International Journal of Basic and Applied Science 10 (2) (2021) 51-59

Published by: IOCSCIENCE

International Journal of Basic and Applied Science

Journal homepage: www.ijobas.pelnus.ac.id

Application Of K-Means Clustering Algorithm On Population Growth In Simalungun Regency

Murniyati Rambe[1], M. Safii[1], Irawan[2]

[1]Departement of Information System, STIKOM Tunas Bangsa, Pematangsiantar

[2]Departement of Computerized Accounting, AMIK Tunas Bangsa, Pematangsiantar

Article Info

Article history:

Received Jun 20, 2021
Revised Jul 22, 2021
Accepted Sep 17, 2021

Keywords:

Data Mining;
Social Problem;
Population Growth;
Clustering;
K-means.

ABSTRACT

Population growth is a condition when the population increases from previous years. Population growth has several variables, namely birth, death and migration rates. Positive population growth indicates an increase in population and vice versa. Population growth is caused by a high birth rate with a decrease in the death rate. The high rate of population growth and occurs in a fast period of time is what triggers a population explosion which is closely related to an increase in poverty, unemployment, crime, slum settlements, hunger and other social problems. An increase in the poverty rate occurs when high population growth is not matched by good economic growth accompanied by equitable distribution of income. An increase in unemployment occurs if the increase in population with reduced availability of adequate employment can lead to an increase in criminal cases. By knowing these problems, Data Mining is needed to classify aid receipts, build jobs. by using the K-Means method in clustering the population growth rate. The K-Means method can assist the Government in making decisions and the information needed to solve the problem of population growth and record all densely populated areas in an appropriate way.

Corresponding Author:

M.Safii,
Department of Information System,
Department of Computerized Accounting,
AMIK dan STIKOM Tunas Bangsa Pematangsiantar,
Jl. Jend. Sudirman Blok A-B No.1-3 Pematangsiantar, North Sumatera, Indonesia.
Email: m.safii@amiktunasbangsa.ac.id

1. INTRODUCTION

Population growth is a condition when the population increases from previous years. Population growth has several variables, namely birth, death and migration rates.[1] Positive population growth indicates an increase in population and vice versa. Population growth is caused by a high birth rate with a decreasing death rate and occurs in a fast period of time can trigger a population explosion

which is closely related to an increase in poverty, unemployment, crime, slum settlements, hunger and other social problems. [2][3] Poverty occurs when high population growth is not matched by good economic growth accompanied by equitable distribution of income.[1] An increase in unemployment occurs when the population increases with the reduced availability of adequate employment which can lead to an increase in criminal cases.[4][5] The problem of population growth rate is a very complex problem with an impact which of course causes social problems due to unstable economic limitations [1][6].

This limitation can hinder the welfare of the population in carrying out their daily activities. Circumstances like this can also lead to a crime-prone situation marked by a lack of necessities of life that triggers crime that comes from bad influences in the neighborhood and causes a decrease in self-confidence, self-acceptance to adjustment to the social environment. Population growth shows major developments in each area, especially Simalungun district. Because of this great development, there are impacts that require special actions in order to get the right to a more prosperous life. Data from BPS Simalungun Regency changes every year, the population growth rate in Simalungun Regency which consists of 32 sub-districts, 27 villages, and 386 villages with an area of 4,369.00 km² with a population density of 235 people/km².[7] Therefore, in an effort to improve the welfare of the population, research is carried out based on population growth in the district of Simalungun Regency so that it can be seen that regional groups have high clusters, medium clusters, and low clusters on population growth. Several branches of computer science can solve complex problems, one of which is data mining. Data mining is a logical process used in searching and finding patterns through large amounts of data. The aim of this technique is to find previously unknown patterns. In data mining, this pile of past data is considered a mine that can be processed to produce valuable knowledge [8].

Before the data mining process will be carried out, it is necessary to carry out a cleaning process on the data that is the focus of Knowledge Discovery in Database (KDD). An enrichment process is also carried out, namely the process of "enriching" existing data with other relevant data or information needed for Knowledge Discovery in Database (KDD), such as other required external data or information. [9]. There are several clustering algorithms, one of which is k-means which is the K-Means algorithm is a clustering algorithm that groups data based on the cluster center point (centroid) closest to the data. [10]. K-Means is one of the non-hierarchical data clustering methods that can partition existing data into one or more clusters or groups so that data with the same characteristics are grouped into the same cluster and data with different characteristics are grouped into in another group. [11][12]. Distance-based clustering algorithm will divide data into a number of clusters and this algorithm only works for numeric attributes [13]. The K-means clustering algorithm will be applied to the population data in Simalungun Regency so that it is known the grouping of areas that have high cluster population rates, medium clusters and low clusters obtained based on these data[14][15].

2. RESEARCH METHOD

This study aims to collect existing information and manage data to solve the problems to be studied using the Data Mining technique method with the k-means clustering algorithm. The data used in this study is secondary data originating from the Central Statistics Agency (BPS) Simalungun Regency. The results of this study are to determine the high cluster, low cluster, low cluster, from population growth in Simalungun Regency [16]. The workflow that the author did in this study is presented in the activity diagram in Figure 1 the following:

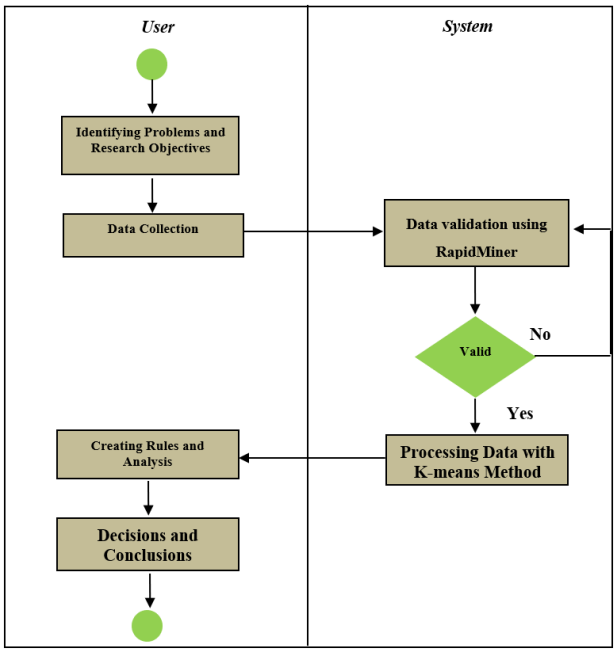

Figure 1. Research Activity Diagram

Figure 1 is the flow of activities carried out in research. Figure 1 explains that the author starts by identifying the problem and the purpose of his research, collecting data from the Central Statistics Agency (BPS) Simalungun Regency. Then the data that has been collected is validated using the RapidMiner application. If the data is valid, it will be processed in the RapidMiner application, then from the RapidMiner application it will be processed with K-Means, then the author or user will make the analysis carried out. Then, the author or user makes a decision and concludes the research that has been done.

3. RESULTS AND DISCUSSIONS

The solutions for grouping population growth data using the k-means clustering algorithm are as follows [17][18]:

$$V_{ij} = \frac{1}{Ni} \sum^{Ni} Xkj$$

Information:
Vij = the average centroid of the i-th cluster for the j-th variable
Ni = number of i-th cluster members
i, k = index of cluster
j = index of variable
Xkj = k-th data value of the j-th variable in the cluster

3.1 Determination of data in the cluster

The population growth data used for the clustering process is population growth data obtained from the Simalungun Central Agency in 2010-2019 which has 32 sub-districts data. [19][20]Below is how to find the average value:

R1 = 13,611 + 14,269 + 14,396 + 14,535 + 15,114 + 15,452 + 15,777 + 16,083 + 16,376 + 16.656 /10 =15,227

R2 = 10,334 + 10,486 + 10,516 + 10,547 + 10,692 + 10,765 + 10,834 + 10,898 + 10,959 + 11,016 /10 = 10,705

R3 = 21,830 + 22,504 + 22,635 + 22,773 +23,373 + 23,708+ 24,027 + 24,325 +24,608 + 24,878 /10 = 23,466

The average search continues until R-32, until the following results are obtained:

Table 1. Average Value of Population Growth

No	Districts	Average
1	Silimakuta	15,227
2	Pamatang Silimahuta	10,705
3	Purba	23,466
4	Haranggaol Horison	5,057
5	Dolok Pardamean	15,163
6	Sidamanik	27,509
7	Pamatang Sidamanik	16,559
8	Girsang Sipangan Bolon	14,740
9	Tanah Jawa	47,359
10	Hatonduhan	21,299
11	Dolok Panribuan	18,251
12	Jorlang Hataran	15,574
13	Panei	21,997
14	Panombeian Panei	19,451
15	Raya	29,947
16	Dolog Masagal	2,939
17	Dolok Silou	14,231
18	Silou Kahean	31,115
19	Raya Kahean	17,716
20	Tapian Dolok	40,356
21	Dolok Batu Nanggar	40,323
22	Siantar	58,800
23	Gunung Malela	34,230
24	Gunung Maligas	27,473
25	Hutabayu Raja	29,628
26	Jawa Maraja Bah Jambi	21,493
27	Pamatang Bandar	31,578
28	Bandar Huluan	26,279
29	Bandar	67,584
30	Bandar Masilam	24,726
31	Bosar Maligas	40,170
32	Ujung padang	41,081

3.2 Determine the Centroid value (cluster center)

The value of the cluster center is determined randomly from the value of the data variable that will be in the cluster as much as has been determined. Where the highest cluster value is obtained from the highest value, while the medium cluster is obtained from the average value and the low cluster is obtained from the smallest value. Below is the initial data centroid value for iteration 1:

Table 2. Initial Data Centroid (Iteration 1)

C1 = Maximum	67,584
C2 = Average	26,626
C3 = Minimum	2,939

3.3 Calculating Centroid jarak distance

Calculation of the distance of each calculated data at the center of the cluster. When the cluster center value is obtained, then calculate the distance of each data based on the cluster center with the following Euclidean Distance:

By calculating the distance to the center point (centroid) in the following first cluster:

$$D\ (1.1) = \sqrt{(67{,}584 - 15{,}227)^2} = 52{,}357$$

$$D\ (1.2) = \sqrt{(67{,}584 - 10{,}705)^2} = 56{,}979$$

$$D\ (1.3) = \sqrt{(67{,}584 - 23{,}466)^2} = 44{,}118$$

$$D\ (1.4) = \sqrt{(67{,}584 - 5{,}057)^2} = 62{,}527$$

$$D\ (1.5) = \sqrt{(67{,}584 - 15{,}163)^2} = 52{,}421$$

Next up to D (1.32)

Until next with D (3.32) to produce the shortest distance from the center point (centroid). [21]Below is a table of the calculation results for centroid 1, centroid 2, and centroid 3 and the shortest distance.

Table 3. Iteration Centroid Distance 1

No	Districts	C1	C2	C3	Shortest Distance
1	Silimakuta	52,357	11,398	12,288	11,398
2	Pamatang Silimahuta	56,879	15,920	7,766	7,766
3	Purba	44,118	3159	20,527	3159
4	Haranggaol Horison	62,527	21,568	2,118	2,118
5	Dolok Pardamean	52,421	11,462	12,224	11,462
6	Sidamanik	40,075	883	24,570	883
7	Pamatang Sidamanik	51,025	10,006	13,620	10,006
8	Girsang Sipangan Bolon	52,844	11,885	11,801	11,801
9	Tanah Jawa	20,225	20,733	44,420	20,225
10	Hatonduhan	46,285	5,326	18,360	5,326
11	Dolok Panribuan	49,333	8,374	15,312	8,374
12	Jorlang Hataran	52,010	11,051	12,635	11,051
13	Panei	45,587	4,628	19,058	4,628
14	Panombeian Panei	48,133	7,174	16,512	7,174
15	Raya	37,637	3,321	27,008	3,321
16	Dolog Masagal	64,645	23,686	0,000	0,000
17	Dolok Silou	53,353	12,394	11,292	11,292
18	Silou Kahean	36,469	4,489	28,176	4,489
19	Raya Kahean	49,868	8,909	14,775	8,909
20	Tapian Dolok	27,228	13,730	37,417	13,730
21	Dolok Batu Nanggar	27,261	13,697	37,384	13,697
22	Siantar	8,784	32,174	55,861	8,784
23	Gunung Malela	33,354	7604	31,291	7604
24	Gunung Maligas	40,111	847	24,534	847
25	Hutabayu Raja	37,956	3002	26,689	3002
26	Jawa Maraja Bah Jambi	46,091	5,132	18,554	5,132
27	Pamatang Bandar	36,006	4952	28,639	4952
28	Bandar Huluan	41,305	346	23,340	346
29	Bandar	0,000	40,958	64,645	0,000
30	Bandar Masilam	42,858	1899	21,790	1899
31	Bosar Maligas	27,414	13,544	37,231	13,544
32	Ujung padang	26,503	14,455	38,142	14,455

If the distance from the center point (centroid) has been calculated until the 7th iteration, then the placement of the cluster position is by grouping the values as before. The following is the position of the cluster in the 7th iteration:

Table 4. Grouping of 7 Iteration Cluster Positions

No	Districts	C1	C2	C3
1	Silimakuta			1
2	Pamatang Silimahuta			1
3	Purba			1
4	Haranggaol Horison			1
5	Dolok Pardamean			1
6	Sidamanik		1	
7	Pamatang Sidamanik			1
8	Girsang Sipangan Bolon			1
9	Tanah Jawa	1		
10	Hatonduhan			1
11	Dolok Panribuan			1
12	Jorlang Hataran			1
13	Panei			1
14	Panombeian Panei			1
15	Raya		1	
16	Dolog Masagal			1
17	Dolok Silou			1
18	Silou Kahean		1	
19	Raya Kahean			1
20	Tapian Dolok		1	
21	Dolok Batu Nanggar		1	
22	Siantar	1		
23	Gunung Malela		1	
24	Gunung Maligas		1	
25	Hutabayu Raja		1	
26	Jawa Maraja Bah Jambi			1
27	Pamatang Bandar		1	
28	Bandar Huluan		1	
29	Bandar	1		
30	Bandar Masilam		1	
31	Bosar Maligas		1	
32	Ujung Padang		1	

Table 5. Results of Cluster Iteration 7

C1	3
C2	13
C3	16

After viewing the table grouping the positions of cluster 1, cluster 2 and cluster 3, they have the same value with no change. In this study, the calculation of the k-means clustering process stops at the 7th iteration, because the 7th iteration has the same results as the previous iteration. Rapid Miner is a software created by Dr. Markus Hofmann from the Blanchardstown Institute of Technology and Ralf Klinkenberg from rapid-i.com with a GUI (Graphical User Interface) [22]RapidMiner is a software platform for machine learning, deep learning, text mining (text mining), and predictive analytics. [23][24]

Furthermore, according to the results of the cluster position, C1 has 3 data, C2 is 13 data and C3 is 16 data, until the following conclusions are obtained:

a) High Cluster (C1) with a total population growth data of 3 sub-districts, namely Tanah Jawa, Siantar, Bandar.
b) Medium Cluster (C2) with population growth data of 13 districts, namely Sidamanik, Raya, Silou Kahean, Tapian Dolok, Dolok Batu Nanggar, Mount Malela, Mount Maligas, Hutabayu Raja, Pamatang Bandar, Bandar Huluan, Bandar Masilam, Bosar Maligas, Ujung Padang.
c) Low Cluster (C3) with population growth data of 16 districts, namely Silimakuta, Pematang Silimahuta, Purba, Haranggaol Horison, Dolok Pardamean, Pamatang Sidamanik, Girsang Sipangan Bolon, Hatonduhan, Dolok Panribuan, Jorlang Hataran, Panei, Panombeian Panei, Dolog Masagal, Dolok Silou, Raya Kahean, Java Maraja Bah Jambi.

According to the results of the data that has been obtained, it is concluded that the data used is valid. It is proven that the final results of manual calculations and the RapidMiner 5.3 application get the same final results. Below is a view of the cluster model in the form of text in the RapidMiner 5.3 application which explains that the cluster has been formed:

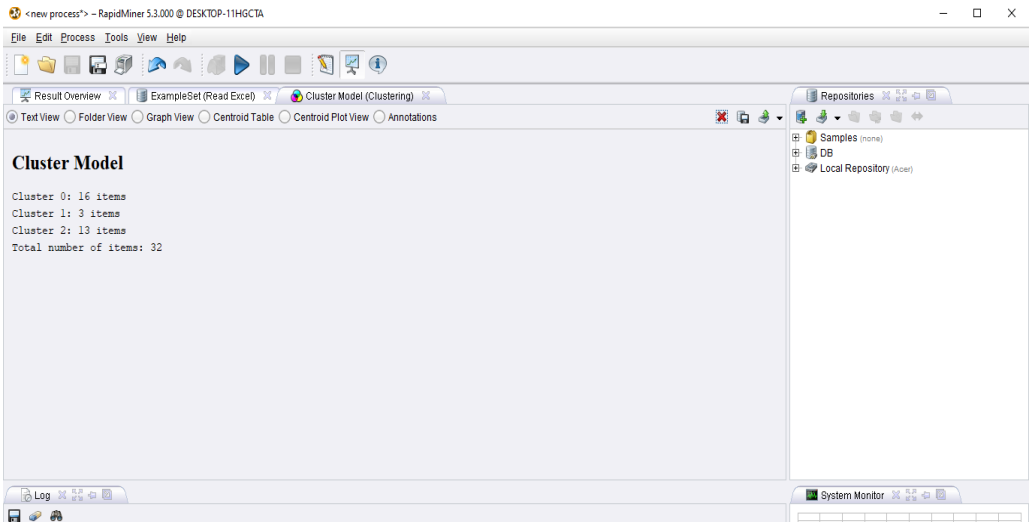

Figure 2. Population Growth Model Cluster

From the picture above, it can be explained that Cluster 0 (low) has 16 items, Cluster 1 (high) has 3 items, Cluster 2 (medium) has 13 items.

4. CONCLUSION

Based on the results of the research that the author has done about population growth in Simalungun district using the k-means clustering algorithm, it can be concluded that the Highest Cluster (C1) with a population growth of 3 districts in Simalungun, Medium Cluster (C2) with a population growth of as many as 3 districts. 13 sub-districts, Low Cluster (C3) with a population growth of 16 districts in Simalungun. Based on the results of the research that the author has done, the distribution of sub-districts in Simalungun Regency has an effect on determining government policy. With the application of the K-means algorithm data mining in the grouping of population growth in Simalungun Regency, it can help the government in improving the welfare of the population based on the number of residents in each Simalungun Regency.

5. ACKNOWLEGMENT

Thanks to the supervisors who are lecturers at AMIK and STIKOM Tunas Bangsa Pematangsiantar who have helped in the process of compiling this research, especially the Central Statistics Agency (BPS) which has provided data. This research is an outcome for completing undergraduate education majoring in informatics engineering at STIKOM Tunas Bangsa, suggestions and constructive

criticism for the improvement of this research are highly expected so that knowledge, especially in the data mining field, can be implemented and developed.

REFERENCES

[1] E. W. F. Peterson, "The role of population in economic growth," *SAGE Open*, vol. 7, no. 4, 2017, doi: 10.1177/2158244017736094.

[2] F. B. Wietzke, "Poverty, Inequality, and Fertility: The Contribution of Demographic Change to Global Poverty Reduction," *Popul. Dev. Rev.*, vol. 46, no. 1, pp. 65–99, 2020, doi: 10.1111/padr.12317.

[3] "Kemiskinan di Kabupaten Simalungun," pp. 1–7, 2014.

[4] M. Sangadji, "Analisis Faktor-Faktor Yang Mempengaruhi Kemiskinan Di Provinsi Maluku," *Media Trend*, vol. 9, no. 2, pp. 162–180, 2014, [Online]. Available: https://docplayer.info/55399035-Analisis-faktor-faktor-yang-mempengaruhi-kemiskinan-di-provinsi-maluku-maryam-sangadji-universitas-pattimura-ambon-abstract.html.

[5] M. Ikbal, "the Implementation of Discretion on Criminal Settlement in the Theft Cases," *IJCLS (Indonesian J. Crim. Law Stud.*, vol. 2, no. 1, pp. 90–101, 2017, doi: 10.15294/ijcls.v2i1.10818.

[6] F. R. Pratiwi, "the Effect of Population Growth and Gross Regional Domestic Product (Grdp) on the Level of Unemployment in the City of Makassar," vol. 3, no. 1, pp. 13–21, 2020.

[7] B. P. Statistik, "Provinsi Sumatera Utara Dalam Angka Tahun 2020," p. 368, 2020, [Online]. Available: https://sumut.bps.go.id/publication/2020/04/27/317f98717fcca50650c40477/provinsi-sumatera-utara-dalam-angka-2020.html.

[8] H. Amalia and E. Evicienna, "Komparasi Metode Data Mining Untuk Penentuan Proses Persalinan Ibu Melahirkan," *J. Sist. Inf.*, vol. 13, no. 2, p. 103, 2017, doi: 10.21609/jsi.v13i2.545.

[9] Y. Mardi, "Data Mining : Klasifikasi Menggunakan Algoritma C4.5," *J. Edik Inform.*, vol. 2, no. 2, pp. 213–219, 2017.

[10] R. A. Asroni, "Penerapan Metode K-Means Untuk Clustering Mahasiswa Berdasarkan Nilai Akademik Dengan Weka Interface Studi Kasus Pada Jurusan Teknik Informatika UMM Magelang," *Ilm. Semesta Tek.*, vol. 18, no. 1, pp. 76–82, 2015.

[11] Y. D. Darmi and A. Setiawan, "Penerapan Metode Clustering K-Means Dalam Pengelompokan Penjualan Produk," *J. Media Infotama*, vol. 12, no. 2, pp. 148–157, 2017, doi: 10.37676/jmi.v12i2.418.

[12] M. Z. Hossain, M. N. Akhtar, R. B. Ahmad, and M. Rahman, "A dynamic K-means clustering for data mining," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 13, no. 2, pp. 521–526, 2019, doi: 10.11591/ijeecs.v13.i2.pp521-526.

[13] W. Dhuhita, "Clustering Menggunakan Metode K-Mean Untuk Menentukan Status Gizi Balita," *J. Inform. Darmajaya*, vol. 15, no. 2, pp. 160–174, 2015.

[14] W. Purba, S. Tamba, and J. Saragih, "The effect of mining data k-means clustering toward students profile model drop out potential," *J. Phys. Conf. Ser.*, vol. 1007, no. 1, 2018, doi: 10.1088/1742-6596/1007/1/012049.

[15] E. M. Sipayung, H. Maharani, and B. A. Paskhadira, "Designing Customer Target Recommendation System Using K-Means Clustering Method," *IJITEE (International J. Inf. Technol. Electr. Eng.*, vol. 1, no. 1, 2017, doi: 10.22146/ijitee.25155.

[16] A. F. Sallaby and E. Suryana, "Penerapan Data Mining untuk Menentukan Jumlah Pencari Kerja Terdaftar Berdasarkan Umur dan Pendidikan Menggunakan K-Means Clustering (Studi Kasus di Dinas Tenaga Kerja Dan Transmigrasi Provinsi Bengkulu)," *J. Technopreneursh. Inf. Syst.*, vol. 1, no. 1, pp. 35–38, 2018, doi: 10.36085/jtis.v1i2.28.

[17] U. B. Mulia, "Jumlah Ritel," vol. XI, no. 1, pp. 32–44.

[18] 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.

[19] K. Simalungun, "RPI2JM Kabupaten Simalungun 2015 - 2019," vol. 2, pp. 1–37, 2019.

[20] S. K. Dini and A. Fauzan, "Clustering Provinces in Indonesia based on Community Welfare Indicators," *EKSAKTA J. Sci. Data Anal.*, vol. 1, no. 1, pp. 56–63, 2020, doi: 10.20885/eksakta.vol1.iss1.art9.

[21] R. Oktavia, J. T. Hardinata, and I. Irawan, "Penerapan Metode Algoritma K-means Dalam Pengelompokan Angka Harapan Hidup Saat Lahir Menurut Provinsi," *Kesatria J. Penerapan ...*, vol. 1, no. 4, pp. 154–161, 2020, [Online]. Available: http://tunasbangsa.ac.id/pkm/index.php/kesatria/article/view/41.

[22] S. Haryati, A. Sudarsono, and E. Suryana, "Implementasi Data Mining Untuk Memprediksi Masa Studi

Mahasiswa Menggunakan Algoritma C4.5 (Studi Kasus: Universitas Dehasen Bengkulu)," *J. Media Infotama*, vol. 11, no. 2, pp. 130–138, 2015.

[23] R. Nofitri and N. Irawati, "Analisis Data Hasil Keuntungan Menggunakan Software Rapidminer," *JURTEKSI (Jurnal Teknol. dan Sist. Informasi)*, vol. 5, no. 2, pp. 199–204, 2019, doi: 10.33330/jurteksi.v5i2.365.

[24] *Uska, M. Z., Wirasasmita, R. H., Usuluddin, U., & Arianti, B. D. D. (2020). Evaluation of Rapidminer-Aplication in Data Mining Learning using PeRSIVA Model. Edumatic: Jurnal Pendidikan Informatika, 4(2), 164-171.*

[25] M. A. W. K. MURTI, "Penerapan Metode K-Means Clustering Untuk Mengelompokan Potensi Produksi Buah – Buahan Di Provinsi Daerah Istimewa Yogyakarta," *Skripsi*, 2017.