

D. Analysis Stage

In this step, the weight of each index is calculated, and the data is analyzed with the received and processed data. After determining the number of clusters in the previous step, analyze the result.

IV. RESULTS AND DISCUSSION

A. Data Processing

Packages prepared for the K-Means Algorithm:

packages("factoextra")
Library("factoextra")
packages("tidyverse")
Library("tidyverse")

The implementation of the K-Means Algorithm in R-Studio has several stages, the first stage is to install and load the package, then data preparation, then look for the number of clusters using the elbow plot method, the last is the execution of K-Means :

Data Preparation

This stage is to prepare the dataset that will be used in the Rstudio process such as entering data into Rstudio, cleaning the data, and selecting the data needed in the clustering process at Rstudio. Because the data used has different units and ranges, it is necessary to standardize the data first before entering into cluster analysis [15]. To standardize the data into standard normal in R, we use the scale function (). The result of standardized data can be seen in Table 1.

Table 1. Display of Standardized Result Data

AHH	RP	RRS	PHBS	SL	BBLR	ASI	DIARE
-0.3497	0.9378	-0.5358	-1.0647	-0.1952	1.4049	0.5036	0.3176
0.0013	0.4256	-0.4127	0.1147	0.5566	0.5085	1.1529	0.3177
0.3231	0.1675	-0.8676	-1.468	-0.3077	0.2695	0.4491	0.3177
0.4621	0.1213	0.064	-1.3152	0.3571	-0.3281	0.2799	0.3176
0.2646	-0.5572	-0.4096	-0.8508	0.5873	0.0305	1.4257	0.3177
-0.006	-0.3653	-0.4537	-0.5208	-0.2616	-0.4476	0.8037	0.3177
-0.0535	-0.9324	-0.2434	-0.7042	-0.0059	-0.6268	0.4382	0.3177
-1.0117	-0.3377	-0.5053	-1.9141	-0.4918	0.6878	1.142	0.3177
-1.3627	-0.5722	-0.5418	0.2858	-0.8089	1.5244	1.1856	0.3177
-0.7593	-0.0776	-0.2604	-1.0097	0.3571	-0.3878	0.9237	0.3177

B. Clustering Stage

In the k-means method, the optimum number of clusters to group objects properly can be detected using the Elbow Plot method. Elbow Plot is a plot between many clusters with a total within-cluster variation (total of deviations per cluster). Many of the selected clusters are part of the "elbow" or point where there is a sharp decline before that point and followed by a not sharp decline after that point. This is because the addition of the number of clusters does not have much effect on the variations that exist within the cluster.

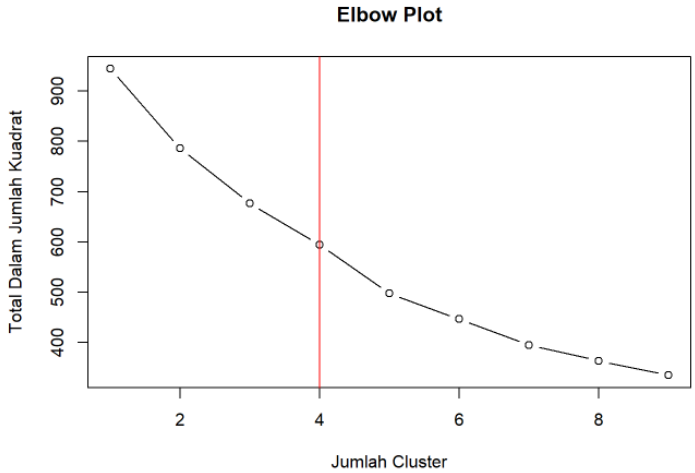

Fig 1. Display of the Number of Clusters with Elbow Plot

Figure 1 shows that the Elbow plot is truncated to the number of clusters of 4, this means that the optimum cluster for the grouping of Health profiles in Java can be grouped into four.

C. Analysis Stage

1. Cluster Analysis

Cluster analysis in R can be done with the kmeans() function and using standardized data. The results of the implementation of this function kmeans() be seen in Fig. 2

```
## K-means clustering with 4 clusters of sizes 1, 53, 52, 13
##
## Cluster means:
##          AHH         RP        RRS       PHBS         SL        BBLR        ASI
## 1  0.5718000 -1.3599000  1.5470000  0.4996000 -1.3642000 -1.69050000 -0.3508000
## 2  0.5887849 -0.2675434 -0.1618604  0.7283509  0.3466830 -0.09940377 -0.6497566
## 3 -0.6949077  0.1299365 -0.4348115 -0.6140654 -0.5387519  0.01485000  0.4986577
## 4  0.3352308  0.6756692  2.2801462 -0.5515462  0.8465077  0.47590000  0.6813308
##         DIARE
## 1  9.72220000
## 2 -0.26004906
## 3  0.08207308
## 4 -0.01596923
##
## Clustering vector:
##   [1] 3 3 3 3 3 3 3 3 3 3 3 3 3 2 3 3 3 3 3 2 3 3 2 2 3 3 3 3 4 4 4 3 3 4 4 2
##  [38] 4 2 2 2 2 2 2 3 3 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 3 4 2 2 2 2 2 2
##  [75] 3 3 2 3 3 3 2 2 3 3 2 2 3 3 2 3 3 2 3 2 4 2 2 3 3 3 3 3 3 3 2 2 3 2 4 4 3
## [112] 4 4 1 2 2 2 3 4
##
## Within cluster sum of squares by cluster:
## [1]   0.00000 228.19076 276.63763  69.41555
##  (between_SS / total_SS =  39.2 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
## [6] "betweenss"    "size"         "iter"         "ifault"
```

Fig 2. Display the output of the function '*kmeans'*

The output of the k means() function consists of the following information:

- cluster : vector containing the cluster location of each object.
- centers : matrix containing the centroid/average value of each cluster.
- withinss: vector that contains the deviation of each cluster that is formed.
- tot.withinss : the total deviation of each cluster that is formed. Usually used to make Elbow Plots to find out how many clusters to choose.
- size : number of objects in each cluster

2. Returns the Centroid Value

The contained means/centroid value is still in the standardized value, so the value needs to be returned to the initial value in order to know the true characteristics of each cluster. The implementation of this function and the results be seen in Fig. 3

```
jawa_data %>%
  mutate(Klaster = kmeans_clustering$cluster) %>%
  group_by(Klaster) %>%
  summarise(Mean_AHH = mean(AHH), Mean_RP = mean(RP), Mean_RRS = mean(RRS), Mean_PHBS = mean(PHBS), Mean_SL = mean(SL), Mean
_BBLR = mean(BBLR), Mean_ASI = mean(ASI), Mean_DIARE = mean(DIARE))
```

```
## # A tibble: 4 x 9
##   Klaster Mean_AHH Mean_RP Mean_RRS Mean_PHBS Mean_SL Mean_BBLR Mean_ASI
##     <int>    <dbl>   <dbl>    <dbl>     <dbl>   <dbl>     <dbl>    <dbl>
## 1       1     73.8   0.260    3.04       71.3    50.0      0.92     55.5
## 2       2     73.9   0.755    0.990      75.0    83.5      3.58     50.1
## 3       3     70.4   0.935    0.663      53.1    66.2      3.77     71.1
## 4       4     73.2   1.18     3.92       54.1    93.3      4.55     74.5
## # ... with 1 more variable: Mean_DIARE <dbl>
```

Fig 3. Display of Centroid Value Return Output

3. Grouping Objects into Clusters

Grouping objects into clusters that are formed can be seen from the cluster output in Fig 4.

```
jawa_data %>%
  mutate (Klaster = kmeans_clustering$cluster) %>%
  select(`Kota/Kab`, Klaster) %>%
  arrange(Klaster)
```

```
## # A tibble: 119 x 2
##    `Kota/Kab`          Klaster
##    <chr>                 <int>
##  1 KOTA JAKARTA PUSAT        1
##  2 KAB. SIDOARJO             2
##  3 KAB. NGAWI                2
##  4 KAB. LAMONGAN             2
##  5 KAB. GRESIK               2
##  6 KOTA SURABAYA             2
##  7 KAB. CILACAP              2
##  8 KAB. BANYUMAS             2
##  9 KAB. PURBALINGGA          2
## 10 KAB. BANJARNEGARA         2
## # ... with 109 more rows
```

Fig 4. Output Display of Grouping Objects Into Clusters

4. Cluster Interpretation

To make it easier to understand and interpret, a table is made as follows:

Table 2. Interpretation of Regency/City Cluster

Cluster 1	Central Jakarta City
Cluster 2	Sidoarjo Regency, Ngawi Regency, Lamongan Regency, Gresik Regency, Surabaya City, Cilacap Regency, Banyumas Regency, Purbalingga Regency, Banjarnegara Regency, Kebumen Regency, Purworejo Regency, Boyolali District, District. Klaten, Sukoharjo Regency, Wonogiri Regency, Karanganyar Regency, Sragen Regency, Grobogan Regency, Blora Regency, Rembang Regency, Pati Regency, Kudus Regency, Jepara Regency, Demak Regency, Semarang Regency, Temanggung Regency, Kendal Regency, Batang Regency, Pekalongan Regency, Pemalang Regency, Tegal Regency, Surakarta City, Salatiga City, Semarang City, Pekalongan City, Tegal City, Bogor Regency, Bandung Regency, Kuningan Regency, Cirebon Regency, Indramayu Regency, Subang Regency, Bekasi Regency, Bogor City, Bandung City, Bekasi City , Depok City, Tangerang City, Cilegon City, South Tangerang City, North Jakarta City, West Jakarta City, South Jakarta City
Cluster 3	Pacitan Regency, Ponorogo Regency, Trenggalek Regency, Tulungagung Regency, Blitar Regency, Kediri Regency, Malang Regency, Lumajang Regency, Jember Regency, Banyuwangi Regency, Bondowoso Regency, Situbondo Regency, Probolinggo Regency, Pasuruan Regency, Mojokerto Regency, Jombang Regency, Nganjuk Regency , Madiun Regency, Magetan Regency, Bojonegoro Regency, Tuban Regency, Bangkalan Regency, Sampang Regency, Pamekasan Regency, Sumenep Regency, Probolinggo City, Pasuruan City, Wonosobo Regency, Magelang Regency, Brebes Regency, Sukabumi Regency, Cianjur Regency, Garut Regency, Regency Tasikmalaya, Ciamis Regency, Majalengka Regency, Sumedang Regency, Purwakarta Regency, Karawang Regency, West Bandung Regency, Pangandaran Regency, Sukabumi City, Cimahi City, Tasikmalaya City, Banjar City, Lebak Regency, Pandeglang Regency, Serang Regency, Tangerang Regency, Serang City , Gunung Kidul Regency, East Jakarta City
Cluster 4	Kediri City, Blitar City, Malang City, Mojokerto City, Madiun City, Batu City, Magelang City, Cirebon City, Kulon Progo Regency, Bantul Regency, Sleman Regency, Yogyakarta City, Seribu Islands

Table 2 The effect of the variables on average life expectancy, ratio of health centers, ratio of PHBS Hospital = Percentage of RT PHBS, Percentage of RT with proper sanitation, percentage of infants with low birth weight, percentage of infants receiving exclusive breastfeeding, diarrhea pain rate can group health profiles in Java Island as follows:

- Cluster 1 contains 1 Cities/Districts with high health quality in Java Island.
- Cluster 2 contains 53 Cities/ Districts with a fairly high quality of health in Java
- Cluster 3 contains 52 cities/districts with fairly low health quality in Java.
- Cluster 4 contains 13 Cities/Districts with low health quality in Java.

5. Cluster Plot

We can display the visualization of 4 clusters obtained from the previous step using the fviz_cluster function. The results can be seen in Fig. 6.

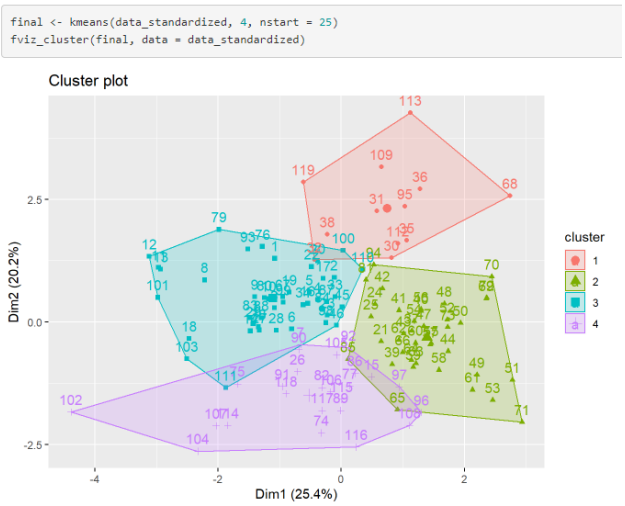

Fig. 6. Plot of Health Profile in Java Island

V. CONCLUSION

Based on the results of the analysis of the K-means method to determine the degree of health based on 8 parameters, there were 4 clusters of health profiles in Java Island with 1 district/city high health quality cluster, 53 districts/cities with moderately high health quality cluster 2, 52 districts/municipalities with health quality is quite low, and the quality of health is low there are 13 districts/cities. This can be an input for the government to pay more attention to each area that has low health quality through improving public health services so that the population of each area becomes healthier.

REFERENCES

[1] Nishom, M., Sharfina Febbi Handayani, and Dairoh Dairoh. "Pillar Algorithm in K-Means Method for Identification of Health Human Resources Availability Profile in Central Java." JUITA: Journal of Informatics 9.2 (2021): 145-152.

[2] IH Rifa, H. Pratiwi, and R. Respatiwulan, "Clustering of Earthquake Risk in Indonesia Using K-Medoids and K-Means Algorithms," Media Stat., vol. 13, no. 2, pp. 194–205, 2020.

[3] Islami, Bagus Muhammad, Cepy Sukmayadi, and Tesa Nur Padilah. "Clustering of Health Facilities Based on Districts in Karawang With the K-Means Algorithm." BINA INSANI ICT JOURNAL 8.1 (2021): 83-92.

[4] M. Harahap, AM Husein, S. Aisyah, FR Lubis, and BA Wijaya, "Mining association rule based on the diseases population for recommendation of medicine need," J. Phys. conf. Ser., vol. 1007, no. 1, pp. 0–11, 2018, doi:10.1088/1742-6596/1007/1/012017.

[5] Z. Ceylan, S. Gürsev, and S. Bulkan, "An Application of Data Mining in Individual Pension Savings and Investment System," no. January, pp. 7–11, 2018.

[6] M. Mardalius, "Utilization of Rapid Miner Studio 8.2 for Grouping of Accessories Sales Data Using the K-Means Algorithm," Jurteksi, vol. 4, no. 2, pp. 123–132, 2018, doi:10.33330/jurteksi.v4i2.36.

[7] Sugianto, Castaka Agus, Ayu Hendrati Rahayu, and Aditia Gusman. "K-means algorithm for grouping patient's disease at Puskesmas Cigugur Tengah." Journal of Information Technology 2.2 (2020): 39-44.
[8] Bastian, Ade. "The application of the k-means clustering analysis algorithm on human infectious diseases (case study in Majalengka district)." Journal of Information Systems 14.1 (2018): 28-34.
[9] Purba, Ninaria, Poningsih Poningsih, and Heru Satria Tambunan. "Application of the K-Means Clustering Algorithm in the Spread of Acute Respiratory Infections (ARI) in Riau Province." Journal of Information Systems Research (JOSH) 2.3 (2021): 220-226.
[10] Ministry of Health of the Republic of Indonesia, "Indonesian Health Profile 2017," Available: https://pusdatin.kemkes.go.id/resources/download/pusdatin/profil-kesehatan-indonesia/Profil-Kesehatan-Indonesia-tahun-2017.pdf, 2018. [Online]. [Accessed: 16-Dec-2021].
[11] Central Bureau of Statistics, "Profile of Health Statistics 2019," Available: https://www.bps.go.id/publication/2019/12/30/9d583b7e2bd81fada82375e0/profil-statistik-kesehatan-2019.html, 2019. [Online] . [Accessed: 23-Dec-2021].
[12] Siregar, M. Hasyim. "Data Mining Clustering Sales of Building Tools Using the K-Means Method (Case Study at Adi Building Stores)." Journal of Technology And Open Source 1.2 (2018): 83-91.
[12] Farissa, Riva Arsyad, Rini Mayasari, and Yuyun Umaidah. "Comparison of K-Means and K-Medoids Algorithm for Grouping Drug Data with Silhouette Coefficient at Karangsambung Health Center." Journal of Applied Informatics and Computing (JAIC) 5.2 (2021): 109-116.
[13] Agus Nur Khormarudin, "Data Mining Techniques: K-Means Clustering Algorithm," Available: https://ilmucomputer.org/wp-content/uploads/2018/05/agus-k-means-clustering.pdf, 2018. [Online]. [Accessed: 23-Dec-2021].
[14] Nishom, M., Sharfina Febbi Handayani, and Dairoh Dairoh. "Pillar Algorithm in K-Means Method for Identification of Health Human Resources Availability Profile in Central Java." JUITA: Journal of Informatics 9.2 (2021): 145-152.
[15] Rofiqo, Nurul, Agus Perdana Windarto, and Dedy Hartama. "Application of Clustering on Residents Who Have Health Complaints With K-Means Datamining." KOMIK (National Conference on Information and Computer Technology) 2.1 (2018).