CLUSTER ANALYSIS OF SALES TRANSACTION DATA USING K-MEANS CLUSTERING AT TOKO USAHA MANDIRI

Fauzia Allamatul Fithri [1*]; Sukma Wardhana[2]

Informatics Engineering
Mercu Buana University
www.mercubuana.ac.id
[1*]41517310015@student.mercubuana.ac.id, [2]sukma@mercubuana.ac.id
(*) Corresponding Author

Abstract—Data mining is a process to obtain useful information from a database warehouse in the form of knowledge. Data transaction history of sales can be information for a business decision. Toko Usaha Mandiri has a problem with the stock of goods, and there are passive goods that settle in the warehouse for a relatively long period. Previous research conducted data analysis to form data clustering into useful information. This study aims to analyze sales data by applying the K-Means Clustering algorithm to form sales clusters. The results of data clustering form cluster1, cluster2 and cluster3 with percentage values of 62% (11 data), 8% (56 data) and 30% (25 data), respectively. Cluster validation of K-Means Clustering algorithm with Davies Bouldin Index produces a value of 0.2. The information of sales clustering can be an alternative solution, input for stock management and marketing strategies.

Keywords: cluster validation, data mining, davies bouldin index, k-means clustering, sales

Abstrak—*Data mining merupakan proses untuk mendapatkan informasi yang berguna dari gudang basis data yang berupa ilmu pengetahuan. Histori data transaksi penjualan dapat menjadi sebuah informasi untuk suatu keputusan bisnis. Toko Usaha Mandiri mempunyai permasalahan pada stok barang terdapat produk barang pasif mengendap di gudang dalam jangka waktu relatif lama. Penelitian sebelumnya melakukan analisis data membentuk clustering data menjadi suatu informasi yang bermanfaat. Penelitian ini bertujuan melakukan analisis data penjualan menerapkan algoritma K-Means Clustering membentuk cluster-cluster penjualan. Hasil clustering data membentuk cluster1, cluster2 dan cluster3 dengan nilai persentase masing-masing 62% (11 data), 8% (56 data) dan 30% (25 data). Validasi cluster algoritma K-Means Clustering dengan Davies Bouldin Index menghasilkan nilai 0.2. Informasi clustering penjualan dapat menjadi alternatif solusi, bahan masukan untuk manajemen stok dan strategi pemasaran.*

Kata Kunci: *data mining, davies bouldin index, k-means clustering, penjualan, validasi cluster*

INTRODUCTION

Data mining is essential for companies to dig some data into databases into knowledge they never knew before. Data stored in data warehouses such as sales transaction data can be utilized information in it as a means of business decision making. K-Means Clustering is one of the famous mining techniques for analyzing data widely used in scientific research. At Toko Usaha Mandiri, a packaged beverage agent has problems with the stock of goods. Some goods passively settle in the warehouse for a relatively long period. The problem is the sale of goods that are not smooth or less sold. Previous research conducted data analysis to form data clustering into helpful information. This study analyzes sales transaction data using the K-Means Clustering algorithm to create a sales data cluster. Sales data clustering information can be an alternative solution to problems and input for business decisions.

Some studies apply methods—K-Means Clustering to solve existing problems. Researchers (Noviyanto, 2020) conducted data grouping the number of deaths of Covid-19 sufferers in several Asian countries using K-Means. This information is a knowledge of the spread of death rates due to the coronavirus in Asian countries. Researchers (Adiya & Desnelita, 2019) conducted data grouping of drug use at Pekanbaru Hospital using K-Means to determine the level of drug use. Thus the hospital can control the need and procurement of drugs effectively and efficiently. Researchers (Yunita, 2018) grouped data on new student admissions to study programs at the Indragiri Islamic University using the K-Means algorithm clustering technique. This cluster information can help promote the study program according to the interests and talents of prospective students. This research explains that K-Means Clustering is a method that can classify data as information that can be input in making business decisions.

Datasets or a set of data can be grouped into several parts with different characteristics using clustering algorithms (Noviyanto, 2020). Non-hierarchical data clustering is a K-Means method that partitions objects into two or more groups based on their factors (Adiya & Desnelita, 2019). Grouping data by K-Means Clustering into several groups have the same properties in one group and traits that are different from other groups (Yunita, 2018). The data characteristics appear through clustering techniques that form data clusters from a set of data (Setiawan, 2018).

MATERIALS AND METHODS

A. Sales Report Data

Dataset

Dataset Used in this study is the Toko Usaha Mandiri sales report for one year period January - December 2020 excel file. The sales report contains 92 packaged beverage product data and daily sales transactions with variable attributes date, invoice number, code, product name, quantity, unit, price, and amount.

Data Selection

Data selection aims to select the necessary data according to the needs of analysis in research (Handoko, 2018). The selected data is numerical variable sales transaction data to be processed by the K-Means Clustering algorithm. The variable used is the Quantity attribute.

Data Changes

Data changes are made from the source in the form of classification or numeric so that the algorithm can process the data (Yaumi et al., 2020). Data changes are adjusted to the analysis needs, forming a sales cluster with variables number of transactions and sold. Data changes made to this sales report are discarding unnecessary data, blank data, renaming quantity attributes to Sold attributes, and creating new attributes Transaction Amount. The data change dataset contains 92 data, as shown in Table 1.

Table 1. Dataset of Data Change Results

No.	Code	Transaction Amount	Sold
1	AL001	770	1447
2	AQ001	332	654
.....			
5	AQ004	810	1627
.....			
15	CA001	72	143
.....			
80	TB001	662	1267
.....			
91	VI003	770	1497
92	YC001	717	1424

B. Research Methods

Type of research

This type of research is applied research, which produces an object to find a solution to a problem—for example, applied sciences such as informatics engineering, engineering, chemistry, medicine, and others.

Data Collection

The data collection method used in this research is a documentation study by collecting the necessary documents related to the problem under study for intensive research. Data collection in related research reference journals and sales transaction data report files from Toko Usaha Mandiri.

Problem Analysis

Researchers conducted a study and found the existing problems. That is, there are several stocks of goods. That passively settles in the warehouse, in other words, the less saleable items. Intense low sales cause existing problems. Analyze the situation by forming sales groups or clusters into the highest, lowest, and medium sales categories. The store owner can take the proper steps to manage the stock of goods and their marketing strategies.

Data Mining

Data mining is a term or concept used to find new information hidden in a large piece of data. Data mining has the primary function that is Descriptive and predictive (Siregar, 2018). The description function is an understanding of the observed data. In contrast, the prediction function is the process of finding a particular pattern of data, extracting and identifying data mining information using statistical techniques, mathematics, artificial intelligence, and machine learning. Several known data mining methods include classification, clustering, association, regression, forecasting, and others (Rofiqo et al., 2018).

K-Means Clustering

K-Means is a data clustering method that groups data into one or more groups with a partition system (Sukamto et al., 2018). This algorithm aims to find groups on the data represented by the variable k, the number of clusters formed. The selection of the centroid starting point can affect clustering because there are differences in cluster results in different experiments (Hutabarat & Sindar, 2019).

The following are the steps for forming clusters in the K-Means method (Indriyani & Irfiani, 2019):

1. Determine the number of groups (k) and early centroid done randomly.

2. Count distance data with centroid using the formula Euclidean equation (1).

$$d(x_i, \mu_j) = \sqrt{\sum(x_i, \mu_j)^2} \quad \text{..... (1)}$$

d = document point
x_i = criteria data
μ_j = centroid in the jth cluster.

3. Group the data closest to the cluster formed and update the centroid value with the object's location from the cluster's center using equation (2).

$$\mu_j(t+1) = \frac{1}{N_{sj}} \sum_{j \in sj} X_j \quad \text{..... (2)}$$

μj(t+1) = new centroid in iteration to (t+1)
N_{sj} = the amount of data in the sj cluster.

4. Repeat steps 2 to 3 so that each cluster does not change, then the process stops.

Davies Bouldin Index

Davies-Bouldin Index (DBI) is a term to evaluate the performance of the K-Means Clustering algorithm. In 1979, David L. Davies and Donald W. Bouldin were the ones who introduced DBI.

Calculation of the value of the Davies Bouldin Index (DBI) through several stages of formulation requirements. The steps for calculating the Davies Bouldin Index are as follows (Butsianto & Saepudin, 2020):

1. Sum Of Square Within-Cluster (SSW)
To find out the matrix of cluster member attachment in one cluster (cohesion).

$$SSW_i = \frac{1}{m_i} \sum_{j=i}^{m_i} d(x_j, c_i) \quad \text{..... (3)}$$

m_i = the amount of data in the ith cluster
x_j = data on the cluster
c_i = centroid cluster i
$d(x_j, c_i)$= distance every data to centroid

2. Sum Of Square Between-Cluster (SSB)
To find out the difference between one cluster and another cluster (separation).

$$SSB_{i,j} = d(c_i, c_j) \quad \text{..... (4)}$$

c_i = cluster one
c_j = other clusters
$d(c_i, c_j)$ = distance centroid between clusters

3. Ratio
Ratio to find out how good the comparison value between one cluster with another cluster is.

$$R_{i,j} = \frac{SSW_i + SSW_j}{SSB_{i,j}} \quad \text{..... (5)}$$

$R_{i,j}$ = ratio between clusters
SSW_i = cluster 1
SSW_j = cluster 2
$SSB_{i,j}$ = separation of clusters 1 and 2

4. Davies Bouldin Index (DBI)
Value DBI gets better when it's close to zero and isn't negative.

$$DBI = \frac{1}{k} \sum_{i=1}^{k} max_{i \neq j}(R_{i,j}) \quad \text{..... (6)}$$

k = number of clusters
$R_{i,j}$ = ratio between clusters i and j
Max = search for the ratio between the most significant clusters

RESULTS AND DISCUSSIONS

Calculation of K-Means Clustering

1. Determine Many Clusters and Early Centroids.
Randomly determining the initial centroid of the 5th, 15th, and 80th data referred to in Table 1 by forming cluster1, cluster2, and cluster3 as shown in Table 2.

Table 2. Early Centroid

Cluster	Data	Code	Transaction Amount	Sold
Cluster1	5	AQ004	810	1627
Cluster2	15	CA001	72	143
Cluster3	80	TB001	662	1267

2. Calculating Distance Object.
Table 3 shows the calculation of data distance to centroid with distance formula Euclidean equation (1).

Table 3. Iteration 1 of Distance Calculation

Object Spacing	Euclidean Distance	Result
Data 1 - C1	$\sqrt{(770-810)^2 + (1447-1627)^2}$	184.39
Data 1 - C2	$\sqrt{(770-72)^2 + (1447-143)^2}$	1479.06
Data 1 - C3	$\sqrt{(770-662)^2 + (1447-1267)^2}$	209.91
.....		
Data 92 - C1	$\sqrt{(717-810)^2 + (1424-1627)^2}$	223.29
Data 92 - C2	$\sqrt{(717-72)^2 + (1424-143)^2}$	1434.22
Data 92 - C3	$\sqrt{(717-662)^2 + (1424-1267)^2}$	166.36

3. Cluster Formation.
The distance calculation results for the formation of clusters based on proximity to the centroid are as shown in Table 4.

Table 4. Formation Cluster Iteration 1

Data	Code	C1	C2	C3	Cluster
1	AL001	184.39	1479.06	209.91	1
2	AQ001	1084.07	573.34	696.18	2
.....					
91	VI003	136.01	1523.33	254.09	1

4. Cluster Grouping and Centroid Update.

Group the data closest to the cluster formed and update the centroid value with the object's location from the collection center by calculating the average using equation (2). Tables 5 and 6 show the closest data grouping of each cluster and the results of the new centroid.

Table 5. Grouping Cluster Iteration 1

Data	Code	Cluster1		Cluster2		Cluster3	
		x	y	x	y	x	y
1	AL001	770	1447	0	0	0	0
2	AQ001	0	0	332	654	0	0
3	AQ002	851	1644	0	0	0	0
.....							
91	VI003	770	1497	0	0	0	0
92	YC001	0	0	0	0	717	1424
Sum of Objects		9565	18777	9005	17863	9941	19498
Sum of Data		10	10	63	63	19	19
Average / Centroid		956.50	1877.70	142.94	283.54	523.21	1026.21

Source: (Fithri & Wardhana, 2021)

Table 6. New Centroid Iteration 1

Cluster	Transaction Amount	Sold
Cluster1	956.50	1877.70
Cluster2	142.94	283.54
Cluster3	523.21	1026.21

Source: (Fithri & Wardhana, 2021)

5. Calculation of the Distance of the Next Iteration. After getting a new centroid in the calculation of iteration 1, then calculate the distance for the next iteration with the same steps as in the first iteration. So that there is no displacement of the cluster formation and the centroid value, in this case, the process stops at iteration 6 of cluster formation, and the center point does not change. The results of the centroid calculation for each iteration are as shown in Table 7.

Table 7. Centroid Each Iteration

Iteration	CLuster1		CLuster2		CLuster3	
	x	y	x	y	x	y
0	810.00	1627.00	72.00	143.00	662.00	1267.00
1	956.50	1877.70	142.94	283.54	523.21	1026.21
2	956.50	1877.70	139.92	277.34	513.55	1008.30
3	956.50	1877.70	130.49	258.17	489.00	962.13
4	934.73	1836.45	123.95	245.16	465.17	915.13
5	934.73	1836.45	120.80	239.20	458.56	901.68
6	934.73	1836.45	120.80	239.20	458.56	901.68

Source: (Fithri & Wardhana, 2021)

End of K-Means Calculation

Table 8 shows the results of K-Means Clustering calculations that end in the 6th iteration forming cluster1 with the amount of 11 data, cluster2 with the amount of 56 data, and cluster3 with the amount of 25 data.

Table 8. Final Results of K-Means Calculation

Cluster	Sum of Nearby Objects	Sum of Data	Centroid	
			Transaction Amount	Sold
Cluster1	3772.21	11	934.73	1836.45
Cluster2	6928.22	56	120.8	239.2
Cluster3	6368.93	25	458.56	901.68

Source: (Fithri & Wardhana, 2021)

Cluster Validation

Evaluate algorithm testing to determine the performance of the clustering process by calculating the Davies Bouldin Index (DBI) value (Triyansyah & Fitrianah, 2018). Data used for algorithm evaluation K-Means Clustering referring to Table 8.

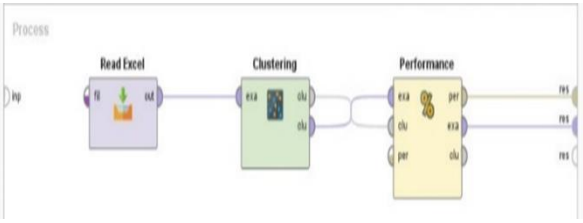
The process of calculating the cluster validation results from the Davies Bouldin Index validation, respectively, uses equations (3), (4), (5), and (6) as follows.

1. Sum Of Square Within-Cluster = 721.41
2. Sum Of Square Between-Cluster = 1195.12
3. Ratio = 0.604
4. Davies Bouldin Index (DBI) = 0,201

Evaluation of algorithm testing with DBI 0.2 is close to zero and harmless, which means the classification process shows good performance and accuracy. Clustering can be said to work well if the test value is minimal or getting smaller near-zero (Syahidatul Helma et al., 2019).

Internal Validation Davies Bouldin Index

Perform internal validation of the Davies Bouldin Index to compare the similarity of the calculation results manually with RapidMiner. RapidMiner is an *open-source* application used for data mining analysis, text mining, and predictive analysis (Sibuea & Safta, 2017).



Source: (Fithri & Wardhana, 2021)
Figure 1. K-Means Process

Figure 1 shows the K-Means Clustering process model on RapidMiner starting from reading the dataset on the Read Excel operator entered through Microsoft Excel, classified as a numerical variable. The Clustering operator is the operator model for the K-Means method to calculate the distance from each data cluster. The Performance operator measures cluster validation using the Davis Bouldin Index method. Figure 2 shows the results of cluster formation and K-Means performance evaluation on RapidMiner.

Source: (Fithri & Wardhana, 2021)
Figure 2. DBI Value & Cluster Model

Research Results

This study researched a problem applying K-Means Clustering to get a fact that could solve the problem at Toko Usaha Mandiri. The study results analyzed sales transaction data forming sales data patterns into sales clusters as helpful information for the company. Here is the essence of the study results, as seen in Tables 9, 10, and 11.

Table 9. Clustering Data

Cluster	Sum Of Data	Transaction Amount	Sold	Percentage
Cluster1	11	934.73	1836.45	62 %
Cluster2	56	120.80	239.20	8 %
Cluster3	25	458.56	901.68	30 %
Sum	92			100 %

Source: (Fithri & Wardhana, 2021)

Table 10. Cluster Formation

Cluster	Sum of Data	Name Product
ster1	11	Ale-Ale, Aqua 1500 ml/ Large, Aqua 240 ml/ Glass, Aqua 330 ml/ Mini, Aqua 600 ml/ Tanggung, Kopiko Bottle 78º C, Le Mineral 600 ml, Nescafe UHT 200 ml, Cup Cup Tea, Vit 600 ml, You C-1000.
CLuster2	56	Aquaria 240 ml/ Glass, Aquaria 330 ml/ Mini, Aquaria 600 ml/ Responsibility, Big Cola 3.1 lt, Big Cola 535 ml, Buavita 250 ml, Capucino, Club Gallon 19 liters, Club Glass 150 ml/ Fill 20 glasses, Club Glass 150 ml / Fill 48 glasses, Club Glass 240 ml, Club Bottle 330 ml, Club Bottle 600 ml, Club Bottle 1500 ml, Cimory Fresh Milk 950 ml, Cimory Fresh Milk 250 ml, Cimory Fresh Milk Full Cream 950 ml, Cimory Fresh Milk MILK UHT 250 ml, Cleo 550 ml, Coca-cola/ Fanta/ Sprite Seru, Coca-cola/ Fanta/ Slim, Coca-cola/Fanta/Sprite Mini, Coolant, Frestea, Fruit Tea Bottle, Good Day Bottle, Golda, Isoplus, Rhino Solution Bottle 330 ml, Rhino Solution Bottle 500 ml, Canned Rhino Solution, Le Mineral 1500 ml, Le Mineral 330 ml, Love Juice 300 ml, Luwak Bottle, Mirai Ocha, Moun Tea Glass, Mytea 500 ml, Mizone, Oasis 19 lt, Oasis 600 ml, Oasis+ 350 ml, Oasis+ 500 ml, Okky Jelly Drink, Pino Ice Cup, Pocari Sweat Bottle 2000 ml, Pocari Sweat Bottle 350 ml, PulpyAloe Vera, Q Guava 350 ml, Sari Kacang Ijo Ultrajaya 150 ml, Ijo Ultrajaya Nut Juice 200 ml, Ijo Ultrajaya Nut Juice 250 ml, Fragrant/ Less Sgr 350 ml, Sosro Box Tea 250 ml, Ultrajaya Box Tea 300 ml, Vit 1500 ml.
CLuster3	25	Aqua Gallon 19 lt, Aquaria 1500 ml/ Large, Buavita 125 ml, Fatigon Hydro, Floridina Orange 360 ml, Kopikap Capucino, Green Kiko, Red Kiko, Rhino Solution Bottle 200 ml, Lemon / Orange Water, Nu Green Tea 500 ml, Oasis 240 ml, Oasis 330 ml, Oasis 1500 ml, Okky Jelly Cup Fill 10, Okky Jelly Cup Jumbo Contents 2, Pocari Sweat Bottle 500 ml, Pocari Sweat Can, Pulpy Orange / Mango, Sosro Tea Bottle 450 ml, Eco Cup Tea, Fragrant Shoot Tea 480 ml, Javana Tea 350 ml, Zegar Tea 2 Tang, Vit 240 ml /Glass.

Source: (Fithri & Wardhana, 2021)

Table 11. Validation Results and Cluster Models

Validation of DBI		Model Cluster			
Manual	**RapidMiner**	**Manual**		**RapidMiner**	
0.201	0.222	Cluster1	11	Cluster0	56
		Cluster2	56	Cluster1	11
		Cluster3	25	Cluster2	25
Amount of Data			92		92

CONCLUSION

The application of K-Means Clustering analyzes sales transaction data forming clusters of sales data providing helpful information for the company. At least this research can help companies to outline existing problems, become an alternative

input material for stock management and marketing strategies. The solution for selling products in cluster2 is low sales, aka goods that are not selling well; the marketing strategy is to sell goods with product packages, on sale, or give discounts on goods. However, this clustering method is susceptible to random initial centroid generation, so that there are differences in cluster results in experiments with different centroids. As a note for further research, to analyze sales problems, for optimal results, try using other methods or a combination of algorithms such as classification, association, regression, basket market analysis, and others.

REFERENCE

Adiya, M. H., & Desnelita, Y. (2019). National Journal of Technology and Information Systems Application of K-Means Algorithm For Clustering Drug Data at Pekanbaru Hospital. *National Technology and Information Systems*, *01*, 17–24.

Butsianto, S., & Saepudin, N. (2020). Application of Data Mining To Students' Interest in Mathematics Subjects With K-Means Method. *National Journal of Computing and Information Technology (JNKTI)*, *3*(1), 51–59. https://doi.org/10.32672/jnkti.v3i1.2008

Fithri, F. A., & Wardhana, S. (2021). *Cluster Analysis Of Sales Transaction Data Using K-Means Clustering At Toko Usaha Mandiri*. *17*(2), 1–7. https://doi.org/10.33480/pilar.v17i2.2273

Handoko, K. (2018). Grouping Mining Data on The Number of Passengers at Hang Nadim Airport. *Computer-Based Information System Journal*, *6*(2), 60. https://doi.org/10.33884/cbis.v6i2.708

Hutabarat, S.M., & Sindar, A. (2019). Data Mining Of Motorcycle Parts Sales Using K-Means Algorithm. *National Journal of Computing and Information Technology (JNKTI)*, *2*(2), 126. https://doi.org/10.32672/jnkti.v2i2.1555

Indriyani, F., & Irfiani, E. (2019). Clustering Sales Data at Outdoor Supply Stores Using the K-Means Method. *JUITA: Journal of Informatics*, *7*(2), 109. https://doi.org/10.30595/juita.v7i2.5529

Noviyanto. (2020). Application of Data Mining in Grouping The Number of Deaths. *Paradigm-Journal of Informatics and Computers*, *22*(2).

Rofiqo, N., Windarto, A. P., & Hartama, D. (2018). Application of Clustering in Residents Who Have Health Complaints With K-Means Datamining. *KOMIK (National Conference on Information and Computer Technology)*, *2*(1), 216–223. https://doi.org/10.30865/komik.v2i1.929

Setiawan, S. (2018). Utilization of K-Means Method in Determining Inventory of Goods. *PIXELS: Research in Computer Science embedded systems and logic*, *6*(1), 41–48. https://doi.org/10.33558/piksel.v6i1.1398

Sibuea, M. L., & Safta, A. (2017). Mapping Outstanding Students Using the K-Means Clustering Method. *Jurteksi*, *4*(1), 85–92. https://doi.org/10.33330/jurteksi.v4i1.28

Siregar, M. H. (2018). Data Mining Clustering of Building Tools Using K-Means Method (Case Study In Adi Building Store). *Journal of Technology and Open Source*, *1*(2), 83–91. https://doi.org/10.36378/jtos.v1i2.24

Sukamto, S., Id, I. D., & Angraini, T. R. (2018). Determination of Fire-Prone Areas in Riau Province Using Clustering K-Means Algorithm. *JUITA: Journal of Informatics*, *6*(2), 137. https://doi.org/10.30595/juita.v6i2.3172

Syahidatul Helma, S., Rustiyan, R. R., Normala, E., Information Systems Studies Faculty of Science and Technology, P., State Islam Sultan Syarif Kasim Riau, U., Soebrantas No, J., & Baru, S. (2019). Clustering on Pekanbaru City Health Care Facility Data Using K-Means Algorithm. *Puzzle Research Data Technology (Predatech) Faculty of Science and Technology*, *1*(November), 4.

Triyansyah, D., & Fitrianah, D. (2018). Data Mining Analysis Uses K-Means Clustering Algorithms to Determine Marketing Strategies. *Journal of Telecommunications and Computers*, *8*(3), 163. https://doi.org/10.22441/incomtech.v8i3.4174

Yaumi, A. S., Zulfiqkar, Z., & Nugroho, A. (2020). Clustering of Consumer Characters Against Product Selection Tendencies Using K-Means. *JOINTECS (Journal of Information Technology and Computer Science)*, *5*(3), 195. https://doi.org/10.31328/jointecs.v5i3.1523

Yunita. (2018). Application of Data Mining Uses K-Means Clustering Algorithm on Admission of New Students (Case Study: Indragiri Islamic University). *Journal of Systemization*, *7*(September), 238–249.