Journal of Artificial Intelligence and Engineering Applications
Website: https://ioinformatic.org/

15th October 2021. Vol. 1. No. 1

Application of K-Means Algorithm in Grouping Households Accessing the Internet by Province

Zahra Syahara[1]*, Saifullah[2*], Jalaluddin[3]

[1] STIKOM Tunas Bangsa Pematangsiantar, Sumatera Utara, Indonesia
[2,3] AMIK Tunas Bangsa Pematangsiantar, Sumatera Utara, Indonesia
Jln. Sudirman Blok A No. 1-3 Pematangsiantar, Sumatera Utara
[1]zahrasyahara20@gmail.com, [2]saifullah@amiktunasbangsa.ac.id, [3]jalaluddin@amiktunasbangsa.ac.id

Abstract

The study aims to group households that access the Internet according to the province. As for the data source in this study was made from BPS (the statistical center body) and the data used in the study was in 2017 to 2019 isolated from 34 provinces in Indonesia. The method of artificially synthesizing the research is using a k-means algorithm. According to the data, groups of households that access the Internet according to the provinces are grouped into 2 clusters of high clusters (c1) and low cluster (c2). It is hoped that this study will provide more attention to the government for the provinces that have low Internet access, It could also lead to programs that would seek to improve people's access to the Internet via e-government, telencenter, smart villages, or smart city, and indonesian-to pursue their relationship with more advanced Internet countries such as Europe and America.

Keywords: *Data Mining, K-means Algorithm, Internet Access*

1. Introduction

The internet is a network of interconnected computers for communication and information purposes. A computer in an internet network can be anywhere or even in the whole world [1],[2]. The internet is a product of information and communication technology, and its use is increasing, from children, teenagers, adults, men, women, poor and rich to slum dwellers. In line with the times and with advances in information and communication technology, all activities must use current information technology, one of which is the use of the internet [3],[4]. Not all households in Indonesia have good internet access, so there is a need for grouping to find out which areas of the household still lack good internet access, therefore internet access must be improved so that the needs of households in Indonesia regarding internet access can be met. well fulfilled.

In this study, the author will make a study by classifying households that access the internet by province using the K-Means method [5]. The data obtained is sourced from the Central Statistics Agency (BPS). Where obtained are households that access the internet by province [6]. Data mining is a process that uses statistical techniques, mathematics, artificial intelligence, and machine learning to extract and identify useful information and related knowledge from large databases [7],[8]. Data mining methods have several parts, one of which is Clustering. Clustering has various methods used to group one of them is the K-means method [9],[10]. K-Means is a non-hierarchical data clustering method that tries to partition existing data into one or more clusters [11],[12],[13]. With the advantages of the K-means method, many researchers use the K-means method to group data on a small or large scale [14],[15].

Based on these problems, a research was conducted to classify households that access the internet by province using the K-Means Clustering Data Mining Algorithm, namely by clustering households that access the internet into 2 clusters, namely high clusters and low clusters.

2. Results and Discussion

In performing clustering, the data obtained will be calculated first. Determination of this cluster point is done by taking the largest (maximum) value for the highest cluster (C1), the smallest value for the low cluster (C2) as follows:

2.1. Data Processing Using K -Means Algoritma Algorithm

Table 1: Internet User Data by Province

Province	Urban			Rural			Urban+rural		
	2017	2018	2019	2017	2018	2019	2017	2018	2019
Aceh	64,57	75,99	78,58	36,45	48,21	58,68	44,83	56,89	65,16
North Sumatra	75,99	78,58	79,77	38,95	47,05	56,30	52,15	60,70	68,91
West Sumatra	71,33	76,77	82,43	43,25	53,91	58,52	54,91	64,00	69,67
Riau	75,29	82,98	85,41	47,48	59,08	69,55	58,41	68,73	76,00
Jambi	69,07	76,15	84,26	43,86	56,19	64,51	51,49	62,43	70,81
South Sumatra	71,57	77,05	81,66	38,05	49,56	58,10	49,73	59,41	66,42
Bengkulu	77,27	79,94	86,00	36,64	48,36	58,19	49,76	58,49	67,36
Lampung	65,93	76,20	82,62	38,03	54,30	62,85	45,25	60,41	68,68
Kep. Bangka Belitung	67,02	74,83	81,67	41,30	55,42	66,26	54,76	65,78	74,80
Kep. Riau	78,10	81,98	90,39	45,67	56,29	66,59	73,33	78,41	87,96
Dki Jakarta	85,70	89,04	93,33	-	-	-	85,70	89,04	93,33
West Java	69,76	77,53	82,53	43,56	52,28	62,52	62,04	70,61	77,55
Central Java	66,17	74,39	80,81	49,23	58,85	69,40	57,48	66,73	75,16
Di Yogyakarta	78,81	85,11	87,24	53,46	62,30	72,91	71,71	79,10	83,68
East Java	67,03	74,70	80,82	45,33	54,50	64,59	56,36	65,01	73,24
Banten	74,81	84,08	89,61	38,27	53,53	63,00	64,11	75,39	82,25
Bali	76,18	81,90	86,58	50,27	58,41	63,82	67,10	74,15	79,59
West Nusa Tenggara	52,23	61,83	74,40	35,61	45,48	56,94	42,95	53,03	62,25
East Nusa Tenggara	73,05	78,20	82,79	25,87	31,06	39,33	36,18	42,21	49,83
West Kalimantan	76,49	80,56	85,95	32,01	42,49	53,35	45,81	54,99	64,71
Central Kalimantan	74,08	76,64	86,41	41,25	50,44	62,49	52,92	60,31	71,84
South Kalimantan	72,18	79,95	84,71	42,81	55,16	65,04	55,66	66,67	74,35
East Kalimantan	77,14	85,34	88,83	53,25	66,16	74,38	69,06	78,98	84,17
North Kalimantan	74,89	82,79	88,83	25,87	31,06	39,33	36,18	42,21	49,83
North Sulawesi	75,34	78,92	82,79	49,45	56,21	64,50	61,78	67,60	74,06
Central Sulawesi	78,76	76,97	83,30	37,00	44,09	52,60	47,77	53,42	61,66
South Sulawesi	75,54	80,67	84,42	43,65	54,77	63,86	55,95	65,22	72,62
Southeast Sulawesi	68,30	80,32	86,27	42,52	50,64	61,90	50,85	61,95	71,21
Gorontalo	73,01	74,87	81,77	43,53	56,55	65,83	54,52	63,76	72,68
West Sulawesi	62,91	70,56	73,74	36,04	44,33	54,69	41,31	50,44	59,09
Maluku	72,15	79,50	80,78	30,13	36,67	40,32	47,81	55,16	58,52
North Maluku	72,15	80,78	83,31	26,11	35,66	41,17	39,23	49,06	53,61
West Papua	69,70	79,22	85,06	36,39	50,45	53,71	49,18	61,95	66,62
Papua	68,22	72,36	77,95	12,94	14,69	14,35	27,33	29,50	31,31

Determination of this cluster point is done by taking the largest (maximum) value for the highest cluster (C1), the smallest value for the low cluster (C2). The point value can be known as follows:

Table 2: Initial Cluster Center

Cluster	2017	2018	2019
C1	85,70	89,04	93,33
C2	27,33	31,31	31,31

The cluster process to calculate the distance between the data and the center of the cluster uses equation (2). The process of searching the distance of data grouping in iteration 1 and data clustering can be described in the following table:

Table 3: Results of Iteration 1

Province	Hight	Low	Distance	C1	C2
Aceh	59,1399	46,92851	46,928505		1
North Sumatra	50,25032	54,80176	50,250318	1	
West Sumatra	46,20413	58,50125	46,204126	1	
Riau	38,17812	67,0981	38,178123	1	
Jambi	48,84226	56,81849	48,842262	1	
South Sumatra	53,81831	51,27456	51,274557		1
Bengkulu	53,83633	51,41136	51,411356		1
Lampung	55,34891	51,70175	51,701754		1
Kep. Bangka Belitung	42,91471	62,92872	42,9147107	1	
Kep. Riau	17,17122	87,84879	17,171217	1	
Dki Jakarta	0	103,9161	0,000000	1	
West Java	33,88907	70,94331	33,889068	1	
Central Java	40,30203	64,9455	40,302027	1	
Di Yogyakarta	19,68873	84,68979	19,688733	1	
East Java	42,91718	62,14359	42,917183	1	
Banten	27,84272	77,80452	27,842719	1	
Bali	27,50381	76,85203	27,503812	1	
West Nusa Tenggara	63,95505	41,89187	41,891872		1
East Nusa Tenggara	80,85499	24,14243	24,142432		1
West Kalimantan	59,74713	45,90001	45,90005		1
Central Kalimantan	48,59796	56,98057	48,597957	1	
South Kalimantan	41,98879	63,53455	41,988795	1	
East Kalimantan	21,49416	83,56939	21,494157	1	
North Kalimantan	80,85499	24,14243	24,142432		1
North Sulawesi	37,45895	66,82795	37,458950	1	
Central Sulawesi	60,91353	43,71593	43,715950		1
South Sulawesi	43,37464	61,65662	43,374635	1	
Southeast Sulawesi	49,37292	56,55266	49,372918	1	
Gorontalo	45,14082	60,20399	45,140816	1	
West Sulawesi	68,06475	37,49203	37,392031		1
Maluku	61,60554	42,64094	42,640944		1
North Maluku	73,04478	31,96081	31,960814		1
West Papua	52,73521	52,69935	52,699346		1
Papua	103,9161	0	0,000000		1

Based on the matrix obtained in the table above, the following groupings are obtained:

C1 = 2,3,4,5,9,10,11,12,13,14,15,16,17,21,22,23,25,27,28,2

C2 = 1,6,7,8,18,19,20,24,26,30,31,32,33,34

After getting the results from each cluster, then the new cluster center is calculated based on the member data of each cluster that has been obtained using a formula that corresponds to the center member cluster. Here are the cluster centers for iteration 2:

Table 4: New Cluster Center

Cluster	2017	2018	2019
C1	60,51	69,23	76,69
C2	43,62	52,21	59,20

Calculations for cluster C1:

$C_{aceh,C1}=\sqrt{(44{,}83-85{,}70)^2+(56{,}89-89{,}04)^2+(65{,}16-93{,}33)^2}$= 59,1399

$C_{North\ Sumatra,C1}=\sqrt{(52{,}15-85{,}70)^2+(60{,}70-89{,}04)^2+(68{,}91-93{,}33)^2}$= 50,25032

$C_{West\ Sumatra,C1}=\sqrt{(54{,}91-85{,}70)^2+(64{,}00-89{,}04)^2+(69{,}67-93{,}33)^2}$= 46,20413

…….

Furthermore, the calculation of the other provinces is carried out in the same way as the example above. The following is the calculation for cluster C2 is:

$C_{aceh,}C_2=\sqrt{(44{,}83-27{,}33)^2+(56{,}89-29{,}50)^2+(65{,}16-31{,}31)^2}$= 46,92851

$C_{North\ Sumatra,}C_2=\sqrt{(52{,}15-27{,}33)^2+(60{,}70-29{,}50)^2+(68{,}91-31{,}31)^2}$= 54,80176

$C_{West\ Sumatra,}C_2=\sqrt{(54{,}91-27{,}33)^2+(64{,}00-29{,}50)^2+(69{,}67-31{,}31)^2}$= 58,50125

……...

Furthermore, the calculation for other provinces is carried out in the same way as the example above. The following are the results of the calculation of the iteration 2 cluster distance:

Table 5: Calculation Results of Iteration 2

Province	Hight	Low	Distance	C1	C2
Aceh	23,04941	7,672567	7,672567		1
North Sumatra	14,25801	15,46368	14,258009	1	
West Sumatra	10,39363	19,39294	10,396355	1	
Riau	2,271356	27,81878	2,2713560	1	
Jambi	12,73902	17,35408	12,739021	1	
South Sumatra	17,83994	11,88673	11,886727		1
Bengkulu	17,83531	11,98845	11,988445		1
Lampung	19,36479	12,63842	12,638424		1
Kep. Bangka Belitung	6,970915	23,48597	6,9709151	1	
Kep. Riau	19,37678	48,95183	19,376783	1	
Dki Jakarta	36,10487	65,51383	36,104867	1	
West Java	2,228997	31,85246	2,2289971	1	
Central Java	4,219461	25,64465	4,2194606	1	
Di Yogyakarta	16,47999	45,94995	16,479985	1	
East Java	6,854614	22,87507	6,8546136	1	
Banten	9,04217	38,5803	9,0421700	1	
Bali	8,716467	38,05831	8,7164665	1	
West Nusa Tenggara	27,9201	3,227275	3,2272747		1
East Nusa Tenggara	45,20887	15,59375	15,593753		1
West Kalimantan	23,71863	6,548377	6,5483772		1
Central Kalimantan	12,6798	17,66002	12,679802	1	
South Kalimantan	5,967112	24,15703	5,9671122	1	
East Kalimantan	14,96689	44,57937	14,966893	1	
North Kalimantan	45,20887	15,59375	15,593753		1
North Sulawesi	3,345364	28,06195	3,3453753	1	
Central Sulawesi	25,26563	4,975441	4,9754413		1
South Sulawesi	7,314384	22,39155	7,3143836	1	
Southeast Sulawesi	13,28358	17,06945	13,283575	1	
Gorontalo	9,052664	20,83073	9,0526636	1	
West Sulawesi	32,12042	2,911706	2,9117057		1
Maluku	26,26064	5,169923	5,1699228		1
North Maluku	37,31849	7,774036	7,7740356		1
West Papua	16,82061	13,4467	13,446699		1
Papua	68,84237	39,48414	39,484136		1

Based on the matrix obtained in the table above, the following groupings are obtained:

C1 = 2,3,4,5,9,10,11,12,13,14,15,16,17,21,22,23,25,27,28,29

C2 = 1,6,7,8,18,19,20,24,26,30,31,32,33

2.2. Testing With RapidMiner

The following is an explanation of how to enter new data to be executed further, in this case the data to be executed is in the form of excel data. The stages can be seen in the image below :

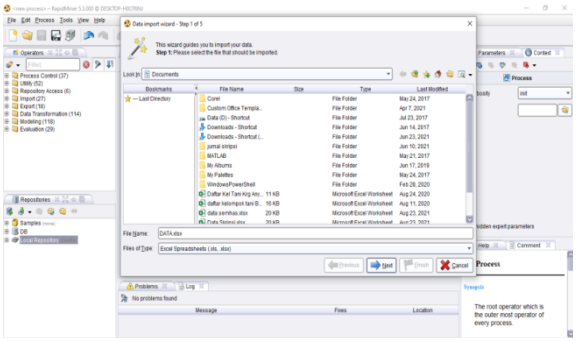

Figure 1: Excel Data Import Process

To input excel data, you can use two ways, namely through the filter section by typing read excel or through the repositories section and then Import Excel Sheet. In the look in section we can find where the excel data file that we saved is located. As in the picture, the data is stored in the data folder with the data name as shown in the picture. Next, a screen like Figure 2.

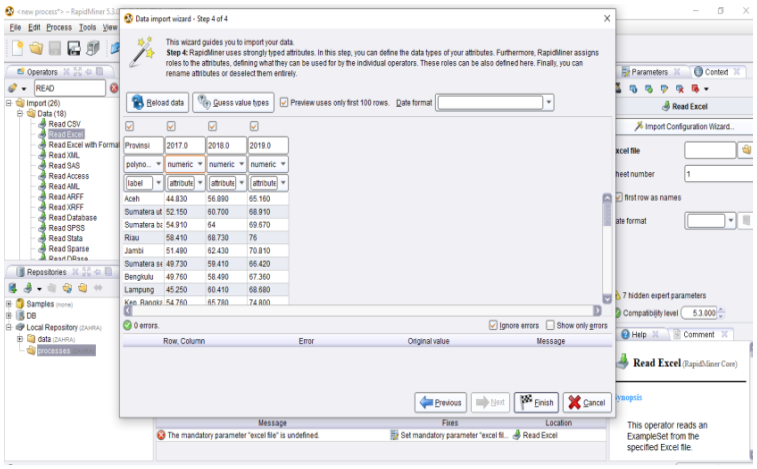

Figure 2: Import Excel data

At this stage the selection of data types is carried out where at this stage chooses the part that will be given the type of "label" which will be the determinant of the formation of the grouping. To drag and drop the selected data into the process view. The following is the flow of the data import process carried out.

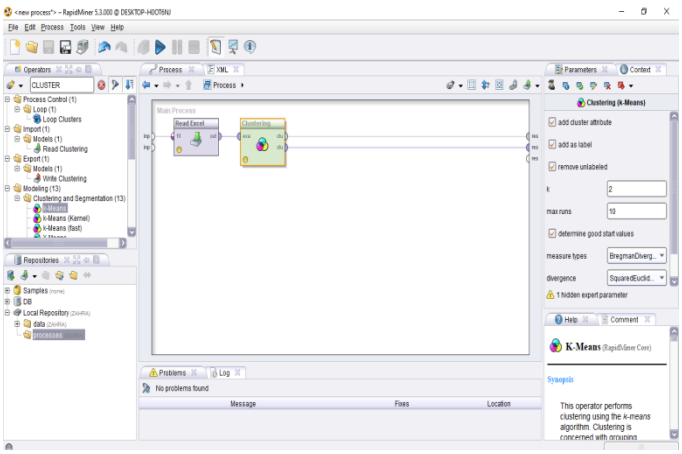

Figure 3: System Process

After the selected data is dragged into the main process, then in the filter section, type k-means. Then the k-means operator will appear then select and drag it into the main process page. Next, connect the read excel data with the k-means oper-

ator and then click run tools. After performing the system process as shown in Figure 4, the results from Rapidminer will appear as follows:

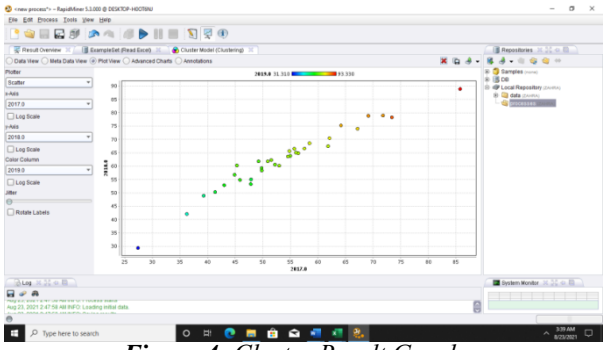

Figure 4: *Cluster Result Graph*

Based on Figure 4. it can be seen that cluster 0 is a high cluster and cluster 1 is a low cluster.

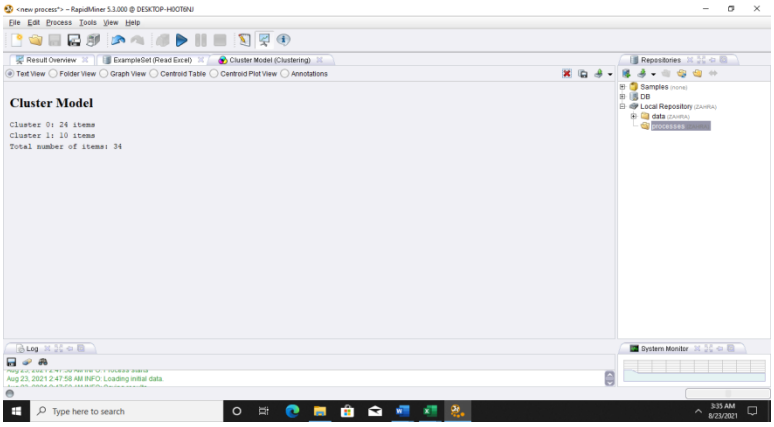

Figure 5: *Cluster Model*

From Figure 5. it can also be seen that there is 1 province in the high cluster and 33 provinces in the low cluster. To validate the data, the algorithm calculation must produce a final result in the form of grouping with 2 clusters, and the data used is valid data and is the same as that used in the Rapidminer tools. The following shows the results obtained from algorithm calculations and tests on rapidminer.

Table 6: Comparison of Manual and RapidMiner Results

No	Province	Manual Calculation Results	*RapidMiner* Results
1	Aceh	Cluster 1	Cluster 1
2	North Sumatra	Cluster 0	Cluster 0
3	West Sumatra	Cluster 0	Cluster 0
4	Riau	Cluster 0	Cluster 0
5	Jambi	Cluster 0	Cluster 0
6	South Sumatra	Cluster 0	Cluster 0
7	Bengkulu	Cluster 0	Cluster 0
8	Lampung	Cluster 0	Cluster 0
9	Kep. Bangka Belitung	Cluster 0	Cluster 0
10	Kep. Riau	Cluster 0	Cluster 0
11	Dki Jakarta	Cluster 0	Cluster 0
12	West Java	Cluster 0	Cluster 0
13	Central Java	Cluster 0	Cluster 0
14	Di Yogyakarta	Cluster 0	Cluster 0
15	East Java	Cluster 0	Cluster 0
16	Banten	Cluster 0	Cluster 0
17	Bali	Cluster 0	Cluster 0

No	Province	Manual Calculation Results	*RapidMiner* Results
18	West Nusa Tenggara	Cluster 1	Cluster 1
19	East Nusa Tenggara	Cluster 1	Cluster 1
20	West Kalimantan	Cluster 1	Cluster 1
21	Central Kalimantan	Cluster 0	Cluster 0
22	South Kalimantan	Cluster 0	Cluster 0
23	East Kalimantan	Cluster 0	Cluster 0
24	North Kalimantan	Cluster 1	Cluster 1
25	North Sulawesi	Cluster 0	Cluster 0
26	Central Sulawesi	Cluster 1	Cluster 1
27	South Sulawesi	Cluster 0	Cluster 0
28	Southeast Sulawesi	Cluster 0	Cluster 0
29	Gorontalo	Cluster 0	Cluster 0
30	West Sulawesi	Cluster 1	Cluster 1
31	Maluku	Cluster 1	Cluster 1
32	North Maluku	Cluster 1	Cluster 1
33	West Papua	Cluster 0	Cluster 0
34	Papua	Cluster 1	Cluster 1

Based on the comparison table above, it can be seen that the results of manual calculations are the same as those of rapidminer. Where the high cluster is 14 provinces, namely (Aceh, South Sumatra, Bengkulu, Lampung, West Nusa Tenggara, East Nusa Tenggara, West Kalimantan, North Kalimantan, Central Sulawesi, West Sulawesi, Maluku, North Maluku, West Papua, Papua), and clusters There are 20 low, namely (North Sumatra, West Sumatra, Riau, Jambi, Bangka Belitung Islands, Riau Islands, DKI Jakarta, West Java, Central Java, Yogyakarta, East Java, Banten, Bali, Central Kalimantan, South Kalimantan, Kalimantan East, North Sulawesi, South Sulawesi, Southeast Sulawesi, Gorontalo).

3. Conclusion

Based on the previous discussion, it can be concluded that the application of data mining using the k-means algorithm on the grouping of households that access the internet by province can be applied. Sources of data used in this study is data obtained from BPS (Central Bureau of Statistics). The amount of data used is 34 provinces consisting of 2017-2019. From the results of the grouping obtained two clusters, namely high and low. The high cluster consists of 14 provinces and the low cluster consists of 20 provinces. Data testing on Rapidminer 5.3 using the k-means algorithm can display the accuracy of the data between manual and system calculations.

REFERENCE

[1] M. Z. Aminy, "Pemanfaatan Media Internet Sebagai Sumber Belajar Mahasiswa Program Studi Pendidikan Matematika Di STKIP BIMA Semester Ganjil Tahun Pelajaran 2013/2014," *J. Kip*, vol. 4, no. 2, pp. 927–932, 2015, [Online]. Available: http://journals.ukitoraja.ac.id/index.php/jkip/article/view/59.

[2] lestari dan Chariri, "Mempengaruhi Pelaporan Keuangan Melalui Internet (Internet Financial Reporting) Dalam Website Perusahaan," *http://eprints.undip.ac.id/2398/1/IFR_research.pdf*, pp. 0–27, 2011, [Online]. Available: http://eprints.undip.ac.id/2398/1/IFR_research.pdf.

[3] M. Rustam and M. Rustam, "INTERNET DAN PENGGUNAANNYA (Survei di Kalangan Masyarakat Kabupaten Takalar Provinsi Sulawesi Selatan) (Survey Among the People of Takalar Town , South Sulawesi Province)," pp. 13–24, 2017.

[4] R. Dyah, "Hubungan Antara Kontrol Diri Dengan Kecanduan Internet Pada Siswa Sekolah Menengah Pertama (Smp)," *J. Hum.*, 2018.

[5] A. F. Yang, "Jawa Timur Dalam Mengakses Internet Tahun 2017," 2018.

[6] M. Apriliani, E. Antriyandarti, and W. Rahayu, "Media Trend," *Media Trend*, vol. 15, no. 2, pp. 217–226, 2020.

[7] A. N. Khomarudin, "Teknik Data Mining : Algoritma K-Means Clustering," pp. 1–12, 2016.

[8] N. Rofiqo, A. P. Windarto, and D. Hartama, "Penerapan Clustering Pada Penduduk Yang Mempunyai Keluhan Kesehatan Dengan Datamining K-Means," *KOMIK (Konferensi Nas. Teknol. Inf. dan Komputer)*, vol. 2, no. 1, pp. 216–223, 2018, doi: 10.30865/komik.v2i1.929.

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

[10] T. Alfina and B. Santosa, "Analisa Perbandingan Metode Hierarchical Clustering, K-Means dan Gabugan Keduanya dalam Membentuk Cluster Data (Studi Kasus : Problem Kerja Praktek Jurusan Teknik Industri ITS)," *Anal. PerbandinganMetode Hierarchical Clust. K-means dan Gabungan Keduanya dalam Clust. Data*, vol. 1, no. 1, pp. 1–5, 2012.

[11] A. Aditya, I. Jovian, and B. N. Sari, "Implementasi K-Means Clustering Ujian Nasional Sekolah Menengah

Pertama di Indonesia Tahun 2018/2019," *J. Media Inform. Budidarma*, vol. 4, no. 1, p. 51, 2020, doi: 10.30865/mib.v4i1.1784.

[12] Yudi Agusta, "K-Means – Penerapan, Permasalahan dan Metode Terkait," *J. Sist. dan Inform.*, vol. 3, no. Februari, pp. 47–60, 2007.

[13] I. Parlina, A. P. Windarto, A. Wanto, and M. R. Lubis, "Memanfaatkan Algoritma K-Means Dalam Menentukan Pegawai Yang Layak Mengikuti Asessment Center," *Memanfaatkan Algoritm. K-Means Dalam Menentukan Pegawai Yang Layak Mengikuti Asessment Cent. Untuk Clust. Progr. Sdp*, vol. 3, no. 1, pp. 87–93, 2018.

[14] M. Nishom and M. Y. Fathoni, "Implementasi Pendekatan Rule-Of-Thumb untuk Optimasi Algoritma K-Means Clustering," *J. Inform. J. Pengemb. IT*, vol. 3, no. 2, pp. 237–241, 2018, doi: 10.30591/jpit.v3i2.909.

[15] E. Muningsih and S. Kiswati, "Penerapan Metode K-Means Untuk Clustering Produk Online Shop Dalam Penentuan Stok Barang," *J. Bianglala Inform.*, vol. 3, no. 1, pp. 10–17, 2015.