Application of K-Means Algorithm for Costumer Grouping

Joanna Ardhyanti Mita Nugraha

***Abstract*—Sales fluctuations are risks that must be faced by business people, PT. Gunung Hijau Success experienced this in 2016, with GAP being quite high. By giving rewards to loyal customers, it is expected to stabilize sales in the next period. So the company needs customer grouping based on customer loyalty to reward. The application of data mining can be used as an analysis to determine the loyal customer inventory according to the total purchase. In the data mining method, the clustering algorithm is one of the most popular to use where the data belonging to the same cluster will be close to each other and will be far from the data about another cluster. The results obtained in the form of customer information with criteria not loyal, loyal, and very loyal based on sales data in 2016. Also, customer criteria information from the clustering process can be used as a reference to determine the reward for customers.**

***Index Terms*—Clusttering, k-means algoritm, customer loyalty, marketing.**

I. Introduction

Clustering groups datasets into several small data groups called clusters. This is done in such a way that the similarity of data groups in one cluster is high, and the similarity between groups of data between clusters is low. In addition, clustering is the primary task of various fields, such as artificial intelligence, machine learning, and pattern recognition [1].

Clustering has been used in various domains with excellent performance in recent years. Clustering has been used in grouping energy loads to analyze load profiles on warships. Clustering is also proposed to improve wind power operations by classifying classifications on wind power loads [2]. Clustering using the k-means algorithm is also used in research to analyze aircraft fuel consumption. In this study, the application of clustering can reduce fuel consumption by an average of 19.3% per unit of time [2].

PT. Gunung Hijau Sukses or often also called PT. GHS is a company engaged in the production of snacks. Established since May 1, 2009, and located on Jl. Slamet Riyadi 234 Gumpang, Kartasura, Central Java, with a home industry scale and using the name "Raja Rasa." The development of various sectors made "Raja Rasa" change its name to PT. Gunung Hijau Sukses with a larger scale of production. The location of the factory and the PT office. Gunung Hijau Sukses moved on Jl. Indronoto No.8 Ngabeyan - Kartasura - Central Java.

Manuscript received August 9, 2019; revised November 2, 2019.

Joanna Ardhyanti Mita Nugraha is with the Department of Informatics Engineering, Universitas Atma Jaya Yogyakarta, Babarsari Street No. 43, Yogyakarta, Indonesia (e-mail: joanna.mita@uajy.ac.id).

Based on the sales recap report in 2016 as shown in Table I, the sale of PT. Gunung Hijau Sukses fluctuations with a reasonably large GAP. This causes the company's operational activities to be hampered. To overcome this, the company took steps to reward loyal customers. It is expected that with a reward for loyal customers, it will have an impact on customers who usually make small purchases to increase because of the award, the customer will get a more significant profit. For that company need customer grouping based on the criteria of customer loyalty.

TABLE I: Recap Sales 2016

Month	Sales
January 2016	14,866,713.45
February 2016	47,509,709.42
March 2016	29,761,991.60
April 2016	73,368,541.49
May 2016	28,257,298.21
June 2016	87,237,701.01
July 2016	61,020,248.30
August 2016	27,467,440.73
September 2016	31,949,599.83
October 2016	14,664,717.41
November 2016	36,642,347.69
December 2016	31,810,651.09

II. Literature Review

A. Clustering

Clustering included in the category of unsupervised learning, whose goal is to partition the data that does not have a label into the same group. Data belonging to the same cluster will be close to each other and will be far from the data about different groups [3]. Various distance criteria can be used to evaluate how close the data is.

There are three essential elements, namely proximity distance (similarity, difference or distance measure), function to evaluate the quality of grouping and the third is an algorithm used for the computational cluster.

In particular, similarity measurements are taken from a measure of proximity that has a large value when point 1 and point 2 are similar. Conversely, a measure of inequality (or means of distance) is a measure of proximity that gets a small value when point 1 and point 2 are similar. The function of evaluating the quality of grouping must be able to distinguish between functional grouping and lousy grouping. Thus, the algorithm used to calculate clusters is based on the optimization of the evaluation function [4].

Grouping problems can be classified as Euclidean and

Non-Euclidean. Euclidean size is based on the concept of Euclidean space, which is characterized by several dimensions and specific solid points. The average of two or more points can be evaluated in the Euclidean space, and the proximity size can be calculated according to the location of the points in the area. The three Euclidean measures that have been used for grouping in many domains are Euclidean distance, Manhattan distance, and Minkowski distance. Euclidean distance is defined in Equation (1).

$$d(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \quad (1)$$

B. K-Means Algorithm

K-means clustering is the most widely used grouping method based on data partitions. The main idea is to collect original data into cluster k so that samples with the same attributes are in the same cluster. The central processing procedures are as follows: (1) Select sample k randomly from the original data. (2) Each sample is taken as a group center k. (3) Calculate the distance between samples using the Euclidean distance formula. (4) The center k sample is calculated separately, and each sample is divided into the nearest center. (5) Clusters that are sampled are clusters from the central sample. (6) Iterations are repeated until the sample group no longer changes. Square error is usually used as a criterion for convergence of functions [5], [6].

III. Methodology

A. Understanding of Data

In this phase collecting and studying data can be used to analyze what can be done on the data to be processed. In addition to the data understanding phase, there is a stage in evaluating data quality and completeness of data [7], [8]. Missing values often occur, especially if data is collected over a long period of time. Check for missing or empty attributes, spelling values, and whether attributes with different values have the same meaning.

B. Data Preparation

In the data preparation phase, data sets are set up, which will be ready to be processed using modeling tools by selecting the required attributes. The attributes that will be used are taken from the sales recap report, namely, customer ID, customer name, repeat orders, and total purchases. Where later, with repeat orders and high total purchases, indicates that the customer is a loyal customer and vice versa. The selection of the new attribute will be stored in the .xls file, which will be ready to be included in the modeling tool.

The sales report recapitulation data that will be used is sales data in 2016 and will be divided into two datasets, namely semester one dataset (Data period for January 2016 to June 2016) and semester two dataset (Data period for July 2016 to December 2016). So that of each half will be produced three clusters, clusters are not loyal customers, loyal and very loyal.

IV. Result

The stages of the clustering method use the K-means algorithm:

1. Determination of the Number of Clusters

In this study, the data will be divided into 3 clusters, namely cluster 0, cluster 1, and cluster 2.

2. Determination of The Initial Cluster Center

The decision of the center of the initial cluster (centroid) obtained itself not by specifying a new point that is with the initial central dimension of the data as shown in Table II.

TABLE II: The Initial Centroid of Each Cluster

Cluster Center	Name Customer	Repeat Order	Total Purchases
Cluster 0	Dewa Surya, Tk	3	24,754,699
Cluster 1	Barokah, Tk - Blitar	3	42,729,530
Cluster 2	Setia 5 Jaya	12	241,466,352

3. Calculation of Distance With The Center of The Cluster

Calculation of distance using Equation (1).

Take data values and center values the cluster then calculates using the Euclidian Distance formula with each cluster center. For example, the distance of the first customer data will be calculated with cluster 0.

$$d(1,0) = \sqrt{(6-3)^2 + (36{,}881{,}767 - 24{,}754{,}699)^2}$$
$$= 12{,}127{,}067.84$$

From the calculation results, the result is that the distance of the first customer with cluster 0 is 12,127,067.84. Then the distance of the first customer will be calculated with cluster 1 with the equation.

$$d(1,1) = \sqrt{(6-3)^2 + (36{,}881{,}767 - 42{,}729{,}530)^2}$$
$$= 5{,}847{,}762.76$$

From the calculation results, the results show that the distance of the first customer with cluster 1 is 5,847,762.76. Then the distance of the first customer will be calculated with cluster 2 with the equation.

$$d(1,2) = \sqrt{(6-12)^2 + (36{,}881{,}767 - 241{,}466{,}352)^2}$$
$$= 204{,}584{,}585.3$$

Based on the calculation of the distance from the first customer with cluster 0 and the first customer with cluster 1 shows that the closest distance to the center of the cluster is the first customer. So that the first customer is in cluster 0. The calculation will continue until the last data. So that it will be known the distance of each data with the nearest cluster center.

Based on the results of the calculation of the distance from customer 1 with each cluster, the shortest distance is created with cluster 1. So that customer 1 enters cluster 1. Forecasts will continue to be carried out on the latest data.

4. Data Grouping

The distance from the calculation will be done and the closest distance between the data and the cluster center, this distance indicates that the data is in one group with the nearest cluster center as shown at Table III.

TABLE III: DATA DISTANCE AND THE CLOSEST DISTANCE TO THE CENTER OF THE CLUSTER

Name	RO	Total Purchases	Distance to			Closest Distance to Cluster
			Cluster0	Cluster1	Cluster2	
Ali Jaya,Tk	6	36,881,767	12,127,067.8	5,847,762.7	204,584,585.2	cluster _1
Barokah,Tk - Blitar	3	42,729,530	17,974,830.6	0.0	198,736,822.5	cluster _1
Dewa Surya,Tk	3	24,754,699	0.00	17,974,830.6	216,711,653.1	cluster _0
Ganesha, Tk	9	82,041,022	57,286,322.9	39,311,492.3	159,425,330.1	cluster _1
Hari,UD	3	6,218,508	18,536,190.8	36,511,021.4	235,247,843.9	cluster _0
Jaya Mulya, Tk	4	40,949,190	16,194,490.5	1,780,340.0	200,517,162.5	cluster _1
Setia 5 jaya	12	241,466,352	216,711,653.1	198,736,822.5	0.0	cluster _2
Marjono	3	16,180,581	8,574,118.1	26,548,948.7	225,285,771.2	cluster _0
Sukiman	14	280,973,136	256,218,436.6	238,243,606.0	39,506,783.5	cluster _2
MAKMUR,TK(SMRANG)	2	1,254,018	23,500,680.7	41,475,511.3	97,891,436.2	cluster _0

5. Determination of New Cluster Centers

To get a new cluster center can be calculated from the average value of members cluster and the cluster center. The new cluster center is used to do the next iteration if the results obtained have not been converged. Example calculation of the new cluster center in cluster 0 is by looking at data that has the closest distance to cluster 0 or data included in cluster 0 divided by the amount of data entered in cluster 0, for example:

$$\textbf{Cluster 0} = \left(\frac{3+3+3}{3}; \frac{2{,}475{,}699 + 6{,}218{,}508 + 16{,}180{,}581}{3}\right)$$

$$= (3 ; 8{,}291{,}596)$$

From the calculation results, the new centroid is obtained in cluster 0, namely (3; 8,291,591). Then also calculated the new centroid in cluster 1, namely:

$$\textbf{Cluster 1} = \left(\frac{6+3+4+14}{4}; \frac{36{,}881{,}767 + 42{,}729{,}530 + 40{,}949{,}190 + 280{,}973{,}136}{4}\right)$$

$$= (6.75 ; 97{,}883{,}405.75)$$

And the new centroid in cluster 1 is (6.75 ; 97,883,405.75). Then also calculated the new centroid in cluster 2, namely:

$$\textbf{Cluster 2} = \left(\frac{9+12+2}{3}; \frac{82{,}041{,}022 + 241{,}466{,}352 + 1{,}254{,}018}{3}\right)$$

$$= (7.6 ; 201{,}493{,}503.3)$$

And the new centroid in cluster 2 namely (7.6; 201,493,503.3). So that in the next calculation, the centroid used is the new centroid as shown in Table IV.

TABLE IV: THE INITIAL CENTROID OF EACH CLUSTER

Attribute	Cluster 0	Cluster 1	Cluster 2
Repeat Order	3	6.75	7.6
Purchases	8,291,591	97,883,405.75	201,493,503.3

The iteration will be done again to find out whether the data is moved or not. Calculations will be carried out like the second stage, which is knowing the distance of data with each cluster using the new centroid.

1) Apply to model tools

The modeling tool used is Rapidminer 5.3, which can be used to facilitate the calculation of the k-means algorithm and C4.5 decision tree algorithm. Also, Rapidminer can also calculate the accuracy of the data that has been processed. The clustering process will divide the data into 3 clusters. Division 3 clusters can be seen in Fig. 1, which is circled in red. K = 3 is the division of clusters into 3, according to the expected process. After the division of clusters, the process is executed by selecting the Run button on the taskbar.

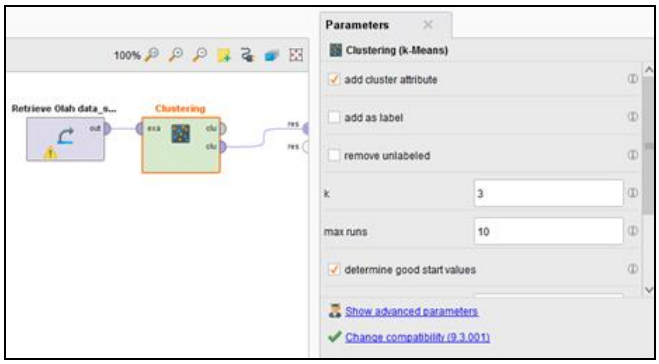

Fig. 1. Clustering process k-means using Rapidminer.

Results obtained from the clustering process in semester one has divided 3 clusters are cluster 0, cluster 1, and cluster 2, which are worth 878, 145 and 7 respectively as shown in Fig. 2. After the results are obtained, the data is analyzed based on the attributes used. So that the result is that cluster 0 is a customer not loyal, cluster 1 is a customer loyal, and cluster 2 are very loyal customers.

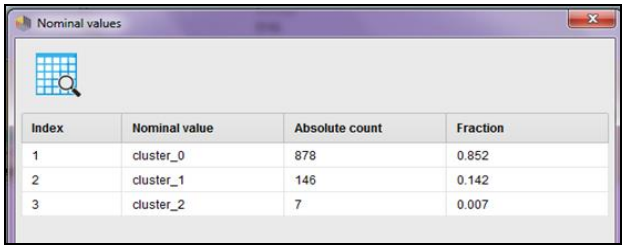

Fig. 2. Results of the clustering process k-means use Rapidminer.

2) Evaluation

In the evaluation phase, the model will be assessed

whether the results obtained from the clustering process have fulfilled the stated objectives in the business understanding stage.

At the stage of understanding the business has been determined, the goal is to find out customer loyalty to be used by PT. Gunung Hijau Sukses for giving rewards to customers as a means of increasing sales. And the results of the clustering process obtained 3 clusters, namely non-loyal customers, loyal customers, and very loyal customers. After the results obtained are the same as those aimed at understanding the business, a checking process will be carried out which serves to ensure that all stages have been carried out in the data processing processor that no critical factors have been missed.

Furthermore, it will be ensured that all the critical stages/factors that have been done by processing the data have not been missed. Thus the next process can be carried out in the data processing because it has fulfilled the purpose of clustering.

At the evaluation stage, the performance of the clustering process will be evaluated using cross-validation, and the results show the results of -0.449, which are quite good results in the application of clustering as shown in Fig. 3.

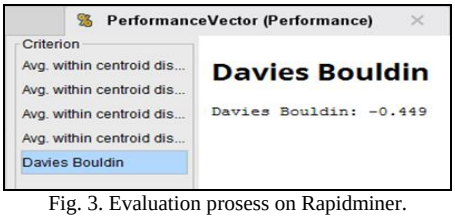

Fig. 3. Evaluation prosess on Rapidminer.

3) Deployment

This phase is the phase of the application of clustering methods that have been formulated at first. The goal achieved is to be able to find out information on the types of customers who are not loyal, loyal, and very loyal based on the data obtained in the sales report. So that later, with the information collected from this clustering method can be used by the PT. Gunung Hijau Sukses as a reference in determining the reward for customers for future purchases.

Conflict of Interest

The author declares no conflict of interest.

References

[1] Q. Liu, R. Zhang, R. Hu, G. Wang, Z. Wang, and Z. Zhao, "An improved path-based clustering algorithm," *Knowledge-Based Syst.*, vol. 163, pp. 69–81, 2019.

[2] Y. Du, B. Sun, R. Lu, C. Zhang, and H. Wu, "A method for detecting high-frequency oscillations using semi-supervised k-means and mean shift clustering," *Neurocomputing*, vol. 350, pp. 102–107, 2019.

[3] Q. Zhu, J. Pei, X. Liu, and Z. Zhou, "Analyzing commercial aircraft fuel consumption during descent: A case study using an improved K-means clustering algorithm," *J. Clean. Prod.*, vol. 223, pp. 869–882, 2019.

[4] A. Amelio and A. Tagarelli, "Data mining: Clustering," *Encycl. Bioinforma. Comput. Biol.*, pp. 437–448, 2018.

[5] J. Melton *et al.*, *Data Mining: Concepts and Techniques*, 1999.

[6] K. B. Chimwayi and J. Anuradha, "Clustering West Nile Virus spatio-temporal data using ST-DBSCAN," *Procedia Comput. Sci.*, vol. 132, no. Iccids, pp. 1218–1227, 2018.

[7] H. Yu, G. Wen, J. Gan, W. Zheng, and C. Lei, "Self-paced learning for k-means clustering algorithm," *Pattern Recognition Letters*, 2018.

[8] A. N. Khormarudin, "Teknik data mining : Algoritma k-means clustering," *IlmuKomputer.Com* , pp. 1–12, 2016.

Joanna Ardhyanti Mita Nugraha was born in 1991 and received her master of computer in information system from Universitas Atma Jaya Yogyakarta, Indonesia in 2017. Her area of interest includes artificial intelligence and data mining. She is a lecturer of Information Engineering Department in Universitas Atma Jaya Yogyakarta since 2018.