DATA MINING CUSTOMER CLUSTERING USING K-MEANS METHOD

Agus Iskandar
Universitas Nasional, Indonesia
Correspondence author email: Iskandaragus1005@gmail.com

Rifqi Aldy Al Hafizh Harahap
Universitas Nasional, Indonesia
Rifqialdy80@gmail.com

Achmad Gilang Ramadhan
Universitas Nasional, Indonesia
ramdhang701@gmail.com

ABSTRACT

The company recognizes the crucial role of customers in achieving business success and as the main source of revenue. Therefore, it is important for companies to understand the needs and desires of customers in order to build a mutually beneficial relationship. Customers have functional and emotional needs that they want to fulfill through the products or services they buy. Customer experience, both positive and negative, has a significant impact on satisfaction, loyalty and corporate image. This research faces the challenge of decreasing the number of customers who make purchases at these companies or service providers. To overcome this problem, companies need to adopt an effective market strategy to improve operational efficiency and better understand customer needs. One approach used is to understand customer needs through grouping, so that companies can develop products or services that are more suitable for each customer group. This helps improve the product's relationship with customer needs and provides services that match their expectations. Customer grouping was performed using the K-means algorithm, with 47 customers grouped based on relevant attributes. Determining the optimal number of clusters is done by comparing the performance of the clusters that are formed, and the results produce two new clusters with different numbers of customers. The K-means algorithm is implemented using the RapidMiner application to simplify the process. The final analysis shows that the second cluster has more customers than the first cluster. This research confirms the importance of understanding customer needs, classifying them appropriately, and taking effective actions to maintain customer satisfaction. The K-means algorithm and the RapidMiner application prove to be very useful in this process, enabling companies to strengthen customer relationships and create significant added value. The final results of this study indicate that the first cluster (cluster 0) contains 22 customers, while the second cluster (cluster 2) contains 25 customers. Therefore, the second cluster has a larger number of customers compared to the first cluster.

Keywords: Data Mining, K-Means Clustering, Customer Grouping.

INTRODUCTION

Customers refer to individuals or organizations that obtain products or services from a company or service provider. They play a very important role in business success because they are the main source of revenue for the company. Understanding customers well is key to building mutually beneficial relationships between companies and customers. Customers have specific needs and wants that they want to fulfill through the products or services they buy. These needs can be functional, such as the need for food or clothing, or emotional, such as the desire to feel valued or recognized. Knowing the needs and wants of customers helps companies design and offer appropriate solutions. Customer experience includes all interactions and contacts that occur between customers and companies, from the purchase process to after-sales service. A positive experience can increase customer satisfaction, strengthen loyalty, and encourage them to recommend the company to others. Conversely, a bad experience can cause customers to switch to competitors and harm the company's image.

The problem that occurs in this study is the reduction in customers shopping at a company or service provider where some customers argue that they are less interested in shopping at that place because there is no special behavior for customers such as differentiating discounts obtained between loyal customers and customers who only buy at that time, besides that restocks of goods that customers usually buy are sometimes not available so they are forced to move to another store or service provider, this certainly makes a big loss for the sales service provider so that a good market strategy is needed to increase operational efficiency and must also understand customer needs. By gaining a deep understanding of customer needs and preferences through clustering, companies have the opportunity to create new products or services that are more suitable for each customer group. In this way, companies can improve the connection between their products and customer needs, thereby creating significant added value for them, in addition to providing more suitable services and better meeting customer expectations. In addition, customer categorization also helps companies identify customers who may be dissatisfied or potentially switching to competitors. With this knowledge, companies can take corrective actions or more effective retention strategies to maintain customer satisfaction and prevent them from moving to competitors.

K-means is one of the popular Data Mining algorithms in data analysis used to cluster data into groups that have similarities based on their attributes and is the easiest algorithm to understand so that this method is very suitable for use in grouping customers in order to distinguish customers who are easy to move and those who stay so that they can easily increase sales turnover. The basic idea of the K-means algorithm is to partition the data into K groups, where K is the number of groups specified by the K-means algorithm.[1][2].

Some research has been done, namely previous research by Ridha Maya Faza Lubis, et al in 2023, in the research they made containing efforts to reduce the high mortality rate due to cervical cancer which continues to increase, clustering techniques are used to group data based on similar characteristics. K-Means algorithm with rapidminer tester application is used in this process. The final results show that cluster 1 has a larger amount of data. Of the total 72 cervical cancer data analyzed, only 28 of them were classified as cervical cancer patients, while the other 44 data did not fall into the category [3]. The next research was conducted by Rizki Muliono and Zulfikar Sembiring in 2019, they investigated issues related to the provision of allowances to lecturers who have compiled and completed teaching documents such as Syllabus, Course Contract, RPS, and RPP. The assessment is carried out by LP2MP by applying clustering techniques using the K-Means algorithm. This approach allows data to be grouped based on values that are closest to the relevant characteristics, making the process more effective. In the results of the study, the level of accuracy obtained had a difference of 53.33%[4]. Research using the same method was also conducted in 2019 by Dewinta Marthadinata Sinaga and her colleagues focusing on clustering the consumer price index using the K-Means algorithm. In the results of the study, it can be seen that there are 14 cities that are members of Cluster 1, while 29 cities are grouped into Cluster 2, and Cluster 3 consists of 23 cities.[5]. In 2022 a study was conducted by Ayu Pangestu and Taufik Ridwan, the study aims to classify water usage based on the volume of use in cubication. The goal is to provide information about water user groups based on water sales data at PAM Kerta Raharja. The method used in this research is using the K-Means algorithm. Data that has been processed using the Weka application will be grouped into saving, medium, and wasteful categories. From the calculation results with the K-Means algorithm, the centroid 0 value is obtained (46.6), centroid 1 is (13.6), and centroid 2 is (25.4). Cluster group 0 is a group of customers who fall into the wasteful category, consisting of 9 people. Cluster group 1 is a group of customers with moderate usage, while cluster group 2 is a group of customers who are economical.[6].

RESEARCH METHODS

Stages of Research

A series of steps or procedures that must be carried out in a study are referred to as research stages. The aim is to ensure the smoothness and success of the research and to ensure that all aspects of the research have been properly considered. The following is an illustration of the research stages used.

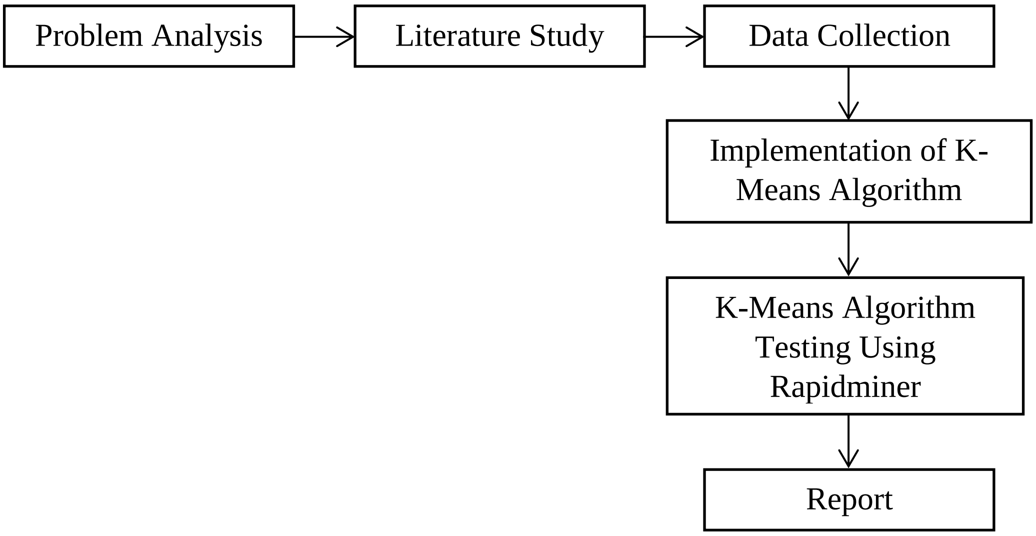

Problem Analysis
Literature Study
Data Collection
Implementation of K-Means Algorithm
K-Means Algorithm Testing Using Rapidminer
Report

Figure 1: Research Stages

a. Problem Analysis Stage

The problem analysis stage is carried out to understand and identify the root of the problem at hand. The purpose of this stage is to find the main cause of the problem in order to find an appropriate and effective solution. The problem analysis process involves data collection, data analysis, and data interpretation to identify the factors causing the problem and its impact on the affected system. Problem analysis aims to gain a thorough understanding of the problem at hand and the reasons for the problem. With an understanding of the root of the problem, relevant steps can be taken to resolve or address the issue at hand. Problem analysis plays an important role in directing problem-solving efforts appropriately and effectively by focusing on addressing the source of the problem, not just the symptoms.

b. Problem Analysis Stage

The problem analysis stage is carried out to understand and identify the root of the problem at hand. The purpose of this stage is to find the main cause of the problem in order to find an appropriate and effective solution. The problem analysis process involves data collection, data analysis, and data interpretation to identify the factors causing the problem and its impact on the affected system. Problem analysis aims to gain a thorough understanding of the problem at hand and the reasons for the problem. With an understanding of the root of the problem, relevant steps can be taken to resolve or address the issue at hand. Problem analysis plays an important role in directing problem-solving efforts appropriately and effectively by focusing on addressing the source of the problem, not just the symptoms.

c. Literature Study Stage

The next stage is a literature study or literature review which has an important role in carrying out quality research. This process involves collecting, analyzing, evaluating, and compiling various sources of information or relevant literature, such as books, journals, articles, reports, and documents related to the topic or problem under study. The purpose of the literature study stage is to gain a deeper understanding of the topic or problem being researched, as well as identify areas of knowledge that are still lacking or require further research.

d. Data Collection

At the data collection stage, researchers will carry out data acquisition in accordance with the methods that have been previously compiled. Methods used for data collection include interviews, surveys, observations, experiments, or document analysis, depending on the type of research being conducted..

e. K-Means Algorithm Implementation Stage

After analyzing the problem to collect data, the next step is to apply the algorithm to the data that has been collected. Algorithms are used to process data using predetermined methods. In this research, the K-Means clustering algorithm is used to solve existing problems.

f. K-Means Algorithm Testing Stage

After applying the K-Means algorithm, the next step is to evaluate the performance of the algorithm. The evaluation is done using RapidMiner platform. If the evaluation results are consistent with the previous algorithm implementation, then the evaluation is considered successful.

g. Report

In the final stage, the researcher will complete writing a research report that succinctly describes all stages and findings of the research. The research report should include an introduction, methods, results, analysis, and conclusion sections. The report should also be organized according to the format and writing style set by the institution or relevant research journal.

Data Mining

Data mining involves analyzing data using various machine learning techniques to automatically extract valuable knowledge. This process has great benefits for the future, as data mining forms patterns that can be used for decision making. The data used in data mining usually has a large volume (big data), making it important to use this technique. Data mining has different classifications, where each type of problem requires an appropriate approach. Some of the groups used in data mining are as follows[7][8][9][10][11]:

Clustering

Clustering is a method in data mining that is used to group data into groups or clusters based on certain characteristics or attributes. The goal is to identify natural patterns or hidden structures in the data without the use of prior categories or labels. The clustering process involves the use of algorithms that analyze the similarities or differences between data points. These algorithms attempt to group data into clusters so that data within a cluster has a high degree of similarity, while data between clusters has significant differences. Some commonly used methods in clustering are K-Means, AHC, K-Medoids, and others.[12][13][14]

Prediction

Prediction involves forecasting or estimating future values or events based on current data. The goal is to provide a reasonable or high-probability estimate of what might happen in the future. In data mining, prediction involves using statistical techniques, machine learning, or other predictive algorithms to identify patterns or trends in data that can be used to predict future values or events. Some commonly used algorithms for prediction are KNN, Naive Bayes, C4.5, Rough Set, SVM, and others.[15][10][16][17].

Association

Association is a concept in data mining that relates to the linkages or relationships between items or attributes in a dataset. Association analysis aims to uncover hidden patterns in data that show certain relationships or trends between the items. The process of association analysis involves finding itemsets, which are sets of items that frequently co-occur in transactions or events. Some of the methods used in association analysis are Apriori, Fp-Growth, and others.[11][18][19].

Classification

Classification is a process in data mining that is used to group data into predefined categories or classes. The goal is to identify patterns that distinguish data into various classes based on relevant attributes. Classification models are used to predict the appropriate class or label for data that does not yet have a class. Some commonly used classification methods are Cart, C4.5, ID3, K-NN, Naive Bayes, etc.[20][21][22].

Estimation

Estimation involves calculating or estimating an unknown value or amount based on available information or data. The goal is to provide an estimate that is close to the true value, although there is no absolute certainty. In data mining, estimation involves using statistical techniques or algorithms to calculate or predict unknown values based on available data. Estimation can be applied to various variables or parameters, such as income, population, expenditure, and others. Estimation is also widely used in various fields, such as economics, finance, social science, and other sciences. Some algorithms used in the estimation process include Expectation Maximization, Multiple Linear Regression, Simple Linear Regression, and others.[23][24].

K-Means Clustering

K-Means is one of the methods in data analysis and clustering that aims to group data into groups that have similarities based on given attributes. The K-Means algorithm operates by finding a group center called centroid, where data with certain similarities will be grouped together. The initial step involves determining the desired number of groups (k) and randomly determining the initial position of the centroid. After that, the algorithm will calculate the distance between each data and the centroid and group the data to the nearest centroid. Next, the centroid position is updated by taking the average of the data in each group. This process repeats until there is no change in the placement of the data or it reaches a pre-set stopping criterion. K-Means can be applied in various fields, such as market analysis, customer segmentation, pattern recognition, and data clustering in various applications.[25][26]. The following steps can be followed in clustering applying the K-Means method[27][28]:

1. Determine the number of clusters
2. In the first iteration, calculate the centroid center by using a randomly drawn data value as the initial centroid center.

   $$Ki = \frac{1}{M}\sum_{j=1}^{M} X_j \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (1)$$

   Use the initial centroid to calculate the closest distance

   $$d_{Euclidean}(X,Y) = \sqrt{\sum_{i=1}^{n}(X_i - Y_i)^2} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (2)$$

   Description t:

   d(x,y) = distance of data x to cluster center y
   Xi = i-th data on the nth data attribute
   Yi = jth data on nth data attribute

1. Data with the closest values will be grouped together in one cluster, while data that has a greater distance will be placed in a different cluster.
2. Perform the next iteration step using the new centroid position, which is determined based on the cluster with the minimum distance to the data. Keep repeating this process from step one until reaching the last iteration. If there is a

shift in the cluster centroid position, continue the iteration. However, if there is no displacement of the cluster centroid position, then the iteration process will be terminated.

Customer

Customers refer to individuals, organizations, or other entities that utilize the products or services offered by a company or service provider. They are people or groups who make purchases or have direct involvement with a particular business. Customers have the ability to directly acquire products or utilize services provided by companies with the aim of meeting their needs, desires, or overcoming problems they face. Customers have a significant role in maintaining company revenue and contribute importantly to maintaining business continuity. In addition, customers also have the ability to provide feedback, recommend products or services to others, and have an influence on the overall reputation of the company. In this case, maintaining a good relationship with customers is a key factor for the company's success in retaining existing customers and attracting new customers.[6][29][30][31][32].

RESULTS AND DISCUSSION

The process of customer clustering involves analyzing customer data to categorize them into different segments based on similar characteristics. One of the methods used for this clustering is the k-means method, where the first step is to determine the initial centroid randomly until the final process of forming clusters determined based on the first iteration to the nth iteration (the clustering search process stops if the cluster location in the previous iteration and the iteration after does not move, if it moves then continue the next penetration process). In this study, 2 clusters were formed, namely loyal customers and unfaithful customers. In order to deal with unfaithful customers, companies have the option to implement various successful marketing and customer retention strategies. This involves improving the quality of the product or service provided, providing additional value to customers who have been loyal, increasing the level of customer satisfaction, and developing loyalty programs that appeal to customers in order to retain them. The following is a table of customers with a total of 47 data that will be grouped with K-Means utilizing the rapidminer application.

Table 1. One Month Customer Data

Customer Code	Age	Gender	Number of Transactions
PLG0001	20	Woman	63
PLG0002	35	Woman	6
PLG0003	28	Man	40
PLG0004	40	Man	15

PLG0005	17	Woman	8
PLG0006	21	Woman	71
PLG0007	51	Man	42
PLG0008	33	Woman	55
PLG0009	30	Woman	90
PLG0010	26	Woman	22
PLG0011	29	Man	52
PLG0012	30	Woman	94
PLG0013	44	Man	25
PLG0014	31	Woman	11
PLG0015	24	Man	82
...	...	...	...
...	...	...	...
PLG0046	27	Man	16
PLG0047	23	Woman	53

Based on table 1, there are 4 attributes used in the clustering process, namely the age attribute which will be used as a determinant of products that are in accordance with the age range of customers, the second attribute is gender where at this attribute of course customers buy needs according to their gender such as cosmetic tools that are more dominantly purchased by female customers, attributes of female gender are given a value of 1 while male gender is given a value of 2. The last attribute is the number of transactions that customers have made during one month.

Application of K-Means Algorithm
Iteration 1

1. Number of clusters C=2 (C1 and C2)
2. Initial cluster centroid

Table 2. Initial Cluster Center (Initial Centroid)

Code	Age	Gender	Number of Transactions
PLG0002	35	1	6
PLG0019	26	2	70

3. Calculate the distance between the data and the cluster center with euclidian distance.

$$d_{Euclidean}(X,Y) = \sqrt{\sum_{i=1}^{n}(X_i - Y_i)^2}$$

Data 1:

$$C_1 = \sqrt{(20-35)^2 + (1-1)^2 + (63-6)^2 = 3264}$$

$$C_2 = \sqrt{(20-26)^2 + (1-2)^2 + (63-70)^2 = 56}$$

Data 2:

$$C_1 = \sqrt{(35 - 35)^2 + (1 - 1)^2 + (6 - 6)^2} = 0$$

$$C_2 = \sqrt{(35 - 26)^2 + (1 - 2)^2 + (6 - 70)^2} = 4106$$

Data 3:

$$C_1 = \sqrt{(28 - 35)^2 + (2 - 1)^2 + (40 - 6)^2 = 1164}$$

$$C_2 = \sqrt{(28 - 26)^2 + (2 - 2)^2 + (40 - 70)^2 = 902}$$

Data 4:

$$C_1 = \sqrt{(40 - 35)^2 + (2 - 1)^2 + (15 - 6)^2 = 87}$$

$$C_2 = \sqrt{(40 - 26)^2 + (2 - 2)^2 + (15 - 70)^2 = 3039}$$

Data 5:

$$C_1 = \sqrt{(17 - 35)^2 + (1 - 1)^2 + (8 - 6)^2 = 22}$$

$$C_2 = \sqrt{(17 - 26)^2 + (1 - 2)^2 + (8 - 70)^2 = 3854}$$

Perform calculations for customer data starting from the 6th data (code PLG0006) to the 47th data (code PLG0047) using the calculation to find the distance between the Euclidean distance clusters above. After the calculation is complete, here are the results of the closest distance found to the 47th data.

Table 3. Closest Distance (Iteration 1)

Code	C1	C2	Closest Distance	Cluster	
PLG0001	3264	56	56	C2	
PLG0002	0	4106	0		C1
PLG0003	1164	902	902	C2	
PLG0004	87	3039	87		C1
PLG0005	22	3854	22		C1
PLG0006	4239	7	7	C2	
PLG0007	1313	809	809	C2	
PLG0008	2403	233	233	C2	
PLG0009	7061	405	405	C2	
PLG0010	265	2305	265		C1
PLG0011	2123	327	327	C2	
PLG0012	7749	581	581	C2	
PLG0013	371	2043	371		C1
PLG0014	29	3487	29		C1
PLG0015	5788	146	146	C2	
...	...	...	...	...	...
...	...	...	...	...	...
PLG0046	109	2917	109		C1

PLG0047	2221	293	293	C2

4. Data that has the shortest distance will be grouped together in one cluster, while data that has a farther distance will be placed in another cluster.

5. Carry out the next iteration using the new centroid value, where the new centroid value is determined based on the cluster position with the minimum distance to the data obtained in table 3.$Ki = \frac{1}{M}\sum_{j=1}^{M} X_j$

C1

$Age = \frac{1}{21}(35 + 40 + 17 + 26 + 44 + 31 + 22 + 19 + 38 + 34 + 39 + 20 + 31 + 38 + 41 + 18 + 23 + 19 +$

$46 + 42 + 27 = 30{,}952$

$Gender = \frac{1}{21}(1 + 2 + 1 + 1 + 2 + 1 + 2 + 1 + 1 + 2 + 1 + 1 + 1 + 2 + 1 + 1 + 2 + 2 + 2 + 1 + 2$

$= 1{,}429$

$Number\ of\ Transactions = \frac{1}{21}(6 + 15 + 8 + 22 + 25 + 11 + 31 + 21 + 24 + 31 + 17 + 37 + 20 + 18 + 34 + 20 +$

$32 + 13 + 13 + 5 + 16 = 19{,}952$

Do the calculation according to the search for the new initial centroid C1 (age, gender and number of transactions) above against the new initial centroid in C2. After calculating the search for the new initial centroid on C2, the following initial centroid table is formed.

Table 4. New Initial Cluster Center (Iteration 2)

Cluster	Age	Gender	Number of Transactions
C1	30,952	1,429	19,952
C2	30,692	1,462	70,385

Table 4 is used as the initial cluster center which will function as the centroid in the second iteration (2). The calculation process is carried out by finding the closest distance value, following the same steps as in the first iteration (1). If a change occurs in the cluster grouping, the process will be repeated. However, if there is no change between the grouping results in the previous iteration and the next iteration, the process will stop.

Testing the K-Means Algorithm Using Rapidminer

RapidMiner provides a user-friendly and easy-to-use user interface with a drag-and-drop method, allowing users to create data analysis processes without the need to write code manually. With such an intuitive interface, even users without a programming background can easily perform data analysis. Apart from that, RapidMiner also provides a variety of powerful data analysis algorithms, including regression, classification, clustering, association analysis, and others. Users have the flexibility to choose the algorithm that best suits their needs and can easily compare the performance of different algorithms. To prepare data before conducting analysis, RapidMiner provides various preprocessing tools. These tools allow users to clean data, remove missing values, normalize data, merge and split columns, and perform other data transformations. With this efficient preprocessing process, users can quickly prepare data for further analysis. Apart from that, RapidMiner also provides a variety of powerful visualization features to help users analyze and understand data. Graphs, diagrams, and other visualizations make it easy for users to explore patterns and relationships in data in an intuitive way. This contributes to a better understanding of the data and can support better decision making as well. The following is a picture of the operator input process used and the performance assessment in determining the number of clusters to be formed.

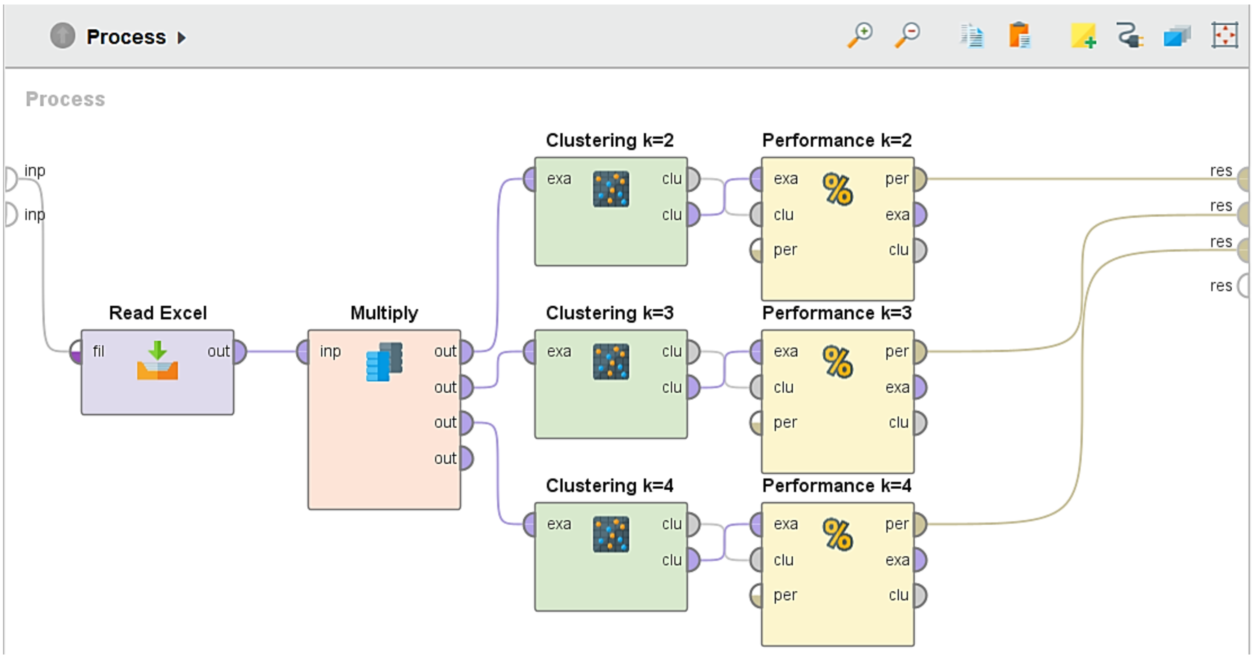
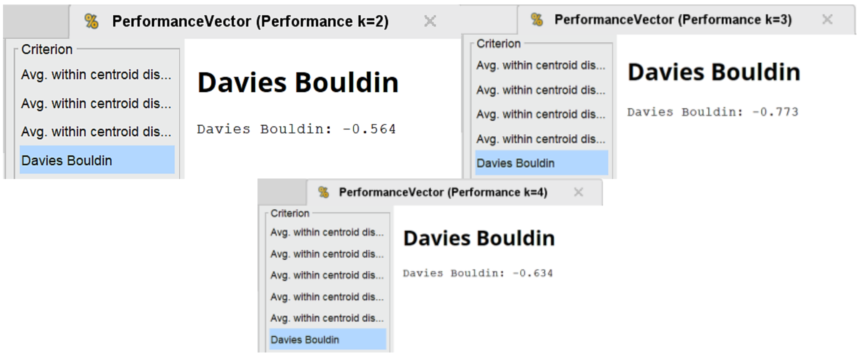


Figure 2. Clustering Operator Input

Based on Figure 2, it can be explained that before continuing the grouping process, the input data to be processed and the required operators need to be determined. One step is to determine the most effective number of clusters using three K-means clustering approaches (k=2 clusters, k=3 clusters, and k=4 clusters). To connect the three clustering operators with the input data, the multiply operator is used as seen in Figure 2. Next, the performance operator, namely Cluster Distance Performance, is used to see the most suitable and accurate number of clusters.

Connect the three performances according to the number of clusters that have been set in the clustering parameters, with the conditions k=2 (cluster k=2), k=3 (cluster k=3), and k=4 (cluster k=4). The three performance parameters (k=2, k=3, and k=4) in the main criterion section were selected using Davies Bouldin. After all operators are connected properly, run the process and the performance results can be seen in the image presented next.



PerformanceVector (Performance k=2)

Criterion
- Avg. within centroid dis...
- Avg. within centroid dis...
- Avg. within centroid dis...
- Davies Bouldin

Davies Bouldin

```
Davies Bouldin: -0.564
```

PerformanceVector (Performance k=3)

Criterion
- Avg. within centroid dis...
- Avg. within centroid dis...
- Avg. within centroid dis...
- Avg. within centroid dis...
- Davies Bouldin

Davies Bouldin

```
Davies Bouldin: -0.773
```

PerformanceVector (Performance k=4)

Criterion
- Avg. within centroid dis...
- Avg. within centroid dis...
- Avg. within centroid dis...
- Avg. within centroid dis...
- Avg. within centroid dis...
- Davies Bouldin

Davies Bouldin

```
Davies Bouldin: -0.634
```

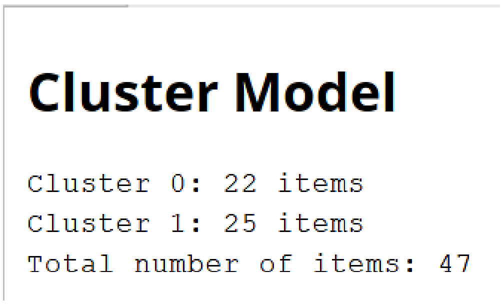
Figure 3. Performance for each number of clusters (K=2, K=3 and K=4)

In Figure 3, it can be seen that this research is more suitable for using 2 clusters, with a Davies Bouldin value of -0.564. Meanwhile, for performance with the formation of 3 clusters, the Davies Bouldin value is only -0.773, and performance with the formation of 4 clusters has a higher value than the performance for cluster 3, namely -0.634. Therefore, in applying the k-means algorithm for customer grouping, 2 clusters are used. The following is a table of the centroids for the final cluster formation obtained after running rapidminer.

Table 5. Table Centroid

Attribut	Cluster 0	Cluster 1
Age	29,318	32,12
Gender	1,455	1,44
Number of Transactions	75,273	23,72

Table 5 shows the final centroid value which indicates the iteration process has stopped with the centroid value in cluster 0 for the age section being 29.318, gender being 1.455 and the number of transactions being 75.273. Meanwhile, the final centroid for cluster 1 in the age section was 32.12, gender was 1.44 and number of transactions was 23.72. So a customer grouping is formed with 2 clusters, each of which can be seen in the following cluster model image.



Cluster Model

```
Cluster 0: 22 items
Cluster 1: 25 items
Total number of items: 47
```

Figure 4. Cluster Model

Figure 4 explains that 22 items (customers) are grouped into the first cluster (cluster 0) and 25 items (customers) are grouped into the second cluster (cluster 2), so that the second cluster accommodates more customer data (number of customers). In more detail, which customers are grouped into cluster 1 and cluster 2 can be seen in the following table.

Table 6. Clusters Formed

Customer Code	Age	Gender	Number of Transactions	Cluster
PLG0001	20	1	63	0
PLG0002	35	1	6	1
PLG0003	28	2	40	1
PLG0004	40	2	15	1
PLG0005	17	1	8	1
PLG0006	21	1	71	0
PLG0007	51	2	42	1
PLG0008	33	1	55	0
PLG0009	30	1	90	0
PLG0010	26	1	22	1
PLG0011	29	2	52	0
PLG0012	30	1	94	0
PLG0013	44	2	25	1
PLG0014	31	1	11	1
PLG0015	24	2	82	0
...	...	...	...	...
...	...	...	...	...
PLG0046	27	2	16	1
PLG0047	23	1	53	0

From observations in Table 6, it can be seen that the objects have been successfully grouped into groups or clusters that have similar attributes or similar characteristics. The table also provides information regarding the comparison of attributes of objects in one group, so that it can be seen to what extent the objects in the group are similar or homogeneous. To see the percentage of each cluster, see the following picture.

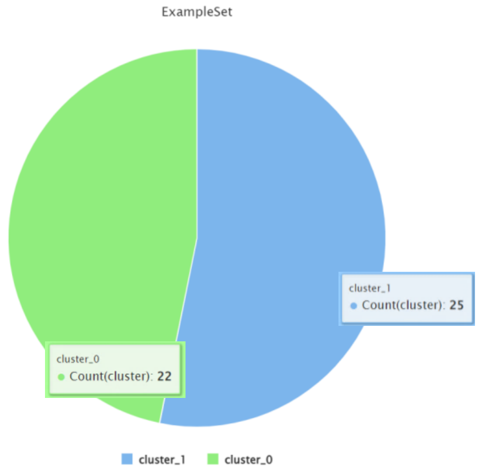


Figure 5. Cluster Visualization

From observations in Figure 5, it can be seen that cluster 0 is shown in green, while cluster 1 is shown in blue. Cluster 1, the first, consists of 22 customers, which accounts for approximately 46.81% of the total 47 customers in the data. On the other hand, cluster 1 has the largest number of customers, namely 25 customers, which covers around 53.19% of the total data.

CONCLUSION

Conclusions were made after applying the K-Means algorithm in grouping 47 customers with each customer grouped based on the similarity of the attributes used (there were 3 attributes used). The manual calculation process is carried out in only a few iterations and then continues using the rapidminer application. The process of determining the number of clusters that will be grouped has been successfully carried out using the rapidminer application by comparing the performance of each cluster that will be formed (k=2, k=3 and k=4), in this determination process, grouping customers that match the Davies Boulsing value the biggest is forming 2 new clusters. The final result after carrying out all the steps in implementing both the algorithm and using the application, was that 22 items (customers) were placed in the first cluster (cluster 0) with a percentage of 46.81%, while 25 items (customers) were placed in the second cluster (cluster 1) with a percentage of 53.19%. So the final result obtained is that the second cluster has more customer data (number of customers) compared to the first cluster.

REFERENCES

[1] M. Marsono, D. Saripurna, and M. Zunaidi, “Analisis Data Mining Pada Strategi Penjualan Produk PT Aquasolve Sanaria Dengan Menggunakan Metode K-Means Clustering,” *J-SISKO TECH (Jurnal Teknol. Sist. Inf. dan Sist. Komput. TGD)*, vol. 4, no. 1, p. 127, 2021, doi: 10.53513/jsk.v4i1.60.

[2] A. Sulistiyawati and E. Supriyanto, “Implementasi Algoritma K-means Clustring dalam Penetuan Siswa Kelas Unggulan,” *J. Tekno Kompak*, vol. 15, no. 2, pp. 25–36, 2021.

[3] R. M. F. Lubis, J.-P. Huang, P.-C. Wang, K. Khoifin, M. Sigiro, and J. Panjaitan, “Data Clustering Mining Applying the K-Means Algorithm, Cervical Cancer Behavior Risk,” *J. MEDIA Inform. BUDIDARMA*, vol. 7, no. 2, pp. 819–827, 2023.

[4] R. Muliono and Z. Sembiring, “Data Mining Clustering Menggunakan Algoritma K-Means Untuk Klasterisasi Tingkat Tridarma Pengajaran Dosen,” *CESS (Journal Comput. Eng. Syst. Sci.*, vol. 4, no. 2, pp. 272–279, 2019.

[5] D. M. Sinaga, A. P. Windarto, D. Hartama, and S. Saifullah, “Pengelompokkan Indeks Harga Konsumen Menurut Kota Dengan Datamining Clustering,” in *Seminar Nasional Sains dan Teknologi Informasi (SENSASI)*, 2019, vol. 2, no. 1.

[6] 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. 12, no. 3, pp. 67–71, 2022.

[7] A. S. L. T. T. H. Hafizah, “Data Mining Estimasi Biaya Produksi Ikan Kembung Rebus Dengan Regresi Linier Berganda,” *J. Sist. Inf. Triguna Dharma (JURSI TGD)*, no. Vol 1, No 6 (2022): EDISI NOVEMBER 2022, pp. 888–897, 2022, [Online]. Available: https://ojs.trigunadharma.ac.id/index.php/jsi/article/view/5732/1938

[8] Y. L. Nainel, E. Buulolo, and I. Lubis, “Penerapan Data Mining Untuk Estimasi Penjualan Obat Berdasarkan Pengaruh Brand Image Dengan Algoritma Expectation Maximization (Studi Kasus: PT. Pyridam Farma Tbk),” *JURIKOM (Jurnal Ris. Komputer)*, vol. 7, no. 2, p. 214, 2020, doi: 10.30865/jurikom.v7i2.2097.

[9] M. Azhari, Z. Situmorang, and R. Rosnelly, “Perbandingan Akurasi, Recall, dan Presisi Klasifikasi pada Algoritma C4.5, Random Forest, SVM dan Naive Bayes,” *J. Media Inform. Budidarma*, vol. 5, no. 2, p. 640, 2021, doi: 10.30865/mib.v5i2.2937.

[10] S. Widaningsih, “Perbandingan Metode Data Mining Untuk Prediksi Nilai Dan Waktu Kelulusan Mahasiswa Prodi Teknik Informatika Dengan Algoritma C4,5, Naïve Bayes, Knn Dan Svm,” *J. Tekno Insentif*, vol. 13, no. 1, pp. 16–25, 2019, doi: 10.36787/jti.v13i1.78.

[11] H. Maulidiya and A. Jananto, “Asosiasi Data Mining Menggunakan Algoritma Apriori dan FP-Growth sebagai Dasar Pertimbangan Penentuan Paket Sembako,” *Proceeding SENDIU 2020*, vol. 6, pp. 36–42, 2020.

[12] F. Harahap, “Perbandingan Algoritma K Means dan K Medoids Untuk Clustering Kelas Siswa Tunagrahita,” *TIN Terap. Inform. Nusant.*, vol. 2, no. 4, pp. 191–197, 2021.

[13] M. A. Rofiq, A. Qoiriah, S. Kom, and M. Kom, “Pengelompokan Kategori Buku Berdasarkan Judul Menggunakan Algoritma Agglomerative Hierarchical Clustering Dan K-Medoids,” *J. Informatics Comput. Sci.*, vol. 2, no. 03, pp. 220–

227, 2021.

[14] B. Harli Trimulya Suandi As and L. Zahrotun, “PENERAPAN DATA MINING DALAM MENGELOMPOKKAN DATA RIWAYAT AKADEMIK SEBELUM KULIAH DAN DATA KELULUSAN MAHASISWA MENGGUNAKAN METODE AGGLOMERATIVE HIERARCHICAL CLUSTERING (Implementation Of Data Mining In Grouping Academic History Data Before Students And Stud,” *J. Teknol. Informasi, Komput. dan Apl.*, vol. 3, no. 1, pp. 62–71, 2021, [Online]. Available: http://jtika.if.unram.ac.id/index.php/JTIKA/

[15] M. M. Effendi, “Menentukan Prediksi Kelulusan Siswa Dengan Membandingkan Algoritma C4. 5 Dan Naive Bayes Studi Kasus SMKN. 1 Cikarang Selatan,” *J. SIGMA*, vol. 11, no. 3, pp. 143–148, 2020.

[16] S. U. Putri, E. Irawan, and F. Rizky, “Implementasi Data Mining Untuk Prediksi Penyakit Diabetes Dengan Algoritma C4. 5,” *Kesatria J. Penerapan Sist. Inf. (Komputer dan Manajemen)*, vol. 2, no. 1, pp. 39–46, 2021.

[17] S. Widaningsih, “Perbandingan Metode Data Mining Untuk Prediksi Nilai Dan Waktu Kelulusan Mahasiswa Prodi Teknik Informatika Dengan Algoritma C4, 5, Naïve Bayes, Knn Dan Svm,” *J. Tekno Insentif*, vol. 13, no. 1, pp. 16–25, 2019.

[18] H. Maulidiya and A. Jananto, “Asosiasi Data Mining Menggunakan Algoritma Apriori Dan Fpgrowth Sebagai Dasar Pertimbangan Penentuan Paket Sembako,” 2020.

[19] K. Erwansyah, B. Andika, and R. Gunawan, “Implementasi Data Mining Menggunakan Asosiasi Dengan Algoritma Apriori Untuk Mendapatkan Pola Rekomendasi Belanja Produk Pada Toko Avis Mobile,” *J. Teknol. Sist. Inf. dan Sist. Komput. TGD*, vol. 4, no. 1, pp. 148–161, 2021.

[20] A. Damuri, U. Riyanto, H. Rusdianto, and M. Aminudin, “Implementasi Data Mining dengan Algoritma Naïve Bayes Untuk Klasifikasi Kelayakan Penerima Bantuan Sembako,” *JURIKOM (Jurnal Ris. Komputer)*, vol. 8, no. 6, pp. 219–225, 2021.

[21] I. A. Nikmatun and I. Waspada, “Implementasi Data Mining untuk Klasifikasi Masa Studi Mahasiswa Menggunakan Algoritma K-Nearest Neighbor,” *Simetris J. Tek. Mesin, Elektro dan Ilmu Komput.*, vol. 10, no. 2, pp. 421–432, 2019.

[22] H. Hozairi, A. Anwari, and S. Alim, “Implementasi Orange Data Mining Untuk Klasifikasi Kelulusan Mahasiswa Dengan Model K-Nearest Neighbor, Decision Tree Serta Naive Bayes,” *Netw. Eng. Res. Oper.*, vol. 6, no. 2, pp. 133–144, 2021.

[23] A. Rivandi, E. Bu’ulolo, and N. Silalahi, “Penerapan Metode Regresi Linier Berganda Dalam Estimasi Biaya Pencetakan Spanduk (Studi Kasus: PT. Hansindo Setiapratama),” *Pelita Inform. Inf. dan Inform.*, vol. 7, no. 3, pp. 263–268, 2019.

[24] P. Purwadi, P. S. Ramadhan, and N. Safitri, “Penerapan Data Mining Untuk Mengestimasi Laju Pertumbuhan Penduduk Menggunakan Metode Regresi Linier Berganda Pada BPS Deli Serdang,” *J. SAINTIKOM (Jurnal Sains Manaj. Inform. dan Komputer)*, vol. 18, no. 1, pp. 55–61, 2019.

[25] . F., F. T. Kesuma, and S. P. Tamba, “Penerapan Data Mining Untuk Menentukan Penjualan Sparepart Toyota Dengan Metode K-Means Clustering,” *J. Sist. Inf. dan Ilmu Komput. Prima(JUSIKOM PRIMA)*, vol. 2, no. 2, pp. 67–72, 2020, doi: 10.34012/jusikom.v2i2.376.

[26] S. A. Rahmah, “KLASTERISASI POLA PENJUALAN PESTISIDA MENGGUNAKAN METODE K-MEANS CLUSTERING (STUDI KASUS DI TOKO JUANDA TANI KECAMATAN HUTABAYU RAJA),” vol. 1, no. 1, pp. 1–5, 2020.

[27] M. A. K-means, “1 , 2 , 3 1,” vol. 1, no. 2, pp. 161–166, 2021.

[28] W. Purba, W. Siawin, and . H., “Implementasi Data Mining Untuk Pengelompokkan Dan Prediksi Karyawan Yang Berpotensi Phk Dengan Algoritma K-Means Clustering,” *J. Sist. Inf. dan Ilmu Komput. Prima(JUSIKOM PRIMA)*, vol. 2, no. 2, pp. 85–90, 2019, doi: 10.34012/jusikom.v2i2.429.

[29] H. A. Wijaya, W. Suharso, and Y. Azhar, “PENERAPAN FREQUENCY, RECENCY, MONETERY MODEL DAN ALGORITMA K-MEAN PADA SISTEM PENGELOMPOKAN PELANGGAN”.

[30] D. K. Gultom, M. Arif, and M. Fahmi, “Determinasi kepuasan pelanggan terhadap loyalitas pelanggan melalui kepercayaan,” *Maneggio J. Ilm. Magister Manaj.*, vol. 3, no. 2, pp. 171–180, 2020.

[31] I. C. Saragih, D. Hartama, and A. Wanto, “Prediksi Perkembangan Jumlah Pelanggan Listrik Menurut Pelanggan Area Menggunakan Algoritma Backpropagation,” *Build. Informatics, Technol. Sci.*, vol. 2, no. 1, pp. 48–53, 2020.

[32] A. Syafii, G. Dwilestari, and A. Ajiz, “KOMPARASI ALGORITMA NAÏVE BAYES DAN ALGORITMA C4. 5 DALAM KLASIFIKASI PELANGGAN PRODUK INDIHOME”.