Rice Productivity Analysis by Province Using K-Means Cluster Algorithm

A Supriyatna1,*, I Carolina1, W Widiati1, C Nuraeni1

1Universitas Bina Sarana Informatika, Indonesia

*Corresponding author e-mail : adi.asp@bsi.ac.id

Abstract. Rice (Oryza sativa L.) is a very important food crop in the world after wheat and corn. It is also a staple food for most of the world's population, especially in Asia, like in Indonesia until now. In 2014 to 2018, rice productivity tended to change dynamically. In 2018, rice productivity in Indonesia was 51.92 (Ku/Ha). This research was conducted to classify rice productivity in 34 provinces in Indonesia in 2018. The data used were sourced from Statistics Indonesia. The method or approach used in this study is the K-Means cluster algorithm to classify rice productivity data by province in 2018. The results of the research are; (1) There are 19 provinces included in cluster 0 (Medium), (2) There are 4 provinces included in cluster 1 (High), and (3) There are 11 provinces that are included in cluster 2 (Low). Based on the results of the study, it was proven that there were 4 provinces in cluster 1 (High) they are West Java, Central Java, East Java and South Sulawesi, with the highest rice productivity. Three of them were on Java Island. It shows that Java still dominates the productivity of rice plants.

1. Introduction

Indonesia, as an agricultural country, enacts the agriculture sector as one of the main contributors in people's lives and becomes a major component in the government's strategy to reduce poverty. Rice (Oryza sativa L.) is a food crop that plays an important role because it produces rice which is a source of staple food. As in Indonesia, rice is the main commodity in supporting people's food in daily needs. Based on this background, various efforts were made to increase the productivity of rice plants in provinces in Indonesia. One way is to use an irrigation system. However, the fact that rice productivity has not met national needs. Figure 1 below shows a graph of the amount of rice productivity from 2014 to 2018, that is 51.92 (Ku / Ha).

![Source: Statistics Indonesia](image)

**Figure 1.** Number of Rice Productivity

Agronomically, the increase in rice production is pursued through increased productivity rather than in rice planting area. Meanwhile, institutionally the program among others was realized through the implementation of Integrated Crop Field Schools / Sekolah Lapangan Tanaman Terpadu (SL PTT) with the aim of improving the quality of rice cultivation techniques carried out by farmers, increasing...
cropping intensity and rice productivity[1]. Based on data from Statistics Indonesia (Badan Pusat Statistik/BPS) rice productivity increased by 0.52% from 2017 to 2018[2]. In this study, the researchers conducted a study in 34 provinces in Indonesia with the purpose of classifying rice productivity in 2018 and data processed from the Statistics Indonesia. The technique that the researchers used was cluster analysis. It is a multivariate technique with the main objective of grouping objects based on the characteristics they possess[3][4]. Clustering analysis is a technology which can classify the similar sample points into the same group from a data set[5][6]. The K-Means algorithm is proposed based on the hybrid rice optimization algorithm with the aim to optimize the initial center of clustering as a good start for clustering. At the same time the optimization algorithm can improve the accuracy of clustering and the efficiency of the K-Means clustering algorithm[7]. The same study was conducted using data from 1993-2015, the results of clustering rice by province using Rapidminer, obtained high production clusters of 23 provinces, normal production clusters of 8 provinces and low production clusters of 3 provinces. The 3 provinces with the highest rice production are Central Java, East Java and West Java[8].

2. Research Method
2.1. Stages of the Research
In preparing the research, the researchers did the stages; they are:
1. Data collection.
   The data used in this study are data on harvested area, rice production and productivity according to 34 provinces in Indonesia. Data obtained from Statistics Indonesia.
2. Data Normalization.
   Normalised data used to scale the attribute values of the data so that they fit within a certain range. The formula used in data normalisation is the Z-Score which is a normalisation method based on the mean (average value) and standard deviation of the data.
3. Implement of the K-Means algorithm using Rapidminer tools.
   The normalised dataset is then processed using the K-Means Cluster algorithm with Rapidminer Studio.
4. Data analysis.
   Rice productivity data processing generated by Rapidminer tools in the form of cluster models and performance vectors, based on these results will group rice productivity by 34 provinces into three clusters.

2.2. K-Means Cluster
This study used the K-means algorithm to determine the best cluster. K-means clustering is a well-known partitioning method[7]. In these objects are classified as belonging to one of K-groups. The results of Partitioning method is a set of K clusters, each object of data set belonging to one cluster. In each cluster, there may be a centroid or a cluster representative. In a case where we consider real-valued data, the arithmetic mean of the attribute vectors for all objects within a cluster provides an appropriate representative; alternative types of centroid may be required in other cases[9]. Another expert stated that the K-Means clustering method, groups the data based on their closeness to each other according to the Euclidean distance[10].

The basic algorithm in K-means is:

a. Specify the number of clusters (k), set the cluster centre to arbitrary
b. Calculate the distance of each data centre cluster using the equation:

\[ d_{ik} =\sqrt{\sum_{j=1}^{m} c_{ij} - c_{kj}}^2 \]  

(1)
c. Group data into cluster with the shortest distance using the equation:

\[ \text{Min} \sum_{k=1}^{k} d_{ik} = \sqrt{\sum_{j=1}^{m} c_{ij} - c_{kj}}^2 \]  

(2)
d. Calculate the new cluster centre using the equation:

\[ c_{kj} = \frac{\sum_{i=1}^{p} x_{ij}}{p} \]  

(3)
e. Repeat steps two through four until no more data is moved to another cluster.

3. Result and Discussions

3.1. Dataset

The dataset used in this study is data on harvested area, rice production and productivity based on 34 provinces in Indonesia, as shown in table 1.

**Table 1. Dataset of harvested area, production and productivity of rice**

| No. | Province            | Harvested Area (ha) | Rice Productivity (ku/ha) | Production (ton) |
|-----|---------------------|---------------------|---------------------------|-----------------|
| 1   | Aceh                | 297294              | 57.11                     | 1697756         |
| 2   | North Sumatra       | 360716              | 52.89                     | 1907725         |
| 3   | West Sumatra        | 318579              | 47.45                     | 1511538         |
| 4   | Riau                | 93755               | 38.96                     | 365293          |
| 5   | Jambi               | 118408              | 42.23                     | 500021          |
| 6   | South Sumatra       | 513209              | 41.51                     | 241398          |
| 7   | Bengkulu            | 59455               | 42.76                     | 254218          |
| 8   | Lampung Bangka Belitung Island | 8618 | 47.83 | 1901041 |
| 9   | West Nusa Tenggara  | 281013              | 49.8                      | 1399495         |
| 10  | East Nusa Tenggara  | 200877              | 39.87                     | 800980          |
| 11  | West Kalimantan Central | 214877 | 28.95 | 622041 |
| 12  | South Kalimantan    | 202142              | 36.74                     | 742758          |
| 13  | Central Kalimantan  | 278853              | 40.76                     | 1136511         |
| 14  | West Kalimantan     | 58151               | 41.51                     | 241398          |
| 15  | North Kalimantan    | 14133               | 32.07                     | 45323           |
| 16  | South Sulawesi      | 82051               | 44.69                     | 366722          |
| 17  | West Kalimantan     | 204158              | 46.77                     | 954794          |
| 18  | Central Sulawesi    | 1145319             | 50.12                     | 5740730         |
| 19  | South Sulawesi      | 131399              | 37.98                     | 499007          |
| 20  | Southeast Sulawesi  | 51765               | 46.74                     | 241948          |
| 21  | Gorontalo           | 67835               | 48.08                     | 326169          |
| 22  | West Sulawesi       | 10343               | 34.19                     | 35360           |
| 23  | North Maluku        | 23317               | 38.98                     | 90892           |
| 24  | Papue               | 31133               | 41.99                     | 130718          |

3.2. Normalisation Data

Normalisation of data is carried out to minimise the magnitude of the variable to avoid difficulties in the grouping or clustering process. The equation used for data normalisation is:

\[ Normalization = \frac{(InitialValue - MinValue)}{(MaxValue - MinValue)} \]  

(4)

The normalised dataset is shown in table 2 below.

**Table 2. Data on Normalization of Harvested Area, Production and Rice Productivity**

| No. | Province            | Harvested Area (ha) | Rice Productivity (ku/ha) | Production (ton) |
|-----|---------------------|---------------------|---------------------------|-----------------|
| 1   | Aceh                | 0.1624              | 0.9848                    | 0.161           |
| 2   | North Sumatra       | 0.1971              | 0.8613                    | 0.181           |
| 3   | West Sumatra        | 0.1741              | 0.7021                    | 0.1434          |
| 4   | Riau                | 0.0511              | 0.4536                    | 0.0346          |
| 5   | Jambi               | 0.0646              | 0.5493                    | 0.0474          |
| 18  | West Nusa Tenggara  | 0.1535              | 0.7709                    | 0.1327          |
| 19  | East Nusa Tenggara  | 0.1097              | 0.4802                    | 0.0759          |
| 20  | West Kalimantan     | 0.1173              | 0.1607                    | 0.059           |
| 21  | Central Kalimantan  | 0.1104              | 0.3886                    | 0.0704          |
| 22  | South               | 0.1523              | 0.5063                    | 0.1078          |
3.3. Centroid Data

Based on the data that has been normalised in table 2, then determine the number of clusters to be selected. The initial clusters selected are 3 (three) clusters based on Euclidean distances, as shown in Table 3 below:

| Cluster 1 (High) | 1 | 1 | 1 |
|------------------|---|---|---|
| Cluster 2 (Medium) | 0,0281 | 0,6813 | 0,0229 |
| Cluster 3 (Low) | 0 | 0 | 0 |

Determination of initial centroid data points is done by taking the highest value as cluster 1 (C1) from the province of East Java, the middle value as cluster 2 (C2) from Gorontalo province and the lowest value as cluster 3 (C3) from the Riau Islands province.

3.4. K-Means Clustering Data Using Rapid Miner

Dataset of harvested area, production and rice productivity based on provinces with initial predetermined centroid data, then processed the data to get the results of grouping with K-Means cluster algorithm using Rapid miner Studio software.

Data processing with K-Means algorithm.

![Figure 2. Design of K-Means Algorithm Operator](image)

The design of data processing with the K-Means algorithm in Rapidminer studio uses 4 operators. Read Excel is used for reading the harvested area, production and rice productivity dataset based on the province that has been normalised. Next, the dataset is clustered or grouped using the K-Means clustering operator with a value of K = 3 and max runs = 10. The Apply Model operator is used to be able to read a clustered dataset based on data that has been studied. The operator of Cluster
Performance Distance is used to measure the performance of the K-Means clustering algorithm towards the dataset being read. The results of data processing with the K-Means algorithm in Rapidminer studio can be seen in the figure below:

Figure 3. Result of K-Means Cluster Model

The figure 3 above shows the results of processing rice productivity data using the K-Means cluster algorithm which is divided into 3 clusters, namely, 19 provinces included in the medium cluster, four provinces included in the high cluster and 11 provinces included in the low cluster.

Performance Accuracy of K-Means Cluster. The parameters used to measure the performance of the K-Means algorithm are done by counting `avg_within_centroid_distance` and `davies_bouldin`. `avg_within_centroid_distance` is the average in cluster distance calculated by the average distance between centroids and all cluster samples. Whereas `davies_bouldin` is an algorithm that produces clusters with low intra-cluster spacing (high level of intra-cluster similarity) and high inter-cluster spacing (low level of inter-cluster similarity) will have a low Davies-Bouldin index. The grouping algorithm that produces clusters with the smallest Davies-Bouldin index is considered as the best algorithm based on criteria. The results of the performance measurement of the K-Means cluster algorithm in grouping rice productivity data using the operator cluster performance distance in Rapidminer studio can be seen in Figure 5 below:

Figure 4. Performance Vector K-Means Cluster

Figure 4. above shows PerformanceVector average within centroid distance of -0.039, average within centroid distance cluster 0 of -0.038, average within centroid distance cluster 1 of -0.058, average within centroid distance cluster 2 of -0.035 and PerformanceVector Davies Bouldin of -0.615.

3.5. Data Analysis
Grouping rice productivity data based on province in 2018 with the K-Means Cluster algorithm using Rapidminer studio has resulted in 3 clusters which can be seen in Figure 6 below. The following data is a grouping of 34 provinces into groups that are formed.

Based on the results of data processing with the K-Means Cluster algorithm, 19 provinces are included in cluster 0. They are Aceh, North Sumatera, West Sumatra, Jambi, South Sumatra, Bengkulu, Lampung, DKI Jakarta, DI Yogyakarta, Banten, Bali, West Nusa Tenggara, South Kalimantan, East Kalimantan, North Sulawesi, Central Sulawesi, Gorontalo, West Sulawesi and Papua. There are four provinces included in cluster 1, i.e. West Java, Central Java, East Java and South Sulawesi. While 11 provinces are included in cluster 2, they are Riau, Bangka Belitung Island, Riau Island, East Nusa Tenggara, West Kalimantan, Central Kalimantan, North Kalimantan, Southeast Sulawesi, Maluku, North Maluku and West Papua.

4. Conclusion
Based on the research results, classification of rice productivity data based on provinces in Indonesia with the K-Means cluster algorithm using Rapidminer Studio produces 3 clusters, namely high clusters there are 4 Provinces, medium clusters there are 19 provinces and low clusters there are 11 provinces. 4 provinces in cluster 1 with the highest rice productivity, 3 of them are located on the island of Java, this shows that the island of Java still dominates the productivity of rice plants.

References
[1] V. D. Sari and B. M. Sukoko, “Analisa Estimasi Produksi Padi Berdasarkan Fase Tumbuh Dan Model Peramalan Autoregressive Integrated Moving Average (Arima) Menggunakan Citra Satelit Landsat 8,” Geoid, vol. 10, no. 2, p. 194, 2015.
[2] Badan Pusat Statistik, “Luas Panen, Produksi dan Produktivitas Padi,” Jakarta, 2018.
[3] 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.
[4] A. P. Windarto, “Implementation of Data Mining on Rice Imports by Major Country of Origin Using Algorithm Using K-Means Clustering Method,” Int. J. Artif. Intell. Res., vol. 1, no. 2, pp. 26–33, 2017.
[5] C. Y. Chen and F. Ye, “Particle swarm optimization algorithm and its application to clustering analysis,” 2012 Proc. 17th Conf. Electr. Power Distrib. EPDC 2012, no. 1, pp. 789–794, 2012.
[6] Ediyanto, M. N. Mara, and N. Satyahadewi, “Pengklasifikasian Karakteristik Dengan Metod K-Means Cluster Analysis,” Bul. Ilm. Mat. Stat. dan Ter., vol. 02, no. 2, pp. 133–136, 2013.
[7] C. Liu, C. Wang, J. Hu, and Z. Ye, “Improved K-means algorithm based on hybrid rice optimization algorithm,” Proc. 2017 IEEE 9th Int. Conf. Intell. Data Acquis. Adv. Comput. Syst. Technol. Appl. IDAACS 2017, vol. 2, no. 2, pp. 788–791, 2017.
[8] Sudirman, A. P. Windarto, and A. Wanto, “Data mining tools | rapidminer: K-means method on clustering of rice crops by province as efforts to stabilize food crops in Indonesia,” *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 420, no. 1, p. 012089, Oct. 2018.

[9] N. Kaur, J. K. Sahiwal, and N. Kaur, “Efficient K-Means Clustering Algorithm Using Ranking Methode In Data Mining,” *Int. J. Adv. Res. Comput. Eng. Technol.*, vol. 1, no. 3, pp. 85–91, 2012.

[10] P. Arora, Deepali, and S. Varshney, “Analysis of K-Means and K-Medoids Algorithm for Big Data,” *Int. Conf. Inf. Secur. Priv.*, vol. 78, pp. 507–512, 2016.