Jurnal Mantik
Volume 3 Number 4, February 2020, pp. 55-63 E-ISSN 2685-4236
https://iocscience.org/ejournal/index.php/mantik/index

Application To Classify The Population Census The Total Population By Level Of Education Using K-Means Clustering Algorithm

Dewi Janetta Az Zahra[1], Agung Triayudi[2], Ira Diana Sholihati[3]

[1]Informatika,
Fakultas Teknologi Komunikasi dan Informaika, Universitas Nasional, Jl. Sawo Manila, Pasar Minggu,
Jakarta Selatan, Indonesia

E-mail: dewi.janetta.az.zahra@gmail.com, agungtriayudi@civitas.unas.ac.id, *iradiana2803@gmail.com*

ARTICLE INFO

Articlehistory:
Received: 11/01/2020
Revised: 12/01/2020
Accepted: 01/02/2020

***Keywords**:*
algorithm, census, clustering, k-means, population

ABSTRACT

The current population census is still Carried out by visiting each house. With a process like this, it will take a lot of time so that it will have an impact on the number of costs incurred by the government. For this reason, an online application is needed to Facilitate officers in conducting population censuses. Also, through this population census, the Researchers conducted a grouping of education levels intending to find out the population based on their level of education. The method used is the k-means clustering algorithm. There are three groupings items, namely C1 = population at a high level, C2 = population at a moderate level and C3 = population at a moderate level. The final centroid value used at C1 = (4.8000000,959.20000), C2 = (4.00000,854.000000) and C3 = (9.00000,668.00000). So that the results Obtained grouping C1 = RW2, RW7 and RW8,

1. Introduction

Implementation of the census is still conducted by visiting every home residents. In the data collection process requires substantial time, consequently the government issued a greater cost. For that, we need a census online application for ease in data retrieval officer. In addition, the results of the population census can also do the grouping of the total population by level of education by the method of k-means clustering algorithm.

Based on previous research to develop a data collection system in urban population hero by using the programming language PHP and MySQL that can help employees to perform data collection[1], Subsequent research by Teguh Hananto, Oky Dwi and Ike Earth that is designing the system in the village census wulunggunung using the Delphi programming language, with this system can help employees manage population census[2],

Previous studies of population census design applications that can then be used by the locals in the process of population census. The system is designed using software MyEclipse and JSP as programming language[3],

Based on previous research conducted by Fithri et al namely the application of k-modes algorithm to classify the number of elderly people using population census data in 2010. From these results obtained four clusters where each cluster has a different pattern[4], Research was also conducted by Eric Famamaldo and Lukman Hakim is mengelempokkan family welfare level for Indonesia Smart card program using clustering method, the results of this test is 60% earn Indonesia smart card from a data sample of 200 family data[5],

K-means algorithm is a method that is the most simple and common to be used in clustering. K-means will partition the data into groups so that the data that has similarities will be grouped together and if no resemblance will dikempokkan others[6], Research conducted Gunawan et al, namely the application of mining in the grouping of the potential use of water potential new customers by using k-means clustering. The results of these tests show the system can run 95.80%[7],

Based on previous research, k-means clustering algorithm can be made to the grouping number peduduk by level of education through the census.

2. Literature review

A. Clustering

clustering which is a process that classifies data according to the characteristics of the group. Therefore, clustering very useful and could find a group or unknown group in the data. Clustering many diguakan in various applications such as for example in the areas of governance[8], It is important when clustering is declared a bunch of data into useful groups to determine the similarities and differences that can lead to the conclusion that more clear.

B. K-Means Algorithm

K-Means algorithm is a method of grouping data nonhierarki which group the data in the form of one or more groups. Data that have the same characteristics will be grouped into one cluster. K-means algorithm group basically do two processes, namely the cluster center location detection process and the search process for each cluster member[9],

Excess k-means algorithm capable of classifying objects such as large so quickly that the process of grouping is made very easy. While the weakness of the k-means algorithm is very weak in determining Babysitting random cluster center, the result of the grouping is always changing and the workmanship fast.

Here is the first picture in the method flowchart k-means clustering algorithm.

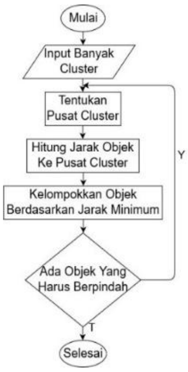

Mulai
Input Banyak Cluster
Tentukan Pusat Cluster
Hitung Jarak Objek Ke Pusat Cluster
Kelompokkan Objek Berdasarkan Jarak Minimum
Ada Objek Yang Harus Berpindah
Y
T
Selesai

Figure 1. Flowchart k-means algorithm

The stages in the process with the k-means method is as follows[10]:

1) Determine the number of clusters a
2) Determining the center point (centroid) in a random way.
3) Categorize all of the data based on the distance the two data stretcher. When the process is carried out is required to calculate the distance of each data kesetiap cluster. To calculate the distance of all data to any point of the cluster center can use the formula Eucliden Distance to equation (1):

$$E\ (i, j) = \sqrt{(X1i - X1j)^2 + (X2i - X2j)^2 + (Xai - Xaj)^2} (1)$$

Information:
E (i, j) : Within the data center cluster i
Xai : Data into i on attribute to a
Xaj : Data to j on attribute to a

4) Recalculating the center distance of the cluster with the new cluster membership. Center of the cluster is the average of all the data in the same cluster. With the formula (2) as follows:

$Ck\frac{1}{nk}\sum di$ (2)

Information :
nk K = number of cluster data
in = Data in cluster k

5) Any data recalculate the central point of the new cluster. If the center of the cluster there are changes in the clustering process has been completed and if there is a center cluster changes, then return to stage C until the optimal cluster center worth.

3. Method

A. Research methods

The stage of the research can be seen in Figure 2:

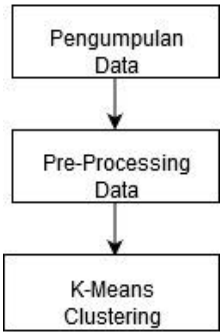


Figure 2. Stages Research

1) Data collected by retrieving data from population censuses conducted in the village of Pondok China with a total sample of 100 family card data.
2) *Pre-processing* ie where do the sorting of data attribute data to be used in the process of grouping data of the population by education level.
3) *K-Means Clustering*, After the data attribute has been set, it can proceed to the next process is clustering. The algorithm used is K-Means.

B. Application System Design

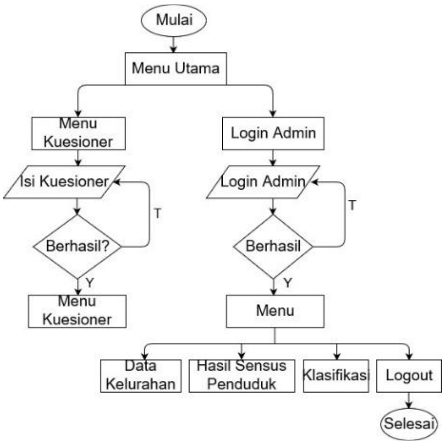


Figure 3. Application System Design

Figure 3 describes the system design process. Where access the app will display a menu which form census questionnaire and admin login menu. When accessing a resident directed to fill out a questionnaire population and admin login menu devoted to admin wherein when logged in, first fill in the username and password. In the login phase the system will do the matching, if appropriate then appeared admin menu form, if not going back to the login menu. In the admin homepage there are several menu is a menu of data villages, population censuses menu, menu classification and logout menu.

4. Results and Discussion

A. System Requirements Analysis

In designing the census application clustering, it takes the form of hardware support device (hadware) and software (software) consisting of:

1. Hadware needs: Laptop Intel® CoreTM i5-82500U 1.80GHZ CPU @ 1.60 GHz 64-bit RAM to 8GB.
2. Software Requirements: Draw.io, XAMPP, MySQL, Text Editor Bracket, Mozilla Firefox and Microsoft Excel.

B. Calculation Process Manual

In Table 1 represents data of the population by level of education that will do the calculation process of grouping by using the k-means algorithm. In this calculation process, occurring three times iteration of grouping by attributes RW and population.

Table 1. Level of Education

RW	SD	SMP	High School	D-3	S-1	S-2	amount
1	143	178	156	120	140	126	863
2	160	241	137	132	153	142	965
3	125	169	245	98	146	118	901
4	115	189	197	99	178	115	893
5	170	168	129	89	196	87	839
6	117	153	190	77	176	149	862
7	134	268	210	85	164	80	941
8	230	190	185	130	229	132	1096
9	120	124	98	60	168	98	668

The stages of the process of calculating the k-means algorithm is as follows:

1) Determining the center point that will be in the input cluster randomly divided into 3 pieces of data centroid ie RW3, RW6 and RW9 using RW and population attribute with a value of C11 = (3.901), C12 = (6.862) and C13 = (9.668).

2) After determining the value of the center point of the cluster and then calculates the distance to each centroid using Distance Eucliden equation (1). Centroid distance calculation process at each data RW 1 is as follows:

$$C11 = \sqrt{(X1i - X1j)^2 + (X2i - X2j)^2 + (Xai - Xaj)^2}$$
$$= \sqrt{(1-3)^2 + (863-901)^2}$$
$$= 38.052952$$

$$C12 = \sqrt{(X1i - X1j)^2 + (X2i - X2j)^2 + (Xai - Xaj)^2}$$
$$= \sqrt{(1-6)^2 + (863-862)^2}$$
$$= 5.0990195$$

$$C12 = \sqrt{(X1i - X1j)^2 + (X2i - X2j)^2 + (Xai - Xaj)^2}$$
$$= \sqrt{(1-9)^2 + (863-668)^2}$$

= 195.1640336

For full results of the calculation of the shortest distance to the centroid presented in Table 2 below:

Table 2. Iteration 1st

RW	SD	SMP	High School	D-3	S-1	S-2	amount	centroid 1	centroid 2	centroid 3	C1	C2	C3
1	143	178	156	120	140	126	863	38.0525952	5.0990195	195.1640336		ok	
2	160	241	137	132	153	142	965	64.0078120	103.0776406	297.0824801	ok		
3	125	169	245	98	146	118	901	0	39.1152144	233.0772404	ok		
4	115	189	197	99	178	115	893	8.0622577	31.0644491	225.0555487	ok		
5	170	168	129	89	196	87	839	62.0322497	23.0217289	171.0467772		ok	
6	117	153	190	77	176	149	862	39.1152144	0	194.0231945		ok	
7	134	268	210	85	164	80	941	40.1995025	79.0063289	273.0073259	ok		
8	230	190	185	130	229	132	1096	195.0640920	234.0085469	428.0011682	ok		
9	120	124	98	60	168	98	668	233.0772404	194.0231945	0			ok

In Table 2 shows the results of the process of iteration-1 which has been found to value the closest distance to each cluster. After finding the closest distance to each cluster then the resulting grouping of the total population by level of education with as many as five data C1 consists of RW2, RW3, RW4, RW7 and RW8 which means the number of inhabitants are at a high level. At C2, there are three data that RW1, RW5 and RW6 which means the population is in the medium level. While in the C3 are the data that RW9 meaning into fairly moderate level of its population.

3) The next step is if there is data that changes the cluster or no change in the numbers in the centroid then be recalculated latest value centroid to the center of the cluster so as to find the numbers optimum by calculating the average value of each cluster to perhitungsn process will be stopped after iteration -3 to the cluster center in table 3. to find the value of the center point of the cluster use the formula (2):

$$C1RW = \frac{di}{nk} = \frac{2+3+4+7+8}{5} = 4.800000$$

$$\mathrm{C1}_{\mathrm{Jumlah}} = \frac{di}{nk} = \frac{2+3+4+7+8}{5} = 959.200000$$

For the new value of the next cluster center presented in Table 3 below:

Table 3. The new cluster centers

C11	4.800000	959.200000
C12	4.00000	854.00000
C13	9.00000	668.00000

4) After getting a new cluster center value. Furthermore recalculated using the new centroid value Eucliden Distance equation (1). The process of calculating the distance to the new centroid of each cluster in RW 1 Diman stopped at the 3rd iteration process as contained in Table 4 are the following:

$$C11 = \sqrt{(X1i - X1j)^2 + (X2i - X2j)^2 + (Xai - Xaj)^2}$$
$$= \sqrt{(1 - 4{,}800000)^2 + (863 - 959{,}200000)^2}$$
= 96.2750227

$$C12 = \sqrt{(X1i - X1j)^2 + (X2i - X2j)^2 + (Xai - Xaj)^2}$$
$= \sqrt{(1 - 4{,}00000)^2 + (863 - 854{,}00000)^2}$
= 8.8568868

$$C13 = \sqrt{(X1i - X1j)^2 + (X2i - X2j)^2 + (Xai - Xaj)^2}$$
$= \sqrt{(1 - 9{,}00000)^2 + (863 - 668{,}00000)^2}$
= 195.1640336

Results fuller than the distance calculation process into a new centroid presented in Table 4 below:

Table 4. Iteration 3rd

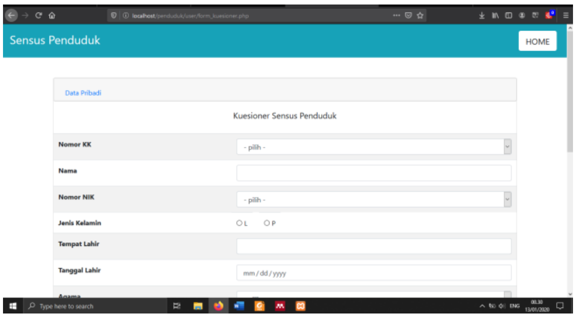
RW	SD	SMP	High School	D-3	S-1	S-2	amount	centroid 1	centroid 2	centroid 3	C1	C2	C3
1	143	178	156	120	140	126	863	96.2750227	8.8568868	195.1640336		ok	
2	160	241	137	132	153	142	965	6.444969	110.3514587	297.0824801	ok		
3	125	169	245	98	146	118	901	58.2278284	46.3441234	233.0772404		ok	
4	115	189	197	99	178	115	893	66.2048337	38.3333333	225.0555487		ok	
5	170	168	129	89	196	87	839	120.2001664	15.6985492	171.0467772		ok	
6	117	153	190	77	176	149	862	97.2074071	7.6011695	194.0231945		ok	
7	134	268	210	85	164	80	941	18.3324848	86.3854411	273.0073259	ok		
8	230	190	185	130	229	132	1096	136.8374218	241.3664802	428.0011682	ok		
9	120	124	98	60	168	98	668	291.2302869	186.7336190	0			ok

In Table 4 are the end result of the calculation process that occurs in the 3rd iteration process, where there is no change on the previous centroid. As there is no change, then found the grouping result of population by education level as shown in Table 4. With the acquisition of data for three C1, C2 and C3 as many as five of data as a single data. Data obtained from C1, that RW2, RW7 and RW8 meaning into high population groups. C2 Data obtained on the pillars of citizens consisting of RW1, RW3, RW4 and RW6 meaning its population clustered into being. While the data C3 consisting of RW9 included into kelompon quite moderate.

C. Implementation of Application Systems

The application system designed in this study is based websites.

1. Menu Census Questionnaire



localhost/penduduk/user/form_kuesioner.php

Sensus Penduduk

Data Pribadi

Kuesioner Sensus Penduduk

Nomor KK	- pilih -
Nama	
Nomor NIK	- pilih -
Jenis Kelamin	○ L ○ P
Tempat Lahir	
Tanggal Lahir	mm / dd / yyyy
Agama	

Figure 4. Display Form Census Questionnaire

In Figure 4 is a view form questionnaire that includes the home button and charging the population census which will fill all the questions that have been provided.

2. Menu Login Admin

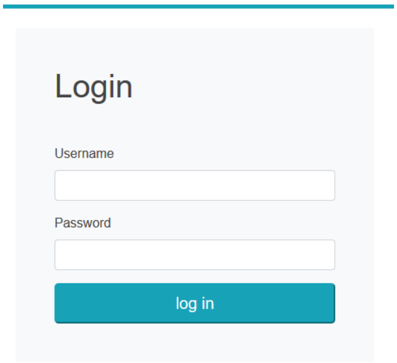


Figure 5. Menu Login Admin

In Figure 5 is an admin login menu to see where the first entering ussername and password in order to access the admin menu.

3. Menu Import File

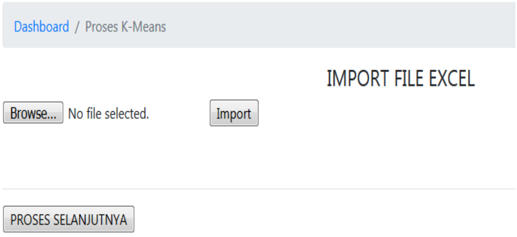
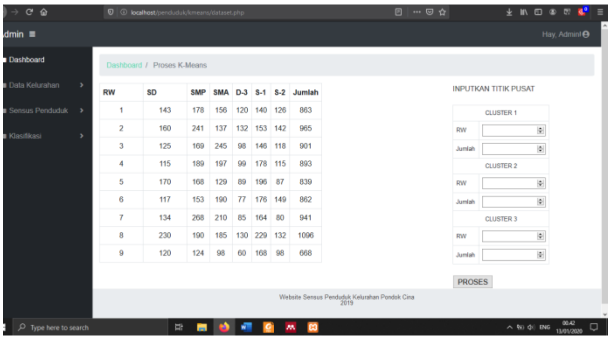


Figure 6. Display Form Import File

In Figure 6 is the form in which the file import process admin asked to import the excel file containing the amount of data to be processed using the k-means clustering.

4. Input Process Cluster

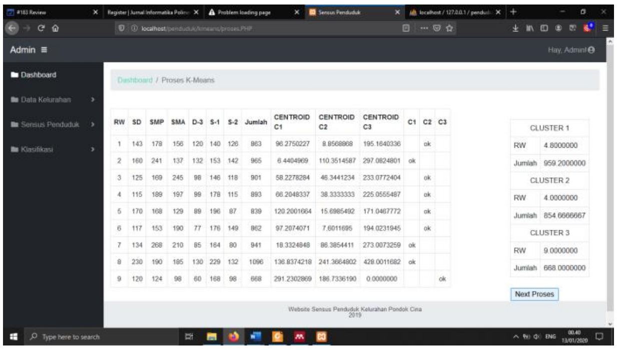


RW	SD	SMP	SMA	D-3	S-1	S-2	Jumlah
1	143	178	156	120	140	126	863
2	160	241	137	132	153	142	965
3	125	169	245	98	146	118	901
4	115	189	197	99	178	115	893
5	170	168	129	89	196	87	839
6	117	153	190	77	176	149	862
7	134	268	210	85	164	80	941
8	230	190	185	130	229	132	1096
9	120	124	98	60	168	98	668

Figure 7. Form Input Process Cluster

In figure 7 is an input display cluster, where guards cluster random input values contained in the table above on inputting values after pressing the button cluster process.

5. Results Process K-Means



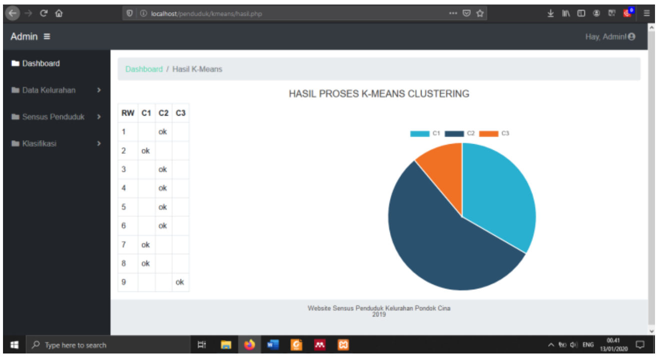
RW	SD	SMP	SMA	D-3	S-1	S-2	Jumlah	CENTROID C1	CENTROID C2	CENTROID C3	C1	C2	C3
1	143	178	156	120	140	126	863	96.2750227	8.8568868	195.1640336		ok	
2	160	241	137	132	153	142	965	6.4404969	110.3514587	297.0824801	ok		
3	125	169	245	98	146	118	901	58.2278284	46.3441234	233.0772404		ok	
4	115	189	197	99	178	115	893	66.2048337	38.3333333	225.0555487		ok	
5	170	168	129	89	196	87	839	120.2001664	15.6985492	171.0467772		ok	
6	117	153	190	77	176	149	862	97.2074071	7.6011695	194.0231945		ok	
7	134	268	210	85	164	80	941	18.3324848	86.3854411	273.0073259	ok		
8	230	190	185	130	229	132	1096	136.8374218	241.3664802	428.0011682	ok		
9	120	124	98	60	168	98	668	291.2302869	186.7336190	0.0000000			ok

CLUSTER 1	
RW	4.8000000
Jumlah	959.2000000
CLUSTER 2	
RW	4.0000000
Jumlah	854.6666667
CLUSTER 3	
RW	9.0000000
Jumlah	668.0000000

Figure 8. Display Process Results Form K-Means

In Figure 8 is the result of a process of inputting the k-means cluster value.

6. Results Graph



localhost/penduduk/kmeans/hasil.php
Admin
Hay, Admin!
Dashboard
Data Kelurahan
Sensus Penduduk
Klasifikasi
Dashboard / Hasil K-Means
HASIL PROSES K-MEANS CLUSTERING

RW	C1	C2	C3
1		ok	
2	ok		
3		ok	
4		ok	
5		ok	
6		ok	
7	ok		
8	ok		
9			ok

C1
C2
C3
Website Sensus Penduduk Kelurahan Pondok Cina
2019

Figure 9. Results Graph

In Figure 9 is the result of the K-Means Clustering graph, the system displays the data grouping of K-Means process where C1 there are three data that RW2, RW7 and RW8. At C2, there are five data that RW1, RW3, RW4, RW5 and RW6. While in the C3 are the data that RW9.

5. Conclusion

Based on the discussion, the following conclusions to be drawn:

1. Applications created census may assist officers in making the retrieval of data population.
2. K-Means algorithm can be implemented for the process of grouping the total population by level of education contained in three groups: C1 = number of people are at a high level, C2 = total population at the level of medium and C3 = total population was at quite moderate.
3. From the test results K-means algorithm obtained groups of population by level of education with final grades centroid C1 = (4.8000000,959.20000), C2 = (4.00000,854.000000) and C3 = (9.00000,668.00000). Thus obtained results of grouping C1 = RW2, RW7 and RW8, C2 = RW1, RW3, RW4, RW5 and RW6 and C3 = RW9.

6. Reference

[1] A. Ibrahim, A. Rifai, and L. Oktarina, “Rancang Bangun Aplikasi Pencatatan Data Kependudukan Kelurahan Pahlawan Berbasis Web,” *J. Sist. Inf.*, vol. 8, no. 1, pp. 947–957, 2016.

[2] T. H. Widodo, O. D. Nurhayati, and I. P. Windasari, “Pembuatan Aplikasi Sensus Penduduk Untuk Desa Wulunggunung,” *J. Teknol. dan Sist. Komput.*, vol. 4, no. 1, p. 9, 2016, doi: 10.14710/jtsiskom.4.1.2016.9-16.

[3] U. Waziri, M. J. Usman, A. Garba, and W. A. Gadzama, “Design and Implementation of Secured Online Census Information Management System Based on B / S Structure,” vol. 3, no. 5, pp. 12345–12354, 2014.

[4] D. P. Fithri Selva Jumeilah, “Klasterisasi Penduduk Lanjut Usia Sumatera Selatan Menggunakan Algoritma K-Modes,” *J. TAM (Technol. Accept. Model)*, vol. 8, no. 2, pp. 85–89, 2017.

[5] Eric Fammaldo dan Lukman Hakim, “Penerapan Algoritma K-Means Clustering Untuk Pengelompokan Tingkat Kesejahteraan Keluarga Untuk Program Kartu Indonesia Pintar,” *J. Ilm. Teknol. Inf. Terap.*, vol. V, no. 1, pp. 24–32, 2018.

[6] F. I. Sri Rahayu, Dodon T. Nugrahadi, “Clustering Penentuan Potensi Kejahatan Daerah Di Kota Banjarbaru Dengan Metode K-Means,” *Kumpul. J. Ilmu Komput.*, vol. 01, no. 01, pp. 33–45, 2014.

[7] G. Abdillah *et al.*, “Penerapan Data Mining Pemakaian Air Pelanggan Untuk Menentukan Klasifikasi Potensi Pemakaian Air Pelanggan Baru Di Pdam Tirta Raharja Menggunakan Algoritma K-Means,” *Sentika 2016*, vol. 2016, no. Sentika, pp. 18–19, 2016.

[8] D. F. Pramesti, Lahan, M. Tanzil Furqon, and C. Dewi, “Implementasi Metode K-Medoids Clustering Untuk Pengelompokan Data,” *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 1, no. 9, pp. 723–732, 2017, doi: 10.1109/EUMC.2008.4751704.

[9] L. Maulida, “Kunjungan Wisatawan Ke Objek Wisata Unggulan Di Prov . Dki Jakarta Dengan K-Means,” *JISKa*, vol. 2, no. 3, pp. 167–174, 2018.

[10] Z. Aras and Sarjono, “Analisis Data Mining Untuk Menentukan Kelompok Prioritas Penerima Bantuan Bedah Rumah Menggunakan Metode Clustering K-Means(Studi Kasus: Kantor Kecamatan Bahar Utara),” *J. Manaj. Sist. Inf.*, vol. 1, no. 2, pp. 159–170, 2016..