Implementation of K-Means Clustering in Food Security by Regency in East Java Province in 2022

Tuslaela[1]*, Rusdiansyah[2], Hendra Supendar [3], Nining Suharyanti [4]
[1]Universitas Nusamandiri, [2,3,4]Universitas Bina Sarana Informatika, Indonesia
[1] tuslaela.tll@nusamandiri.ac.id, [2] rusdiansyah.rds@bsi.ac.id, [3]hendra.hds@bsi.ac.id,
[4]nining.nni@bsi.ac.id

Abstract: Food is the main need that society must fulfill. If food security is disrupted, it will have a negative impact on the nation's life. The agricultural sector has an important role in West Java Province. This province has a large area of agricultural land, so it has high potential to produce abundant agricultural production. However, knowing the adequate number of farmers is very important. Therefore, the implementation of K-Means Clustering can make a significant contribution to the East Java Provincial Agriculture Service in grouping farmers by district. To achieve optimal results, determining the best K value needs to be considered carefully. K-Means Cluster Analysis is a method of non-hierarchical Cluster Analysis that groups data into one or more groups. Data with the same characteristics is grouped into one cluster and data with different characteristics is grouped into another cluster. The data used in this research are land area and rice production in the Regency of East Java Province in 2022. Based on the results of research with the object of Food Security, it can be concluded that, the results of the analysis of the application of manual data mining calculations in Excel Software using the K-Means Clustering method, resulted in two types of clustering in the form of C0, namely the Highest Land Area and Production group with 4 districts: Jember Regency, Ngawi Regency, Bojonegoro Regency and Lamongan Regency, for C1 clustering, namely the Lowest Land Area and Production group with 25 districts in East Java Province.

Keywords: *Land area, Production, Clustering, K-Means, Grouping*

INTRODUCTION

Because the majority of the population works in the agricultural sector, Indonesia was previously known as an agricultural country. However, we often see agricultural land converted into housing, roads, and so on(Kriswibowo, 2020). This will have a negative impact on the environment and society itself, including food security. Food security is defined as when a country is able to meet food needs in quantity, quality and sustainability.(Helin G Yudawisastra, 2023). Therefore, food plays an important role in the life of a country. If food security is disrupted, it will cause economic, social and political problems(Widyawati Boediningsih, 2023).

Based on data released by the Central Statistics Agency (BPS), the contribution of the agricultural sector to Indonesia's Gross Domestic Product (GDP) was 13.65% in 2020(Bembok, 2020). Farmers are also very important in maintaining the quality of the food they make to meet human food needs. Indonesia's horticultural production in 2020 reached 39.17 million tonnes, with the largest vegetable production at 12.97 million tonnes, according to BPS data(Nurzaman, 2023). The agricultural sector is very important in East Java Province because it has extensive agricultural land and great potential to

produce high agricultural production (Firmansyah, 2021). But the role and abilities of farmers are very important to manage agricultural land well. It is very important to carry out research that divides the number of farmers by district in East Java Province. This study can help determine the number of farmers in each district in East Java Province and determine which sub-districts have different numbers of farmers. As a result, appropriate action can be taken to increase the number of farmers in districts that have low numbers of farmers. The decline in the agricultural sector also has the potential to affect future food security in Indonesia(Rasmikayati, 2019), especially because the number of farmers is decreasing. There is no interest from the younger generation in agriculture in the current millennial era. This research can also provide statistical data that can help make appropriate policies regarding agricultural development in East Java Province. Therefore, this research aims to ensure progress in the agricultural sector in East Java Province to achieve optimal levels of agricultural production.(Syairozi, 2021).

LITERATURE REVIEW

Previous research that the author uses as a reference includes the following

1. A study conducted by Joseph Stenica Hariyono et al. aims to simplify the processing of harvest data by building a geographic information system for shallot farming in Nganjuk Regency. This system can manage data faster and in more detail by using the K-Means clustering method. Nganjuk Regency data from 2016 to 2020 consists of 19 sub-districts each year, for a total of 95 data for 5 years. Based on the results of data processing, the data for 2017, 2018, 2020 ended in the 6th iteration. In addition, the data for the 2016 and 2019 shallot harvests were completed in the 5th and 7th iterations, so that the final results of data processing used The clustering method can solve the problem of grouping shallot harvest data in Nganjuk Regency(Stenica Hariyono, 2021).

2. The research of Muhammad Yamin Nurzaman (1) and his colleagues is very important in knowing the sufficient number of farmers. As a result, the use of K-Means clustering can really help the West Java Provincial Agriculture Service in grouping farmers based on sub-districts. Determining the best K value must be considered carefully to achieve optimal results. The evaluation results using the silhouette coefficient score show that the cluster with a value of 0.7374355377121188 is the best in this study. Therefore, the use of K-Means Clustering was successfully used to group the farmer population most efficiently based on sub-districts in West Java Province(Nurzaman, 2023).

3. Arif Rohmatullah et al.'s study. examine the area of agricultural land and agricultural production, so agricultural data must be clustered. The purpose of clustering is to identify a group of data from a data population so that the characteristics of the data itself can be produced. This research will use two algorithms, namely the K-Means algorithm and the Fuzzy C Means (FCM) algorithm. Different sub-districts in Lamongan Regency can be classified using this algorithm based on the area of agricultural land and agricultural production results. The FCM algorithm changes the degree of membership, or membership level, so as to produce a minimum objective function value. On the other hand, the K-Means algorithm changes the cluster center point so that it produces the minimum euclidean distance. The simulation results show that the two methods can cluster several sub-districts in Lamongan Regency based on agricultural land area and agricultural production results(Rohmatullah, 2020).

Classification is a data processing process that has several applications, but the main goal is to identify a group of data from a population of data to produce the properties of the data itself. Classification is the process of grouping a collection of data objects into one or more groups so that the data collected in these groups has a high level of similarity(Zen, 2023).

The K-Means algorithm is based on the centroid point value, which is the center point of each cluster. This algorithm can be designed as follows: Let N be the number of cluster centroid points and D be the amount of data(Cesar, 2023). The authors of this study used the K-means algorithm to model clustering. One technique popularized by James B MacQueen is K-Means. Using the k-means method, objects are grouped into k groups (k < n) where the k value is predetermined. Facts(Manihuruk, 2020). To start the clustering process, data Xij (i = 1,...,n), j = 1,...,m) is selected, where n is the amount of data to be clustered, and m is an attribute (variable). Ckj(k = 1,...,k) ; j = 1,...,m. At the start of the

iteration, the center of each cluster is chosen freely. Next, the distance between each cluster, called the centroid, is calculated. This is calculated using the Euclidian distance measure, i.e(Ali, 2019):

$$d(x, y) = \sqrt{\sum_{i=1}^{n}(xi - yi)^2} \qquad (1)$$

Information :
d = distance of object to centroid di-k
i = number of attributes (data dimensions)
y = centroid
x = data

K-means clustering may be successful with small data sets. Large data sets must be grouped so that each entity or data point in the set is the same as every other entity in the set. The advantages of the k-means method are as follows(Muningsih, 2018):
   1) The process is easy to use and easy to implement
   2) The algorithm can run quickly
   3) Has the ability to break down a collection of patterns into similar groups.

METHOD

Is the plan of the researcher that will be used as a guide for the progress of the research(Ediansa, 2018):
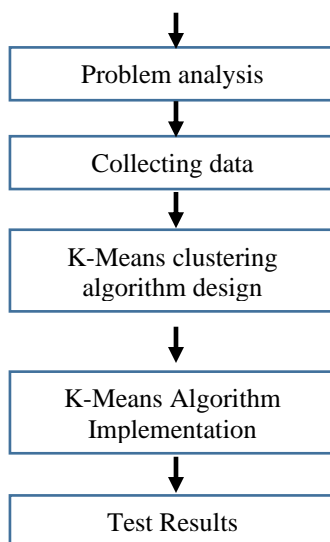

Figure 1. Research Design

1. Identification of problems
    Identification of the problem is the first step in knowing the problem to be studied so that the right solution can be found. In this study, the identification of the problem is that there are still beneficiaries who do not meet the requirements of the recipients of assistance and that there are no criteria to determine who are truly worthy of receiving assistance.
2. Problem analysis
    Problem analysis is a step to identify problems that have been determined in the scope or boundaries. The problem required analysis is how to classify and provide the basis for the decisions of beneficiaries who meet the requirements of beneficiaries.
3. Collecting data
    In order to clarify the current issues, observations, interviews, and questionnaires were distributed while collecting data. The data that was collected was related to the conditions of the current.
4. K-Means clustering algorithm design
    Cluster centers are randomly selected from the beneficiary data, and the Euclidian distance formula is used to design the process from the system model to the k-means algorithm.

5. K-Means Algorithm Implementation

For its implementation, the k-means algorithm stage is used which uses the entire K-means clustering process. Starting by finding the cluster center then using the Euclidian distance formula to find the closest distance to the center until the iterative process finds a new cluster value from the previous

6. Stage Test Results

This is the testing and analysis stage of the results produced by the Rapidminer software. Steps and models are used to test the system.

Researchers collected data from East Java Regency. This data was taken from the East Java Province Central Statistics Agency in 2022. This research selected 29 districts in East Java Province based on land area and production. Population is the entire collection of elements that have some of the same characteristics, which consists of the fields to be studied. This research utilizes land area and rice production as the research population, and the samples come from 29 districts throughout Java Province in 2022. Data analysis is an effort to process, organize, sort it into controllable parts, look for and find patterns, find what is important and learning, and deciding what to say to others. In this research, quantitative data analysis methods were used. This is a type of research that is carried out to apply a hypothesis that has been created initially and is based on a theory that has been built.

The data used in this research is data on land area and production in 2022, as follows :

Table 1. District Land Area and Production Data

Number	Regency	Land area	Production
1	Pacitan	17.988,72	90.955,25
2	Ponorogo	64.543,87	370.435,11
3	Trenggalek	22.293,31	117.346,67
4	Tulungagung	40.611,23	221.637,05
5	Blitar	34.255,13	217.566,97
6	Kediri	29.851,89	172.474,88
7	Malang	47.821,21	283.895,29
8	Lumajang	56.064,69	303.468,63
9	Jember	119.808,13	613.237,38
10	Banyuwangi	77.952,96	462.584,81
11	Bondowoso	49.262,33	246.388,27
12	Situbondo	27.172,07	141.914,27
13	Probolinggo	34.231,53	187.277,08
14	Pasuruan	47.999,47	254.578,42
15	Sidoarjo	31.431,32	196.839,63
16	Mojokerto	49.281,71	287.251,32
17	Jombang	55.229,58	343.427,84
18	Nganjuk	71.223,24	387.897,28
19	Madiun	74.141,77	419.977,93
20	Magetan	42.606,10	254.578,42
21	Ngawi	129.474,04	785.037,99
22	Bojonegoro	133.739,17	715.198,84
23	Tuban	85.194,97	502.136,24
24	Lamongan	151.263,60	920.935,59
25	Gresik	63.241,32	410.323,14
26	Bangkalan	39.691,95	193.329,37
27	Sampang	34.880,84	172.558,93
28	Pamekasan	20.497,42	108.020,20
29	Sumenep	40.558,48	223.000,46

*name of corresponding author

In Table 1, based on the East Java Province Central Statistics Agency, there are 29 districts taking data on land area and production in 2022.

RESULT

This research uses secondary data obtained from the official publication of the government agency, the Central Statistics Agency (BPS), which can be found at www.bps.go.id. This data includes land area and production data for the province of East Java in Indonesia in 2022. Next, the k-means algorithm will be used to convert this data into previous data using the data processing tool Rapidminer Data. This research uses statistical data from the District of East Java Province in 2022.

The results of clustering research carried out using the K-Means algorithm on rice production in East Java in 2022 are as follows. To carry out this calculation, the K-Means algorithm formula is entered manually into Microsoft Excel, and then calculated according to the data. In the process of determining the first cluster, the centroid or center point of the three clusters that have been determined is identified.

Table 2. K-Means Iteration Results

Iteration	Cluster 1		Cluster 2	
	Center Point	Number of data	Center Point	Number of data
1	151263,60; 920935,59	4	17988,72 ; 90955,25	25
2	133571,24; 758602,45	4	45601,54; 263164,31	25

Table 2 shows the results of the iterations required to show stable data. 2. The second iteration has the same results as the first iteration, where no data moves clusters again. In accordance with the K-Means algorithm, when no data moves, the iteration process stops and the clustering process is declared complete. Thus, the data resulting from the second iteration is stable data, as suggested by Davies Bouldin.

The clustering results produced using the K-Means algorithm for rice production in East Java Province are as follows:

Table 3. Data from the 2nd Iteration clustering results

Value Data	Cluster 1		Cluster 2	
	Land area	Production	Land area	Production
Total	534.284,94	3.034.409,80	1.581.027,11	7.192.345,13
Average	133.571,24	758.602,45	46.321,08	263.164,31
Amount	4	4	25	25
Max Value	151.263,60	920.935,59	85.194,97	502.135,24
Min Value	119.808,13	613.237,38	17.988,72	90.955,25

In Table 3, data obtained from the 2nd iteration shows that in cluster 1 there are 4 areas with high production. Next is cluster 2, there are 25 areas with low production

At this point, testing will be carried out using RapidMiner software, which is a solution for data mining, text mining and predictive analysis created using the Java language. to evaluate a food security data set that has land area and production variables. With the number of clusters that have been determined, there are two clusters that meet the Davies Bouldin comparison recommendations, and these two clusters have the smallest value, namely 0.373. The processed dataset includes 29 data related to land area and district production in East Java Province 2022.
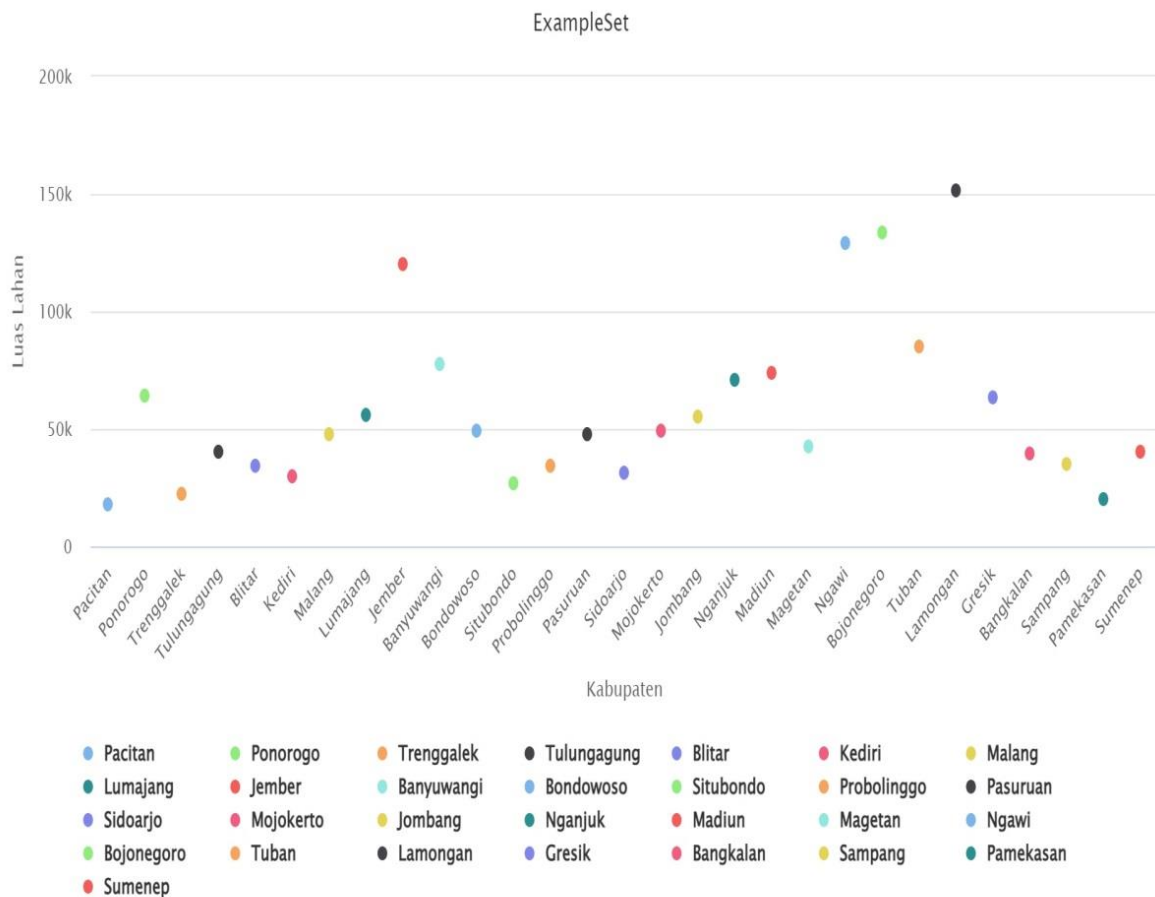
*name of corresponding author

Figure 2. Plot of Food Security Implementation Results

DISCUSSIONS

According to Joseph Stenica Hariyono et al., this research aims to simplify the process of harvest data using the K-Means clustering method. Data from 19 sub-districts in Nganjuk Regency from 2016 to 2020 have the same regional structure, In this research, regions are grouped according to districts in East Java Province. The biggest difference from the research used as a reference is the use of the Davies Bouldien model. Previous studies did not use this model, but previous research used it for two groups and found the smallest value, namely 0.373.

CONCLUSION

Based on the results of research on the subject of Food Security, it can be concluded that two types of clustering in the form of C0—Land Area and Highest Production groups—resulted from analysis of the application of manual data mining calculations in Microsoft Excel using the K-Means Clustering method. This happened in 4 districts: For the C1 categorization, namely the Lowest Land Area and Production group, there are 25 districts, including Jember, Ngawi, Bojonegoro and Lamongan. The resulting test results from the RapidMiner application are comparable to manual calculations in Microsoft Excel. Therefore, it can be used for processing type groupings and their distribution appropriately, effective and efficient for the 2022 East Java Province Food Security Program, so that the fertilizer subsidy program and soil restoration program are improved so that they are of good quality and provide the best rice seeds, this must be supported by the regional government of East Java Province.

*name of corresponding author

REFERENCES

Ali, A. (2019). Clustering of Patient Medical Record Data Using the K-Means Clustering Method at Anwar Medika Hospital Balong Bendo, Sidoarjo. *MATRIK : Jurnal Manajemen, Teknik Informatika Dan Rekayasa Komputer*, *19*(1), 186–195. https://doi.org/10.30812/matrik.v19i1.529

Bembok, N. (2020). Kontribusi Sektor Pertanian Dalam Perekonomian Di Kabupaten Minahasa. *Agri-Sosioekonomi*, *16*(3), 333. https://doi.org/10.35791/agrsosek.16.3.2020.30313

Cesar, W. (2023). 4593-Article Text-15458-1-10-20230613 (1). *Teknik Informatika Dan Sistem Informas*, *10*(2), 340–354.

Ediansa, O. (2018). Analisis Dan Perancangan Sistem Penunjang Keputusan Warga Miskin Penerima Sktm Pada Pemerintah Kota Jambi. *Manajemen Sistem Informasi*, *3*(1). https://stt-pln.e-journal.id/petir/article/view/418

Firmansyah, F. (2021). Strategi Pengendalian Alih Fungsi Lahan Sawah di Provinsi Jawa Timur. *Jurnal Penataan Ruang*, *16*(1), 47. https://doi.org/10.12962/j2716179x.v16i1.8726

Helin G Yudawisastra. (2023). Budikdamber akuaponik sebagai strategi ketahanan pangan dan stimulus kewirausahaan saat pandemi covid-19. *BEMAS: Jurnal Bermasyarakat*, *3*(2), 162–170. https://doi.org/10.37373/bemas.v3i2.258

Kriswibowo, S. I. P. A. (2020). Prosiding Konferensi Nasional Administrasi Negara Sinagara 2020. In *Researchgate.Net*. https://www.researchgate.net/profile/Yohanes_Kristianto10/publication/346495787_Meningkatkan_Kinerja_Industri_Kecil_Kampung_Batik_melalui_Keunggulan_Bersaing/links/5fc51f9b299bf1a422c3093a/Meningkatkan-Kinerja-Industri-Kecil-Kampung-Batik-melalui-Keunggul

Manihuruk, N. A. (2020). Penerapan Data Mining Dalam Mengelompokkan Calon Penerima Beasiswa Dengan Menggunakan Algoritma K-Means. *KOMIK (Konferensi Nasional Teknologi Informasi Dan Komputer)*, *4*(1), 29–34. https://doi.org/10.30865/komik.v4i1.2575

Muningsih, E. (2018). Sistem Aplikasi Berbasis Optimasi Metode Elbow Untuk Penentuan Clustering Pelanggan. *Joutica*, *3*(1), 117. https://doi.org/10.30736/jti.v3i1.196

Nurzaman, M. (2023). Implementasi K-Means Clustering Dalam Pengelompokkan Banyaknya Jumlah Petani Berdasarkan Kecamatan Di Provinsi Jawa Barat. *Teknik Informatika Dan Sistem Informasi*, *10*(3).

Rasmikayati, E. (2019). Potensi Dan Kendala Yang Dihadapi Petani Mangga Dalam Mengakses Lembaga Pemasaran. *Paradigma Agibisnis*, *2*(1), 22–30.

Rohmatullah, A. (2020). Klasterisasi Data Pertanian Di Kabupaten Lamongan Menggunakan Algoritma K-Means Dan Fuzzy C Means. *Jurnal Ilmiah Teknosains*, *5*(2), 86–93. https://doi.org/10.26877/jitek.v5i2.4254

Stenica Hariyono, J. (2021). Perancangan Sistem Informasi Geografis Hasil Produksi Pertanian Bawang Merah Di Kabupaten Nganjuk Menggunakan Metode K-Means. *JATI (Jurnal Mahasiswa Teknik Informatika)*, *5*(2), 487–494. https://doi.org/10.36040/jati.v5i2.3737

Syairozi. (2021). ANALISIS KEMISKINAN DI SEKTOR PERTANIAN (Studi Kasus Komoditas Padi di Kabupaten Malang). *Media Ekonomi*, *28*(2), 113–128. https://doi.org/10.25105/me.v28i2.7169

Widyawati Boediningsih. (2023). Jurnal Pro Hukum: Vol . 12, No. 3, Maret 2023 Jurnal Pro Hukum: *Pro Hukum*, *12*(3), 1002–1018.

Zen, M. A. (2023). *Aplikasi Pengelompokan Data Runtun Waktu dengan Algoritma K-Medoids*. *6*(2), 117–123. https://doi.org/10.12962/j27213862.vxix.xxxx