JOURNAL OF SOFTWARE ENGINEERING & INTELLIGENT SYSTEMS
ISSN 2518-8739
31 August 2019, Volume 4, Issue 2, JSEIS, CAOMEI Copyright © 2016-2019
www.jseis.org

FINDING OF NON-OPTIMAL PRODUCTION RESULTS FOR PALM OIL DATA PROCESSING USING K-MEANS ALGORITHM

YANA SHINTYA, ONNY MARLEEN

Gunadarma University, Depok, Indonesia
Email: yanashintya@gmail.com, onny_marleen@staff.gunadarma.ac.id

ABSTRACT

Academic Editor: Hikmat Ullah Khan, COMSATS University, Wah, Pakistan.

Published: 31st August, 2019

Funding: Authors have no funds.

Article Type: Research Article

Indonesia is the largest producer of Crude Palm Oil (CPO) in the world with an average growth of 8% every year since 2004 with an area of 12.30 million hectares. Even though Indonesia is the number one palm oil producer in the world, the results of productivity of each province have not been evenly distributed. This study aims to find out which provinces have not been optimally produced. A mapping method for productivity results is needed to categorize the data so that there is an even distribution of oil palm production in each province in Indonesia. The source of research data was collected based on documents produced by the Central Bureau of Statistics from 2012-2017 in the form of plant area data and crop production data for oil palm plantations consisting of 24 provinces. Data is processed by applying the clustering method and the K-Means algorithm to map the results of oil palm productivity by dividing into 3 clusters, namely clusters less than the target, clusters according to the target and cluster more than the target. The results of the mapping found that the Province was not optimal at 67%, which was optimal at 25% and more than optimal at 8%.

Keywords: palm oil; data mining; clustering method; k-means algorithm;

1. INTRODUCTION

Crude Palm Oil (CPO) production in Indonesia is expected to continue to increase in the next few decades with a predicted increase in land area of around 5% every year [1-2]. Indonesia is targeted to continue to increase productivity results every year to balance the fulfillment of national needs and fulfill export needs so that the province needs to know that the results are not optimal. Ttherefore it will receive effective attention and handling as it is related to government policy making, which of course must have relevance and be supported by knowledge derived from available data. Based on these constraints, a method is needed to map the production of oil palm plantations in each province in Indonesia.

Previous research has been conducted relating to the use of data mining to process data collection on fruit exports according to the destination country using the K-Means algorithm by Agus Perdana W. [3]. K-Means algorithm is a non-hierarchical grouping method that has a relatively fast computational time. Based on the comparative analysis between K-Means and Fuzzy C-Means (FCM) conducted by Suomi G. and Sanjay Kumar D., the results obtained prove that the K-Means algorithm is faster with the elapsed time of 0.433755 seconds compared to the FCM algorithm has elapsed time of 0.781679 seconds [4]. Data is processed using the Rapid Minner application. Rapid Minner is a machine learning environment, data mining, text mining and predictive analytics. Cluster results can be used as input for Indonesia as a form of mapping of destination countries. Obtained the results of 2 countries with the highest export rates namely India and Pakistan. The 3 export level clusters are in Singapore, Bangladesh and other countries and 6 clusters with the lowest levels are Hong Kong, China, Malaysia, Nepal, Vietnam and Iran [3].

In order to meet the needs of Crude Palm Oil (CPO) to achieve the target, the government will encourage investment in the palm oil sector. It is estimated that a minimum of 1 million ha of additional plantation land is needed in the next two years [1]. A mapping method for productivity results is needed to group data so that oil palm production is evenly distributed in each province in Indonesia. Sources of research data were collected based on documents produced by the Central Bureau of Statistics from 2012-2017 in the form of plant area data and production data of oil palm plantations consisting of 24 Provinces [5]. Data is processed by applying clustering method and K-Means algorithm to map the results of the productivity of oil palm by dividing into 3 clusters, namely cluster less than target, cluster according to target and cluster more than target.

Corresponding Author: Yana Shintya
Email Address: yanashintya@gmail.com

2. RESEARCH BACKGROUND

2.1 Data Mining

Data mining is defined as the process of obtaining useful information from large database warehouses. Data mining can also be interpreted as extracting new information taken from large chunks of data that helps in making decisions. The term data mining is sometimes also called knowledge discovery [6]. The terms data mining and knowledge discovery in databases (KDD) are often used interchangeably to describe the process of extracting hidden information in a large database. Actually the two terms have different concepts, but are related to each other. One of the stages in the whole KDD process is data mining. The KDD process can generally be explained as follows [7]:

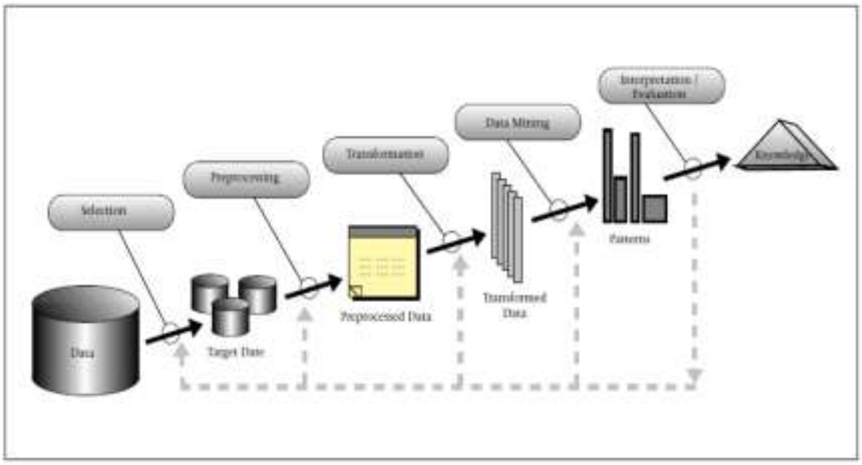
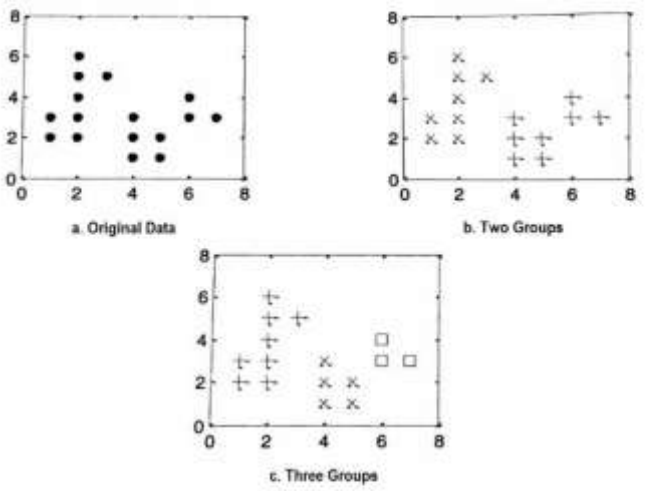


Figure.1. Steps in data mining

1. Data Selection: To select data from a set of operational data which requires to be done before the information extraction stage in KDD.
2. Pre-processing / Cleaning: includes, among other things, removing data duplication, checking for inconsistent data, and correcting errors in data, such as typographical errors.
3. Transformation: the transformation process on selected data, therefore the data is suitable for the data mining process.
4. Data mining: the process of looking for patterns or interesting information in the selected data using certain techniques or methods.
5. Interpretation / Evaluation: the information patterns generated from the data mining process is to be displayed in a form that is easily understood by the parties who needs the data.

2.2 Classification

Cluster analysis is the work of grouping data (objects) based only on information found in data that describes the object and the relationships between them [6].



a. Original Data

b. Two Groups

c. Three Groups

Figure. 2 Different clustering results of the same data

The goal is that the objects that are joined in a group are objects that are similar (or related) to each other and are different (or not related) to objects in another group. The greater the similarity (homogenetic) in the group and the greater the difference between the other groups, this concept will be discussed in the grouping [8]. Figure 2 shows the results of the different clustering results of the same data.

2.3 K-Means Algorithm

K-Means is a non-hierarchical (grouped) data grouping method that attempts to partition existing data into two or more groups. This method of partitioning data into groups so that data with the same characteristics are included in the same group and data with different characteristics are grouped into other groups. Grouping data by the K-Means method is generally done by [9]:

1. Determine the number of groups
2. Allocate data into groups randomly
3. Calculate the center of the group (centroid / average) from the data in each group
4. Allocate each centroid data / closest average
5. Return to step 3, if there is still data that moves between groups or if there is a change in the centroid value above the specified threshold value or if the change in the value of the objective function used is still above the specified threshold value.

K-Means is a group analysis method that leads to the N partitioning of observation objects into K groups (clusters) where each object of observation is owned by a group with the nearest mean (average). The objective function used for K-Means is determined based on the distance and value of membership data in groups. The objective functions used are as follows [10]:

$$C_I = \frac{1}{M}\sum_{J=1}^{M} x_j \qquad \text{Eq(1)}$$

In this step, the centroid location (center point) of each group taken from the average of all data values for each feature must be recalculated. If M states the amount of data in a group, i declares the i feature in a group, and p denotes the data dimension, to calculate the feature centroid. The formula is done as many as p dimensions so i starts from 1 to p.

Measurement of Euclidean distance space using a formula:

$$D(x_2, x_1) = \|x_2 - x_1\| = \sqrt{\sum_{j=1}^{p} |x_{2j} - x_{1j}|^2} \qquad \text{Eq(2)}$$

D is the distance between data x_2 and x_1, and | . | is absolute value.

2.4 MATLAB

MATLAB stands for Matrix Laboratory,a software created by The Mathworks.inc and with the most recent version 7.04. MATLAB provides a special function for cluster analysis with K-Means, with the k-means () function. The syntactic of the utiization is as follows [11]:

1. [IDX,C,sumd,D] = kmeans(X,k)
2. [IDX,C,sumd,D] = kmeans(…,'distance',val)

The first syntactic is the basic syntax, while the second syntactic is used to complete the first syntactic for the 'distance' parameter. There is a Graphical User Interface (GUI) facility in Matlab to design programs in the form of windows so that users can implement programs created with the GUI easily, interactively and attractively. Matlab is equipped with a toolbox, Simulink so that adding power to solving complicated problems becomes easier [11].

3 METHODOLOGY

3.1 Data selection stage

The source of the research data was obtained from data collected based on the data documents of area and production data obtained from the Sub-directorate of Plantation Crop Statistics and other data from other sources such as the Directorate General of Plantation, Ministry of Agriculture and the site https://www.bps.go. id. Attributes used (1) data on the area of oil palm plantations according to Province and plant species in Indonesia and (2) data on oil palm plantation production by Province and plant species in Indonesia. Then the data will be processed by pre-processing or cleaning on the data.

Table. 1 Attributes selection

Attributes		Data Used
Plantation Plant Production by Province and Plant Types in Indonesia	√	Yes
Area of Plantations by Province and Plant Types in Indonesia	√	Yes
Large Plantation Production by Plant Type in Indonesia	X	No
Large Plantation Area by Plant Type in Indonesia	X	No
Number of Large Plantation Companies by Plant Type in Indonesia	X	No
Monthly Production of Plantations in Indonesia	X	No

3.2 Pre-processing/cleaning stage

Before the data mining process can be carried out, it is necessary to do a cleaning process on the data. The cleaning process includes removing data duplication, checking inconsistent data, correcting errors in data such as typographical errors and enrichment processes that are processes that enrich existing data with data or other information that is relevant and needed for KDD, such as existing or external information.

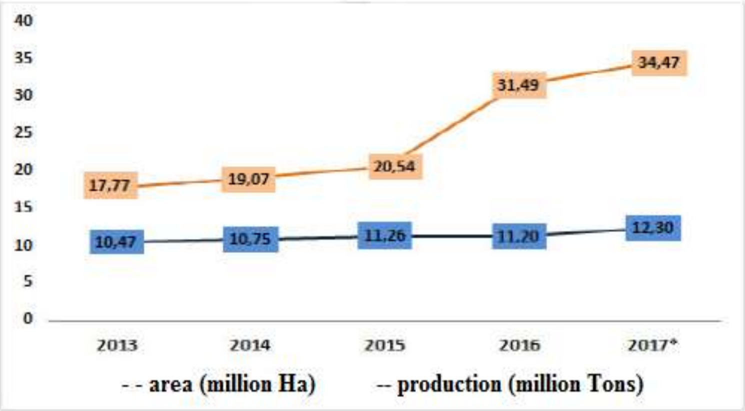


Figure. 3 Area and production of oil palm plantations in Indonesia

3.3 Data transformation stage

Data that has been obtained will be processed first to be clustered. In the previous stage, the data for each plantation area and plantation production per province will be added up, so that at this stage the calculation of the value that will be processed at the clustering stage has been obtained.

Table. 2 Data on extensive accumulation and production of palm oil

Provincial Data	Total Plant Area (Thousand of hectares)	Total Plant Production (Thousand of tons)
Accumulated Data	67907.39	183882.62
Lowest Data	83.09	51.48
Top Data	13814.74	44824.04
Average Data	2829.47	7661.78

3.4 Data Mining – Clustering Stage

The following are the steps in the process of processing data using the clustering method and the K-Means algorithm in figure 4:

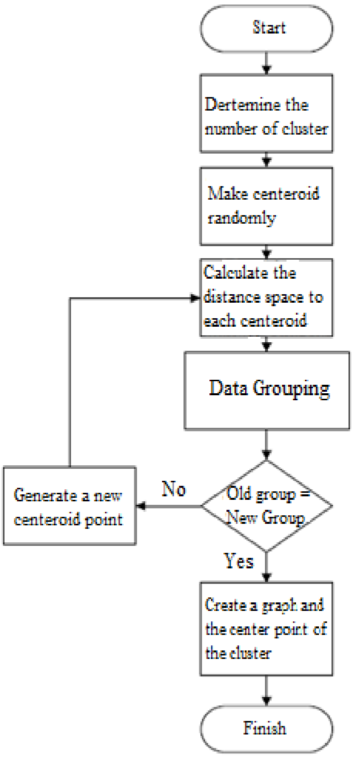

Start
Dertemine the number of cluster
Make centeroid randomly
Calculate the distance space to each centeroid
Data Grouping
Old group = New Group
No
Generate a new centeroid point
Yes
Create a graph and the center point of the cluster
Finish

Figure. 4 K-Means flowchart

4 RESEARCH RESULT AND DISCUSSION

4.1 Data transformation

In the application of the K-Means algorithm a midpoint or centroid value is generated from the data obtained provided that the desired cluster is 3, i.e. the cluster is less than the target (C1), the cluster is targeted (C2) and the cluster is more than the target (C3). Then the middle or centroid value also has 3 points. Determination of cluster points is done by summing the total data first, then taking the smallest value (minimum) for the cluster less than the target (C1), the average value for the cluster according to the target (C2), and the maximum value for the cluster more of target (C3).

Table. 3 Centroid point

Planting Area (Thousand of hectares)	Planting Production (Thousand of tons)	Cluster
150.00	100.00	C1 (less than the target)
2500.00	5000.00	C2 (according to the target)
10000.00	40000.00	C3 (more than the target)

4.2 Clustering data

Using the centroid, data can be clustered into 3 clusters. The cluster process takes the closest distance from each data that is processed. Data on oil palm plantations by province will be grouped in iterations 1 for the 3 clusters. Aceh Province will be used as reference data for clusters less than the target, West Sumatra Province is made into clusters according to target and North Sumatra Province is made more cluster than target. Then the data can be described as follows:

Table. 4 Calculation using centroid

Province	Plant Area	Plant Product	Plant Area	Plant Product	Plant Area	Plant Product
	C1		C2		C3	
	150.00	100.00	2500.00	5000.00	10000.00	40000.00
ACEH	5625.76		198.70		35596.02	
WEST SUMATRA	29473.78		24130.97		11703.09	
NORTH SUMATRA	6466.68		1193.94		34672.54	

Furthermore, the distance of the comparison results will be chosen based on the closest distance between the data and the cluster center. This distance indicates that the data is in a group with the nearest cluster center. Value 1 means that data is entered into the cluster. The following data can be explained in Table 5:

Table 5. Cluster Results

Province	C1	C2	C3
ACEH	1		
WEST SUMATRA	1		
NORTH SUMATRA			1
RIAU			1
JAMBI		1	
SOUTH SUMATRA		1	
BENGKULU	1		
LAMPUNG	1		
KEP. BANGKA BELITUNG	1		
KEP. RIAU	1		
WEST JAVA	1		
BANTEN	1		
WEST KALIMANTAN		1	
CENTRAL KALIMANTAN		1	
SOUTH KALIMANTAN		1	
EAST KALIMANTAN		1	
NORTH KALIMANTAN	1		
CENTRAL SULAWESI	1		
SOUTH SULAWESI	1		
SOUTHEAST SULAWESI	1		
WEST SULAWESI	1		
MALUKU	1		
WEST PAPUA	1		
PAPUA	1		
Total	**16**	**6**	**2**

4.3 Evaluation stag

The K-Means process will continue to iterate until the data grouping is the same as the previous iteration data grouping. In other words, the process will continue to iterate until the data in the last iteration is the same as the previous iteration. The cluster results above will form an easy to understand iteration point. Based on data that has been processed using Matlab R2015a, the final iteration point can be shown in Figure 5:

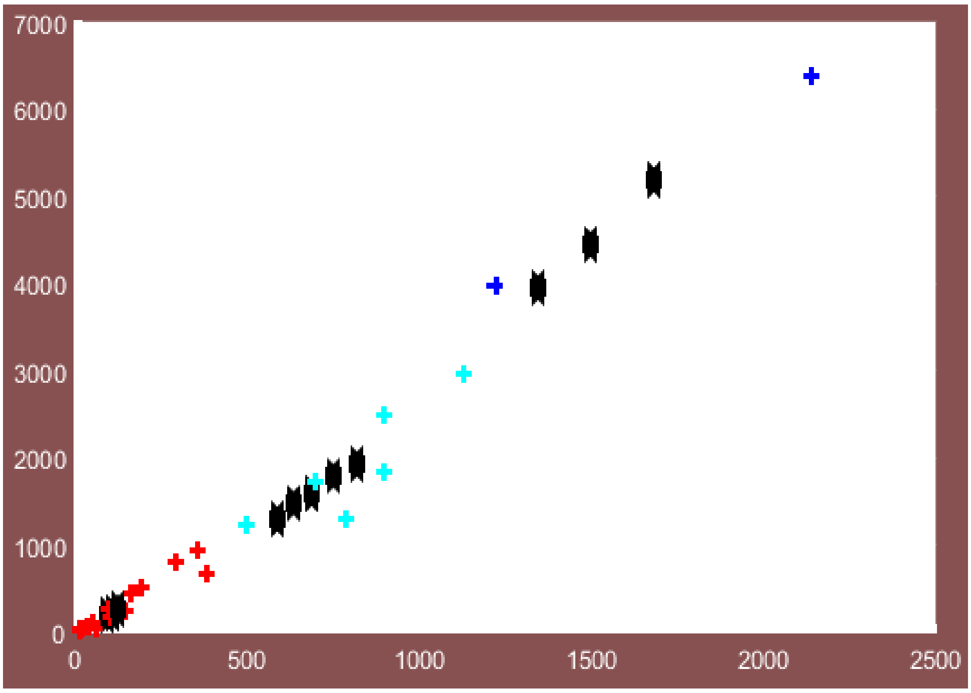


Figure. 5 Final iteration point

4.4 Analysis results

Based on the results of data on plant area and production of oil palm plants that have been processed, it can be described as follows:

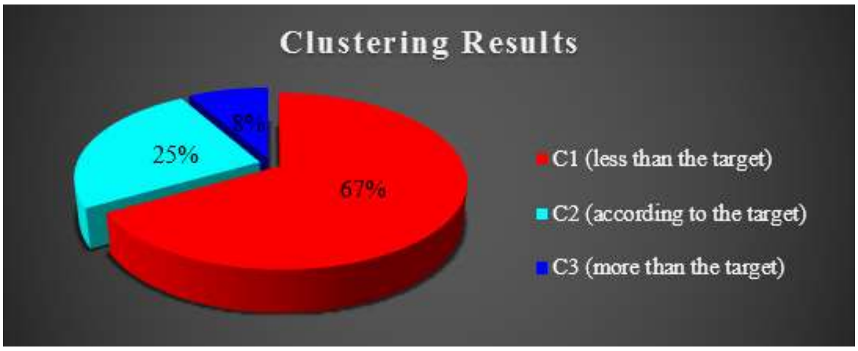


Figure. 6 Final results

Classification of the accuracy of the processed data that enters the cluster:

Table. 6 Classification of the accuracy

Cluster	Plant Area (Thousand of hectares)	Plant Product (Thousand of tons)
C1 (less than the target)	836.71	1885.10
C2 (according to the target)	5396.75	13413.27
C3 (more than the target)	11069.75	36620.74

Then it can be concluded that the provinces that are in clusters C1, C2 and C3 are seen in table 7:

Table. 7 Final conclusions

Final Conclusion	Province
Cluster 1 **(less than the target)**	Aceh, West Sumatra, Bengkulu, Lampung, Kep. Bangka Belitung, Kep Riau, West Java, Banten, North Kalimantan, Central Sulawesi, South Sulawesi, Southeast Sulawesi, West Sulawesi, Maluku, West Papua, Papua.
Cluster 2 **(according to the target)**	Jambi, South Sumatra, West Kalimantan, Central Kalimantan, South Kalimantan, East Kalimantan.
Cluster 3 **(more than the target)**	North Sumatra, Riau.

5 CONCLUSION

The Clustering method with the K-Means Algorithm has been used to process data in order to obtain productivity results from the plant area and oil palm plantation production of each province in Indonesia over the past six years. The productivity data of oil palm plantations can be mapped into three groups consisting of more than the target productivity, according to the target and less than the target with a percentage value that can be seen annually. Clustering the data that has been processed can be concluded that there are 8% of provinces considered to exceed the target of North Sumatra and Riau Provinces. 25% of provinces that are on target are Jambi, South Sumatra, West Kalimantan, Central Kalimantan, South Kalimantan, East Kalimantan and 67% of provinces that have not reached the target namely Aceh, West Sumatra, Bengkulu, Lampung, Kep. Bangka Belitung, Kep Riau, West Java, Banten, North Kalimantan, Central Sulawesi, South Sulawesi, Southeast Sulawesi, West Sulawesi, Maluku, West Papua, Papua. Based on the test results and analysis explained by the Province whose production results are still less than optimal, this needs to get attention and effective handling from the government so that there is an even distribution of productivity results in every province in Indonesia. These results relate to Indonesian government policy making to suggest local governments which are of course relevant and supported based on available data.

REFERENCE

1. Ministry of Industry. (2016). *Prospects and Problems of the Palm Oil Industry*. http://www.kemenperin.go.id/artikel/494/Prospek-Dan-Permasalahan-Industri-Sawit.
2. Hendaryanti, et al. (2016). *Statistics of Indonesian Plantations, Oil Palm 2015-2017*. Directorate General of Plantation, 81 p.
3. A. P. Windarto, "*Application of Data Mining in Exports of Fruits by Destination Country Using K-Means Clustering*" (*Techno.COM, Vol 16, No. 4, November 2017: 348-357*).
4. Ghosh, s. and S. K. Debey, "*Comparative Analysis of K-Means and Fuzzy C-Means Algorithms*," *vol. 4, no 35-39, 2013*.
5. Statistics Indonesia. 2018. *STATISTICS OF PALM OIL INDONESIA 2017*. Jakarta: CV. Dharmaputra.
6. Tan, P. et al. 2006. *Introduction to Data Mining*. Boston: Pearson Education
7. Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth. 1996. *From Data Mining to Knowledge Discovery in Databases.* AI Magazine Volume 17 Number 3.
8. Kusrini and Emha Taufiq Luthfi. 2009. *Data Mining Algorithm.* Yogyakarta: Andi Offset.
9. Agusta, Y. 2007. "K-Means - Applications, Problems and Related Methods". *Journal of Systems and Information, Vol. 3, pp. 47-60*.
10. MacQueen, J.B. 1967. "*Some Methods for Classification and Analysis of Multivariate Observation*". Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, University of California Press, 1:281-297.
11. Budi Halomoan Siregar and Ridwan Abdullah Sani. 2017. "*Introduction to Matlab*". Tangerang: Tsmart Printing.

AUTHORS PROFILE

Yana Shintya S. Kom received a Bachelor Degree in Computer Science from Gunadarma University, Depok, Department of Information System in 2016. Currently continuing her master degree in Gunadarma University at Department of Business Information System.

Dr. Onny Marleen is a lecturer at the Faculty of Computer Science, Department of Information System Gunadarma, University, Depok. She has completed her Doctoral Program in 2012 at Gunadarma University.