International Journal of Informatics and Information System ISSN 2579-7069
Vol. 3, No. 1, March 2020, pp. 12-22

Analysis of Data Mining Using K-Means Clustering Algorithm for Product Grouping

Mohammad Imron [a,*], Uswatun Hasanah [a], Bahrul Humaidi [a]

[a] Information System Program, Faculty of Computer Science, Universitas Amikom Purwokerto
* corresponding author

Abstract

Rizki Barokah Store is one of the stores that every day sell a variety of basic materials of daily necessities such as food, drinks, snacks, toiletries, and so on. However, some problems occur in the Rizki Barokah Store is often a build-up of product stocks that resulted in the product has expired. This is due to an error in making decisions on the product stock. In addition to these problems, with the amount of sales data stored on the database, the store has not done data mining and grouping to know the potential of the product. Whereas data-processing technology can already be done using data mining techniques. To overcome the period of the land, the technique used in data mining with the clustering method using the algorithm K-means. With the use of these techniques, the purpose of this research is to grouping products based on products of interest and less interest, advise on the stock of products, and know the products of interest and less demand.

Keywords: Data mining; K-Means Algorithm; Clustering; Stock.

1. Introduction

The Rizki Barokah Store is a store that is located in Jalan Brigadir 17 No. 47 RT 001 RW 001 Rempoah Village Baturraden District. In its activities, the Rizki Barokah Store sells various basic materials of daily necessities such as food, beverages, snacks, toiletries, and so on. The store was established in February 2018. Because the store is classified into a large store, then the longer a shop stands then, the larger the data owned by the store. Based on the observed results, the number of products sold in November amounted to 1127, then in December 1812, and in January of 2075 products. However, with the amount of sales data stored in the database, the data is still raw data that has not produced useful information.

In sales activities carried out, there is a common problem that is the accumulation of stock products expired. The activity of supplying product stock in the Rizki Barokah Store by using distributors and other stores. The stock of the accumulated product occurs because several expired products cannot be returned to the distributor. Mistakes in determining the product 's stock can be detrimental to the shop because the expired items do not make a profit on the shop. Based on the data contained in the store is known to have expired product as much as 426 from August 2018 to March 2019.

Sales activities carried out by the Rizki Barokah Store at this time not yet utilize their sales data to explore information about the potential of the product, whereas data processing technology can already be done using data mining techniques[1][2]. Data mining is a process to explore useful and previously unknown information from a large stack of data[3]. One of the methods that can be used in the business world is data mining with a clustering technique using the K-means algorithm[4][5].

This research method used is data mining with a clustering technique using the K-means algorithm for product grouping based on the most desirable and less desirable products to determine the stock of products in the Rizki Barokah Store. K-means algorithm is an iterative grouping algorithm that performs the partition of the data set into a number of the preset k clusters[6]. The K-means algorithm is simple to implement and run, comparatively fast, adaptable, common usage in practice. Historically, K-means became one of the most important algorithms in data mining [7].

2. Literature Review

2.1. Stock

According to [8], stock of goods is as an activity that covers the goods of the owner of the organization to be sold at a time or certain business period or stock of goods that are still in the process of production or supplies of raw materials waiting for their use in the production process. Meanwhile, the supply of goods is defined as goods acquired by the company for resale or further processing in order to carry out the company's activities [9][10]. Companies that can precisely control their system of availability will facilitate the company to survive operational activities and maintain the smooth operation of the company. Therefore, the supply of goods is important, because the success of the planning and supervision of supplies will have a significant impact on the success of a company, one of them on the company's profit determination[11].

2.2. Data Mining

Data mining is a process of hiring one or more computer learning techniques (machine learning) to analyze and extract knowledge automatically [14]. Data mining can be defined as induction-based learning, which is a process of forming the definitions of general concepts done by observing specific examples of concepts to be learned. Knowledge Discovery in Database (KDD) is the implementation of scientific methods on data mining[12]. In this context, data mining is a step of the KDD. The process of data Mining that implements the Knowledge Discovery in Databases (KDD) process is in Figure 1 as follows:

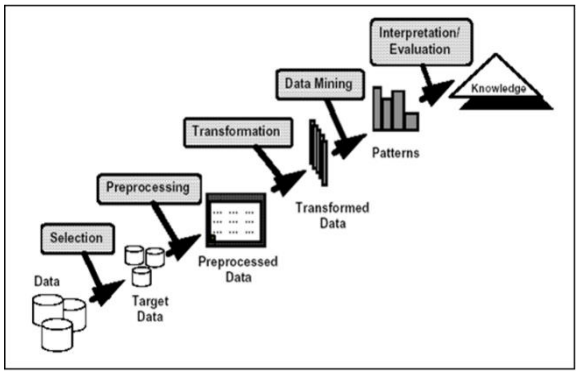

Fig 1. Knowledge Discovery In Databases (KDD) Process

Figure 1 is Knowledge Discovery in Databases (KDD) consisting of several processes such as:

a. Understand your app's domain to know and dig new knowledge and what your users are aiming for.

b. Create a target data set that includes picking data and focusing on sub-sets of data.

c. Data cleaning and transformation include noise elimination, outliers, missing values, and dimensional reduction and feature selection.

d. The use of data mining algorithms consisting of associations, sequences, classification, classifications, etc.

e. Interpretation, evaluation, and visualization of patterns to see if something is new and exciting and done if needed.

2.3. Clustering

Clustering is also known as segmentation. This method is used to identify the natural group of a case based on an attribute group, grouping data that has attribute resemblance [13][15].

2.4. K-Means Algorithm

Algoritma K-means to set the cluster values(k) randomly, for the meantime, the value becomes the center of the cluster or is commonly called centroid, mean or "means." Then the distance of each existing data against each centroid is calculated using the Euclidean formula so that it finds the closest distance on each data with a centroid. Then do the classification based on its proximity to the centroid. Do until the centroid value is not changed [15]. The steps of doing clustering with the method K-means [12] are as follows:

a. Select the number of clusters K.

b. Initiation k Cluster Center This can be done in various ways. However, the most often done is using random. Cluster centers are given an initial value with random numbers.

c. Place each record or object to the nearest cluster. The distance of both objects determines the proximity of two objects. Similarly, the proximity of a record to a particular cluster is determined the distance between the data with the cluster center. The closest distance between a single record and one specific cluster will determine which data is entered in which cluster. To calculate the distance of all the data to each cluster center point can use the distance theory of Euclidean distance formulated in Figure 1.2 as follows:

$$D(i,j) = \sqrt{(X_{1i} - X_{1j})^2 + (X_{2i} - X_{2j})^2 + \cdots + (X_{ki} - X_{kj})^2 \ldots}$$

Fig 2. Euclidean Distance Formula

Where:

D (ij) = data distance to I to the center of Cluster J

Xij = Data to I on attribute data to K

Xkj = center point to J on the attribute to K

d. Recalculate the cluster center with the current cluster membership. The cluster center is the average of all the data or objects in a given cluster. So you are also required to use the median of the cluster. So the average (mean) is not the only size that can be worn.

e. Redefine each object by using the new cluster center. If the cluster center is no longer changed, then the classifying process is complete. Alternatively, go back to step c until the cluster Center is not changed anymore.

3. Research Methods

The research methods used are as follows:

a. Research Time and Place

This research was held at Rizki Barokah Shop, which is located at Jl. Brigadir 17 No. 47 RT 001 RW 001 Rempoah Village Baturraden District. The research time starts from October 2018 to March 2019.

b. Data Collection Methods

The method of data collection used in this research is interviews, documentation, and observation to explore the existing problems and obtained data amounting to 1397 in November 2018 until January 2019.

c. Research concept

In this research, the stages of which are the identification of problems, data collection, stage preprocessing, the use of clustering methods, as well as the withdrawal of conclusions from the results that have been obtained. The research concept is found in Figure 3 below:

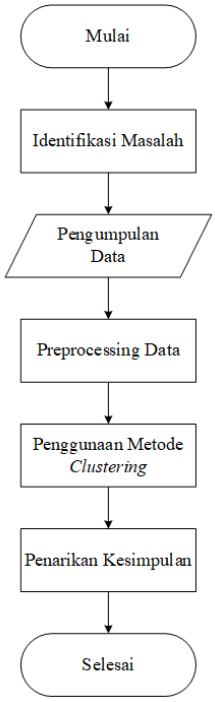

Fig 3. Research Concept

As for the explanation of the concept of research on figure 3 Initial step on the concept of this research is the identification of the problem to know the existing problems, as well as the method used in this research, is clustering with the Algorithna K-means. Then the collection of data obtained for three months with the amount of 1397. The next step is to preprocessing The data that aims to handle the duplication of data as well as the selection of attributes, and the attributes used are item name, amount sold, and stock. The next step is processed using the clustering method with the K-means algorithm and then concludes the results obtained.

4. Results and Discussion

4.1. Problem Identification

In this study, several literary studies were conducted by studying the literature related to the research conducted for the grouping of products as well as conducting the appropriate selection of algorithms for use in this study. Based on the results of identification, the algorithm used is K-means clustering for product grouping based on products that are desirable and less desirable to advise on the Rizki Barokah Store in determining the product stock.

4.2. Data Collection

Data used in this research is obtained from the Rizki Barokah Store, which is a stock data that explains the amount of stock remaining goods and product sales data from November 2018 – January 2019 describing the number of each product sold that month.

4.3. Preprocessing

At this stage is the handling of data that has duplication and selection of attributes. The attributes used are the item name, number sold, and stock. So The result of this stage of the final DataSet consists of 931 number of items and consists of 3 attributes.

4.4. Use of Clustering Methods

Once the preprocessing stage is completed, the next step is processed using the algorithms K-means with the clustering method. The following is a test on 10 samples of manually conducted data to view the results of product groupings based on the algorithms K-means with clustering techniques contained in table 1. The calculation step is as follows:

Table 1. Sample Data
(Source: Rizki Barokah Store)

Item Name	Sold Amount	Stock
NU MILK TEA 330ML	3	3
NU TEH TARIK 330ML	2	4
NUTRI SARI BRAZILIAN ORANGE 11G	9	5
THE NUTRIENTS OF GRASS JELLY 15G	2	3
NUTRIJELL FLAVORED STRAW 15G	1	3
NUTRIJELLYOGHURT BLACKCURRANT 15G	3	3
NUVO FAMILY COOL 80g	2	7
NUVO FAMILY NATURE PROTECT 80g	3	4
NUVO FAMILY PINK 80 GR	2	8
OREO CHOCOLATE CREME 29.4 G	1	10

a. Specify the number of clusters and then specify some problems that occurred in the cluster center. On this research author selects 2 clusters with data 2nd and 4th as Cluster center namely :

   Data 2 (C1) : 2, 4

   Data 4 (C2) : 2, 3

b. Once the cluster center selection is performed, the next step is to calculate the cluster center distance in the 1st iteration.

   1) Calculates distances on the cluster Center.

      To compute distances on any existing data against the cluster Center used the Euclidean distance formula in Figure 4:

$$= \sqrt{(X_{1i} - X_{1j})^2 + (X_{2i} - X_{2j})^2 + \cdots + (X_{ki} - X_{kj})^2 \ldots}$$

Fig 4. *Euclidean Distance* formula

As for the calculation with the formula Euclidean Distance in the 1st iteration is as follows:

The initial step is to calculate the distance on each record with the second Cluster Center (C1):

$$D1,1 = \sqrt{(3 - 2)^2 + (3 - 4)^2} = 1.41$$

$$D1,2 = \sqrt{(2 - 2)^2 + (4 - 4)^2} = 0$$

$$D1,3 = \sqrt{(9 - 2)^2 + (5 - 4)^2} = 7.07$$

$$D1,4 = \sqrt{(2 - 2)^2 + (3 - 4)^2} = 1$$

$$D1,5 = \sqrt{(1 - 2)^2 + (3 - 4)^2} = 1.41$$

$$D1,6 = \sqrt{(3 - 2)^2 + (3 - 4)^2} = 1.41$$

$D1,7 = \sqrt{(2-2)^2 + (7-4)^2} = 3$

$D1,8 = \sqrt{(3-2)^2 + (4-4} = 1$

$D1,9 = \sqrt{(2-2)^2 + (8-4)^2} = 4$

$D1,10 = \sqrt{(1-2)^2 + (10-4)^2} = 6.08$

The next step is to calculate the distance on each data with the center of the second cluster (C2):

$D2, 1 = \sqrt{(3-2)^2 + (3-3)^2} = 1$

$D2, 2 = \sqrt{(2-2)^2 + (4-3)^2} = 1$

$D2, 3 = \sqrt{(9-2)^2 + (5-3)^2} = 7.28$

$D2, 4 = \sqrt{(2-2)^2 + (3-3)^2} = 0$

$D2, 5 = \sqrt{(1-2)^2 + (3-3)^2} = 1$

$D2, 6 = \sqrt{(3-2)^2 + (3-3)^2} = 1$

$D2, 7 = \sqrt{(2-2)^2 + (7-3)^2} = 4$

$D2, 8 = \sqrt{(3-2)^2 + (4-3)^2} = 1.41$

$D2, 9 = \sqrt{(2-2)^2 + (8-3)^2} = 5$

$D2, 10 = \sqrt{(1-2)^2 + (10-3)^2} = 7.07$

Here are the results of the calculations that can be seen in table 2:

Table 2. Result of Cluster Spacing in 1st Iteration
(Source: Data already processed)

Item name to-	C1	C2	Shortest distance
1	1.41	1.00	1.00
2	0.00	1.00	0.00
3	7.07	7.28	7.07
4	1.00	0.00	0.00
5	1.41	1.00	1.00
6	1.41	1.00	1.00
7	3.00	4.00	3.00
8	1.00	1.41	1.00
9	4.00	5.00	4.00
10	6.08	7.07	6.08

2) Grouping the Data

The next step is to group the data from the calculation of the process in Table 2 by entering each object (data) into a cluster (group) based on the minimum distance. The following is the result of the data grouping in the first iteration found in Table 3:

Table 3 Results of 1st Iteration Data Grouping
(Source: Data already processed)

Item Name	C1	C2
1	0	1
2	1	0
3	1	0
4	0	1
5	0	1
6	0	1
7	1	0
8	1	0
9	1	0
10	1	0

Based on the results obtained in table 3, the value of 1 is the closest distance obtained by viewing the minimum value between C1 and C2.

3) Calculating the new cluster center

The next step is to calculate the center of the new cluster, as the way is to calculate the average of the minimum value entered in each cluster. The calculation of the first cluster Center (C1) that is calculated based on the average is the following:

C1 = (2 + 9 + 2 + 3+2 + 1)/6 = 3.1

C1 = (4 + 5 + 7 + 4+8 + 10)/6 = 6.3

A second cluster center calculation (C2) that has been calculated based on average is as follows:

C2 = (3 + 2 + 1 + 3)/4 = 2.2

C2 = (3 + 3 + 3 + 3)/4 = 3

c. 2nd Iteration

1) The next step is to calculate the cluster center distance in the 2nd iteration. The cluster center is extracted from the calculation results of the new cluster in the 1st iteration.
2) The next step is to calculate the distance between the center of the cluster using the Euclidean distance formula just as in the first step. The calculation result can be seen in table 4:

Table 4. Calculation Results for Cluster Spacing in the 2nd Iteration
(Source: Data already processed)

Item name to-	C1	C2	Shortest distance
1	3.34	0.75	0.75
2	2.61	1.03	1.03

Item name to-	C1	C2	Shortest distance
3	5.98	7.04	5.98
4	3.53	0.25	0.25
5	3.98	1.25	1.25
6	3.34	0.75	0.75
7	1.34	4.01	1.34
8	2.34	1.25	1.25
9	2.03	5.01	2.03
10	4.26	7.11	4.26

3) Grouping the Data

The next step is to group the data from the calculation of the process in Table 4 by entering each object (data) into a cluster (group) based on the minimum distance. The following is the result of the data grouping in the 2nd iteration found in Table 5:

Table 5. Results of 2nd Iteration Data Grouping
(Source: Data already processed)

Item Name	C1	C2
1	0	1
2	0	1
3	1	0
4	0	1
5	0	1
6	0	1
7	1	0
8	0	1
9	1	0
10	1	0

Based on the results obtained in table 5, the value of 1 is the closest distance obtained by viewing the minimum value between C1 and C2. Because the results of the grouping in the 1st iteration and the 2nd iteration do not yet have the same result, then the grouping continues in the 3rd iteration.

4) Calculating the new cluster center

The calculation of the new first cluster Center (C1) is calculated based on the average of the following:

C1 = (9 + 2 + 2 + 1)/4 = 3.5

C1 = (5 + 7 + 8 + 10)/4 = 7.5

The calculation of the new cluster (C2) which is calculated based on the average is as follows:

C2 = (3 + 2+ 2 + 1+3 + 3)/6 = 2.3

C2 = (3 + 4 + 3 + 3+3 + 4)/6 = 3.3

d. 3rd Iteration

1) The next step is to calculate the cluster center distance in the 3rd iteration. The cluster center is extracted from the calculation results of the new cluster in the 2nd iteration.

2) The next step is to calculate the distance between the center of the cluster using the Euclidean distance formula just as in the first step. The calculation result can be seen in table 6:

Table 6. Calculation Results for Cluster Spacing in the 3rd Iteration
(Source: Data already processed)

Item name to-	C1	C2	Shortest distance
1	4.53	0.75	0.75
2	3.81	0.75	0.75
3	6.04	6.87	6.04
4	4.74	0.47	0.47
5	5.15	1.37	1.37
6	4.53	0.75	0.75
7	1.58	3.68	1.58
8	3.54	0.94	0.94
9	1.58	4.68	1.58
10	3.54	6.80	3.54

3) Grouping the Data

The next step is to group the data from the calculation of the process in Table 6 by entering each object (data) into a cluster (group) based on the minimum distance. The following is the result of the data grouping in the 3rd iteration found in Table 7:

Table 7. Results of 3rd Iteration Data Grouping
(Source: Data already processed)

Item Name	C1	C2
1	0	1
2	0	1
3	1	0
4	0	1
5	0	1
6	0	1
7	1	0
8	0	1
9	1	0
10	1	0

Based on the results obtained in table 7, the value of 1 is the closest distance obtained by viewing the minimum value between C1 and C2. Because the results of the grouping in the 2nd iteration and the 3rd iteration already have the same results, then the iteration process stops at the 3rd iteration.

Based on the calculation result, the product grouping in the store Rizki Barokah with a sample number of 10 data consists of 2 clusters, namely cluster1 and cluster2. Pada cluster1 known to have 4 less popular products, and on cluster2, 6 products are most in-demand. The Following are the results of the clustering K-means contained in table 8:

Table 8. Results Clustering K-Means
(Source: Data already processed)

Item Name	Sold Amount	Stock	Cluster
NU MILK TEA 330ML	3	3	2
NU TEH TARIK 330ML	2	4	2
NUTRI SARI BRAZILIAN ORANGE 11G	9	5	1
THE NUTRIENTS OF GRASS JELLY 15G	2	3	2
NUTRIJELL FLAVORED STRAW 15G	1	3	2
NUTRIJELLYOGHURT BLACKCURRANT 15G	3	3	2
NUVO FAMILY COOL 80g	2	7	1
NUVO FAMILY NATURE PROTECT 80g	3	4	2
NUVO FAMILY PINK 80 GR	2	8	1
OREO CHOCOLATE CREME 29.4 G	1	10	1

5. Conclusions

Based on the results of the research that is done, it can be concluded that:

a. The method of clustering with the K-means algorithm can be used for product grouping in Rizki Barokah Store, so it can be used in determining the stock of products.

b. Based on the results of the calculations can be found that 6 products are in demand and 4 less desirable products.

c. By knowing the most desirable and less desirable products, Rizki Barokah Store can determine the stock of products by prioritizing the purchase of stock products in the most in-demand products and reduction of purchases against less popular products to reduce product stock buildup.

References

[1] Bauer, J.C., Kotouc, A.J., Rudolph, T., 2012. What constitutes a “good assortment”? A scale for measuring consumers' perceptions of an assortment offered in a grocery category. J. Consum. Serv. 19 (1), 11–26.

[2] Bezawada, R., Balachander, S., Kannan, P.K., Shankar, V., 2009. Cross-category effects of aisle and display placements: a spatial modeling approach and insights. J. Mark. 73 (3), 99–117.

[3] Binninger, A.-S., 2008. Exploring the relationships between retail brands and consumer store loyalty. Int. J. Retail Distrib. Manag. 36 (2), 94–110.

[4] Bond, C., Thilmany, D., Keeling Bond, J., 2008. Understanding consumer interest in product and process-based attributes for fresh produce. Agribusiness 24 (2), 231–252.

[5] Brunk, K.H., 2010. Exploring origins of ethical company/brand perceptions - a consumer perspective of corporate ethics. J. Bus. Res. 63 (3), 255–262.

[6] Charton-Vachet, F., Lombart, C., 2015. New conceptual and operational approach to the link between individual and region: regional belonging. Rech. Appl. Mark. 30 (1), 50–75.

[7] Chernev, A., Hamilton, R., 2009. Assortment size and option attractiveness in consumer choice among retailers. J. Mark. Res. 46 (3), 410–420.

[8] Chin, W.W., Dibbern, J., 2010. An introduction to a permutation based procedure for multi-group PLS analysis: results of tests of differences on simulated data and a cross cultural analysis of the sourcing of information system services between Germany and the USA. In: Vinzi, V.E., Chin, W.W., Henseler, J., Wang, H. (Eds.), Handbook of Partial Least Squares: Concepts, Methods and Applications. Springer Handbooks of Computational Statistics, New York, pp. 171–193.

[9] Coelho do Vale, R., Matos, P.V., Caiado, J., 2016. The impact of private labels on consumer store loyalty: an integrative perspective. J. Retail. Consum. Serv. 28, 179–188.

[10] Collins-Dodd, C., Lindley, T., 2003. Store brand and retail differentiation: the influence of store image and store brand attitude on store own brand perceptions. J. Retail. Consum. Serv. 10 (6), 345–352.

[11] de Wulf, K., Odekerken-Schröder, G., Goedertier, F., van Ossel, G., 2005. Consumer perceptions of store brands versus national brands. J. Consum. Mark. 22 (4), 223–232.

[12] Deselnieu, O.C., Constanigro, M., Souza-Monteiro, D.M., McFadden, D.T., 2013. A metaanalysis if geographical indication food valuation studies: What drives the premium for origin-based labels? J. Agric. Resour. Econ. 38 (2), 204–219.

[13] Dick, A.S., Basu, K., 1994. Customer loyalty: toward an integrated conceptual framework. J. Acad. Mark. Sci. 22 (2), 99–113.

[14] Fernández-Ferrín, P., Bande-Vilela, B., 2015. Attitudes and reactions of Galician (Spanish) consumers towards the purchase of products from other regions. Glob. Bus. Econ. Rev. 17 (2), 131–150.

[15] Lee, W.J.T., Cheah, I., Phau, I., Teah, M., Elenein, B.A., 2016. Conceptualising consumer regiocentrism: examining consumers' willingness to buy products from their own region. J. Retail. Consum. Serv. 32, 78–85.