Implementation of K-Means Algorithm for Clustering Corn Planting Feasibility Area in South Lampung Regency

A A Aldino¹, D Darwis², A T Prastowo³, C Sujana⁴
¹ Department of Informatics, Faculty of Engineering and Computer Science, Universitas Teknokrat Indonesia, Jl. ZA. Pagar Alam No.9 -11, Bandar Lampung, Indonesia
² Department of Accounting Information System, Faculty of Engineering and Computer Science, Universitas Teknokrat Indonesia, Jl. ZA. Pagar Alam No.9 -11, Bandar Lampung, Indonesia
³ Department of Computer Engineering, Faculty of Engineering and Computer Science, Universitas Teknokrat Indonesia, Jl. ZA. Pagar Alam No.9 -11, Bandar Lampung, Indonesia

email: aldino@teknokrat.ac.id¹

Abstract. South Lampung is a regency with the capital of Kalianda which has an area of 2,007.01 km² that dominates the agricultural area. Based on the data of corn crops in the South Lampung Regency Agriculture Office through BPS (Central Bureau of Statistics), showing several areas with corn crops that vary in number. Therefore, a grouping of potential corn-producing regions is required to know which areas produce large or small amounts of corn. The distribution of crops is usually done based on the name of the corn-producing sub-district. The K-Means clustering method is one of the data mining methods that is non-hierarchical clustering that groups data in the form of one or more clusters. Data that have the same characteristics are grouped in one cluster and the remaining is grouped into another cluster so that the data that is in one cluster has a small degree of variation. So the authors tried to apply the K-Means clustering method from the corn crop data of the last 2 years to produce feasibility information from each sub-district.

Keyword: Data Mining, K-Means Clustering, Corn Planting Feasibility, South Lampung Regency

1. Introduction

Land feasibility is the level compatibility of a land for a particular use that is more specific than land capability. The purpose of the land feasibility evaluation is to provide an assessment of the suitability of the land for the purposes that have been considered such as corn planting feasibility. The benefits of evaluating land feasibility are to provide an understanding of the relationships between land conditions and their use, as well as provide community/investors with various comparisons and alternative use of the land. [1]

Based on BPS (Central Bureau of Statistics) data, South Lampung District produced corn in 51.9 kw/ha in 2018 and 51.3 kw/ha 2019 which is slightly under the national corn productivity with 52.41 kw/ha. The data shows several areas with corn crops that vary in number. Therefore, a clustering of potential corn-producing regions is required to understand which areas need more concern to improve their production. The distribution of crops is usually done based on the name of the corn-producing sub-district. [2]

Data mining is a process with artificial intelligence, statistics, mathematics, and machine learning to identify and extract a useful information and related knowledge from a large database. Clustering is one of the methods of data mining and it has become a valid instrument for solving complex problems of computer science and statistics. Clustering is the process of grouping data points into two or more groups so that the data points included in the same group are more similar to each other than in different groups, based solely on the information available with data points. [3]
K-Means is one of the methods of grouping nonhierarchical data that can split data into the form of two or more groups. The method will split the data into a group where the same characteristic data will be entered into one group together while the data that has different characteristics will be grouped into another group. The purpose of grouping is to minimize the objective functions set in the grouping process, generally speaking it is to minimize variations within a group and maximize variation between groups. [4]

With the K-Means clustering approach, it is expected that the division of regional groups can be done based on the area of harvest (Ha), production (tons) and harvest year. In this study, the clustering of potential corn-producing areas was conducted using the K-Means algorithm. Using K-Means aims to facilitate the grouping of an area by high or low corn production. The result is an overview that shows the grouping of regions based on corn planting feasibility.

2. Literature Review

2.1 Data Mining

Data mining is the process of analyzing data from different perspectives and transform it into important information that can be used to increase profits, reduce expenses, or even both. Technically, data mining can be referred to as a process for finding correlations or patterns from hundreds or thousands of fields from a large relational database. [5]

Data mining's ability to search for valuable business information from a very large database can be used to predict trends and business properties, where data mining automates the process of searching for predictive information in a large database. As well as discovery of previously unknown patterns, where data mining handle the database, then identifies previously hidden patterns in one stroke. Data mining and knowledge discovery in databases (KDD) are often used interchangeably to explain the process of extracting hidden information in a large database. [6]

Data mining is a large-scale data processing method therefore data mining has an important role in the fields of industry, finance, weather, science and technology. In general, data mining studies discuss methods such as clustering, classification, regression, variable selection, and market analysis. [7]

2.2 Data Mining Process

Data mining processes are as follows:

1. Data Selection
   Data selection is used to determine the input variables so that there are no redundancies and unnecessary iteration in the data mining process.

2. Data Preprocessing
   There are two stages in the data preprocessing, namely the following:
   a. Data Cleaning
      Eliminates unnecessary data such as handling missing value, noise data, inconsistent and irrelevant data.
   b. Data Integration
      Performed against attributes that identify unique entities.

3. Transformation
   Change the data according to the appropriate format in data mining processing because some methods in data mining require a special format before it can be processed in data mining.

4. Data mining
   The main process on the method applied to obtain new knowledge of the processed data. In this study, clustering techniques were applied, namely the K-Means Clustering method.

5. Evaluation/ Interpretation
   Identify interesting patterns into the identified knowledge base. At this stage, the typical patterns and prediction models evaluated to assess existing studies meet the desired targets.

6. Knowledge
The resulting patterns will be presented to the user. At this stage the resulting new knowledge can be understood by everyone who will be the reference of decision making. [8]

2.3 Clustering

One of the techniques known in data mining is clustering. Clustering is a grouping of a number of data or objects into clusters so that each in that cluster will contain data that is as similar as possible and different from objects in other clusters. There are two clustering methods that is well known, namely hierarchical clustering and partitioning. The hierarchical clustering method itself consists of complete linkage clustering, single linkage clustering, average linkage clustering and centroid linkage clustering. While the partitioning method itself consists of k-means and fuzzy k-means. [9]

Clustering is a method or technique of data grouping. Clustering differs from classification i.e. the absence of target variables in clustering. Clustering does not attempt to classify or predict the value of the target variable. However, this process tries to divide the entire data into homogeneous groups. Clustering techniques are widely applied in various fields. For example, in the medical field, clustering can be used to group disease types based on the characteristics and symptoms experienced by patients. [10]

2.4 K-Means Algorithm

K-Means Clustering is a method of data analysis or data mining method that performs an unsupervised modeling process and is one of the methods that performs data grouping with the partition system. The K-Means method attempts to group existing data into groups, where data in one group has the same characteristics as each other while different characteristics will be grouped into the other group. [11] The K-means algorithm is an algorithm that needs the input parameters by as much as k and divides a set of n objects into k clusters so that the level of similarity between members in a cluster is high while the level of resemblance to members in other clusters is very low. The similarity of cluster towards members is measured by the object's proximity to the mean value on the cluster or can be referred to as a centroid cluster or mass center, the steps of clustering with the K-Means method are as follows:

a. Select the number of clusters k.
b. Initialization of the center of this cluster can be done in a variety of ways, but the most common thing is to do so in the way the cluster centers are initially rated with random numbers.
c. Allocate all data/objects to the nearest cluster. The proximity of the two objects is determined based on the distance of the two objects. Likewise, the proximity of a data to a particular cluster is determined by the distance between the data and the center of the cluster. At this stage it is necessary to calculate the distance of each data to each cluster center. The distance between one data and one specific cluster determines which data is entered in which cluster. To distance all data to each cluster center point can use the Euclidean distance theory.
d. Recalculate the center of the cluster with the current cluster member. The center of the cluster is the average of all data/objects in a given cluster. You can also use the median of the cluster. So mean isn't the only size you can use.
e. Task each object again using the new cluster center. If the center of the cluster is not changed then the clustering process is complete. Or, go back to step 3 until the center of the cluster is unchanged. [12]

3. Research Methods

3.1 Research Stage

Below is the the stages of research that will be conducted:

a. Crop Data Stage

At this stage, the authors prepare corn crop data from each sub-district that is ready to be
processed for research purposes. The data taken by the authors is data sourced from the Central Bureau of Statistics (BPS) of South Lampung.

b. Data Analysis Stage
At this stage it is divided into two, namely:
1. Analyzing crop data based on the year in which the data will be examined is the last 2-year crop data from 2018-2019.
2. Analyzing the data based on sub-district. The data taken for this study is the data of corn crops from each sub-district in the South Lampung region.

c. K-Means Method Calculation Stage
At this stage the data obtained will be calculated with K-Means method, as the following:
1. Determine the number of clusters
2. Determining Centroid (central point)
3. Calculate the distance of each object to the center point (centroid) with K-Means formula
4. Get results with many or low harvest category.

3.2 Data Collection Methods
The methods of data collection carried out in this study, as the following:
1. Interview
The method of collecting data or information is done by asking directly to the Central Static Agency (BPS) of South Lampung.
2. Observation
Researchers made direct observations to the location of the corn farm area, observing the location directly studied. It is intended to obtain a clear picture of the corn-producing area.
3. Literature Review
The method of data collection is done by studying various scientific reports and document or reading sources as well as books related to or related to the topic of research proposals.

4. Result and Discussion

4.1 K-Means Clustering Process
The following will be calculated from corn crop data taken from the Central Bureau of Statistics (BPS) of South Lampung which can be seen in the Table 1.

Table 1. Harvested Area and Production of Crop in South Lampung Regency

| No | District   | Harvested Area (Ha) | Year 2018 Production (Ton) | Harvested Area (Ha) | Year 2019 Production (Ton) |
|----|------------|----------------------|----------------------------|----------------------|----------------------------|
| 1  | Natar      | 15146                | 84313                      | 11568                | 59175                      |
| 2  | Jati Agung | 6180                 | 30725                      | 6000                 | 28040                      |
| 3  | Tanjung Bintang | 7234           | 40000                      | 7762                 | 40223                      |
| 4  | Tanjung Sari | 5810              | 29968                      | 4781                 | 24457                      |
| 5  | Katibung   | 5603                 | 54365                      | 4934                 | 35470                      |
| 6  | Merbau Mataram | 1289              | 6737                       | 3247                 | 16972                      |
| 7  | Way Sulan  | 1734                 | 9063                       | 1387                 | 7250                       |
| 8  | Sidomulyo  | 8819                 | 46096                      | 8675                 | 44376                      |
| 9  | Candipuro  | 3919                 | 16826                      | 3273                 | 16743                      |
| 10 | Way Panji  | 3816                 | 18889                      | 3816                 | 19946                      |
| 11 | Kalianda   | 10401                | 54300                      | 10401                | 54365                      |
| 12 | Rajabasa   | 324                  | 715                        | 142                  | 726                        |
The steps in K-Means Clustering are as follows:

a. determine the number of clusters
   C0 : High
   C1 : Low

b. specify centroid (central point)
   C0 : 14,325 : 73,278 : 84,313
   C1 : 142 : 726 : 715

c. calculate the distance of each object to the center point (centroid) with the formula (1)

\[
D(i,j) = \sqrt{(x_{1i} - x_{1j})^2 + (x_{2i} - x_{2j})^2 + \cdots + (x_{ki} - x_{kj})^2}
\]  

(1)

Where:

\( D (i,j) \)  = Data distance to i to cluster center  j
\( x_{ki} \)  = Data to i on data attribute to k
\( x_{kj} \)  = Center point to j on attribute to k

1) Calculation distance of Natar to center centroid 1 is:

\[
d_{10} = \sqrt{(11568 - 14,325)^2 + (59175 - 73,278)^2 + (84313 - 84,313)^2}
\]

\[= 14369.95678\]

2) Calculation Distance of Natar to center centroid 2 is:

\[
d_{11} = \sqrt{(11568 - 142)^2 + (59175 - 726)^2 + (84313 - 715)^2}
\]

\[= 102642.4117\]

After the distance of each object to the centroid of each sub-district's crop, the result can be seen as in Table 2 below:

**Table 2. Centroid calculation results**

| Centroid 1 | Centroid 2 |
|------------|------------|
| 14,325     | 102642.4117|
| 70621.92303| 40999.64463|
| 55671.78876| 56226.22728|
| 73674.67002| 37952.84826|
| 48795.02484| 64277.51784|
| 96494.29963| 17602.0004 |

\[d_{i,j} = \sqrt{(x_{1i} - x_{1j})^2 + (x_{2i} - x_{2j})^2 + \cdots + (x_{ki} - x_{kj})^2}\]
d. after obtaining the results of each centroid, the minimum value of the centroid is taken as shown in the Table 3.

**Table 3. Results of the K-Means Iteration 1**

| Centroid 1 | Centroid 2 | Results |
|------------|------------|---------|
| 100943.9853 | 10667.61667 | ok      |
| 48247.14699 | 63541.87399 |         |
| 88729.07132 | 22932.76632 |         |
| 85059.12257 | 26705.73302 |         |
| 35691.41906 | 76509.78785 |         |
| 111595.5904 | 0           |         |
| 56746.15414 | 56916.41198 |         |
| 83858.52146 | 27804.82134 |         |
| 10727        | 103803.3373 |         |
| 22901.66061 | 88765.48811 |         |
| 75109.88938 | 36499.96501 |         |

| Number of Centroid Members | 8 | 9 |

Afterwards, the new centroid is shown as in the Table 4

**Table 4. New Centroid**

| Centroid 1 | Centroid 2 |
|------------|------------|
| 9,905      | 51,151     |
| 57043.8    | 3495.22    |
| 17689.4    | 17867.9    |
The next step is to calculate the second iteration as shown in Table 5

**Table 5. Results of the K-Means Iteration 2**

| Centroid 1 | Centroid 2 | Results |
|------------|------------|---------|
| 9,905      | 51,151     | 3,495   | 17,689 | 17,867 | Centroid 1 | Centroid 2 |
| 28474.60351 | 78748.53929 | ok |
| 35242.08379 | 16695.69975 | ok |
| 20358.72005 | 31872.52946 | ok |
| 38365.0966 | 13924.57615 | ok |
| 16183.08642 | 40744.25955 | ok |
| 61182.04512 | 11155.84281 | ok |
| 65589.32194 | 13817.83836 | ok |
| 12932.53007 | 39190.62145 | ok |
| 53341.36272 | 1424.036868 | ok |
| 49664.50861 | 2498.201994 | ok |
| 4254.435467 | 52155.4351 | ok |
| 76228.8684 | 24355.227 | ok |
| 22498.70837 | 32893.80055 | ok |
| 48468.90614 | 3452.125432 | ok |
| 27978.76656 | 79448.30257 | ok |
| 12831.54984 | 64526.96824 | ok |
| 39731.02855 | 12210.52996 | ok |

| Number of Centroid Members | 8 | 9 |

4.2 **Testing using Rapid Miner 5.3 app**

Rapid Miner is one of the software for data mining processing. The work done by Rapid Miner is to range from text analysis, extract patterns from large data sets and combine them with statistical methods, artificial intelligence, and databases. The purpose of this analysis is to obtain the highest quality information from the processed data.

The steps of using Rapid Miner are as the following:

a. Insert data from excel into an application
   The first step is to insert the harvest data file from excel into the rapid miner application. The data included in this application is the sub-district name data, land area, 2018 production and 2019 production.

b. Determine the Cluster
   At this stage determines the number of clusters to be created in the processing of crop data by inputing the number of K =2 and max runs = 10.

c. Random selected center point
   At this stage the central point on each cluster is randomly selected according to the input data that will be continued to the centroid calculation stage.
d. Clustering Results
At this stage the application brings up centroids from each sub-district. The class determination is taken from the smallest value of the centroid result.

e. New Centroid
Once centroid results are obtained from each sub-district, then you can get a central point value on the new cluster for further iterations.

f. Plot View Results

![Plot View](image)

**Figure 1.** Plot View

It can be seen from Figure 1 that cluster_0 is regions with high number of harvest while cluster_0 is the opposite. While the Table 6 shows the clustering result of every region

**Table 6.** The Clustering result of corn planting feasibility per region

| Regions      | Centroid 1   | Centroid 2   | Feasibility |
|--------------|--------------|--------------|-------------|
| Natar        | 28474.60351  | 78748.53929  | High        |
| Jati Agung   | 35242.08379  | 16695.69975  | Low         |
| Tanjung Bintang | 20358.72005  | 31872.52946  | High        |
| Tanjung Sari | 38365.0966   | 13924.57615  | Low         |
| Katibung     | 16183.08642  | 40744.25955  | High        |
| Merbau Mataram | 61182.04512  | 11155.84281  | Low         |
| Way Sulan    | 65589.32194  | 13817.83836  | Low         |
| Sidomulyo    | 12932.53007  | 39190.62145  | High        |
| Candipuro    | 53341.36272  | 1424.036868  | Low         |
| Way Panji    | 49664.50861  | 2498.201994  | Low         |
| Kalianda     | 4254.435467  | 52155.4351   | High        |
| Rajabasa     | 76228.8684   | 24355.227    | Low         |
| Palas        | 22498.70837  | 32893.80055  | High        |
| Sragi        | 48468.90614  | 3452.125432  | Low         |
| Penengahan   | 27978.76656  | 79448.30257  | High        |
| Ketapang     | 12831.54984  | 64526.96824  | High        |
| Bakauheni   | 39731.02855  | 12210.52996  | Low         |
The Table 6 tells that there are 8 regions (cluster_0) having high corn harvest. These regions are Natar, Tanjung Bintang, Katibung, Sidomulyo, Kalianda, Palas, Penengahan, and Ketapang. In contrast, there are 9 regions (cluster_1) having low corn harvest. These regions are Jati Agung, Tanjung Sari, Merbau Mataram, Way Sulam, Candipuro, Way Panji, Rajabasa, Sragi, and Bakauheni.

5. CONCLUSION & SUGGESTION

Based on the results of the research, the conclusions are:
1. The region with the highest number of corn harvest is Penengahan with 79448.30257 centroid 2 value.
2. The region with the lowest number of corn harvest is Candipuro with 1424.036868 centroid 2 value.

Authors suggest that stakeholder can give more support in the regions having small number of corn harvest so that they can improve the production.

REFERENCES

[1] M. S. A. Pasaribu, Z. Nasution, “Evaluasi Kesesuaian Lahan Untuk Tanaman Jagung (Zea Mays L.) Dan Ubi Jalar (Ipomoea Batatas L.) Di Kecamatan Kualuh Selatan Kabupaten Labuhanbatu Utara,” J. Agroekoteknologi, vol. 6, no. 4, Okt, pp. 779–786, 2018.
[2] M. R. Islam, Z. Zeng, J. Mao, A. E. Encji, X. Xue, and Y. Hu, “Feasibility of summer corn (Zea mays L.) production in drought affected areas of northern China using water-saving superabsorbent polymer,” Plant, Soil Environ., vol. 57, no. 6, pp. 279–285, 2019.
[3] D. Aggarwal and D. Sharma, “Application of clustering for student result analysis,” Int. J. Recent Technol. Eng., vol. 7, no. 6, pp. 50–53, 2019.
[4] X. Yao, S. Ge, H. Kong, and H. Ning, “An Improved Clustering Algorithm and Its Application in WeChat Sports Users Analysis,” Procedia Comput. Sci., vol. 129, pp. 166–174, 2018.
[5] K. Govindasamy and T. Velmurugan, “Analysis of Student Academic Performance Using Clustering Techniques,” Int. J. Pure Appl. Math., vol. 119, no. 15, pp. 309–323, 2018.
[6] 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.
[7] S. Bano and M. N. A. Khan, “A Survey of Data Clustering Methods,” Int. J. Adv. Sci. Technol., vol. 113, no. April, pp. 133–142, 2018.
[8] R. Pung et al., “Investigation of three clusters of COVID-19 in Singapore: implications for surveillance and response measures,” Lancet, vol. 395, no. 10229, pp. 1039–1046, 2020.
[9] S. M. Kim, M. I. Peña, M. Moll, G. Giannakopoulos, G. N. Bennett, and L. E. Kavraki, “An evaluation of different clustering methods and distance measures used for grouping metabolic pathways,” Proc. 8th Int. Conf. Bioinforma. Comput. Biol. BICOB 2016, no. BICOB, pp. 115–122, 2016.
[10] A. Khaleghi, D. Ryabko, J. Mary, and P. Preux, “Consistent algorithms for clustering time series,” J. Mach. Learn. Res., vol. 17, pp. 1–32, 2016.
[11] A. V. D. Sano and H. Nindito, “Application of K-Means Algorithm for Cluster Analysis on Poverty of Provinces in Indonesia,” ComTech Comput. Math. Eng. Appl., vol. 7, no. 2, p. 141, 2016.
[12] L. Scrucca, M. Fop, T. B. Murphy, and A. E. Raftery, “Mclust 5: Clustering, classification and density estimation using Gaussian finite mixture models,” R J., vol. 8, no. 1, pp. 289–317, 2016.