Jurnal TAM (Technology Acceptance Model)
Jurnal TAM, Volume 14, Number 1, July 2023
E ISSN: 2579-4221; P ISSN: 2339-1103, pp. 91-97
Accredited SINTA 4 Number 225/E/KPT/2022
https://jurnal.ftikomibn.ac.id/index.php/JurnalTam/index

PRODUCT CLUSTERING USING K-MEANS METHOD IN CV. JAYA ABADI

Ika Ramadhaniati
Study Program in Informatics Engineering at STMIK Dharma Wacana, Metro
Kenanga Street, No.3 Mulyojati, Metro West, Metro City
E-mail: Ikaramadhaniati99@gmail.com

Article history:
Received: February 7, 2023
Revised: July 4, 2023
Accepted: July 7, 2023

Corresponding authors
Ikaramadhaniati99@gmail.com

Keywords:
Sale;
Stock of goods;
Clustering;
K-means.

Abstract
CV. Jaya Abadi is a company engaged in the distribution of goods, or, to be more precise, a distributor of raw materials for making bread. The fluctuating number of requests from consumers results in stock that must be prepared to be unstable. In addition, many types of products make stock management inaccurate. Sometimes we do not want to have a shortage of stock, goods, or certain products when consumer demand is high. The purpose of this research is to analyse sales data mining through the clustering method with the k-means algorithm. The method used for this analysis uses clustering with the k-means algorithm. So that this research is not subjective, the authors also use research methods in the form of observation, interviews, and documentation. The results of this study were carried out by analysing sales by applying data mining through the clustering method with the k-means algorithm. Data mining that has been going on for a long time can be used as a reference in the management of CV Jaya Abadi. Apart from being used for administrative purposes each period, it can also be used as a decision-making solution in order to maximize existing products and sales at CV Jaya Abadi.

1. INTRODUCTION

CV Jaya Abadi is a company engaged in the distribution of goods, or, to be more precise, a distributor of raw materials for making bread. CV Jaya Abadi has branches in Bandar Lampung City and Metro City. CV Jaya Abadi is a company engaged in the field of product sales or distribution services. Distribution is carried out so that its use is in accordance with what is needed (type, quantity, price, place, and when needed). Distributor companies are intermediaries that distribute products from manufacturers to retailers. After a product is produced by the factory, it is sent (and usually sold at the same time) to a distributor. The distributor then sells the product to retailers or customers, who always want success in their future activities.

Based on the results of the interviews conducted, the fluctuating number of requests from consumers resulted in stock that had to be prepared to be unstable. In addition, the various types of products make stock management inaccurate. Sometimes, because we don't want shortages to occur, stocks of certain goods or products are held when consumer demand is high. In addition, inaccurate stock management also results in frequent shortages or excesses of certain products, which ultimately disappoint consumers. The problems that occur in CV Jaya Abadi are that CV Jaya Abadi has difficulty determining the minimum stock of each item that must be met based on consumer interest.

One of the data mining analysis techniques is cluster analysis, better known as clustering. Clustering is a data analysis method whose goal is to group data with the same characteristics into the same area. One of the approaches used in developing the clustering method is the K-Means method, where this method is a method of grouping non-hierarchical data (blocks) that seeks to partition data into the form of

two or more groups (clusters) with the same characteristics, put into one group [1][2]–[4].

In finding a solution to the problem above, the researcher applies data mining through the clustering method with k-means algorithm. In the research conducted, the application of the clustering method with the k-means algorithm was applied to the number of existing sales products at CV Jaya Abadi. By applying the k-means algorithm, you can group data based on distribution patterns, and you can partition existing data into two or more groups by finding interesting relationships between data attributes [5]. Data Mining is the process of extracting information from data sets through the use of algorithms and techniques that involve the fields of statistics, machine learning, and databases.[6], [7] [8]. The use of data mining can help in knowing the goods that are often in demand by customers. Data mining that has been going on for a long time can be used as a reference in the management of CV Jaya Abadi. Besides being used for administrative purposes every period, this can also be used as a decision-making solution in order to maximize existing products and sales at CV Jaya Abadi.

II. RESEARCH METHODS

2.1. Business Understanding

Business Understanding is the stage where the business problem is defined simply and precisely. In relation to this research, the problems experienced by the company are stock instability and the frequent occurrence of vacancies in the warehouse of CV Jaya Abadi. This can result in a decrease in purchase orders from customers and a decrease in turnover at CV Jaya Abadi. Therefore, it is necessary to determine the stock of goods that must be considered so that there is no shortage of goods.

2.2. Data Understanding

Data Understanding is a process where we bring together what data we have and what data we need. It could be that a data analysis project starts with the discovery of existing data, which then directs the analyst to explore the existing knowledge in the data set. The type of data determines the type of algorithm and the goals of data mining to be achieved.

2.3. Data Preparation

The data preparation process is a data treatment process towards a useful quality model. This stage is the one that drains the most resources from the analysis team.[2]–[4], [6], [7] A good and accurate model starts with good data preparation. Some common things to do at this stage are:

1. Re-check the correctness of the data.
   Checking data needs to be done in stages so that accountability can be maintained. Checking is also needed for the consistency of entering data. A good system for collecting data is to use defaults, which can maintain data consistency.
2. Manage data outliers.
   Outlier data needs to be managed properly. Outlier data can be in the form of Univariate Outliers and Multivariate Outliers and can be in the dependent variable or independent variable. The purpose of Data Mining in general will be affected by Outlier data, so it needs to be neutralized. Before doing treatment on Outlier data, it would be nice to first check the data collection and filling.
3. Enact missing and inconsistent data.
   The treatment of missing data must match the objectives of the data mining itself. For example, missing data filled with averages may still be acceptable for prediction and forecasting purposes, but for clustering, it may lead to inaccurate groups. On the other hand, using frequently occurring data to fill in missing data for multi-variable data mining has an effect on results for prediction and forecasting purposes.

2.4. Data Transforms

Data Transformation is an effort made with the main objective of changing the measurement scale of the original data into another form so that the data can meet the assumptions underlying the analysis of variance. The initial sales transaction data needs to be selected to group the attributes according to the information needed from the initial attribute data, the number of items sold and the number of transactions.

2.5 Determining the Best Cluster

1. WCSS

Calculation Evaluation is the result of calculating the amount of data that meets the combination divided by the total number of data multiplied by 100% (Fitriati, 2016). To find accuracy, calculate the Within Cluster Sum of Squares (WCSS) for each cluster value. Because the greater the number of K cluster values, the smaller the WCSS value. The WCSS formula is as follows:

$$WCSS = \sum_{pi\ in\ cluster\ 1}^{k} jarak\ (P_i C_1)^2 + \sum_{pi\ in\ cluster\ 2}^{k} jarak\ (P_i C_1)^2 + \ldots. (n) \ldots\ldots\ldots (1)$$

In this study, the application of data mining uses the elbow method and the k-means cluster technique to find information about product data and sales transactions. The elbow method is implemented by looking at the graph of the K value to be input. The value of the K function that is compared in the elbow method is determined by looking at the WCSS (Within Cluster Sum of Squares) value at the specified cluster value. The results of the best number of K clusters will be used as a basis for carrying out the clustering process using the K-Means method in a case study.

2. Elbow method

The Elbow method is a method used to generate information for determining the best number of clusters by looking at the percentage of the results of the comparison among the number of clusters that make up the elbow at a point. To overcome this, the authors conducted research by finding the best K value using the elbow method. This method has been used for a long time. It looks at the function of the cluster values in a set of data. This method provides ideas by selecting cluster values and then adding the cluster values to be used as a data model in determining the best cluster. And apart from that, the percentage of calculations produced is a comparison among the number of clusters added. The results of different percentages of each cluster value can be shown using graphics as a source of information. If the first cluster value with the second cluster value gives a corner in the graph or the value has the greatest decrease, then the cluster value is the best value.

2.6. Model K-Means

K-Means algorithm is one of the partition algorithm because K-Means is based on determining the initial number of groups by defining the initial centroid value. K-Means algorithm uses an iterative process to get a cluster database. It takes the desired number of initial clusters as input and produces the final number of clusters as output. If the algorithm is required to generate K clusters, then there will be K starts and K ends. K-Means method randomly selects the K pattern as the centroid starting point. The number of iterations needed to reach the cluster centroid is influenced by the initial cluster centroid candidate, if the new centroid position does not change. K value is calculated using the Euclidean Distance formula, which is to find the shortest distance between the centroid point and the data or object. Data that has the shortest or closest distance to the centroid will form a cluster.

K-Means Algorithm

1. Determine K as the number of clusters to be formed.
2. Determine the initial K Centroid randomly.

$$v = \frac{\sum_{i-1}^{n} x_i}{n} ; i = 1,2,3, \ldots . n \ \ldots\ldots\ldots\ldots\ldots\ldots\ldots (2)$$

Where:
v: Centroids in the cluster
xi: Object that is in i-order
n: The number of objects that are members of the cluster

3. Calculate the distance of each object to the centroid of each cluster. To calculate the distance between objects and centroids, use Euclidian Distance.

$$d(x,y) = |x - y| = \sqrt{\sum_{i-1}^{n}(x_i - y_i)} ; 1,2,3, \ldots n \ \ldots\ldots\ldots (3)$$

Where:
x_i: Object that is in i-order
y_i: Data that is in i-order
n: The number of objects

4. Allocate each object to the nearest centroid.
5. Do iterations, then determine the position of the new centroid using the equation.
6. Repeat step 3 if the new centroid position is not the same.

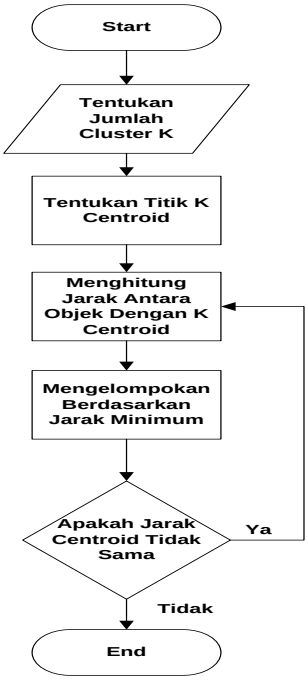

Figure 1. Flowchart of the K-Means Algorithm

IV. RESULTS

3.1. Business Understanding

CV Jaya Abadi is a company engaged in the distribution of goods, or, to be more precise, a distributor of raw materials for making bread. With the increase in competitors in every area of the company, both local and international, it is one of the factors that makes the company must be able to survive in product distribution, where in the process it still has constraints that have not been able to determine which products should prioritize its stock. As a result of this, there is often a shortage of product stock that customers are interested in. Unstable stock and vacancies in the warehouse can result in decreased demand from customers. Therefore, it is necessary to determine the stock of goods that must be considered. To overcome this problem, a system is needed that is able to determine the products that are most in demand or those that are frequently sold. Prioritized product groupings are calculated using the K-Means method, which is able to calculate them accurately and consistently.

3.2. Data Understanding

1. Data Types and Sources

The type of data used is quantitative. Quantitative data is a type of data that can be calculated in the form of numbers or nominal. Sale transaction historical data is a type of quantitative data because it is in the form of numbers or nominal and can be calculated. More specifically, the data used is in the form of matrix data, data types that have objects and attributes. At this stage, researcher needs to master an understanding of the types of quantitative data. The types of data used on the attributes in this study are:

Table 1. Types of Data

Attributes	Types of Data	Description
Code	Nominal	The code is the identity of the product being sold.
Name	Nominal	The name is the name of the product being sold.
Amount	Numeric	The amount is the number of sales of goods.
Unit	Numeric	The unit is the unit of goods sold.
Gross	Numeric	Gross is the identity of products sold wholesale.
Disc	Numeric	Disc is a product discount bonus.
Net	Numeric	Net is the weight of the product sold.
COGS	Numeric	COGS is the sales production price or product capital price.
Profit	Numeric	Profit is the profit from the product sold.
Price	Numeric	The price is the selling price of the product.
Tax	Numeric	Tax is a tax expense borne from the sales price.
Total Bill	Numeric	The total invoice is the total selling price of goods.

3.3. Data Preparation

At this stage, the researcher prepares the data for testing. The data used was obtained from CV Jaya Abadi, which was then transformed using a clustering process. Based on the data that has been transformed into two features, the number of goods sold is the result of sales of goods, and the number of transactions is the total of one incoming item that is managed in sales transactions. The data used is 342 data points, with the number of transactions made as high as 320. The following results of the transformation can be seen in Table 2.

Table 2. Transformation Data

Product Name	Number of Items Sold	Number of Transactions
168 Compound @24x250gr	44	20
African Black 10-12% (2x5kg)	47	39
African Black Powder Fat 10 - 12 % (10x1kg)	50	76
African Red, Fat 20-24% (2x5kg)	30	20
African Red Powder Fat 20% - 24% (10 X 1kg)	40	31
Verlin Salted Caramel Powder @12x1kg	45	16
Verlin Taro Powder @12x1kg	44	15
Verlin Thai Tea @12x1kg	14	3
Verlin Tiramisu Powder @12x1kg	12	4
Verlin Vanilla Powder @12x1kg	16	6

3.4. Model

1. Best Cluster Determination

The determination of the best cluster uses the Within Cluster Sum of Squares (WCSS) as a distance calculator for each data point to the centroid. This was done in ten iterations with values of K = 1 to K = 10. The results of the within-cluster sum of squares can be seen in Table 3. The resulting values from WCSS are visualized in a line graph. In the elbow method, the point that forms the elbow is the optimal cluster value. In the implementation of the elbow method, the optimal cluster value is obtained at K = 3. The results of determining the best cluster can be seen in Table 3.

Table 3. Results of the Within Cluster Sum of Squares (WCSS)

Cluster	Results of Within Cluster Sum of Squares (WCSS)
K1	17281106.27615064
K2	3959812.8017094033
K3	1186125.9655555561
K4	852456.2155555557
K5	544874.0117370893
K6	279637.0341425563
K7	195442.86790725007
K8	144254.0345739167
K9	122996.44685737981
K10	99635.02134428993

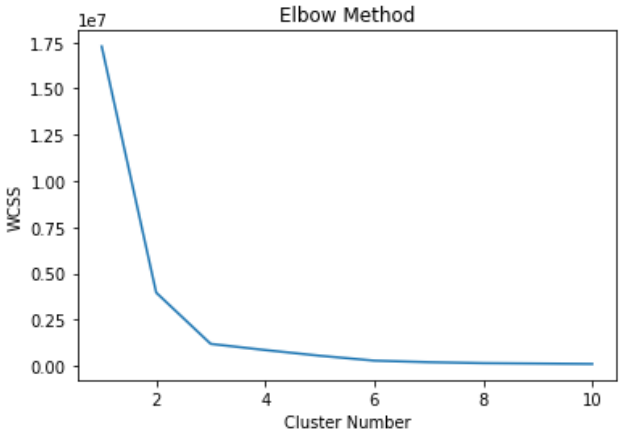

Figure 2. Elbow Method Graph

2. K-Means

In the application of K-Means, it uses the results of the preparation process with a total of 238 data points with two features: the number of goods sold and the number of transactions. The data is applied to the K-Means clustering algorithm for grouping data using the parameter value k = 3. The implementation of the K-Means algorithm uses *Scikit-learn* library in the Python programming language. Table 4 is the result of grouping the data into three groups. In cluster 0, there are 234 items sold and 20 transactions. In cluster 1, there are 4 data points, and in cluster 2, there are 0 data points. The results of K-Means clustering can be seen in Table 4.

Table 4. Cluster Results

No	Product Name	Number of Items Sold	Number of Transactions	Cluster
0	168 Compound @24x250gr	44	20	0
1	African Black 10-12% (2x5kg)	47	39	0
2	African Black Powder Fat 10 - 12 % (10x1kg)	50	76	0
3	African Red, Fat 20-24% (2x5kg)	30	20	0
4	African Red Powder Fat 20% - 24% (10 X 1kg)	40	31	0
5	Alpha Compound Medium @24x250gr	8	3	0
..	…	…	…	…
236	Verlin Thai Tea @12x1kg	14	3	0
237	Verlin Tiramisu Powder @12x1kg	12	4	0
238	Verlin Vannila Powder @12x1kg	16	6	0

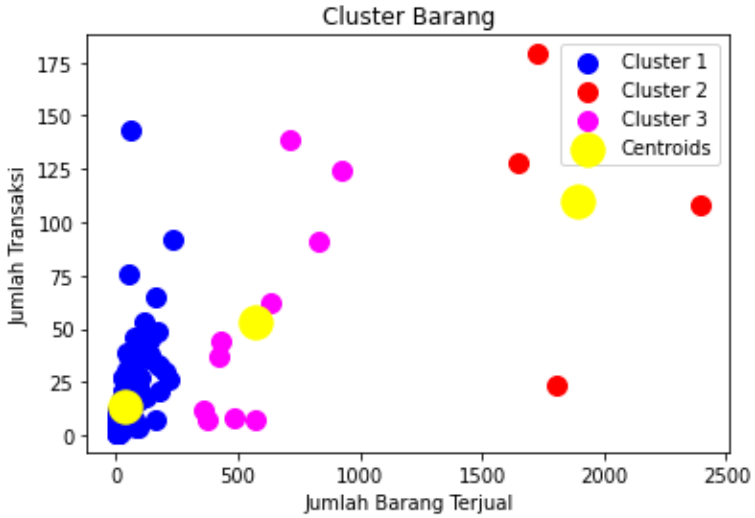

Figure 2. Cluster Results

3 Cluster Analysis

Based on the results of the data obtained, the following is a cluster analysis table to find out items that need priority and items that sell a lot. Data entered in clusters 0 and 2 has fewer transactions and sales volume compared to data entered in cluster 1, where there are more transactions and sales volume. So, the goods that must be prioritized and paid attention to are the goods that are included in cluster 1. Goods with a large number of sales and the number of transactions can be seen in Table 5.

Table 5. Cluster Analysis

Product Name	Qty Sold In Crt	Number of Transactions	Cluster
Amanda Mgr Krim 15kg	1647	128	1
Bimoli B.I.B	2394	108	1
Dunia coklat butir 12kg	1726	179	1
Nesta Irot Coklat @ 10 Kg	1808	23	1

V. CONCLUSION

The K-Means algorithm is used to classify data using two features: the number of sales and the number of transactions. Determining values for parameters in K-Means uses the elbow method by calculating the Within Cluster Sum of Squares (WCSS). The result obtained from the elbow method is K=3. The application of the K-Mean method with a parameter value of K = 3 produces three data clusters. Cluster 0 contains 230 data, Cluster 1 contains four data, and Cluster 2 contains four data. The results of grouping data on sales of goods with data entered in clusters 0 and 2 show fewer transactions and sales volume compared to data entered in cluster 1, where there are more transactions and sales volume. So, the goods that must be prioritized and paid attention to are the data on the goods that are included in cluster 1, the goods that have sold the most, and the number of transactions. The results of this grouping can be used for consideration in determining the priority of determining the stock of goods.

REFERENCES

[1] Y. Darmi and A. Setiawan, "PENERAPAN METODE CLUSTERING K-MEANS DALAM PENGELOMPOKAN PENJUALAN PRODUK," *J. Media Infotama*, vol. 12, no. 2, pp. 148–157, 2016.

[2] S. Mukodimah, M. Muslihudin, D. R. Mustofa, and D. Susianto, "Naive Bayes Classifier Method Analysis and Support Vector Machine (SVM) Student Graduation Prediction," *NEUROQUANTOLOGY*, vol. 20, no. 12, pp. 3522–3533, 2022.

[3] L. J. Anreaja, N. N. Harefa, J. G. P. Negara, V. N. H. Pribyantara, and A. B. Prasetyo, "Naive Bayes and Support Vector Machine Algorithm for Sentiment Analysis Opensea Mobile Application Users in Indonesia," *JISA(Jurnal Inform. dan Sains)*, vol. 5, no. 1, pp. 62–68, 2022.

[4] R. Syahputra, G. J. Yanris, and D. Irmayani, "SVM and Naïve Bayes Algorithm Comparison for User Sentiment Analysis on Twitter," *Sinkron*, vol. 7, no. 2, pp. 671–678, 2022.

[5] M. Dahria, R. Gunawan, and Z. Lubis, "Implementasi K-Means Untuk Pengelompokan Produk Terbaik PT . Koko Pelli," *Semin. Nas. Sains Teknol. Inf.*, pp.

495–498, 2019.

[6] A. Srivastava, E.-H. S. H. E.-H. S. Han, V. Singh, and V. Kumar, "Parallel formulations of decision-tree classification algorithms," *Proceedings. 1998 Int. Conf. Parallel Process. (Cat. No.98EX205)*, vol. 24, pp. 1–24, 1998.

[7] D. A. C, D. A. Baskoro, L. Ambarwati, and I. W. S. Wicaksana, *Belajar Data Mining dengan RapidMiner*. 2013.

[8] J. Han and Kamber, *Data Mining Concepts and Techniques*. San Francisco: Morgan, 2006.