ILKOM Jurnal Ilmiah Vol. 13, No. 3, December 2021, pp. 235-243
Accredited 2nd by RISTEKBRIN No. 200/M/KPT/2020; E-ISSN 2548-7779 | P-ISSN 2087-1716

Research Article Open Access (CC–BY-SA)

K-means algorithm for clustering system of certain plant seeds specialization areas in east aceh

Rozzi Kesuma Dinata [a,1,*]; Novia Hasdyna [b,2]; Sujacka Retno [b,3]; Muhammad Nurfahmi [a,4]

[a] *Universitas Malikussaleh, Jl. Cot Tengku Nie, Reuleut, Aceh Utara, 24351, Indonesia*
[b] *Universitas Islam Kebangsaan Indonesia, Jl. Medan B.Aceh Sp.Elak, Lhokseumawe, 24352, Indonesia*
[1] *rozzi@unimal.ac.id;* [2] *noviahasdyna@gmail.com;* [3] *sujackaretno@gmail.com;* [4]*muhammadnurfahmi990@gmail.com*
** Corresponding author*

Article history: Received June 06, 2021; Revised July 04, 2021; Accepted July 04, 2021; Available online August 08, 2021

Abstract

East Aceh consists of many regions and it has different types of plants. Clustering process is necessary to find the most favorable area for certain plants. This research employed k-means algorithm to categorize the obtained data. The data was obtained from Dinas Peranian Tanaman Pangan dan Hortikultura (Department of Food Crops and Horticulture) of East Aceh. The result of the research showed that the iteration number of the data in 2015-2019 was 8, 7,6,4,3 times in each commodity. The test showed areas where seeds were favorable divided into three: high interest, decent interest, and low interest. The system in this study was developed with web base using PHP programming language.

Keywords: K-means; Clustering; Areas of Interest; Seed plant; East Aceh

Introduction

Many types of agricultural crops require best quality seeds. Indonesian people could maximize the national soil quality by utilizing the potential of agricultural plant nurseries [1]. In short, nurseries are managed and developed in order to produce best quality seeds so that they can grow well and produce abundant production, also these seeds can be marketed to farmers or consumers in all regions in Indonesia [2][3]. With so many types of plants that can be managed in the fertile Indonesian soil, samples of the type of plants that are most in demand by farmers or consumers can be taken. This type of grouping can be divided into several parts, namely areas with high, decent, and low interest.

In the process of grouping the data in machine learning, there are several clustering algorithms to use, such as the K-Means algorithm [4] and K-Medoid [5]. In this research, the author applies K-Means algorithm to classify data on favorable areas of certain plant seeds in East Aceh. Clustering is a process of grouping data and looking for data that has a characteristically similar between one data and the others [6][7]. The purpose of this method is to minimize the objective function in the process rules for grouping favorable areas of certain plant seeds.

Several previous research related to the K-Means algorithm, such as the research conducted by Mustofa [8], applied K-Means to the characters in multiplayer online battle arena game. In the research conducted by Sujacka[9], analyzed the performance of k-means with the Davies Bouldin Index combined with Purity algorithm. Rozzi, et. al [10] in his research applying K-Means in classifying motorcycle data. Hajar [11], in his research applying K-Means to the palm oil export processing according to the target country. Irnanda [12], applied K-Means to find out the proportion of individuals who had ICT skills based on the specified area.

Based on the previous research, the focus of this research is to apply K-Means algorithm in the clustering system of favorable areas of certain plant seeds in East Aceh.

Method

K-Means method that used in the clustering system of areas of interest for certain plant seeds in East Aceh in this research has several steps which can be seen in **Figure 1**.

http://dx.doi.org/10.33096/ilkom.v13i3.863.235-243

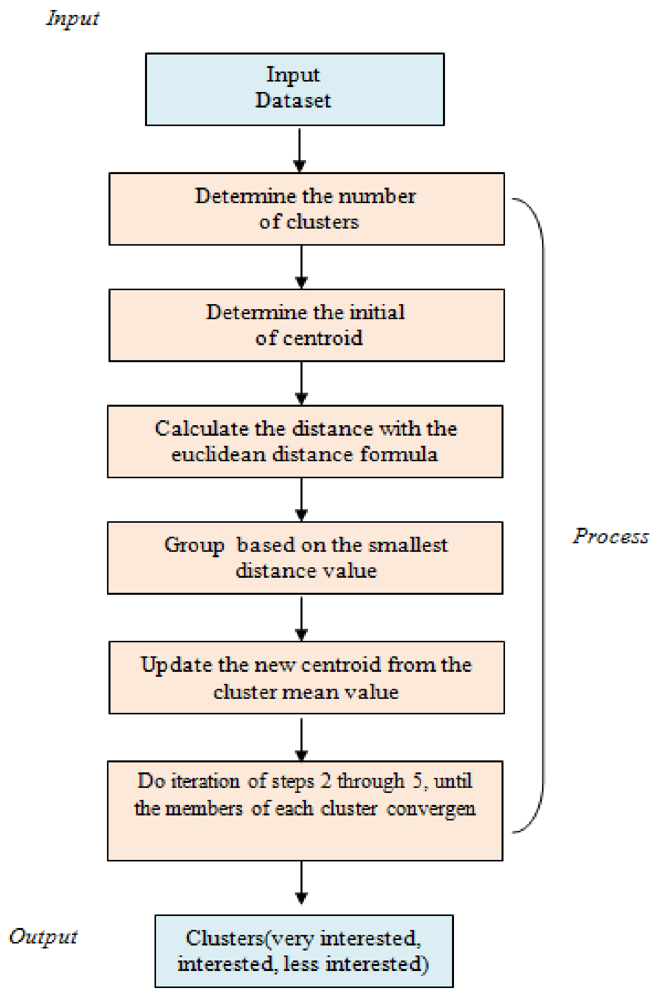

Input
Input Dataset
Determine the number of clusters
Determine the initial of centroid
Calculate the distance with the euclidean distance formula
Group based on the smallest distance value
Update the new centroid from the cluster mean value
Do iteration of steps 2 through 5, until the members of each cluster convergen
Process
Output
Clusters(very interested, interested, less interested)

Figure1. Research Flow

The steps of this research process based on the picture above are as follow:

1. Inputting the dataset.
   The data used in this research are regional and plant species data obtained from Dinas Pertanian Tanaman Pangan dan Hortikultura (Department of Food Crops and Horticulture), East Aceh Regency.
2. Determining the number of clusters.
   The number of clusters in this research are 3 clusters.
3. Determining the initial centroid.
4. Calculating the distance using the Euclidean distance. The formula for Euclidean distance is [13][14]:

$$d(xi,\mu j) \quad = \sqrt{\sum\nolimits_{\mathrm{i=1}}^{\mathrm{n}} (xi-\mu j)^2} \qquad (1)$$

   Where xi is the criterion data, and µj is the centroid of the js cluster.
5. Create a new centroid update from the cluster average results, with the formula [15]:

$$\mu j(t+1) \quad = \frac{1}{Nsj}\sum\nolimits_{j\in sj} xj \qquad (2)$$

Where xi is the criterion data, and µj is the centroid of the js cluster. In order for the system to run as expected, the authors collect data on surveyed plants in each sub-district by the Dinas Pertanian Tanaman Pangan dan Hortikultura in East Aceh Regency. Specifically the data obtained from the following survey:

1. Data on sales results were obtained from 2015 to 2019, they are types of plant, land area and seed in needs.
2. The cluster was divided into 3 clusters: high, decent, and low interest.
3. The initial centroid determination was determined randomly.

Results and Discussion

A. Research Dataset

This research used the data consisting of 24 sub-districts, data on plant types with 5 commodities per year from 2015-2019. The data were obtained from Dinas Pertanian Tanaman Pangan dan Hortikultura in East Aceh Regency.

1) Criteria of Data

The criteria data can be seen in **Table 1**.

Table 1. Criteria of Data

No	Criteria Code	Criteria (Unit)
1	X1	Plant Area (Ha)
2	X2	Harvest Area (Ha)
3	X3	Production (Ha)
4	X4	Productivity (Ton/Ha)
5	X5	Seed in Needs (Kg)

Based on table 1, in this research there are 5 criteria: X1 as planting area, X2 as harvested area, X3 as production, X4 as productivity and X5 as seeds in needs.

2) Rice Intensification Dataset

The dataset of Rice Intensifications consists from 2015 to 2019. **Table 2** is the dataset of Rice Intensifications dataset in 2019.

Table 2. Rice Intensifications Dataset in 2019

No.	Area	X1	X2	X3	X4	X5
1	Serbajadi	1221	1425	6.983	4,90	30525
2	Simpang Jernih	122	380	1.778	4,68	3050
3	Peunaron	2098	1763	10.049	5,70	52450
4	Birem Bayeun	2252	2770	15.512	5,60	56300
5	Rantau Selamat	917	774	3.715	4,80	22925
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
23	Simpang Ulim	3710	3524	24.915	7,07	92750
24	Madat	5706	5253	37.191	7,08	142650

3) Corn Intensification Dataset

Table 3 is the Corn Intensifications dataset in 2019.

Table 3. Corn Intensification Dataset in 2019

No.	Area	X1	X2	X3	X4	X5
1	Serbajadi	969	337	1.769	5,25	19380
2	Simpang Jernih	10	2	8	3,80	200
3	Peunaron	1133	281	1.503	5,35	22660
4	Birem Bayeun	62	37	192	5,20	1240
5	Rantau Selamat	51	10	42	4,15	1020
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
23	Simpang Ulim	55	15	62	4,10	1100

No.	Area	X1	X2	X3	X4	X5
24	Madat	75	1	4	3,85	1500

4) *Soybean Intensification Dataset*
Table 4 is the Soybean Intensifications dataset in 2019.

Table 4. Soybean Intensification Dataset in 2019

No.	Area	X1	X2	X3	X4	X5
1	Serbajadi	600	1000	1.540	1,54	21000
2	Simpang Jernih	30	30	44	1,45	1050
3	Peunaron	1424	720	1.138	1,58	49840
4	Birem Bayeun	170	150	216	1,44	5950
5	Rantau Selamat	70	90	137	1,52	2450
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
23	Simpang Ulim	25	30	47	1,58	875
24	Madat	35	35	58	1,67	1225

5) *Chili Intensification Dataset*
Table 5 is the Chili Intensifications dataset in 2019.

Table 5. Chili Intensification Dataset in 2019

No.	Area	X1	X2	X3	X4	X5
1	Serbajadi	8	10	368	36,80	32
2	Simpang Jernih	5	6	201	33,50	20
3	Peunaron	6	10	318	31,80	24
4	Birem Bayeun	14	10	320	32,00	56
5	Rantau Selamat	7	5	194	38,80	28
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
23	Simpang Ulim	5	5	146	29,20	20
24	Madat	12	10	325	32,50	48

6) *Cayenne Pepper Intensification Dataset*
Table 6 is the Cayenne Pepper dataset in 2019.

Table 6. Cayenne Pepper Intensification Dataset In 2019

No.	Area	X1	X2	X3	X4	X5
1	Serbajadi	4	5	73	14,60	12
2	Simpang Jernih	6	2	98	49,00	18
3	Peunaron	10	13	396	30,46	30
4	Birem Bayeun	4	8	291	36,38	12
5	Rantau Selamat	1	1	1	1,00	3
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
23	Simpang Ulim	1	1	1	1,00	3

No.	Area	X1	X2	X3	X4	X5
24	Madat	2	1	36	36,00	6

B. K-Means Calculation

1) Determining the Initial Centroid

The determining the initial centroid on each commodity can be seen in **Table 7**.

Table 7. The Determine of Initial Centroid

Commodity	Initial centroid as the data-
Rice	5, 11, 17
Corn	5, 11, 17
Soybean	5, 11, 17
Chili	5, 11, 17
Cayenne Pepper	5, 11, 17

2) K-Means Calculation in Rice Commodity

The initial centroid values for rice commodity are described in **Table 8**.

Table 8. Initial Centroid Value in Rice Commodity

Data -	X1	X2	X3	X4	X5
5	917	774	3.715	4,80	22925
11	1366	1184	6.062	5,12	34150
17	2041	1841	9.481	5,15	51025

3) Calculating distance Value with Euclidean Distance

To perform the clustering process on the obtained data that is to calculate the distance equation using the Euclidean distance method. The calculation of the distance from the first data with the first centroid:

$d_{(1.5)}$ = √(1221-917)^2 +(1425-774)^2+(6983-3715)^2+ (4,90-4,80)^2 + (30525-22925)^2 = 8303,701964.

The calculation of the distance from the first data with the second centroid:

$d_{(1.11)}$ = √(1221-1366)^2 + (1425-1184)^2 + (6983-6062)^2 + (4,90-5,12)^2 + (30525-34150)^2 = 3750,587157.

The calculation of the distance from the first data with the third centroid:

$d_{(1.17)}$ = √(1221-2041)^2 + (1425-1841)^2 + (6983-9481)^2 + (4,90-5,15)^2 + (30525-51025)^2 = 20672,17231.

After calculating all the data, it will provide the closest distance and cluster group for each data as shown in **Table 9**.

Table 9. Closest Distance Value and Cluster Group

Area	C1	C2	C3	Closest Distance	Cluster
Serbajadi	8303,70	3750,587157	20672,17231	3750,587157	2
Simpang Jernih	19988,84	31428,55177	48649,25554	19988,8489	1
Peunaron	30236,01	18752,53033	1537,050912	1537,050912	3
Birem Bayeun	35479,87	24150,02858	8068,725979	8068,725979	3
Rantau Selamat	0	11483,82218	28727,30486	0	1
Sungai Raya	8639,3	3077,697326	20178,75558	3077,697326	2
Peureulak	142276,49	130794,1914	113555,9393	113555,9393	3

4) Calculating a new centroid

The new centroid value calculation process conducted by adding up all the values in each of the same clusters and dividing it by the amount of data in the clusters. The new centroid values can be seen in **Table 10**.

Table10. New Centroid Value (2nd Iterations)

Centroid	X1	X2	X3	X4	X5
C1	694,50	664,17	3457,25	5,04	17362,50
C2	1370,43	1294,14	6917,06	5,33	34260,71
C3	3360,18	3136,18	18923,16	5,94	84004,55

5) Comparing the Value of the New centroid with the previous Centroid

After the new centroid is calculated, the next step is to compare it with the previous centroid, if the value is the same then the iteration process is stopped. Besides, if the values are not the same, then step 2 to step 5 is repeated. **Table 11** describes the values of the new centroids.

Table11. New Centroid Value (3rd Iterations)

Centroid	X1	X2	X3	X4	X5
C1	609,6	591,4	2983,0	4,9	15240,0
C2	1633,9	1554,8	8534,8	5,5	40848,1
C3	4405,5	4062,2	25241,9	6,2	110137,5

Table 11 shows the new centroid values in the 3rd iteration process. The centroid value for the 4th iteration process can be seen in **Table 12**.

Table12. New Centroid Value (4th Iterations)

Centroid	X1	X2	X3	X4	X5
C1	694,5	664,2	3457,3	5,0	17362,5
C2	1739,9	1650,8	9276,9	5,5	43498,1
C3	4787,2	4419,4	27194,9	6,1	119680,0

The following is the new centroid value for the 5th iteration process can be seen in **Table 13**.

Table 13. New Centroid Value (6th Iterations)

Centroid	X1	X2	X3	X4	X5
C1	768,0	717,0	3790,6	5,1	19200,0
C2	1784,2	1702,2	9567,5	5,5	44604,2
C3	4787,2	4419,4	27194,9	6,1	119680,0

Table 14. New Centroid Value (6th Iterations)

Centroid	X1	X2	X3	X4	X5
C1	871,2	858,3	4440,0	5,1	21780,6
C2	1894,5	1772,0	10138,4	5,7	47362,5
C3	4787,2	4419,4	27194,9	6,1	119680,0

Table 14 shows the new centroid values in the 6th iteration process. The new centroid value for the 7th iteration process can be seen in **Table 15**.

Table 15. New Centroid Value (7th Iterations)

Centroid	X1	X2	X3	X4	X5
C1	960,1	921,2	4780,2	5,1	24002,3
C2	2028,1	1914,0	11095,1	5,8	50703,1
C3	4787,2	4419,4	27194,9	6,1	119680,0

The following is the new centroid values for the 8th iteration process, it can be seen in **Table 16**.

Table 16. New Centroid Value (8th Iterations)

Centroid	**X1**	**X2**	**X3**	**X4**	**X5**
C1	960,1	921,2	4780,2	5,1	24002,3
C2	2028,1	1914,0	11095,1	5,8	50703,1
C3	4787,2	4419,4	27194,9	6,1	119680,0

This test stops at the 8th iterations. The result of the last iteration will be used for the clustering parameters. The author determines which cluster members are included in the cluster that is high interest, decent interest, and low interest based on the average value of the last centroid. The average value of C1 = 6133,778364, C2 = 13149,225, C3 = 31217,514, the members of C1 are described as low interest, members of C2 are decent interest, and members of C3 are high interest. As for the calculation of other commodities, it is the same as the calculation of the rice commodity.

6) The results of K-Means calculation

Based on the tests that we have researched with the initial centroid 5,7 and 11 from the 2019 data, it shows that the results of Rice commodity in the 5th data, called Rantau Selamat was found in low interest cluster, while in the 11th data, called Idi Rayeuk was found in decent interest cluster, and the 17th data, called Darul Aman was found in high interest cluster.

Corn commodity in the 5th data, Rantau Selamat was found in decent interest cluster, while in the 11th data Idi Rayeuk was found in low interest cluster and the 17th data Darul Aman was found in high interest cluster. Soybean commodity testing in the 5th data, Rantau Selamat was placed in the decent interest cluster, while the 11th data Idi Rayeuk was placed in low interest cluster and the 17th data, Darul Aman was found in high interest cluster.

Chili commodity in the 5th data test, Rantau Selamatwas was clustered in low interest, while the 11th data, Idi Rayeuk was clustered in high interest and the 17th data, Darul Aman was clustered in decent interest.

Cayenne Paper commodity in the 5th data test, Rantau Selamat was categorised in low interest, while the 11th data, Idi Rayeuk was categorised in decent interest and the 17th data, Darul Aman was categorised in high interest.

In calculating each commodity, the iteration process for the K-Means algorithm was different as described in **Figure 2**.

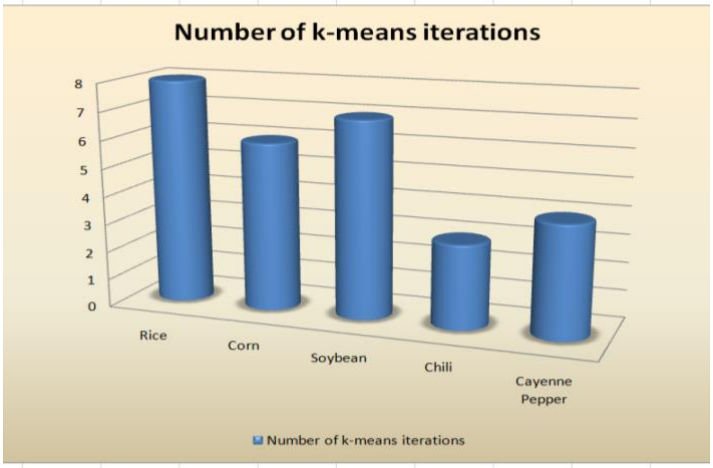
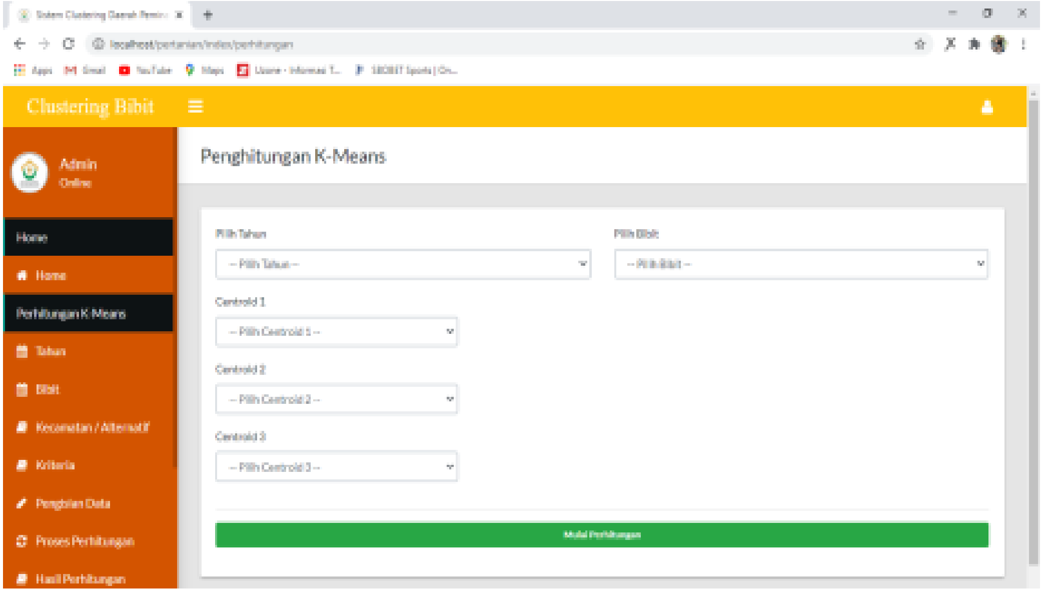


Figure 2. Average number of K-means Iterations for Each Commodity in 2015-2019

Based on **Figure 2**, the average number of iterations in the testing data for 2015-2019, the most iterations are 8 iterations by rice commodity. Furthermore, 7 iterations by soybean commodity. Corn commodity has 6 iterations, Cayenne pepper commodity has 4 iterations and the lowest iteration was Chili commodity as it only has 3 iterations.

C. K-Means Implementation to The System

The following is a display of the clustering system of the area of interest for certain plant seeds as in **Figure 3-5**.



Penghitungan K-Means

Pilih Tahun: -- Pilih Tahun --

Pilih Bibit: -- Pilih Bibit --

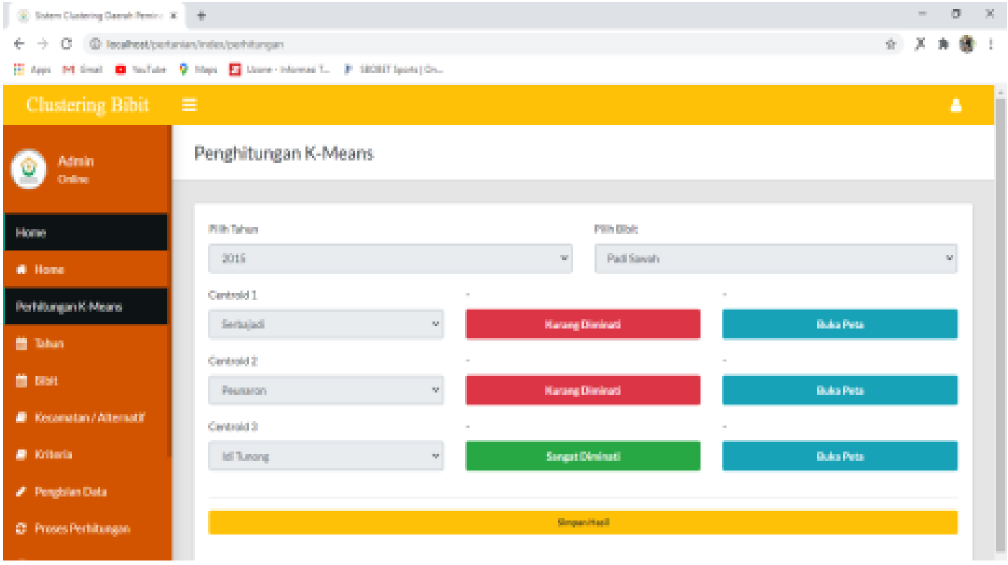
Centroid 1: -- Pilih Centroid 1 --

Centroid 2: -- Pilih Centroid 2 --

Centroid 3: -- Pilih Centroid 3 --

Mulai Perhitungan

Figure 3. K-Means Calculation Menu Form



Penghitungan K-Means

Pilih Tahun: 2015

Pilih Bibit: Padi Sawah

Centroid 1: Sertajadi — Kurang Diminati — Buka Peta

Centroid 2: Peusaron — Kurang Diminati — Buka Peta

Centroid 3: Idi Tunong — Sangat Diminati — Buka Peta

Simpan Hasil

Figure 4. K-Means Calculation Results Form

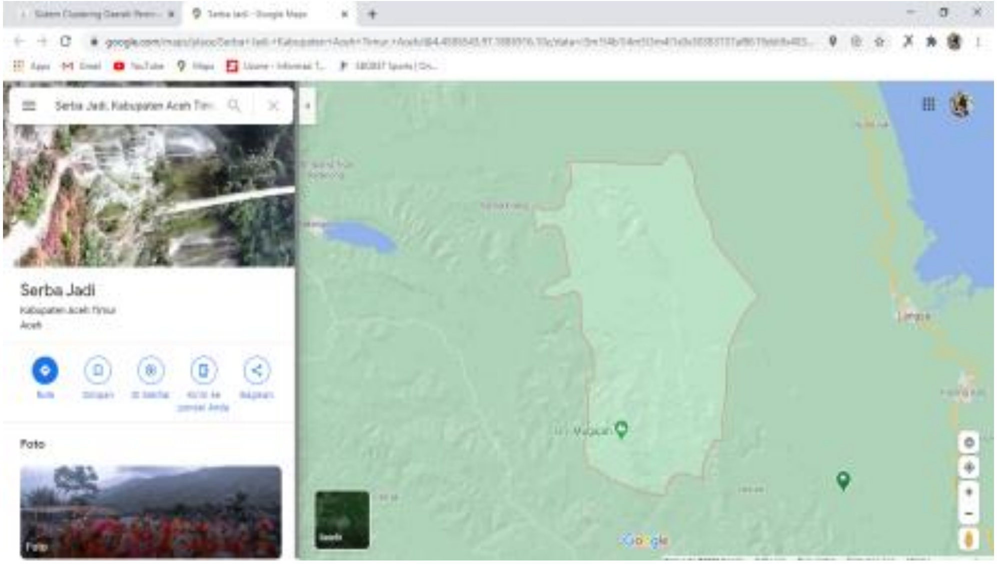


Figure 5. Regional view with Google Map

Conclusion

Based on the research that has been done, the results of random testing for 5 times using K-Means algorithm which is applied to the clustering system for favorable areas of certain plant seeds succeeded in classifying clusters for the areas of most favorable, which are high interest cluster, decent interest and low interest. The average of iteration from the k-means result in the 2015-2019 data, the highest number of iterations is on the Rice commodity with 8 iterations. Furthermore, there are 7 iterations for Soybean commodity. Corn commodity has 6 iterations, Cayenne Pepper commodity has 4 iterations and the lowest was Chili commodity for only 3 iterations. The author suggests that for further research, other algorithms can be used to perform clustering, such as the K-Medoid algorithm.

References

[1] S. R. O. S. Chan “Industri Perbenihan dan Pembibitan Tanaman Hortikultura di Indonesia : Kondisi Terkini dan Peluang Bisnis,” media.neliti.com, vol. 2, no. 1, pp. 26–31, 2021.

[2] M. Irpan, H. Suparto, A. Rizali , “Uji Komposisi Media Tanam dan Pemberian Pupuk,” Agroekotek View, vol. 4, no. 1, pp. 31–38, 2021.

[3] F. Kautsar, A. M. Aqib, A. P. Sari, and A. Sholikhah, “Etnomatematika Pada Aktivitas Petani Padi Kecamatan Ampelgading,” ProSANDIKA UNIKAL (Prosiding Semin. Nas. Pendidik. Mat. Univ. Pekalongan), vol. 2, no. 1, pp. 19–28.

[4] M. Ahmed, R. Seraj, and S. M. S. Islam, “The k-means algorithm: A comprehensive survey and performance evaluation,” Electron., vol. 9, no. 8, pp. 1–12, 2020, doi: 10.3390/electronics9081295.

[5] H. Song, J.-G. Lee, and W.-S. Han, "PAMAE: parallel k-medoids clustering with high accuracy and efficiency,” Proceedings of the 23rd ACM SIGKDD, pp. 1087–1096, 2017, doi: 10.1145/3097983.3098098.

[6] M. G. Johnson et al., “A Universal Probe Set for Targeted Sequencing of 353 Nuclear Genes from Any Flowering Plant Designed Using k-Medoids Clustering,” Syst. Biol., vol. 68, no. 4, pp. 594–606, 2019, doi: 10.1093/sysbio/syy086.

[7] C. Yuan and H. Yang, “Research on K-Value Selection Method of K-Means Clustering Algorithm,” J— Multidisciplinary Scientific Journal, vol. 2, no. 2, pp. 226–235, 2019, doi: 10.3390/j2020016.

[8] M. Mustofa, “Penerapan Algoritma K-Means Clustering pada Karakter Permainan Multiplayer Online Battle Arena,” J. Inform., vol. 6, no. 2, pp. 246–254, 2019, doi: 10.31311/ji.v6i2.6096.

[9] S. Retno, “Peningkatan Akurasi Algoritma K-Means dengan Clustering Purity Sebagai Titik Pusat Cluster Awal (Centroid)," Universitas Sumatera Utara., pp. 4–16, 2019.

[10] R. K. Dinata, S. Safwandi, N. Hasdyna, and N. Azizah, “Analisis K-Means Clustering pada Data Sepeda Motor,” INFORMAL Informatics J., vol. 5, no. 1, p. 10, 2020, doi: 10.19184/isj.v5i1.17071.

[11] S. Hajar, A. A. Novany, A. P. Windarto, A. Wanto, and E. Irawan, “Penerapan K-Means Clustering Pada Ekspor Minyak Kelapa Sawit Menurut Negara Tujuan,” Semin. Nas. Teknol. Komput. Sains, pp. 314–318, 2020.

[12] K. F. Irnanda, A. P. Windarto, I. S. Damanik, and ..., “Penerapan K-Means pada Proporsi Individu dengan Keterampilan (Teknologi Informasi dan Komunikasi) TIK Menurut Wilayah,” … Teknol. Inf. …, no. c, pp. 452–456, 2019, [Online]. Available: http://seminar-id.com/prosiding/index.php/sensasi/article/view/344.

[13] N. Hasdyna, B. Sianipar, and E. M. Zamzami, “Improving the Performance of K-Nearest Neighbor Algorithm by Reducing the Attributes of Dataset Using Gain Ratio,” J. Phys. Conf. Ser., vol. 1566, no. 1, 2020, doi: 10.1088/1742-6596/1566/1/012090.

[14] Hartono, O. S. Sitompul, Tulus, and E. B. Nababan, “Biased support vector machine and weighted-SMOTE in handling class imbalance problem,” Int. J. Adv. Intell. Informatics, vol. 4, no. 1, pp. 21–27, 2018, doi: 10.26555/ijain.v4i1.146.

[15] K. Sirait, Tulus, and E. B. Nababan, “K-Means Algorithm Performance Analysis with Determining the Value of Starting Centroid with Random and KD-Tree Method,” J. Phys. Conf. Ser., vol. 930, no. 1, 2017, doi: 10.1088/1742-6596/930/1/012016.