INFORMATIKA
Universitas Labuhanbatu
Vol. 12 No. 3 / Juli/2024
2615-1855 (E-ISSN)
2303-2863 (P-ISSN)

Application Of Data Mining In Selecting Superior Products Using The K-Means And K-Medoids Algorithm Methods

Eva Hermika[1], Syaiful Zuhri Harahap[2], Irmayanti[3]

Information Systems, Labuhanbatu University[1,2,3]

Email : evahermika@gmail.com[1], syaifulzuhriharahap@gmail.com[2], irmayantiritonga2@gmail.com[3]

Corresponding Author: evahermika@gmail.com

Abstract

As a supermarket, we are committed to always improving everything, including selecting the greatest goods. To evaluate which items are more superior or popular and which are less popular, you will want a sizable amount of information sources. To select products and identify those that belong in the superior product cluster, researchers employed the clustering method. The clustering strategy uses two forms of cluster analysis, k-means and k-medoids, which have related techniques. The research results show that the k-means algorithm's Davies Bouldin value is -0.430, whereas the k-medoids algorithm's Davies Bouldin value is -1.392. This suggests that the Davies Bouldin value of the k-medoids approach is the lowest, showing that the grouping findings of the k-means method are a better method to apply to the issue of choosing better products.

Keywords : *K-Means; K-Medoids; Clustering; Algorithm; Data Mining.*

I. Introduction

Opening up trade between nations and expanding access to other countries' product markets is becoming more and more important due to the evolution of the global economy and the pattern of relations between nations, which typically indicate that the distance between one country and another one is reducing. Financial transparency and Trade presents opportunities as well as challenges all at once. The increasing openness of commerce between nations can provide both problems to the competitiveness of domestic industries selling goods abroad and opportunities for improved domestic product market access. In theory, free trade can yield economic gains by broadening market accessibility and augmenting the aggregate economic surplus. However, free trade will not work provide great benefits if the competitiveness of domestic companies is much lower compared to overseas companies.

As a basic food store that distributes and handles logistics for consumer goods, Yuli Sembako Store is dedicated to constantly improving all elements of its business, including identifying outstanding products. Yuli Sembako Store uses the clustering approach to sort products and determine which ones are part of the superior product cluster. Organizing several records, observations, or other situations

INFORMATIKA
Universitas Labuhanbatu
Vol. 12 No. 3 / Juli/2024
2615-1855 (E-ISSN)
2303-2863 (P-ISSN)

into a distinct class according to comparable items is known as clustering.

The two types of cluster analysis used in the clustering approach are k-means and k-medoids, and they have related methods. Research conducted easier for consumers to get information on the categories of goods to be purchased by applying the k-means and k-medoid algorithms. The difference in the number of clusters in the performance of each algorithm has a mode The calculations are different for each literacy, depending on the data set used and the points centroid which is calculated in the algorithm.

The research was conducted at a moving shop in the consumption sector. Study This takes object data at the Yuli grocery store with several products ranging from rice, oil, sugar and others, these categories are the needs most frequently required by customers so there needs to be the ability to Estimate the sales volume of each product for sale. Customer satisfaction is always a priority meet customer needs. These needs based on the history of sales that have been made so that it can be a reference in improving quality next products in stock planning to target better sales so that it can provide satisfaction for customers .

The k-means and k-medoids algorithm comparison approach was utilized by researchers to handle data. In this research is: using data mining with k-medoids and k-means, where these two algorithms are grouped obtain the same thing but in a different way, k-means is by taking the average value, while k-medoids is by taking the middle value[9]. The research results show that both methods produce the same group, namely recommendations for the taste of authentic Thai milk tea and milk tea based on value research targets centroid of 0.286.

II. Literature Review

Products

The product is the central point of marketing activities because the product is the result of a company that can be offered to the market for consumption and is a tool of a company to achieve the objectives of the company. A product must have advantages over other products in terms of quality, design, shape, size, packaging, service,warranty, and taste in order to attract consumers to try and buy the product.

Product Quality Dimension

According to Mullins, Orville,Larreche, and Boyd (2005:422) when companies want to maintain a competitive advantage in the market, companies must understand what aspects of the dimensions used by consumers to distinguish the products sold by the company with competitor products.

Data Mining

Data mining is the mining or discovery of new information by looking for certain patterns or rules from a very large amount of data. Data mining is also referred to as a series of processes to extract added value in the form of knowledge that has not been known manually from a data set. Data mining is also known as knowledge discovery in databases (KDD).

Pattern Recognition, Data Mining, and Machine Learning

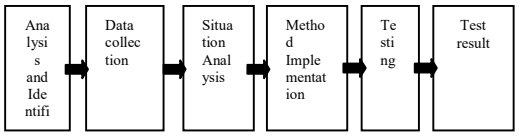
Pattern recognition is a discipline that studies ways of classifying objects into several classes or categories and recognizing data trends. Depending on the

INFORMATIKA
Universitas Labuhanbatu
Vol. 12 No. 3 / Juli/2024
2615-1855 (E-ISSN)
2303-2863 (P-ISSN)

application, these objects can be patients, students, credit applicants, images or signals or other measurementsthat need to be classified or searched for their regression function

III. Method

In the research method there are several work sequences that must be followed, this work sequence is the steps that must be followed and carried out according to the main problem so as not to deviate from the specified problem boundaries. The framework of thought used in carrying out this research is as shown in the following picture:



1. Application of Methodology

Data Collection

The warehouse stock and sales report data from the Yuli Sembako Store were directly collected to supply the data for this investigation. The research data includes product code, product name, transaction date, sales amount and inventory amount. The data used is sales transaction data for 6 months, namely May to October 2023. After the data is obtained, the data is used to find out what products are included in superior products or products that are in high demand and which products are less popular. .

Data Selection

Sales data and warehouse stock data are filtered first and several attributes are taken from the table for analysis. There are three attributes used, namely, item code, average sales for 6 months from May to October 2023 and warehouse stock data for October 2023.

Table 1. Data Selection

Data Ke-	Kode Produk	Rata-Rata Penjualan	Stok Terakhir
1	Pl-0001	104	95
2	Pl-0002	156	116
3	Pl-0003	140	205
4	Pl-0004	124	219
5	Pl-0005	97	115
6	Pl-0006	17	50
7	Pl-0007	7	20
8	Pl-0008	61	38
9	Pl-0009	8	16
10	Hhc-0010	64	80
11	Hhc-0011	6	2
12	Hhc-0012	64	102
13	Hhc-0013	3	5
14	Hhc-0014	8	20
15	Hhc-0015	22	1
16	Enn-0016	65	48
17	Enn-0017	20	20
18	Enn-0018	31	30
19	Enn-0019	25	26
20	Enn-0020	33	14
21	Enn-0021	14	12
22	Enn-0022	58	98
23	Enn-0023	41	154
24	Enn-	27	551

INFORMATIKA
Universitas Labuhanbatu
Vol. 12 No. 3 / Juli/2024
2615-1855 (E-ISSN)
2303-2863 (P-ISSN)

	0024		
25	Enn-0025	1	1
26	Gs-0026	-	-
27	Gs-0027	-	-
28	Gs-0028	-	-
29	Gs-0029	-	-
30	Gs-0030	-	-
31	Gs-0031	-	-
32	Gs-0032	-	-
33	Gs-0033	-	-
34	Gs-0034	-	-
35	Gs-0035	33	81
36	Gs-0036	19	96
37	Gs-0037	4	26
38	Gs-0038	1	2
39	Gs-0039	1	8
40	Gs-0040	1	1
41	Gs-0041	16	51
42	Gs-0042	19	26
43	Gs-0043	20	25
44	Gs-0044	66	112
45	Gs-0045	60	98

3. Data Pre-Processing

After the data has been selected and selected according to the attributes to be used, pre-processing of the data is carried out, so that there is no duplication of data, no missing values and corrects errors in the new data in Excel format. Data that has passed the pre-processing stage will be saved in a new data set using Microsoft Office Excel.

4. Data Conversion

The data change operation is now finished in order to process the data using the k-means and k-medoids algorithms. Data that is not numeric is converted to numeric form. However, no commencement is necessary if the data you currently have is numerical.

5. Data Entry

At this point, the outcomes of the data transformation are used for data modeling. This study used the k-means and k-medoids algorithms as its technique. The selected data will be processed using the clustering method. This method works by grouping data that has similar characteristics in each data. The data has two variables (x and y) for easy visualization in Cartesian coordinates:

a. Variable X = Average sales for 6 months (May – October 2023)

b. Variable Y = Stock of goods in warehouse (30 October 2023)

3.4.3 K-Means Algorithm

Determining the initial centroid, this determination is carried out randomly on the existing data table. The calculation process in the k-means algorithm starts from iteration 1 (one). First, the distance from each data to all existing centroids is calculated. From the results of calculating the distance between each data to all centroids, the smallest distance value to one centroid is obtained, then that centroid is called the closest centroid, and the data will be affiliated into a cluster from the closest centroid.

Table 2. Initial Centroid

Centroid	X	Y
C0	1	1
C1	77	166
C2	27	551

The following formula is used to find the distance between each available piece of data and the centroid value.

INFORMATIKA
Universitas Labuhanbatu
Vol. 12 No. 3 / Juli/2024
2615-1855 (E-ISSN)
2303-2863 (P-ISSN)

$$De = \sqrt{(Mix - Cix)2 + (Miy - Ciy)2}$$

3.4.5 K-Medoids Algorithm

K-Medoids performs grouping by using representative objects (medoids) as the cluster center for each cluster. The k-medoids algorithm goes through the following steps:

a. Determine the desired k (number of clusters) for the previously collected data processed.

The desired cluster consists of 3 clusters, namely cluster 2 is a superior product or a product that is in high demand, cluster 1 is a product that is in moderate demand, and cluster 0 is a product that is in low demand or is not selling well.

b. Select centroid points randomly or sequentially from k initial medoid data.

Table 3. Medoids data

Clusters	Data To-	X	Y
C0	89th data	16	22
C1	71st data	30	52
C2	82nd data	66	112

c. Calculate each data set's distance using k centroid points and the formula [15]. Calculating the separation between the data and the centroid point can be done in a number of ways. The following formula will be used in this study to estimate distances:

$$d1 = \sqrt{(X1 - Xc0)2 + (Y1 - Yc0)2}$$

d. Perform medoids iteration by calculating the distance using one of the distance calculation methods, for all data used as process data.

e. Randomly selecting non-representative objects (non-medoids)

Table 4. Non Medoids Data

Clusters	Data To-	X	Y
C0	25th data	1	1
C1	5th data	97	115
C2	24th data	27	551

Using all of the data's distance computation results, compute the total deviation (S). And do another iteration of the calculation with a new centroid point, using the initial steps again. The new centroid point is drawn randomly from the medoids data. If a is the total calculation of the closest distance using the initial medoids, and b is the total calculation of the closest distance between the object and the new medoids, then the total deviation can be calculated using the equation formula.

S = New total cost – Old total cost

Information:

S=Difference

New total cost = Total cost for non-medoids

Old total cost = Total cost for medoids

So we get:

Because the S value > 0, the clustering process can be stopped.

S= 61,920–19,857=42,063

IV. Results And Discussion

Following k-means algorithm computations, the following product code information are obtained: cluster 0 has 32 products, cluster 1 has 12 items, and cluster 2 has 1 product.

1) Cluster 0 = PL-006, PL-007, PL-008, PL-009, HHC-011, HHC-013, HHC-014, HHC-015, ENN-016, ENN-017, ENN-018, ENN-019, ENN-020, ENN-021, ENN-025, GS-026, GS-027, GS-028, GS-029, GS-030, GS-031, GS-032, GS-033, GS-034, GS-035, GS-037, GS-038, GS-039, GS-040, GS-041, GS-042, GS-043.

INFORMATIKA
Universitas Labuhanbatu
Vol. 12 No. 3 / Juli/2024
2615-1855 (E-ISSN)
2303-2863 (P-ISSN)

2) Cluster 1 = PL-001, PL-002. PL-003, PL-004, PL-005, HHC-010, HHC-012, ENN-022, ENN-023, GS-036, GS-044, GS-045.

3) Cluster 2 = ENN-0024.

Meanwhile, there are 27 products in cluster 0, 6 products in cluster 1, and 12 goods in cluster 2, according to the results of the computation using the k-medoids algorithm. The product codes are as follows:

a. Cluster 0 = PL-007, PL-009, HHC-011, HHC-013, HHC-014, HHC-015, ENN-017, ENN-018, ENN-019, ENN-020, ENN-021, ENN-025, GS-026, GS-027, GS-028, GS-029, GS-030, GS-031, GS-032, GS-033, GS-034, GS-037, GS-038, GS-039, GS-040, GS-042, GS-043.
b. Cluster 1 = PL-006, PL-008, ENN-016, GS-035, GS-036, GS-041.
c. Cluster 2 = PL-001, PL-002, PL-003, PL-004, PL-005, HHC-010, HHC-012, ENN-022, ENN-023 , ENN-024, GS-044, GS-045.

Test Results

The software used to analyze and carry out testing is RapidMiner Studio. Since Rapidminer is standalone software, it is compatible with all operating systems and is developed in Java. The following are the membership outcomes that this approach produced:

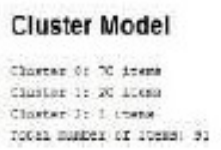


Figure 2. Results of *K-means* Clustering

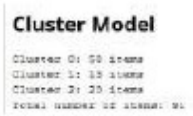


Figure 3. Results of *k-medoids* clustering

cluster or *the cluster* with the least interest is *cluster* 0, judging from the value of the benchmark stock. Then *the cluster* with moderate interest is *cluster* 1 which can be seen from the next distance which is closest to the lowest *cluster* . The highest *cluster* or *the cluster* with the most interest is *cluster* 2, This is depicted in the image.

Figure 4. Results of *K-Means Cluster* Data

Figure 5. Results of *k-medoids cluster* data

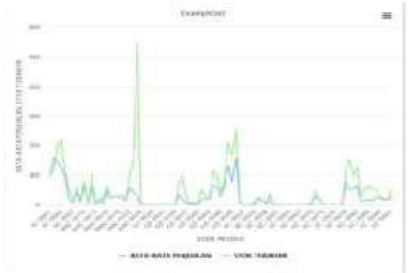


Figure 6. Results of *k-means cluster* graph data

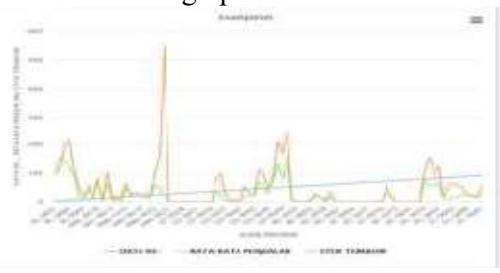


Figure 7. Results of *K-Medoids Cluster* Graph Data

INFORMATIKA
Universitas Labuhanbatu
Vol. 12 No. 3 / Juli/2024
2615-1855 (E-ISSN)
2303-2863 (P-ISSN)

Performance Vector is used to compare the two clustering methods' respective performances in order to decide whether algorithm is better suited for the task of choosing superior products at the Yuli Sembako Store. Based on the cluster centroid, the Performance Vector offers a list of performance criteria values.

Figure 8 . ***Cluster* validation Main criteria of *the Performance Vector***

a. Average inside cluster distance (Avg_within_centroid_distance) is the result of averaging the distances between each cluster instance and the centroid.
b. Davies_bouldin: An method producing clusters with low intra-cluster distance and large inter-cluster distance will have a low Davies Bouldin index. Based on this criterion, the best clustering method is the one that generates a set of clusters with the smallest Davies Bouldin index.

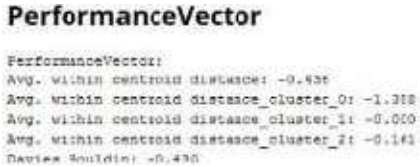


Figure 9. ***Vector K-Means Performance* Results**

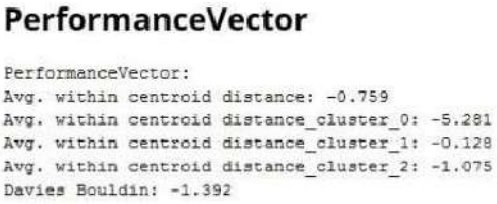


Figure 10. Performance Results *of Vector K-Medoids*

Based on these findings, it is evident that the k-means algorithm's Davies Bouldin value is -0.430, while the k-medoids method's Davies Bouldin value is -1.392. When it comes to selecting superior products at the Yuli Sembako Store, the k-means technique is a better option for grouping results, as seen by the fact that the k-medoids approach's Davies Bouldin value is the lowest.

V. Conclusion

Based on the results and discussion, the conclusion of this research is that with 45 product samples, sales reports and warehouse stock data, two clustering methods can be applied, namely the k-means and k-medoids algorithms at the Yuli Sembako Store, after carrying out the calculations you can The grouping of products that are superior products, products that are in moderate demand, and products that are less popular are known. By implementing the clustering method, the problem of stock shortages in superior products can be overcome, because companies can see the results of grouping products with high, medium and low demand. For those included in the superior product grouping, the company will pay more attention to availability, so that there is no shortage of stock. And for goods that are less popular, companies do not need

INFORMATIKA
Universitas Labuhanbatu
Vol. 12 No. 3 / Juli/2024
2615-1855 (E-ISSN)
2303-2863 (P-ISSN)

to stockpile goods which will result in high and uneconomical storage costs.

VI. Bibliography

R. K. Purba and E. Bu'ulolo, "Implementasi Algoritma K-Medoids dalam Pengelompokan Mahasiswa yang Layak Mendapat Bantuan Uang Kuliah Tunggal," *INSOLOGI J. Sains dan Teknol.*, vol. 1, no. 2, pp. 79–86, 2022, doi: 10.55123/insologi.v1i2.195.

D. A. I. C. Dewi and D. A. K. Pramita, "Analisis Perbandingan Metode Elbow dan Silhouette pada Algoritma Clustering K-Medoids dalam Pengelompokan Produksi Kerajinan Bali," *Matrix J. Manaj. Teknol. dan Inform.*, vol. 9, no. 3, pp. 102–109, 2019, doi: 10.31940/matrix.v9i3.1662.

E. Rahmah, E. Haerani, A. Nazir, and S. Ramadhani, "Penerapan Algoritma K-Medoids Clustering Untuk Menentukan Srategi Promosi Pada Data Mahasiswa (Studi Kasus : Stikes Perintis Padang)," *J. Nas. Komputasi dan Teknol. Inf.*, vol. 5, no. 3, pp. 556–564, 2022, doi: 10.32672/jnkti.v5i3.4355.

I. K. Dan, K. D. Pengelompokan, P. Produksi, and D. Ayam, "Implementasi K-Means Dan K-Medoids Dalam Pengelompokan Wilayah Potensial Produksi Daging Ayam," *J. Teknol. Ind. Pertan.*, vol. 32, no. 158, pp. 239–247, 2022, doi: 10.24961/j.tek.ind.pert.2022.32.3.239.

J. R. S. Penda Sudarto Hasugian, "Penerapan Data Mining Untuk Pengelompokan Siswa Berdasarkan Nilai Akademik dengan Algoritma K-Means," *KLIK Kaji. Ilm. Inform. dan Komput.*, vol. 3, no. 3, pp. 262–268, 2022, [Online]. Available: https://djournals.com/klik

T. L. Afandi, B. Warsito, and R. Santoso, "Implementasi K-Medoids Dan Model Weighted-Length Recency Frequency Monetary (W-Lrfm) Untuk Segmentasi Pelanggan Dilengkapi Gui R," *J. Gaussian*, vol. 11, no. 3, pp. 429–438, 2023, doi: 10.14710/j.gauss.11.3.429-438.

S. Oktarian, S. Defit, and Sumijan, "Clustering Students' Interest Determination in School Selection Using the K-Means Clustering Algorithm Method," *J. Inf. dan Teknol.*, vol. 2, pp. 68–75, 2020, doi: 10.37034/jidt.v2i3.65.

M. Triyana, R. Juita, and C. D. Suhendra, "Penerapan Metode K-Means dalam Pengelompokan Data Penduduk Tidak Mampu di Distrik Oransbari," *INFORMAL Informatics J.*, vol. 7, no. 3, p. 220, 2022, doi: 10.19184/isj.v7i3.34722.

A. Pangestu and T. Ridwan, "Penerapan Data Mining Menggunakan Algoritma K-Means Pengelompokan Pelanggan Berdasarkan Kubikasi Air Terjual Menggunakan Weka," *JUST IT J. Sist. Informasi, Teknol. Inf. dan Komput.*, vol. 11, no. 3, pp. 67–71, 2022, [Online]. Available: https://jurnal.umj.ac.id/index.php/just-it/article/view/11591

R. Bayu Prasetyo, Y. Agus Pranoto, and R. Primaswara Prasetya, "Implementasi Data Mining Menggunakan Algoritma K-Means Clustering Penyakit Pasien

INFORMATIKA
Universitas Labuhanbatu
Vol. 12 No. 3 / Juli/2024
2615-1855 (E-ISSN)
2303-2863 (P-ISSN)

Rawat Jalan Pada Klinik Dr. Atirah Desa Sioyong, Sulteng," *JATI (Jurnal Mhs. Tek. Inform.*, vol. 7, no. 4, pp. 2144–2151, 2023, doi: 10.36040/jati.v7i4.7419.

A. Nugraha, O. Nurdiawan, and G. Dwilestari, "Penerapan Data Mining Metode K-Means Clustering Untuk Analisa Penjualan Pada Toko Yana Sport," *JATI (Jurnal Mhs. Tek. Inform.*, vol. 6, no. 2, pp. 849–855, 2022, doi: 10.36040/jati.v6i2.5755.

G. Gustientiedina, M. H. Adiya, and Y. Desnelita, "Penerapan Algoritma K-Means Untuk Clustering Data Obat-Obatan," *J. Nas. Teknol. dan Sist. Inf.*, vol. 5, no. 1, pp. 17–24, 2019, doi: 10.25077/teknosi.v5i1.2019.17-24.

K. Handoko and L. S. Lesmana, "Data Mining Pada Jumlah Penumpang Menggunakan Metode Clustering," *Snistek*, no. 1, pp. 97–102, 2018.

D. U. Iswavigra, S. Defit, and G. W. Nurcahyo, "Data Mining dalam Pengelompokan Penyakit Pasien dengan Metode K-Medoids," *J. Inf. dan Teknol.*, vol. 3, pp. 181–189, 2021, doi: 10.37034/jidt.v3i4.150.

M. R. Alhapizi, M. Nasir, and I. Effendy, "Penerapan Data Mining Menggunakan Algoritma K-Means Clustering Untuk Menentukan Strategi Promosi Mahasiswa Baru Universitas Bina Darma Palembang," *J. Softw. Eng. Ampera*, vol. 1, no. 1, pp. 1–14, 2020, doi: 10.51519/journalsea.v1i1.10.