ENTHUSIASTIC
INTERNATIONAL JOURNAL OF STATISTICS AND DATA SCIENCE
Volume 2, Issue 1, April 2022, pp. 10-18

# Application of K-Medoids Clustering to Increase the 2020 Family Planning Program in Sleman Regency

Syintya Febriyanti [a,1,*], Jaka Nugraha [a,2]

[a] Department of Statistics, Universitas Islam Indonesia, Jalan.Kaliurang Km14.5, Yogyakarta 55584, Indonesia
[1] syintyaf11@gmail.com*; [2] jaka.nugraha@uii.ac.id
* corresponding author

ARTICLE INFO

**Article history**
Received August 21, 2021
Revised April 20, 2022
Accepted April 27, 2022

**Keywords**
Family Planning Program
K-Medoids Clustering
Sleman Regency

ABSTRACT

Indonesia is a country with a large population. Based on the results of the 2020 census, Indonesia's population ranks fourth in the world. The Indonesian government has made a policy to reduce population growth, namely the Family Planning Program or Keluarga Berencana (KB). One of the areas that did not escape the target was the DIY. Based on BKKBN DIY data, there is a significant difference between the number of active KB participants and the number of couples of childbearing ages, the number of KB equipment and the number of KB health facilities that exist between sub-districts in Sleman Regency. Then the sub-district classification is carried out based on the 2020 KB data in Sleman Regency using the K-Medoids Clustering method. This study aims to see the sub-district grouping used as a reference by the government in increasing active KB participants in the community to overcome the population in Yogyakarta, primarily focusing on Sleman. The categories in each cluster, namely Cluster 1, which consists of 6 sub-districts, have a high level of KB active participants, couples of reproductive ages, KB equipment, and KB health facilities. Then Cluster 2, which consists of 6 sub-districts, has a medium level of KB active participants, couples of reproductive ages, KB equipment, and KB health facilities. While Cluster 3 consists of 5 sub-districts, where KB active participants, teams of reproductive age, KB equipment, and KB health facilities are low level.

## 1. Introduction

Indonesia is a country with a vast population. Based on the results of the population census in 2020, the total population of Indonesia is 270.20 million people and ranks fourth in the world [1]. The rate of population growth continues to increase every year. If not controlled, the following year, there will be a large population explosion and can pose threats and losses such as poverty and hunger.

The government has made a policy to suppress the population growth rate, namely the Family Planning Program or Keluarga Berencana (KB) in Indonesian words. The program to increase family planning participation during the Covid-19 pandemic is an essential step in controlling the


p-ISSN 2798-253X
e-ISSN 2798-3153

population and other social problems [2]. The Province of the Special Region of Yogyakarta, or DIY Province, is one area that does not escape the target of the national family planning program. The family planning program in DIY Province has been implemented for quite a long time. However, in its implementation, there are still obstacles. Based on data from Badan Kependudukan dan Keluarga Berencana Nasional (BKKBN) DIY, it is known that active family planning participants in DIY Province in the last five years have decreased [3]. Likewise, Sleman Regency, which is one of the regencies in DIY, has the largest population [4].

Minimizing population growth can be done in various ways. One way is to classify population and national family planning using clustering analysis. Clustering is one of the data mining techniques that aims to group data based on information found in the data [5]. This method seeks to group/classify objects into the same group. Suppose it has been collected into the same group or cluster. In that case, the results can be used to see the potential in the Sleman Regency that can help the government pay more attention to sub-districts that have superior or fewer numbers to increase further the number of family planning programs in Sleman Regency.

To evaluate the government to improve family planning programs in the community to overcome population density in the DIY Province by focusing on one of the sub-districts, namely Sleman Regency.

## 2. The Proposed Method

### 2.1. Cluster Analysis

Cluster analysis is a classification technique that is used to classify objects or cases (respondents) into relatively homogeneous groups, called clusters. Objects/cases in each group tend to be similar and much different from objects from other clusters. In addition, each object only belongs to one group. There is no overlapping or interaction [6].

### 2.2. K-Medoids Clustering

K-medoids clustering is one of the partitioning or non-hierarchical clustering methods used in this study. K-medoids clustering, also known as Partitioning Around Medoids (PAM), is a variant of the K-Means method. It is based on the use of medoids instead of observing the mean held by each cluster to reduce the sensitivity of the partition concerning the extreme values present in the dataset. K-medoids clustering exists to overcome the weakness of k-means clustering, which is sensitive to outliers because an object with an enormous value may substantially deviate from the data distribution [7].

The steps of K-Medoids clustering are as follows:

1. Initialize *k* cluster centers (number of clusters)
2. Allocate each data (object) to the nearest cluster using the Euclidean Distance measure equation with the following equation:

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

   where $d_{ij}$ is distance between object i and j, $x_{ik}$ is value of object i in the k variable, $x_{jk}$ is object value j in the k variable, and p is number of observed variables.
3. Randomly select the object of each cluster as a candidate for a new medoid
4. Calculate the distance of each object in each cluster with the new candidate medoid.
5. Calculate the total deviation (*S*) by calculating the new total distance minus the old total distance. If $S < 0$, then swap objects with cluster data to form a new set of k objects as medoids.
6. Repeat steps 3 to 5 until there is no medoid change to obtain clusters and their respective cluster members.


p-ISSN 2798-253X
e-ISSN 2798-3153

### 2.3. Cluster Analysis Assumptions

Group analysis has two assumptions, namely a representative sample (population) and no multicollinearity. To determine whether the sample used is representative of the population, it can be seen from the Kaisen Meyer Olkin (KMO) value. KMO is a comparison index of correlation coefficient value to partial correlation.

$$\mathbf{KMO} = \frac{\sum_{i=1}^{p}\sum_{j=1}^{p} r_{ij}^2}{\sum_{i=1}^{p}\sum_{j=1}^{p} r_{ij}^2 + \sum_{i=1}^{p}\sum_{j=1}^{p} a_{ij}^2} \quad (2)$$

where p is number of variables; $r_{ij}$ is correlation coefficient between i and j variables, and $a_{ij}$ is partial correlation coefficient between i and j variables.

If the KMO value is < 0.5, then the sample does not represent the population. While if the KMO value is > 0.5, then the sample represents the population so that it is feasible for cluster analysis to be carried out [8].

Multicollinearity is a linear relationship that exists between independent variables. Multicollinearity can be seen from the value of the Variance Inflation Factor (VIF).

$$\mathbf{VIF} = \frac{1}{(1-R_i^2)} \quad (3)$$

where $R_i^2$ is coefficient of determination of the *i* independent regression.

The data is multicollinear if the VIF value is > 10 and the tolerance value is < 0.1. On the other hand, the data is said to be non-multicollinear if the VIF value is < 10 and the tolerance value is > 0.1 [9].

### 2.4. Measure of Distance

The main objective of group analysis is to divide a set of objects into several groups based on the size of the similarity between the objects used in terms of the characteristics used. The smaller the distance between an individual and another individual, the greater the resemblance of the individual so that the individual will be included in the same group. The most widely used distance measures are Euclidean and Manhattan. Euclidean is used when you want to give the shortest distance between two points (straight distance), while Manhattan offers the furthest distance between two data points. Manhattan is also often used because of its ability to better detect exceptional circumstances [10].

## 3. Method

### 3.1. Data Source

This research obtained secondary data, taken from Badan Kependudukan dan Keluarga Berencana Nasional (BKKBN) of Yogyakarta, namely data on the national family planning program in Sleman Regency in 2020 obtained from the BKKBN Website.

### 3.2. Research Variable

This research analyses the variables presented in Table 1.

**Table 1.** Units of Research Variables

| Variable | Description |
|---|---|
| KB active participant | Couples of childbearing age data are currently using one of the contraceptives without interspersed with pregnancy. |
| Couples of childbearing age | Married couples data which are bound in legal marriages, whose wives are still of childbearing age (aged between 15 to 49 years) or, in other words, have not experienced menopause. |
| KB equipment | The number of available methods and contraceptives at family planning facilities. |
| KB health facilities | The number of health service places organized by the government for the community to obtain family planning services. |


p-ISSN 2798-253X
e-ISSN 2798-3153

### 3.3. Data Analysis Method

This research used R-Studio dan Tableau as the research software to analyse the data. The research analysis was conducted using the following steps:

1. Inputting the data.
2. Formulate the problems.
3. Doing descriptive analysis.
4. Making assumptions from the data (representative sample test and multicollinearity test).
5. Determine the number of clusters.
6. Analysis of K-Medoids clustering.
7. Making conclusions from the results.

## 4. Results and Discussion

### 4.1. Descriptive Analysis

Descriptive analysis aims to describe the data so that the data presented is easy to understand for the reader. The following, in Figure 1, descriptive analysis describes the number of each variable, which consists of KB active participants, couples of childbearing age, KB equipment, and KB health facilities in 2020 in Sleman Regency.

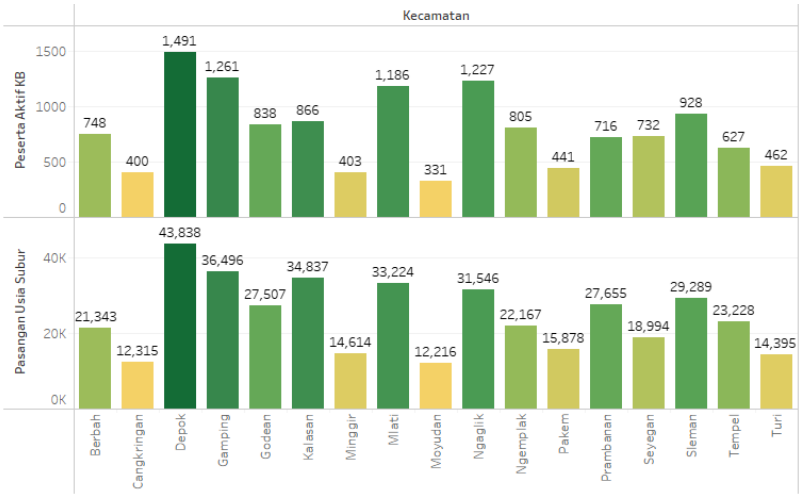


**Fig. 1.** The number of KB active participants and the couple of childbearing age in 2020 in Sleman Regency.

Figure 1 visualizes the number of active KB participants and the number of couples of childbearing age in 2020 in Sleman Regency. The greener the bar chart color, the higher the number, while the yellower, the lower. The bar chart shows that the highest number of active family planning participants was obtained by Depok Sub-District, which was 1491. Moyudan Sub-District, 331, received the lowest number of active family planning participants. Likewise, the number of couples of childbearing age, the highest number of couples of childbearing age was acquired by the Depok Sub-District, which is 43838, and the lowest number of couples of childbearing age is obtained by Moyudan Sub-District, which is 12216. From Fig.1, it can also be seen that there is a significant difference between the number of active family planning participants and the number of couples of childbearing age in each sub-district.

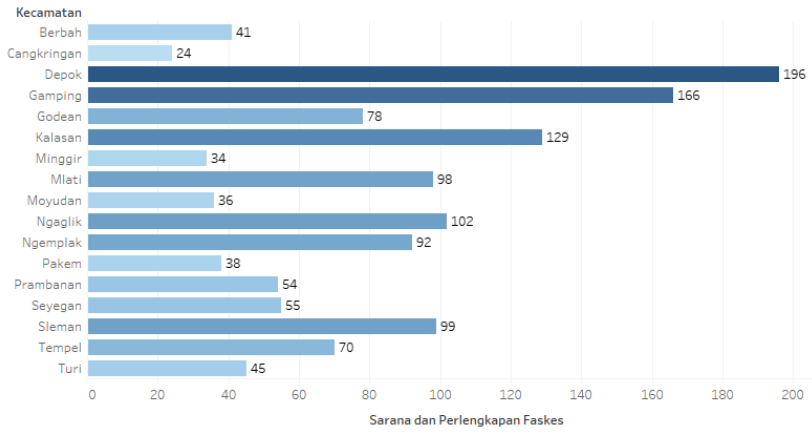


**Fig. 2.** The number of KB equipment in 2020 in Sleman Regency.

The picture in Figure 2 visualizes the number of KB equipment in 2020 in Sleman Regency. The bluer the bar chart, the higher the number. The bar chart shows that the highest number of KB equipment was obtained by Depok Sub-District, which was 196, and the lowest number of health facilities was obtained by Cangkringan Sub-District, which was only 24. From Figure 2, it can also be seen that there was a significant difference between the number of KB equipment that exists between sub-districts in Sleman Regency.

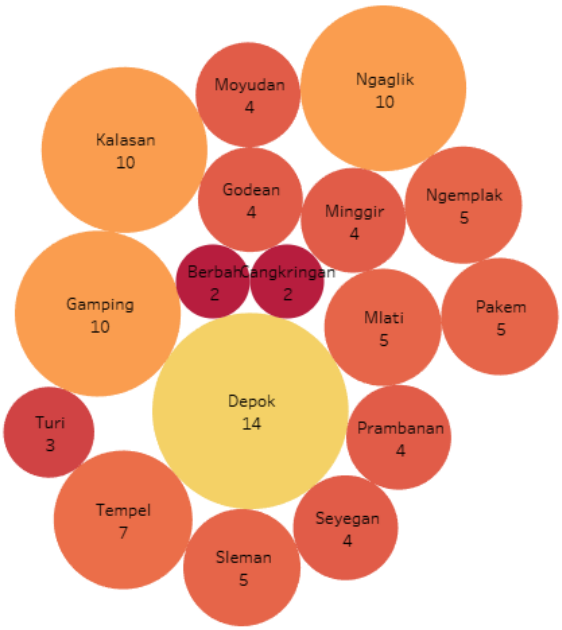


**Fig. 3.** The number of KB health facilities in 2020.

The display in Figure 3 visualizes the KB health facilities in 2020 in Sleman Regency. It is known that the highest number of KB health facilities in Sleman Regency in 2020 was obtained by Depok Sub-District, which was 14 health facilities, and the lowest number of KB health facilities was obtained by Berbah and Cangkringan Sub-District, which both have only 2 KB health facilities. From Figure 3, it can also be seen that there is a significant difference between the number of KB health facilities that exist between sub-districts in Sleman Regency.

### 4.2. Cluster Analysis Assumptions

There are two assumption tests in cluster analysis, namely, the population or sample must be representative, and there is no multicollinearity between variables. To test the population or sample is representative or not, it can use the Kaiser Meyer Olkin (KMO) test. The output obtained is shown in Figure 4 below.


p-ISSN 2798-253X
e-ISSN 2798-3153

```
> kmo(data[,3:6])
$KMO
[1] 0.8194733
```

**Fig. 4.** Kaiser Meyer Olkin test.

From these tests in Figure 4, the results obtained a KMO value of 0.8194733. Then it is accepted that the KMO value is more significant than 0.5, so it can be concluded that the sample can represent the population or is represented to be used for further analysis.

The next step is the multicollinearity test between variables. The multicollinearity test tests whether there is a correlation (strong relationship) between the independent or independent variables. The results of the analysis are listed in Table 2 below.

**Table 2.** Multicollinearity Test

| **Group I** | | | |
|---|---|---|---|
| **Dependent** | **Independent** | **Tolerance** | **VIF** |
| KB active participant | KB equipment | 0.284 | 3.521 |
| | KB health facilities | 0.116 | 8.588 |
| | Couples of childbearing age | 0.165 | 6.049 |
| **Group II** | | | |
| **Dependent** | **Independent** | **Tolerance** | **VIF** |
| Couples of childbearing ages | KB health facilities | 0.283 | 3.529 |
| | KB equipment | 0.150 | 6.682 |
| | KB active participant | 0.225 | 4.441 |
| **Group III** | | | |
| **Dependent** | **Independent** | **Tolerance** | **VIF** |
| KB equipment | KB health facilities | 0.392 | 2.551 |
| | KB active participant | 0.134 | 7.452 |
| | Couples of childbearing age | 0.127 | 7.897 |
| **Group IV** | | | |
| **Dependent** | **Independent** | **Tolerance** | **VIF** |
| KB health facilities | KB active participant | 0.132 | 7.578 |
| | Couples of childbearing ages | 0.107 | 9.345 |
| | KB equipment | 0.158 | 6.329 |

If the value of VIF ≥ 10 or Tolerance ≥ 0.10, then there is multicollinearity, or in other words, there is a correlation or linear relationship between the variables. The output regarding the results of VIF and tolerance is in Table 2. Based on these results, it can be concluded that there is no multicollinearity between predictor variables.

### 4.3. K-Medoids Cluster Analysis

To determine how the results of clustering on population and family planning in Sleman Regency in 2020 using the K-Medoids cluster method, a researcher must first find the number of clusters (k). There are various ways to be able to determine the best number of k, but in this study, the researcher chose to use the Elbow method, which uses the Within Sum of Square (WSS) value approach, as shown in Figure 5 below.


p-ISSN 2798-253X
e-ISSN 2798-3153

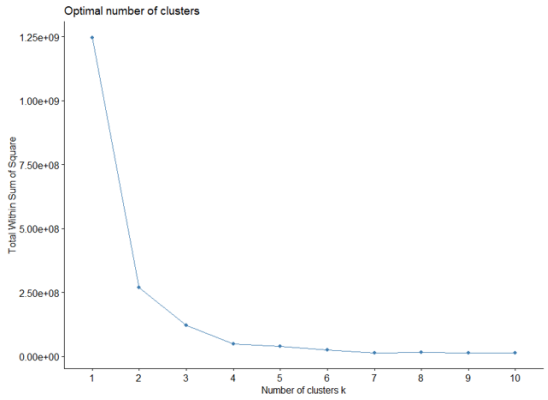


**Fig. 5.** Determination of the number of clusters (k).

Within a Cluster, the Sum of Squares is the sum of the squares of the values of each object in the cluster. From the picture above, it can be seen that the point where the number of clusters starts is said to be sloping at the number of cluster 3. This is different from the last number, which was too steep. Therefore, in determining the number of clusters in this study is to use 3 clusters (k=3).

After the number of k has been determined, the K-Medoids cluster analysis can be performed. The results of the clustering can be seen in Figure 6 below.

```
> print(pam.hasil)
Medoids:
     ID Peserta.aktif.KB Pasangan.usia.subur Sarana.dan.perlengkapan.faskes Faskes.KB
[1,]  6             1186               33224                             98         5
[2,] 11              805               22167                             92         5
[3,] 15              462               14395                             45         3
Clustering vector:
 [1] 1 2 3 3 2 1 1 2 2 1 2 1 1 2 3 3 3
```

**Fig. 6.** Clustering.

In Figure 6, it is known that the medoids or objects that represent to be the midpoint in cluster 1 are the 6th objects, the medoids used in cluster 2 are the 11th objects, and the medoids used in cluster 3 are the objects of the 6th. Thus, objects that have the closest distance to the medoids of a cluster will be included in each of these clusters. The clustering vector is numbering for members of each cluster. The way to read it is according to the order of 17 data. Namely, the 1st data is a member of cluster 1, and the 2nd data is a member of cluster 3, the 4th data is cluster 3, and so on. Members of each cluster can also be shown in Table 3 as follows.

**Table 3.** Clustering Results

| Number | Sub-District | KB Active Participant | Couples of Childbearing Age | KB Equipment | KB Health Dacilities | Pam Cluster |
|---|---|---|---|---|---|---|
| 1. | Gamping | 1261 | 36496 | 166 | 10 | 1 |
| 2. | Godean | 838 | 27507 | 78 | 4 | 2 |
| 3. | Moyudan | 331 | 12216 | 36 | 4 | 3 |
| 4. | Minggir | 403 | 14614 | 34 | 4 | 3 |
| 5. | Seyegan | 732 | 18994 | 55 | 4 | 2 |
| 6. | Mlati | 1186 | 33224 | 98 | 5 | 1 |
| 7. | Depok | 1491 | 43838 | 196 | 14 | 1 |
| 8. | Berbah | 748 | 21343 | 41 | 2 | 2 |
| 9. | Prambanan | 716 | 27655 | 54 | 4 | 2 |
| 10. | Kalasan | 866 | 34837 | 129 | 10 | 1 |
| 11. | Ngemplak | 805 | 22167 | 92 | 5 | 2 |
| 12. | Ngaglik | 1227 | 31546 | 102 | 10 | 1 |
| 13. | Sleman | 928 | 29289 | 99 | 5 | 1 |
| 14. | Tempel | 627 | 23228 | 70 | 7 | 2 |
| 15. | Turi | 462 | 14395 | 45 | 3 | 3 |
| 16. | Pakem | 441 | 15878 | 38 | 5 | 3 |
| 17. | Cangkringan | 400 | 12315 | 24 | 2 | 3 |


p-ISSN 2798-253X
e-ISSN 2798-3153

From the cluster results that have been obtained through the R-studio software, a plot of the visualization results of the K-Medoids that have been received can be displayed.

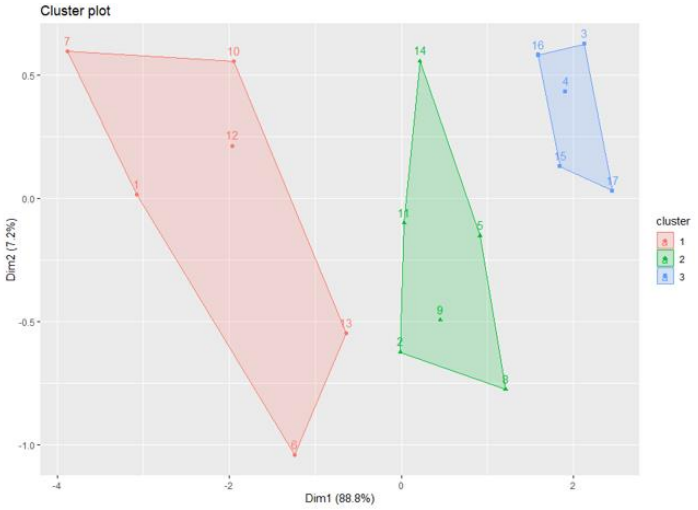


**Fig. 7.** Visualization of Clustering

In Figure 7 above, the plot of cluster results has three different colors showing the results of each cluster in each sub-district. It can be seen that members of cluster 1 are red, members of cluster 2 are green, and members of cluster 3 are blue. Each color has its distinct characteristics or characteristics. The table in Table 4 is the result of sub-district clustering regarding the number of national family planning in Sleman Regency in 2020 made by researchers to make it easier to see the results.

**Table 4.** Clustering Results

| Cluster | Number of Sub-Districts | Sub-District Name |
|---|---|---|
| 1 | 6 | Gamping, Mlati, Depok, Kalasan, Ngaglik, Sleman |
| 2 | 6 | Godean, Seyegan, Berbah, Prambanan, Ngemplak, Tempel |
| 3 | 5 | Moyudan, Minggir, Turi, Pakem, Cangkringan |

Then the results obtained from the average value of clustering are summarized in a table in Table 5 below.

**Table 5.** Clustering Average Value

| Cluster | Level Name | KB Active Participant | Couples of Childbearing Age | KB Equipment | KB Health Dacilities |
|---|---|---|---|---|---|
| **1** | High | 1160 | 34038 | 123 | 9 |
| **2** | Medium | 744 | 23482 | 65 | 4.33 |
| **3** | Low | 407 | 13884 | 35.4 | 3.6 |

Based on the average in each cluster contained in the table, it can be interpreted as follows:

1. Cluster 1 which consists of 6 sub-districts is a group of districts with a high level, where the first cluster has a high mean of KB active participants, couples of childbearing age, KB equipment, and KB health facilities.
2. Cluster 2 which consists of 6 sub-districts is a group of districts with medium level, where the second cluster has the medium mean of KB active participants, couples of childbearing age, KB equipment, and KB health facilities.


p-ISSN 2798-253X
e-ISSN 2798-3153

3. Cluster 3, Cluster 3, which consists of 5 sub-districts, is a group of districts with low levels, where the third cluster has the expected mean of KB active participants, couples of childbearing age, KB equipment, and KB health facilities.

## 5. Conclusion

Based on the research which has been conducted, it can be concluded.

1. There are a significant difference between the number of KB active participants and the number of couples of childbearing age between sub-districts. Likewise, the number of KB equipment and the number of KB health facilities exist between sub-districts in Sleman Regency..
2. The number of clusters that can be formed from the results of the analysis of the K-Medoids Clustering method obtained 3 clusters. By containing 17 sub-districts, Cluster 1 contains six sub-districts, Cluster 2 contains six sub-districts, and Cluster 3 contains five sub-districts.
3. The characteristics possessed by each cluster, namely Cluster 1, is a sub-district group with a high level which has a high mean of KB active participants, couples of childbearing age, KB equipment, and KB health facilities. Then, Cluster 2 is a sub-district with a medium level with a medium mean of KB active participants, couples of childbearing age, KB equipment, and KB health facilities. Meanwhile, Cluster 3 is a sub-district group with a low level that has a common mean of KB active participants, couples of childbearing age, KB equipment, and KB health facilities.

## References

[1] N. Atthina, & L. Iswari, “Klasterisasi Data Kesehatan Penduduk Untuk Menentukan Rentang Derajat Kesehatan Daerah Dengan Metode K-Means”, Seminar Nasional Aplikasi Teknologi Informasi (SNATI), 2014.

[2] Aqmal, Romi, “Pendidikan Keluarga dan Partisipasi Masyarakat pada Program Keluarga Berencana di Masa Pandemi Covid-19 Desa Kerandin Kecamatan Lingga Timur Kabupaten Lingga”, Journal of Education and Teaching, 2020.

[3] Badan Pusat Statistik, “Potret Sensus Penduduk 2020 Menuju Satu Data Kependudukan Indonesia”, Retrieved from Badan Pusat Statistik: https://www.bps.go.id/, on 20 April 2020.

[4] Badan Pusat Statistik Provinsi D.I. Yogyakarta, Retrieved from Badan Pusat Statistik: https://yogyakarta.bps.go.id/, on 25 April 2022.

[5] P. Tan, M. Steinbach, and V. Kumar, “Introduction to data mining,” AddisonWesley, Boston, 2006.

[6] J. Supranto, “Analisis Multivariat: Arti Dan Interpretasi”, Jakarta: PT. Rineka Cipta, 2004.

[7] J. K. Han, “Data Mining: Concept And Techniques”, Waltham: Morgan Kauffman Publisher, 2006.

[8] Imam Ghozali, “Aplikasi Analisis Multivariate Dengan Program IBM SPSS 21 Update PLS Regresi”, Semarang : Badan Penerbit Universitas Diponegoro, 2003.

[9] A. Widarjono, “Analisis Statistika Multivariat Terapan”, Yogyakarta: UPP STIM YKPN, 2010.

[10] W. R. Sukma, “Analisis Algoritma K-Medoids Clustering Dalam Pengelompokan Penyebaran Covid-19 di Indonesia”, Jurnal Teknologi Informasi, vol. 4, 2020.


p-ISSN 2798-253X
e-ISSN 2798-3153