ANALYSIS OF CUSTOMER DATA IN SELECTING POTENTIAL CUSTOMERS USING DATA MINING WITH THE K-MEANS ALGORITHM

Randi Rian Putra[1]*, Sri Handayani[2], Fahmi Kurniawan [3], Cendra Wadisman[4]
Universitas Pembangunan Panca Budi[1,3], Universitas Pembinaan Masyarakat Indonesia[2], Universitas Putra Indonesia[4]
randirian@dosen.pancabudi.ac.id , srihandayani111218@gmail.com, fahmikurniawan@dosen.pancabudi.ac.id, cendra_wadisman@upiyptk.ac.id,

Keywords: Data Mining, K-Means Algorithm, Customers, Data Analysis, Potential Customers, Marketing Strategy

***Correspondence Address:**
Randirian@dosen.pancabudi.ac.id

Abstract: In an increasingly competitive business era, understanding customer characteristics and preferences has become essential. This research aims to analyze customer data to find potential customers using data mining techniques and the K-Means algorithm. The data used comes from the company's database which includes transaction history, demographics and customer interactions. The K-Means algorithm is applied to cluster customer data into several groups based on similar characteristics. The clustering results show that there are customer segments with different profit potential for the company. By identifying potential customer segments, companies can design more targeted marketing strategies, increase the efficiency of allocating resources, and ultimately increase ROI (Return on Investment). This study provides guidance for companies in optimizing their approach to reaching potential customers and achieving marketing success.

INTRODUCTION

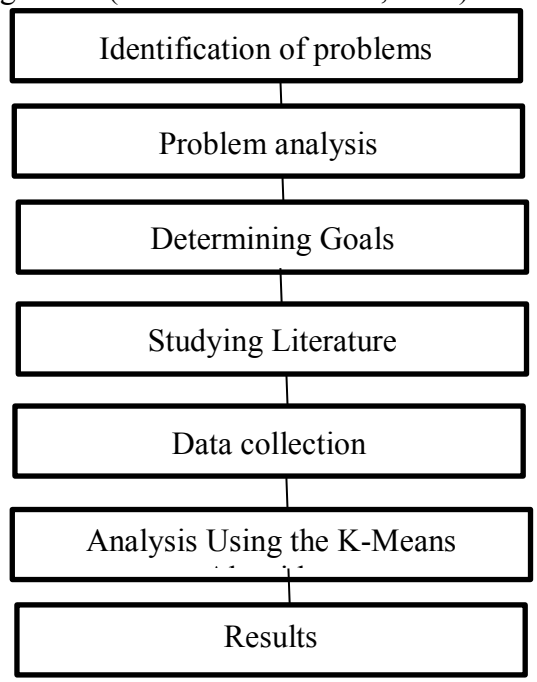
The relationship between companies and customers is very important to support company expansion and sustainability (Romadhona et al., 2022). This data, when analyzed properly, can provide valuable insights into customer behavior, preferences and expectations. Dynamic market conditions and increasing business competition require companies to utilize all the information they have in order to achieve competitive advantage. However, with the large volume of data available, challenges arise in processing and understanding this information for business purposes. Data mining has long been recognized as an effective method for extracting knowledge from large data sets. In the context of customer analysis, this technique allows companies to understand more deeply the characteristics and behavioral patterns of their customers. One data mining technique that has been proven effective in customer analysis is the K-Means clustering algorithm. Clustering is the stage of grouping data into several groups (clusters) (Suarna & Wijaya, 2023).

The K-means Clustering algorithm is a method to assist in grouping into several clusters so as to obtain significant data visualization results (Ahsina et al., 2022). In the context of customer data, this method allows companies to group customers into certain segments based on similar behavior or preferences. In this way, companies can more easily identify potential customers who may provide more value, both in terms of loyalty and revenue potential (Sumadikarta & Abeiza, 2014).

In this research, we will in-depth explore how customer data analysis using data mining and the K-Means algorithm can help companies identify and target potential customers. Data has become important and valuable in the era of information technology because data is needed in formulating strategies and making decisions (Murpratiwi et al., 2021).

RESEARCH METHODS

This research uses a quantitative approach with the aim of identifying and classifying potential customers based on historical data using data mining techniques and the K-Means algorithm (Andi Cuhwanto & R, 2021) with a research framework as below:



Gambar 1. Kerangka Penelitian

Based on the framework in the image above, each step can be described as below:

1. Identify the Problem

   At this stage, a review of the system will be carried out to observe and carry out more

in-depth exploration and explore the problems that exist in the current system. This stage is the first step to determine the problem of the research.

2. Problem Analysis

   The problem analysis step is a step to understand the problem whose scope or boundaries have been determined, where:

   a. Preliminary analysis is limited to descriptive analysis to determine the variable characteristics of each data for the classification process according to the specified categories.
   b. Further analysis which aims to carry out calculations and test hypothesis results.

3. Determine goals

   Based on the problem formulation that has been created, the objective determination stage is useful for clarifying the framework of what is the target of this research.

4. Study Literature

   Through studying the literature, theories related to data mining were studied using the K-Means algorithm. Literature sources were obtained from journals and Proceedings.

5. Collect Data

   In collecting data, supporting methods are used, including:

   a. Observation: This is data collection carried out by making direct observations at the research site.
   b. Literature review: Collecting data and searching for information by reading and studying books and research journals that are related and supportive, both in analyzing data and information, as well as solving problems as a whole.

6. Conduct analysis using the K-Means algorithm

   After collecting data and analyzing the problem, an analysis is then carried out to determine the variables that will be used as input for data processing using the K-Means algorithm.

7. Research Results

   At this stage is the stage of testing research results in the application of customer selection data mining using the K-Means algorithm. The research results tested.

   a) Perform data processing using the K-Means algorithm manually.
   b) Carry out testing on the results of data processing manually using the K-Means algorithm.

RESULTS AND DISCUSSION

1. Test Data

The implementation process and data analysis were carried out using the Tanagra application program, using sample data, where the data source which was initially stored in a database using Microsoft Access was converted into an Excel application program document format. The overall data can be seen in the table below. (Rofianto et al., 2023)

Table 1. Overall Data

No	Name Customer	Amount Transaction	Total transaction
1	Fauzan Junanda Putra	18	260300
2	Ibu Ita	12	262200
3	Olifah	13	602900
4	Nb Ayu	20	486100
5	Riki	24	704300
6	Hhanifa	19	355500
7	Friska	20	810200
8	Putri Nb	11	437100
9	Nanda	12	249200
10	Dr Kiki	31	850900
11	Ibnu	12	182800
12	Lukfi	29	506300
13	Ami	16	448100
14	Andi	19	370100
15	Radi	11	559300
16	Megi	15	416300
17	Effendi	40	692400
18	Reza	14	372700
19	Rahmi Stifarm	26	800200
20	Rikafloo	23	628600
21	Indra	17	356800
22	Sarjan	19	845400
23	Alin	13	335600
24	Adit	16	259700
25	Apin	23	714000
26	Ferdi	11	253500
27	Bu Wat	26	843400
28	Dino Zanur	11	276300
29	Trisya	21	441000
30	Roy	25	566300

2. Potential Customer Selection Input Data

The data used in this application system is sample data from grouping potential customers using K-Means clustering. In this research, input data is taken from the number and total of transactions carried out by customers (Muzaqi et al., 2022), where the input pattern is as follows:

1. Number of Transactions
2. Total Transactions

The data used as input data for selecting potential customers is only taken from 20 customer data in the category of taking the number of transactions above 10 and total transactions below 900,000. All the data that has been obtained is then processed and analyzed for the problems that occur so as to produce information that is useful for overcoming problems and improvements can be proposed. This data processing uses data mining theories with the K-Means clustering algorithm. In the image below, a flowchart of the K-Means clustering algorithm that will be carried out in this research is shown, namely:

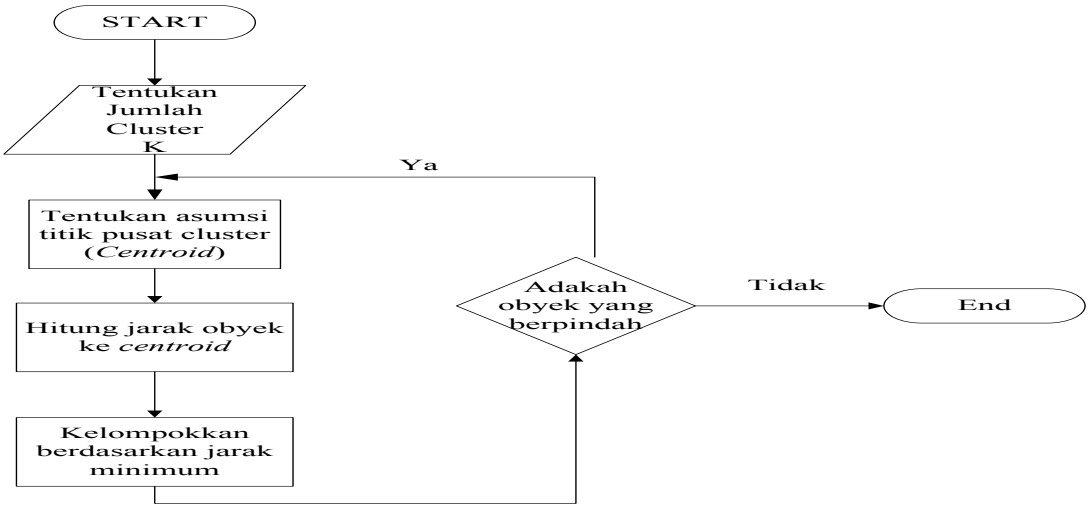


Figure 2. K-Means Clustering Flowchart

3. K-Means Method Computation Steps

To be able to group the data into several clusters, several steps need to be taken (Rianto et al., 2022), namely:

1. Determine the desired number of clusters. In this research, the existing data will be grouped into three clusters.
2. Determine the initial center point of the cluster (centroid). In this study, the initial

center point was determined randomly and the center point obtained from each cluster can be seen in the table below:

Table 2. Initial Center of Each Cluster

Data Ke i	***Centroid***	**Jumlah Transaksi**	**Total Tramnsaksi**
4	1	13	602,900.00
17	2	16	448,100.00

3. After determining the initial centroid, each data will find its closest centroid by calculating the distance of each data to each centroid using the correlation formula between two objects, namely Euclidean Distance. As for the initial manual centroid calculation, only 5 data samples were taken, namely as follows:

$C1_1 = \sqrt{(13-18)^2+(602900-260300)^2}$
$C1_1 = 342600$

$C1_2 = \sqrt{(13-12)^2+(602900-262200)^2}$
$C1_2 = 340700$

$C1_3 = \sqrt{(13-13)^2+(602900-602900)^2}$
$C1_3 = 0$

$C1_4 = \sqrt{(13-20)^2+(602900-486100)^2}$
$C1_4 = 116800$

$C1_5 = \sqrt{(13-24)^2+(602900-704300)^2}$
$C1_5 = 101400$

After the process of calculating the distance of the centroid data for each cluster with a total of 5 records, the distance of each centroid to the center of the group is obtained as shown in the table.

Table 3. Results of Iteration 1

No	Customer's name	Amount Transaction	Total Transaction	Iterasi 1		Distance	
				C1	C2	C1	C2
1	Fauzan J.P.	18	260300	457858.3	88011.11	0	1
2	Ibu Ita	12	262200	455958.3	86111.11	0	1
3	Olifah	13	602900	115258.3	254588.89	1	0
4	Nb Ayu	20	486100	232058.3	137788.89	0	1

5	Riki	24	704300	13858.33	355988.89	1	0
6	Hhanifa	19	355500	362658.3	7188.8907	0	1
7	Friska	20	810200	92041.67	461888.89	1	0
8	Putri Nb	11	437100	281058.3	88788.89	0	1
9	Nanda	12	249200	468958.3	99111.11	0	1
10	Dr Kiki	31	850900	132741.7	502588.89	1	0
11	Ibnu	12	182800	535358.3	165511.11	0	1
12	Lukfi	29	506300	211858.3	157988.89	0	1
13	Ami	16	448100	270058.3	99788.89	0	1
14	Andi	19	370100	348058.3	21788.89	0	1
15	Radi	11	559300	158858.3	210988.89	1	0
16	Megi	15	416300	301858.3	67988.89	0	1
17	Effendi	40	692400	25758.34	344088.89	1	0
18	Reza	14	372700	345458.3	24388.89	0	1
19	Rahmi S	26	800200	82041.67	451888.89	1	0
20	Rikafloo	23	628600	89558.33	280288.89	1	0
21	Indra	17	356800	361358.3	8488.8901	0	1
22	Sarjan	19	845400	127241.7	497088.89	1	0
23	Alin	13	335600	382558.3	12711.11	0	1
24	Adit	16	259700	458458.3	88611.11	0	1
25	Apin	23	714000	4158.33	365688.89	1	0
26	Ferdi	11	253500	464658.3	94811.11	0	1
27	Bu Wat	26	843400	125241.7	495088.89	1	0
28	Dino Z	11	276300	441858.3	72011.11	0	1
29	Trisya	21	441000	277158.3	92688.89	0	1
30	Roy	25	566300	151858.3	217988.89	1	0

In the table above you can see the rows that are blocked. The blocked rows are the rows that show the shortest distance to the center of the group. The next step is to create object membership into a matrix, with the matrix element having a value of 1 if an object is a member of the group. Based on this table, which is taken from the distance to the C1

and C2 values in the table above, an assignment table will be obtained as seen in the table below (Putra, Randi Rian & Nadya, 2022).

Table 4. Distance to Initial Centroid

Customer's name	**K-Means 0**	**K-Means 1**
Fauzan Junanda Putra	0	1
Ibu Ita	0	1
Olifah	1	0
Nb Ayu	0	1
Riki	1	0
Hhanifa	0	1
Friska	1	0
Putri Nb	0	1
Nanda	0	1
Dr Kiki	1	0
Ibnu	0	1
Lukfi	0	1
Ami	0	1
Andi	0	1
Radi	1	0
Megi	0	1
Effendi	1	0
Reza	0	1
Rahmi Stifarm	1	0
Rikafloo	1	0
Indra	0	1
Sarjan	1	0
Alin	0	1
Adit	0	1
Apin	1	0
Ferdi	0	1
Bu Wat	1	0
Dino Zanur	0	1
Trisya	0	1
Roy	1	0

Based on the minimum value that has been generated in the table above when determining the centroid value, the grouping results obtained (Putra et al., 2022) are as in the table below:

Table 5. Grouping Results

Group (Cluster)	Members of the group	Number of members
0	[3,5,7,10,15,17,19,20,22,25,27,30]	12
1	[1,2,4,6,8,9,11,12,13,14,16,18,21,23,24,26,28,29]	18

The next step is to place each object into a cluster based on the centroid value closest to the difference (distance). In the table above, the results obtained are: Cluster 1 members = {3, 5, 7, 10, 15, 17, 19, 20, 22, 25, 27, 30}, and cluster 2 = {1, 2, 4, 6, 8, 9, 11, 12, 13, 14, 16, 18, 21, 23, 24, 26, 28, 29}, so that a new centroid is obtained which can be seen in the table below:

Tabel 6. Nilai Centroid Baru Iterasi 1

Centroid	**Jumlah Transaksi**	**Total Transaksi**
C1	23.41667	718,158.33
C2	15.88889	348,311.11

Untuk kelanjutan nya dilanjutkan dengan iterasi yang ke 2 sama dengan cara iterasi 1.

Table 7. Research Results

Cluster	Research result	
	Centroid end	Customer
Cluster 0	3,5,7,10,15,17,19,20,22, 25,27,30	Number of members = 12 members Consist of: Olifah, Riki, Friska, Dr. Kiki, Radi, Effendi, Rahmi Stifarm, Rikafloo, Sarjan, Apin, Bu Wat and Roy.
Cluster 1	1,2,4,6,8,9,11,12,13,14,16 ,18,21,23,24,26,28,29	Number of members = 18 members Consist of: Fauzan Junanda Putra, Ibu Ita, NB Ayu, HHanifa, Putri NB, Nanda, Ibnu, Lukfi, Ami, Andi, Megi, Reza, Indra, Alin, Adit, Ferdi, Dino Zahur and Trisya.

CONCLUSION

1. The potential customer group with the highest centroid value from both clusters is in cluster 0. The number of customers in cluster 0 is 12 customers consisting of Olifah, Riki, Friska, Dr. Kiki, Radi, Effendi, Rahmi Stifarm, Rikafloo, Sarjan, Apin, Bu Wat and Roy.
2. Grouping data using the K-Means algorithm is done by determining the number of clusters, calculating the closest distance to the cluster center. The data with the closest distance represents the members of the cluster, recalculation is carried out until the

data does not move to another cluster, to minimize the objective function.

3. Potential customer data is obtained after the K-Means algorithm calculation is complete, the data with the largest centroid center is included in the most potential customers.

REFERENCE

Ahsina, N., Fatimah, F., & Rachmawati, F. (2022). Analisis Segmentasi Pelanggan Bank Berdasarkan Pengambilan Kredit Dengan Menggunakan Metode K-Means Clustering. *Jurnal Ilmiah Teknologi Infomasi Terapan*, *8*(3). https://doi.org/10.33197/jitter.vol8.iss3.2022.883

Andi Cuhwanto, Y. N., & R, D. A. (2021). Implementasi Data Mining Pemilihan Pelanggan Potensial Menggunakan Algoritma K-Means. *Petir*, *15*(1), 48–56. https://doi.org/10.33322/petir.v15i1.1358

Murpratiwi, S. I., Agung Indrawan, I. G., & Aranta, A. (2021). Analisis Pemilihan Cluster Optimal Dalam Segmentasi Pelanggan Toko Retail. *Jurnal Pendidikan Teknologi Dan Kejuruan*, *18*(2), 152. https://doi.org/10.23887/jptk-undiksha.v18i2.37426

Muzaqi, K. A., Junaidi, A., & Saputra, W. A. (2022). Klasifikasi Status Gizi Pada Lansia Menggunakan Learning Vector Quantization 3 (LVQ 3). *Journal of Dinda : Data Science, Information Technology, and Data Analytics*, *2*(1), 28–36. https://doi.org/10.20895/dinda.v2i1.272

Putra, Randi Rian & nadya, andhika putri. (2022). Implementasi sistem informasi perpustakaan dalam meningkatkan pelayanan dan struktur perpustakaan pada smp swasta pab 9 1. *Jaringan Sistem Informasi …*, *6*(1), 83–88. http://ojsamik.amikmitragama.ac.id/index.php/js/article/view/136

Putra, R. R., Putri, N. A., & Wadisman, C. (2022). Village Fund Allocation Information System for Community Empowerment in Klambir Lima Kebun Village. *Journal of Applied …*, *3*(2), 98–104. https://journal.yrpipku.com/index.php/jaets/article/view/681%0Ahttps://journal.yrpipku.com/index.php/jaets/article/download/681/467

Rianto, M., Rusdiah, R., & Ichwan, H. (2022). Penerapan Data Mining Dengan Metode Naïve Bayes Dan Learning Vector Quantization Credit Rating Dalam Memprediksi Kelayakan Pemberian Kredit Oleh PT. BPR Lebak Sejahtera. *Respati*, *17*(1), 69. https://doi.org/10.35842/jtir.v17i1.443

Rofianto, D., Arifin, O., & Widyawati, D. K. (2023). *Perbandingan Metode Klasifikasi Learning*

Vector Quantization Dengan Diskriminan Fisher Pada Data Bunga Iris. *7*(1), 44–49.

Romadhona, W., Indarmawan Nugroho, B., & Alim Murtopo, A. (2022). Implementasi Data Mining Pemilihan Pelanggan Potensial Menggunakan Algoritma K-Means. *Jurnal Minfo Polgan*, *11*(2), 100–104. https://doi.org/10.33395/jmp.v11i2.11797

Suarna, N., & Wijaya, Y. A. (2023). *ANALISA PENERAPAN METODE CLUSTERING K-MEANS UNTUK PENGELOMPOKAN DATA TRANSAKSI KONSUMEN (Studi Kasus : Cv . Mitra Indexindo Pratama)*. *7*(2), 1322–1328.

Sumadikarta, I., & Abeiza, E. (2014). Penerapan Algoritma K-Means Pada Data Mining Untuk Memilih Produk Dan Pelanggan Potensial. *Jurnal Satya Informatika*, *1*, 12–22. https://lppm.usni.ac.id/jurnal/Istiqomah-Sumadikarta-Evan-Abeiza.pdf