Journal of Artificial Intelligence and Engineering Applications

Website: https://ioinformatic.org/

15th June 2022. Vol. 1. No. 3

Application Of K-Means Algorithm In Grouping Productive Seed Distribution Data In BPDASHL Asahan Barumun

Dina Patresia Samuana Manurung[1]*, Muhammad Ridwan Lubis[2], Ika Okta Kirana[3], Dedy Hartama[4], Dedi Suhendro[5]

[1,3,4]STIKOM Tunas Bangsa Pematangsiantar, North Sumatra, Indonesia
[2,5]AMIK Tunas Bangsa Pematangsiantar, North Sumatra, Indonesia
**manurungdina153@gmail.com*

Abstract

Preserving the environment is a human effort that must be done immediately so that survival can be maintained properly. One of the human efforts in preserving the environment is planting and maintaining trees in the surrounding environment. Balai Pengelolaan Daerah Aliran Sungai dan Hutan Lindung (BPDASHL) Asahan Barumun has a Permanent Nursery that produces 19 types of productive seeds, where productive seeds have an ecological impact for reforestation and an economic impact to improve people's welfare. BPDASHL Asahan Barumun provides and distributes productive seeds to people who want to participate in preserving the environment. Before distributing productive seeds, the nursery staff of BPDASHL Asahan Barumun conducted data collection which was added to the distribution data for productive seeds to find out to whom and how many seeds were distributed. In the data on the distribution of productive seeds of the Asahan Barumun BPDASHL, it can be seen that almost every day the distribution of productive seeds to the community is carried out, so the addition of data to the distribution data is getting more and more. Data mining is able to process large data into information in the form of patterns that have meaning for decision support. By using K-Means algorithm in classifying the 2019/2020 BPDASHL Asahan Barumun distribution data by type, so that the final results obtained are 3 clusters where there are 6 seeds that are most in demand, including suren, jengkol, mahogany, avocado, durian, coffee, 10 seeds that are quite in demand, including pine, calliandra, macadamia, petai, sugar palm, cempedak, frankincense, mango, africa, trembesi, and 3 seeds that are less desirable, including meranti, jackfruit, macadamia nut.

Keywords: *Productive Seeds, Data Distribution, Data mining, K-Means*

1. Introduction

Preserving the environment is a human effort that must be done immediately for its survival can be maintained properly[1], [2]. Planting and caring for trees around the residential environment is an effort to preserve the environment[3]. Watershed Management Center and Protected Forest (BPDASHL) Asahan Barumun is a technical implementation unit of the Ministry of Environment and Forestry (KLHK). BPDASHL Asahan Barumun has a Permanent Nursery which aims to produce productive seeds. Permanent Nursery produces 19 types of productive seeds which are provided free of charge by BPDASHL Asahan Barumun for people who want to participate in preserving the environment. Productive seeds are planting material that will produce quality plants with ecological impacts for reforestation and economic impacts to improve people's welfare. BPDASHL Asahan Barumun distributes productive seeds to communities who have submitted proposals to the BPDASHL Asahan Barumun Office. Before distributing productive seeds, Nursery Officers conduct data collection first which is added to the distribution data. Distribution data is used to find out to whom and how many seeds are distributed. In the data on the distribution of productive seeds of the BPDASHL Asahan Barumun in 2019/2020, the distribution of productive seeds to the community is carried out almost every day. With the distribution that is carried out almost every day, the addition of data to the distribution data is getting more and more.

Data Mining is a series of processes in searching for patterns, relationships, extracting added value from large data and information in the form of knowledge with the aim of finding relationships and simplifying data in order to obtain understandable and useful information with the help of statistics and mathematics[4],[5],[6]. Clustering is a data mining method that is unsupervised[7]. Clustering is a process of grouping a number of data or objects into clusters (groups) so that each in the same cluster will contain data that is as similar as possible and different from objects in other clusters[8],[9],[10].

K-means is an algorithm that groups data into one or more clusters so that data with the same characteristics are grouped into one cluster and data with different characteristics are grouped into other groups[11].

[12] Applies the K-Means method in determining the stock of goods at Ragam Jogja.. This research was conducted to facilitate online shops in determining the minimum stock of each item that must be met based on consumer interest by grouping online shop products. In grouping online shop products using the K-Means method, it produces C1 with the 3 most desirable products, the stock must be large, C2 with 11 products that are of interest, the stock is moderate, and C3 with 17 less desirable products, the stock is small. [12] Yulianti applies K-Means algorithm to determine customer interest in the Hijab Shop. This study was conducted to create a sales strategy to find out which hijab products are most in demand at the Hijab Shop by grouping hijab products using K-Means algorithm. The conclusion shows that the results obtained using K-Means algorithm are C1 with 3 hijab brands that have low fans, C2 with 4 hijab brands that have low fans, and C3 with 3 hijab brands that have the highest interest [13]. [14] Applies K-Means algorithm to determine reader interest in Dinas Perpustakaan Kabupaten Karawang. This research was conducted to help officers arrange bookshelves according to the level of interest of the readers by grouping the books. The conclusion shows that the research results obtained using K-Means algorithm are 31 books that are most in demand as C1, 3 books that are quite attractive as C2, 4 books that are less attractive as C3. [15]applies K-Means algorithm in classifying flood-prone areas in Indonesia. Based on the results of the study, there were 11 provinces in C0, 2 provinces in C1, 4 provinces in C2, 3 provinces in C3, and 14 provinces in C4. [16] applies the K-Means algorithm in grouping divorce cases in Jambi City. Based on the results of the study, it was found that the divorce rate with 30 villages in the low divorce rate as C0, 20 villages in the medium divorce rate as C1, 11 villages in the high divorce rate as C2.

2. Research methodology

The research methodology is the steps taken by the author in solving research problems. This research was conducted to apply K-Means algorithm in classifying the distribution data of productive seeds by type. The method used in the grouping is the K-Means algorithm. In this study, the research design is a research flow that is used to explain and solve the problems in the research. The research design can be seen in Figure 1 below:

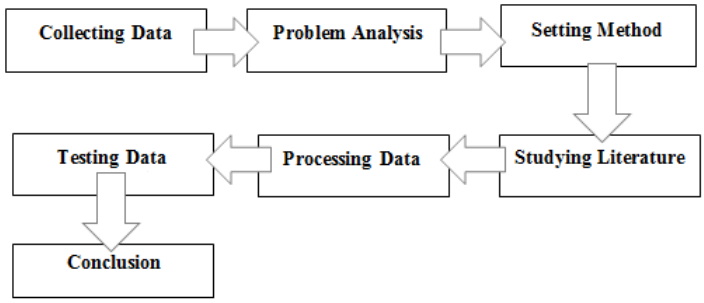

Figure 1. Research Design

Based on Figure 1 above, it is explained that this research was carried out in stages, namely: Collecting Data process is carried out in obtaining the information needed to achieve the research objectives. The following is the procedure for collecting data, specifically Observation, Interview and. Literature Review. Next is Problem Analysis, namely analyzing the problems related to the distribution of productive seeds in knowing the types of productive seeds that are of interest to the community. Setting Method used by the author in solving research problems is the K-Means algorithm. Studying Literature must be based on the references used to obtain information and theories that support the research. Processing Data is carried out using the K-Means algorithm by determining the attributes to be used. Testing Data is carried out using the RapidMiner application to determine the suitability of the final decision results obtained from the K-Means algorithm. Conclusion, In this study the conclusion obtained is the result of grouping the distribution data of productive seeds by type in BPDASHL Asahan Barumun, so that they know the types of productive seeds that are in demand by the community to facilitate the supply of productive seeds according to to the attributes in the distribution data.

Table 1. Data Transformation Results

No	Seed Type	Number of Fans	Number of Seeds
1	Pine	36	55000
2	Suren	71	85000
3	Jengkol	50	65000
4	Mahogany	62	70000

5	Caliandra	38	50000
6	Macadamia	20	40000
7	Avocado	46	65000
8	Petai	36	60000
9	Aren	44	40000
10	Durian	52	85000
11	Coffee	50	90000
12	Meranti	19	30000
13	Cempedak	34	42500
14	Jackfruit	23	25000
15	Macadamia Nut	14	2500
16	Frankincense	17	35000
17	Mango	46	60000
18	Africa	44	50000
19	Trembesi	58	50000

3. Results And Discussion

The data sample used is data on the distribution of productive seeds in 2019/2020 obtained from BPDASHL Asahan Barumun. The following is a calculation using K-Means algorithm:

a. Determine The Data To Be Clustered

Based on the sample data obtained, the data consisted of 19 types of productive seeds. The sample data to be clustered with K-Means algorithm can be seen in table 2:

Table 2. Sample Data for Distribution of Productive Seeds in 2019/2020

No	Seed Type	Number of Customer	Number of Seeds
1	Pine	36	55000
2	Suren	71	85000
3	Jengkol	50	65000
4	Mahogany	62	70000
5	Caliandra	38	50000
6	Macadamia	20	40000
7	Avocado	46	65000
8	Petai	36	60000
9	Aren	44	40000
10	Durian	52	85000
11	Coffee	50	90000
12	Meranti	19	30000
13	Cempedak	34	42500
14	Jackfruit	23	25000
15	Macadamia Nut	14	2500
16	Frankincense	17	35000
17	Mango	46	60000
18	Africa	44	50000
19	Trembesi	58	50000

b. Determine the Number of Clusters
The number of clusters used in the productive seed distribution data is three clusters, namely the most desirable cluster (C1), the most desirable cluster (C2), and the least desirable cluster (C3).

c. Determine the Centroid Value
In determining the value of the centroid, it is necessary to make a provision that the cluster used is three clusters, then the value of the centroid is determined to be 3 centroids. Determination of the centroid is done manually or randomly taken from the sample data. The specified centroid value can be seen in table 3 as follows:

Table 3. Centroid Early

Centroid		
Centroid 1	71	85000
Centorid 2	20	40000
Centroid 3	23	25000

d. Calculating Distance from Centroid
The distance between the centroid point and the point of each object can be calculated using the Euclidean Distance formula as follows:

$$D(i,j) = \sqrt{(X_{1i}^{\ 2} + X_{1j})^2 + (X_{2i}^{\ 2} + X_{2j})^2 + ... + (X_{ki}^{\ 2} + X_{kj})^2}$$

The process of calculating the distance from the 1st centroid to the 1st iteration is as follows:

$$D_{x1,c1} = \sqrt{(36-71)^2 + (55000-85000)^2} = 30000,02042$$

$$D_{x2,c1} = \sqrt{(71-71)^2 + (85000-85000)^2} = 0$$

$$D_{x3,c1} = \sqrt{(50-71)^2 + (65000-85000)^2} = 20000,01102$$

$$D_{x4,c1} = \sqrt{(62-71)^2 + (70000-85000)^2} = 15000,0027$$

$$D_{x5,c1} = \sqrt{(38-71)^2 + (50000-85000)^2} = 35000,01556$$

And so on until $D_{x19,c1}$. Furthermore, the calculation of the distance from the 2nd centroid for the 1st iteration is as follows:

$$D_{x1,c2} = \sqrt{(36-20)^2 + (55000-40000)^2} = 15000,0853$$

$$D_{x2,c2} = \sqrt{(71-20)^2 + (85000-40000)^2} = 45000,0289$$

$$D_{x3,c2} = \sqrt{(50-20)^2 + (65000-40000)^2} = 25000,018$$

$$D_{x4,c2} = \sqrt{(62-20)^2 + (70000-40000)^2} = 30000.0294$$

$$D_{x5,c2} = \sqrt{(38-20)^2 + (50000-40000)^2} = 10000,0162$$

And so on until Dx19,c2 Furthermore, the calculation of the distance from the 3rd centroid for the 1st iteration is as follows:

$$D_{x1,c3} = \sqrt{(36-23)^2 + (55000-25000)^2} = 30000,00282$$

$$D_{x2,c3} = \sqrt{(71-23)^2 + (85000-25000)^2} = 60000,0192$$

$$D_{x3,c3} = \sqrt{(50-23)^2 + (65000-25000)^2} = 40000,0911$$

$$D_{x4,c3} = \sqrt{(62-23)^2 + (70000-25000)^2} = 45000.169$$

$$D_{x5,c3} = \sqrt{(38-23)^2 + (50000-25000)^2} = 25000,0045$$

And so on until Dx19,c2. The distance from the calculation results will be compared and the closest distance between the data and the centroid is selected, the distance will indicate that the data is in one group with the nearest centroid. The following is a table of the closest distance from the 1st iteration centroid, which can be seen in table 4 below:

Table 4. Closest Distance in 1st Iteration

C1	C2	C3	Closest Distance
30000.02042	15000.00853	30000.00282	15000.00853
0	45000.0289	60000.0192	0
20000.01102	25000.018	40000.00911	20000.01102
15000.0027	30000.0294	45000.0169	15000.0027
35000.01556	10000.0162	25000.0045	10000.0162
45000.0289	0	15000.0003	0
20000.01562	25000.01352	40000.00661	20000.01562
25000.0245	20000.0064	35000.00241	20000.0064
45000.0081	24	15000.0147	24
19	45000.01138	60000.00701	19
5000.0441	50000.009	65000.00561	5000.0441
55000.02458	10000.00005	5000.0016	5000.0016
42500.01611	2500.0392	17500.00346	2500.0392
60000.0192	15000.0003	0	0
82500.01969	37500.00048	22500.0018	22500.0018
50000.02916	5000.0009	10000.0018	5000.0009
25000.0125	20000.0169	35000.00756	20000.0169
35000.01041	10000.0288	25000.00882	10000.0288
35000.00241	10000.0722	25000.0245	10000.0722

e. Determine the Cluster or Grouping
In determining the members of each cluster by looking for data values that have the minimum value and are placed in the cluster that matches the minimum value. The following are the members of each cluster in the 1st iteration, which can be seen in table 5. below:

Table 5. Cluster in 1st Iteration

No	Seed Type	C1	C2	C3
1	Pine		1	
2	Suren	1		
3	Jengkol	1		
4	Mahogany	1		
5	Caliandra		1	

6	Macadamia		1	
7	Avocado	1		
8	Petai		1	
9	Aren		1	
10	Durian	1		
11	Coffee	1		
12	Meranti			1
13	Cempedak		1	
14	Jackfruit			1
15	Macadamia Nut			1
16	Frankincense		1	
17	Mango		1	
18	Africa		1	
19	Trembesi		1	

Explanation:
C1: (Suren, Jengkol, Mahogany, Avocado, Durian, Coffee)
C2: (Pinus, Kaliandra, Macadamia, Petai, Aren, Cempedak, Frankincense, Mango, Africa, Trembesi)
C3: (Meranti, Jackfruit, Macadamia Nuts)
In K-Means algorithm, the calculation process stops if the resulting cluster in the next iteration is the same as the cluster in the previous iteration. Then the next is to look for clusters in the next iteration the same as the previous iteration. Here is the process of calculating the 2nd iteration

In determining the members of each cluster by looking for data values that have the minimum value and are placed in the cluster that matches the minimum value. It can be seen the members of each cluster in the 2nd iteration in table 6 as follows:

Table 6. Cluster in 2nd Iteration

No	Seed Type	C1	C2	C3
1	Pine		1	
2	Suren	1		
3	Jengkol	1		
4	Mahogany	1		
5	Caliandra		1	
6	Macadamia		1	
7	Avocado	1		
8	Petai		1	
9	Aren		1	
10	Durian	1		
11	Coffee	1		
12	Meranti			1
13	Cempedak		1	
14	Jackfruit			1
15	Macadamia Nut			1
16	Frankincense		1	
17	Mango		1	
18	Africa		1	
19	Trembesi		1	

From the 1st iteration cluster table and 2nd iteration cluster table have the same cluster value and there is no more movement from one cluster to another. Therefore, the calculation process is stopped in the 2nd iteration and the results obtained from both iterations are:

1. C1 has 6 data which is defined as the most desirable seeds, including Suren, Jengkol, Mahogany, Avocado, Durian, Coffee.
2. C2 has 10 data which are interpreted as seeds that are quite attractive, including Pinus, Kaliandra, Macadamia, Petai, Aren, Cempedak, Frankincense, African Manggam, Trembesi.
3. C3 has 3 data which are interpreted as seeds that are less desirable, including Meranti, Jackfruit, Macadamia Nut.

The output of K-Means algorithm on the distribution data for productive seeds of the BPDASHL Asahan Barumun with rapidminer as shown in Figure 2 below:

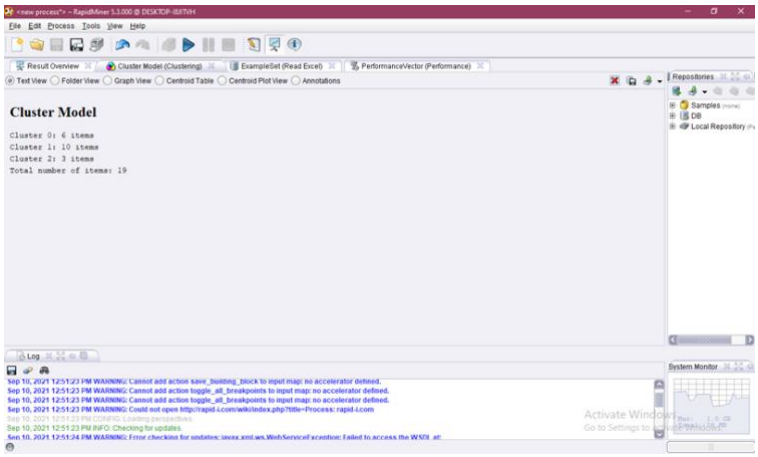

Figure 2. Clustering Result Textview Display

Based on Figure 2 above, it can be explained that cluster 0 has 6 items, cluster 1 has 10 items, and cluster 2 has 3 items

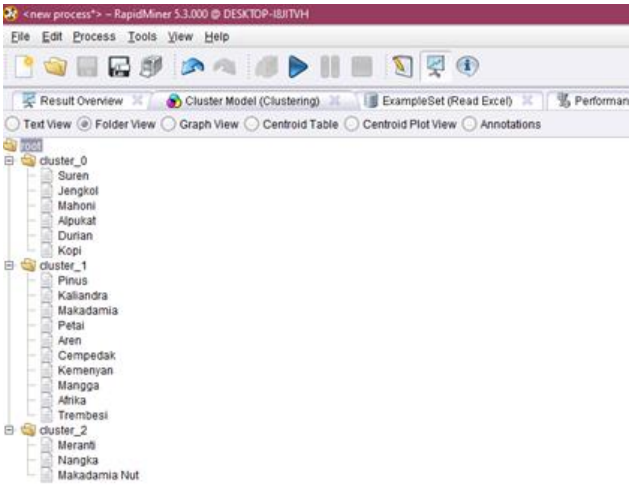

Figure 3. Clustering Results Folder View

Based on the results of K-Means grouping in Figure 3 above, it can be explained that:

1. C0 has 6 items which are defined as the most desirable seeds, including Suren, Jengkol, Mahogany, Avocado, Durian, Coffee.
2. C1 has 10 items which are defined as seeds that are quite attractive, including Pinus, Kaliandra, Makadamia, Petai, Aren, Cempedak, Frankincense, African Manggam, Trembesi.
3. C2 has 3 items which are defined as less desirable seeds, including Meranti, Jackfruit, Makadamia Nut.

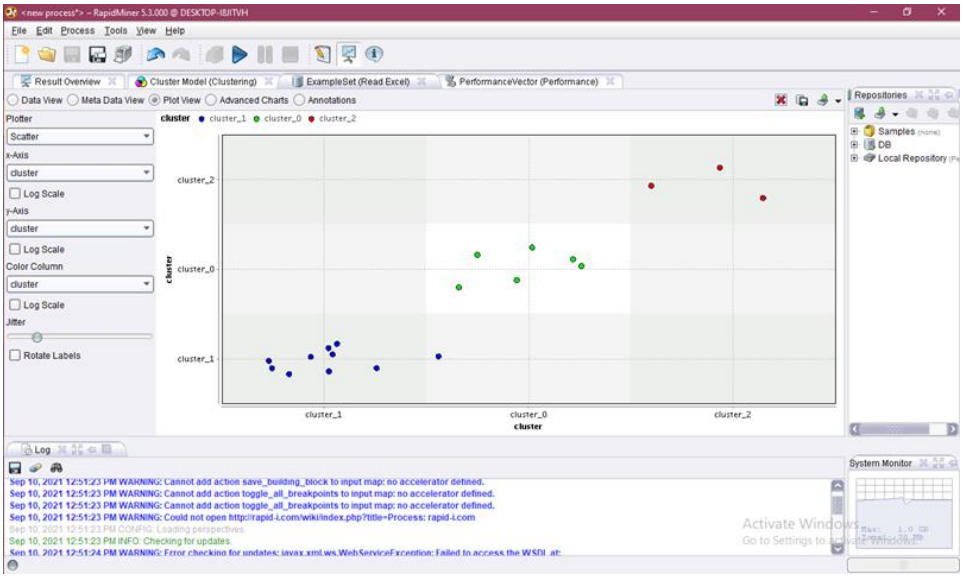

Figure 4. Plot View Clustering Results Display

Based on Figure 4.13, it is explained that the graph obtained from RapidMiner is known that the green dot is C0, the blue dot is C1, and the red dot is C2.

4. Conclusion

The grouping of productive seed distribution data at BPDASHL Asahan Barumun was successfully implemented using K-Means algorithm. The calculations carried out in this study, iteration clustering on the distribution of productive seed data in BPDASHL Asahan Barumun as much as 2 iterations. The results of the calculation of K-Means algorithm and testing the data of K-Means algorithm with RapidMiner have the same value, namely for cluster 0 (C1) it has 6 data which is defined as the most desirable seeds, including suren, jengkol, mahogany, avocado, durian , coffee, for cluster 1 (C2) has 10 data which are interpreted as seeds that are quite attractive, including pine, calliandra, macadamia, petai, sugar palm, cempedak, incense, mango, africa, trembesi, and for cluster 2 (C3) it has 3 data that are interpreted as seeds that are less desirable, including meranti, jackfruit, macadamia nut.

Acknowledgement

Acknowledgments to the supervisors and examiners who are lecturers at AMIK and STIKOM Tunas Bangsa so that this research can be arranged as one of the requirements for completing Bachelor's education (S1) at STIKOM Tunas Bangsa. I hope this research can be a reference for other research related to the methods and algorithms used. I hope for constructive suggestions for the readers for the perfection of this research in the future.

References

[1] A. Taufiq, "Upaya Pemeliharaan Lingkungan Oleh Masyarakat Di Kampung Sukadaya Kabupaten Subang," *J. Geogr. Gea*, vol. 14, no. 2, pp. 124–134, 2016, doi: 10.17509/gea.v14i2.3402.

[2] Istianah, "Upaya Pelestarian Lingkungan Hidup dalam Perspektif Hadis," *Riwayah*, vol. 1, no. 2, pp. 249–270, 2015.

[3] A. N. Lailia, "Gerakan Masyarakat Dalam Pelestarian Lingkungan Hidup (Studi Tentang Upaya Menciptakan Kampung Hijau Di Kelurahan Gundih Surabaya)," *J. Polit. Muda*, vol. 3, no. 3, pp. 283–302, 2014.

[4] K. Cios, W. Pedrycz, R. Swiniarski, and L. Kurgan, *Data Mining: A Knowledge Discovery Approach*. 2007.

[5] F. O. Isinkaye, Y. O. Folajimi, and B. A. Ojokoh, "Recommendation systems: Principles, methods and evaluation," *Egypt. Informatics J.*, vol. 16, no. 3, pp. 261–273, 2015, doi: https://doi.org/10.1016/j.eij.2015.06.005.

[6] A. M. H. Pardede *et al.*, "Implementation of Data Mining to Classify the Consumer's Complaints of Electricity Usage Based on Consumer's Locations Using Clustering Method," in *Journal of Physics: Conference Series*, 2019, vol. 1363, no. 1, doi: 10.1088/1742-6596/1363/1/012079.

[7] R. Rahim, J. Santoso, S. Jumini, G. Bhawika, D. Susilo, and D. Wibowo, "Unsupervised Data Mining Technique for Clustering Library in Indonesia," Feb. 2021.

[8] M. Omran, A. Engelbrecht, and A. Salman, "An overview of clustering methods," *Intell. Data Anal.*, vol. 11, pp. 583–605, Nov. 2007, doi: 10.3233/IDA-2007-11602.

[9] C. Sreedhar, N. Kasiviswanath, and P. Chenna Reddy, "Clustering large datasets using K-means modified inter and intra clustering (KM-I2C) in Hadoop," *J. Big Data*, vol. 4, no. 1, p. 27, 2017, doi: 10.1186/s40537-017-0087-2.

[10] D. Abdullah *et al.*, "Data Mining to Determine Correlation of Purchasing Cosmetics with A priori Method," in *Journal of Physics: Conference Series*, 2019, vol. 1361, no. 1, doi: 10.1088/1742-6596/1361/1/012056.

[11] N. A. Khairani and E. Sutoyo, "Application of K-Means Clustering Algorithm for Determination of Fire-Prone Areas Utilizing Hotspots in West Kalimantan Province," *Int. J. Adv. Data Inf. Syst.*, vol. 1, no. 1, pp. 9–16, 2020, doi: 10.25008/ijadis.v1i1.13.

[12] 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.

[13] A. Rachma, A. Aden, and Y. Rusdiana, "ANALISIS CLUSTER MENGGUNAKAN ALGORITMA K-MEANS CLUSTER UNTUK CULSTERING JENIS PENYAKIT MENULAR PADA PUSKESMAS DI KECAMATAN KOTA TANGERANG," *J. Saintika Unpam J. Sains dan Mat. Unpam*, vol. 2, p. 15, Aug. 2019, doi: 10.32493/jsmu.v2i1.2915.

[14] S. S. Hilabi *et al.*, "TechnoXplore Jurnal Ilmu Komputer & Teknologi Informasi ISSN : 2503-054X Vol 4 No: 1, April 2019," *J. Ilmu Komput. Teknol. Inf.*, vol. 4, no. 1, pp. 28–37, 2019.

[15] S. Fatonah, "Penerapan Deteksi Bencana Banjir Menggunakan Metode Machine Learning," vol. 10, pp. 119–126, 2021.

[16] E. Yanti, "Analisis Algoritma K-Means Dalam Pengelompokan Perkara Perceraian Berdasarkan Kelurahan Di Kota Jambi," *J. Process.*, vol. 16, no. 1, p. 9, 2021, doi: 10.33998/processor.2021.16.1.920.