MOBILE AND FORENSICS

Vol. 8, No. 2, 2026, pp. 93-109 | DOI: 10.12928/mf.v8i2.15807

ISSN: 2714-6685 (Online)
ISSN: 2656-6257 (Print)

Combining RFM-Based Customer Segmentation and K-Means Clustering to Optimize E-Commerce Marketing Strategies

Baiq Nikum Yulisasih[a,1,*], Panggah Widiandana[a,2], Muhammad Immawan Aulia[a,3], Rizky Andhika[a,4], Siti Hartinah[a,5]

[a] Department of Informatics, Faculty of Science, Technology, and Health, Universitas Islam Mulia Yogyakarta, Indonesia
[1] baiqnikumyulisasih@gmail.com*; [2] panggah.widiandana@uim-yogya.ac.id; [3] muhimmawanaulia16@uim-yogya.ac.id; [4] rizky.andhika@uim-yogya.ac.id; [5] siti.hartinah@uim-yogya.ac.id

* Corresponding Author

ARTICLE INFO

ARTICLE HISTORY

Received: Feb 22, 2026
Revised: Jun 24, 2026
Accepted: Jul 05, 2026

KEYWORDS

Customer Segmentation, Customer Profiling, K-Means Clustering, RFM Model, E-commerce

ABSTRACT

This research examines the shift of consumers from offline shopping to online shopping through e-commerce platforms, driven by convenience, product variety, discounts, and rapid technological advancements such as the internet, smartphones, big data, and machine learning. The study utilizes a transactional dataset containing attributes such as InvoiceNo, StockCode, Description, Quantity, InvoiceDate, UnitPrice, CustomerID, and Country to perform customer segmentation. The segmentation process is conducted using the K-means clustering algorithm combined with the RFM (Recency, Frequency, Monetary) model to develop comprehensive customer profiles. This approach aims to identify distinct customer groups and support the creation of more targeted and personalized marketing strategies. The quality of clustering results is evaluated using the Silhouette Score to ensure optimal grouping performance. The entire data processing, analysis, and modeling are implemented using Python as the primary tool. The research methodology includes literature review, data collection, data preprocessing, clustering model development, customer profile analysis, and evaluation and validation of results. The findings are expected to provide valuable insights into customer behavior, enabling e-commerce businesses to enhance marketing effectiveness, improve customer retention, and attract new customers. Ultimately, this study contributes to the development of data-driven marketing strategies that improve customer satisfaction and business performance in the digital commerce environment.

1. INTRODUCTION

The e-commerce ecosystem itself consists of two main categories of transactions, namely business-to-business (B2B) and business-to-consumer (B2C) [1]. B2B transactions involve relationships between companies for the supply of goods or services, while B2C transactions focus more on direct sales to individual consumers, which generally dominate retail activities on e-commerce platforms. Popular marketplaces such as Shopee, Tokopedia, and Alibaba are concrete examples of B2C platforms that have successfully attracted millions of customers through innovative marketing strategies, efficient logistics integration, and easily accessible payment systems. The popularity of these platforms has increased with features such as free shipping, flash sales, and huge discounts that provide added value to customers [2]. However, amid this rapid growth, competition among e-commerce platforms has also become fiercer. Customers now tend to be more flexible in choosing platforms, easily switching from one platform to another, especially if they find more attractive offers or a better user experience elsewhere.

In this increasingly competitive environment, a company's ability to understand its customers in depth is key to achieving competitive advantage. With the increasing volume of customer data available, including transaction data, online shopping behavior, and individual preferences, companies have a great opportunity to create more targeted and effective marketing strategies [3]. One of the main approaches widely used is customer segmentation, which is the process of dividing the market into several segments based on similar characteristics. This segmentation allows companies to target each segment with specific and personalized marketing strategies, providing a relevant experience for customers. These principles of data-driven personalization and segmentation also apply in the world of e-commerce, where determining optimal marketing strategies requires a comprehensive consumer segmentation model one example being the RFM (Recency, Frequency, Monetary) model. In this study, customer segmentation was conducted using the RFM (Recency, Frequency, Monetary) model, a method that has been proven effective in understanding customer value based on three main dimensions [4], [5].

The first attribute is Recency (R), which measures how recent the customer's last transaction was. Recency reflects the level of customer engagement with the platform, providing an indication of their potential to make transactions in the future. The second attribute is Frequency (F), which calculates how often customers make transactions within a certain period. This transaction frequency is used to assess the level of customer loyalty to the platform. The last attribute is Monetary (M), which measures the total value of customer purchases within a certain period, thus providing an indication of the customer's financial contribution to the Company's revenue [6]. By using the RFM model, companies can group customers into categories such as high-value customers who make a significant contribution, new customers who have the potential to be further engaged, or at-risk customers who show a decline in activity and require special attention through retention strategies [7]. To improve the effectiveness of customer segmentation, this study integrates the RFM model with the K-Means algorithm, which is a fast and efficient data clustering method. The K-Means algorithm works by grouping customer data into clusters based on similarities in RFM dimension values.

Previous studies have shown the superiority of k-means in various customer segmentation contexts, demonstrating that this algorithm is faster and more efficient than agglomerative clustering methods in terms of segmentation [8]. This study examines the factors that influence purchasing decisions for online fashion products in Indonesia. The results show that product quality and trust drive loyalty, while Brand Ambassadors have no effect. Loyalty links product quality to purchasing decisions, but not trust to purchasing [9]. This study reveals a positive relationship between ecommerce (B2C, B2B, B2G), innovation, and economic development in European Union countries, with countries such as Bulgaria and Romania requiring more attention on this issue. Additionally, the combination of a new clustering method by integrating the RFM model and formal concept analysis (FCA) provides clearer insights into customer relationships, helping marketers design more effective strategies [10], [11]. This study shows that e-commerce increases corporate innovation, with increased awareness as the main factor (18.44%), followed by income incentives (14.59%) and competition pursuit (11.32%) [12].

However, previous studies on customer segmentation using RFM and K-Means clustering still exhibit several limitations [13]. Most studies rely on the elbow method to determine the optimal number of clusters, which does not adequately evaluate cluster quality. Additionally, many

approaches apply RFM and K-Means without incorporating validation metrics such as the Silhouette Score, resulting in less reliable segmentation outcomes [14], [15]. This study addresses these gaps by integrating RFM analysis with K-Means clustering and evaluating the clustering results using the Silhouette Score. This approach provides a more robust and interpretable customer segmentation model that can better support data-driven marketing strategies in e-commerce [16], [17].

However, previous research still has limitations, particularly in comprehensively examining the relationships among the variables under study within a more specific context. Most studies focus only on a single aspect without considering the interplay among variables, which could yield more in-depth results. Therefore, the research gap in this study lies in the lack of studies that integrate these variables into a single, comprehensive analytical framework. This study offers a novel approach that is more comprehensive by combining several variables simultaneously and applying them to a context more relevant to current conditions.

Based on this, this study aims to analyze to produce more relevant customer segments, provide strategic insights for companies to design more personalized marketing strategies, increase customer loyalty, and ultimately, provide significant added value for e-commerce companies in an increasingly competitive digital era and an effective approach to customer segmentation.

2. METHOD

2.1. Research Object

This study collected data by searching for a customer segmentation dataset on the Kaggle website by Faris Adhi as the research object. The dataset description is presented in Table 1.

Table 1. Customer segmentation dataset details

Attribute	Description
InvoiceNo	Unique invoice number for each transaction
StockCode	Unique code for each product purchased
Description	Product description
Quantity	Number of product unit purchased in each transaction
InvoiceDate	Date and time when the transaction was made
UnitPrice	Unit price of the product
CustomerID	Unique ID for each customer

This dataset has great potential for use in customer profiling and segmentation, given that the available attributes enable in-depth analysis of purchasing behavior. One analysis that can be performed is RFM Analysis (Recency, Frequency, Monetary). The Recency (R) attribute is measured based on the Invoice Date, which shows when the customer last made a purchase. Frequency (F) is calculated from the number of Invoice Nos, which reflects how often a customer makes a purchase. Meanwhile, Monetary (M) is measured based on Quantity and UnitPrice, which describe the customer's total expenditure. In addition, the segmentation method used is K-Means Clustering, which groups customers based on purchasing behavior, such as purchase frequency or total purchase value. The Country attribute can also be used for Geographic Segmentation, dividing customers based on geographic region, which can provide important insights for more targeted regional marketing strategies.

2.2. Research Framework

The research framework as show in Fig. 1, used in this study was designed to describe the main stages in the process of analyzing e-commerce customer data. Starting from customer data acquisition, this framework covers the pre-processing stage to clean and prepare the data, the development of a clustering model using the k-means algorithm, to statistical analysis to understand customer profiles based on various dimensions, such as demographics, geography, and RFM (Recency, Frequency, Monetary) metrics. These stages not only produce more targeted customer segmentation, but also support the optimization of marketing strategies focused on increasing customer satisfaction.

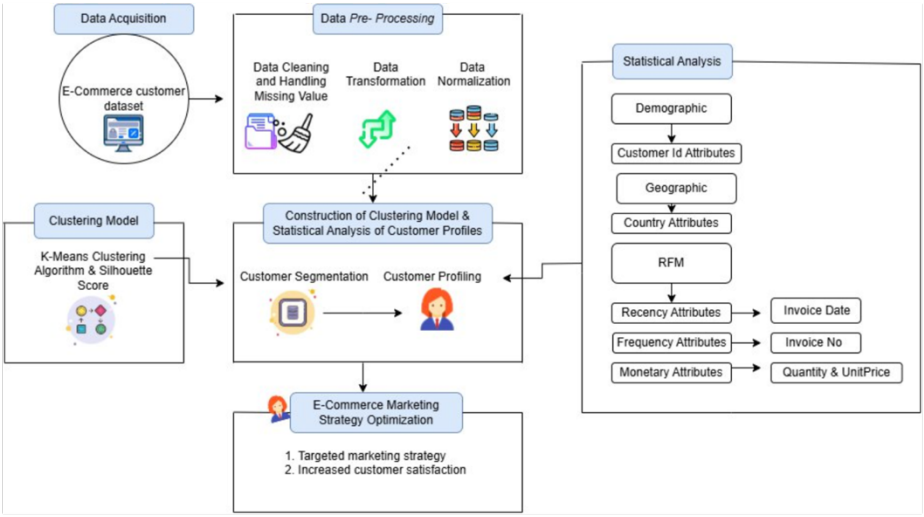

Fig. 1. Framework

This study focuses on e-commerce customer segmentation to optimize marketing strategies through clustering analysis using the K-Means Clustering algorithm. Fig. 1 shows the research framework that focuses on e-commerce customer segmentation to optimize marketing strategies through clustering analysis using the K-Means Clustering algorithm. The research framework begins with the data acquisition process, where e-commerce customer datasets are collected. The data is then processed through data pre-processing stages, which include data cleaning, handling missing values, data transformation, and normalization to ensure the data is ready for analysis. After the data is processed, a clustering model is built using the K-Means algorithm, which aims to group customers based on demographic characteristics and other indicators. This process resulted in customer segmentation that helped identify the unique profiles of each customer group. These customer profiles were further analyzed using statistical analysis, which included demographic, geographic, and RFM (Recency, Frequency, Monetary) factors to enrich the understanding of each cluster. In the RFM model, the Recency (R) attribute is calculated by measuring the difference between the date of the last transaction made by each customer and the date of analysis. The Frequency (F) attribute is calculated by counting the number of transactions (InvoiceNo) made by each customer.

2.3. Data Pre-processing

This pre-processing stage produces RFM values that will be used for further analysis, such as customer segmentation using the K-Means Clustering method. This process involves checking for missing values and normalizing the data. Checking for missing values is the process of identifying and handling empty data in the dataset, while data normalization (min-max) is a step to change the scale of variable values to make them uniform. Both steps are important in the application of KMeans Clustering. Missing value checking ensures data integrity before analysis, while normalization helps overcome scaling issues that can affect clustering results. Min-max normalization is a method of changing the data scale so that it is within a certain range, usually between 0 and 1. To adjust the scale of the RFM attributes so that no single attribute dominates the clustering process due to differences in value ranges, data normalization is conducted using the Min-Max normalization method. The mathematical formula for Min-Max normalization is expressed in Equation (1): [15], [18].

$$X' = X - Xmin \,/\, Xmax - Xmin \quad (1)$$

Where X' represents the normalized data value, X is the original data value, X_min s the minimum value of the attribute, and X_max s the maximum value of the attribute. This process ensures all values are transformed into a standard range between 0 and 1, making them suitable for the subsequent distance calculations.

2.4. Clustering Model Creation

The next process after pre-processing is the application of k-means clustering with the silhouette score method. K-means clustering is the most well-known algorithm in clustering techniques for dividing data into several groups. K-means was first introduced by Macqueen. K-means is also one

of the most prominent clustering techniques in science and technology [19]. Evaluation using the silhouette score method to evaluate the quality of cluster formation. The silhouette score value ranges from -1 to 1 [20]. A higher value indicates that the data point is closer to its cluster centroid than to other cluster centroids. The K-Means Clustering flowchart can be seen in Fig. 2 [21].

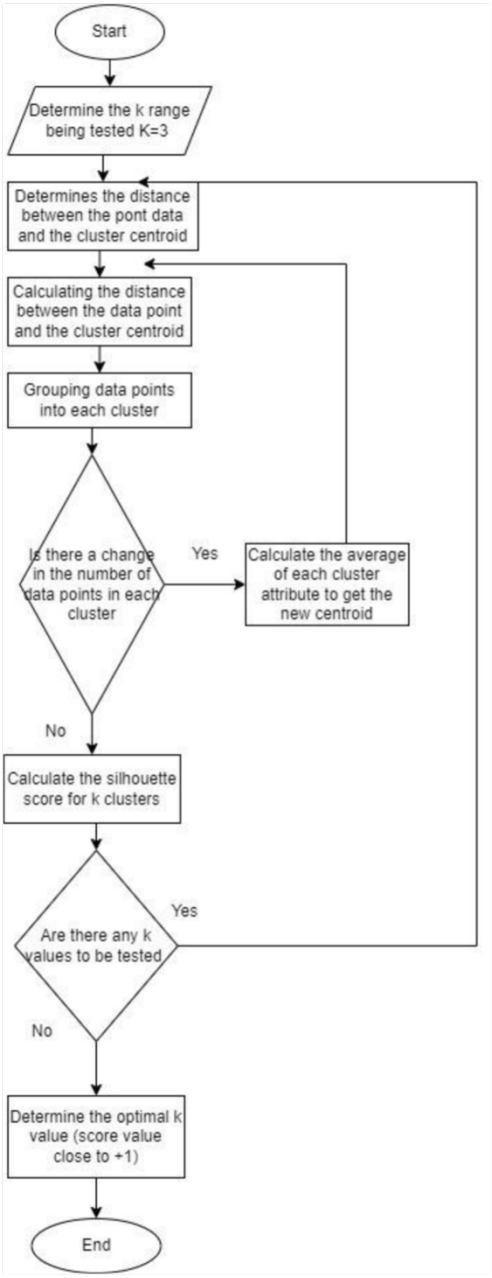

Fig. 2. Flowchart K-Means

Fig. 2 explains that the k-means clustering algorithm begins by determining the number of clusters (k), which in this case is set to 3 clusters. Next, three initial centroids are randomly selected from the available data points, and in this implementation, the initial centroids are taken from data points with customer IDs 17553, 17511, and 18079 To calculate the geometric distance between each customer data point and the cluster centroids, the Euclidean distance method is applied. The mathematical formula for the Euclidean distance is expressed in equation (2) is used [22].

$$d\,(x,y) = \sqrt{\sum_{i=1}^{n}(y_i - x_i)^2} \qquad (2)$$

Where d (x,y) is the distance from data point x to centroid y, y_i is the value of the i-th attribute of centroid y, and x_i is the value of the i-th attribute of data point x, where i is the attribute order and

n is the number of attributes of the data point. After calculating the distance of each data point, they are then grouped based on the closest distance between the first data point and the three cluster centroids. The data points are then grouped into clusters based on the closest distance to the centroid. Next, the grouping results of each cluster in iteration 1 will be compared with each cluster grouping result in iteration 2. However, before iteration 2 grouping, a new centroid is first sought by calculating the average value of all data points from each cluster. After the new centroid value for each cluster is obtained, the Euclidean distance is calculated in the same way as in the previous process.

Then, iteration 2 grouping is continued. The grouping results are compared with the grouping results of iteration 1. Then, iteration 2 grouping is continued. The grouping results are compared with the grouping results of iteration 1. If there are changes in the composition of data points or centroid values, the process continues to the next iteration. Conversely, if no changes occur in the data point composition, the algorithm has converged. To demonstrate the clustering mechanism prior to the complete algorithmic execution, a manual calculation of the distance matrix is performed using the initial centroids, as presented in equation (3). Based on the coordinates of the initial centroids and the data from the first customer, the calculation for the distance matrix is initialized as follows:

$$d(x_i, c_k) = (x_{i1} - c_{k1})^2 + (x_{i2} - c_{k2})^2 + (x_{nm} - c_k)^2 \quad (3)$$

$$d(1,1) = \sqrt{(0 - 0.346)^2 + (0.0097 - 0.00485)^2 + (0.6012 - 0.0053)^2} = d(1,1) = 0.689$$

$$d(1,2) = \sqrt{(0 - 0.0042)^2 + (0.0097 - 0.1504)^2 + (0.6012 - 0.3249)^2} = d(1,2) = 0.310$$

$$d(1,3) = \sqrt{(0 - 0.1194)^2 + (0.0097 - 0.0194)^2 + (0.6012 - 0.0130)^2} = d(1,3) = 0.600$$

The distance calculation (d) of the first data point from the initial centroid consisting of 3 attributes is grouped into cluster 1, because the distance of the first data point from the first centroid of cluster 1 produces the closest distance d(1,1) with a smaller value of 0, while the distance between the first data point and the second centroid of cluster 2, d(1,2), is 1, and the distance between the first data point and the third centroid of cluster 3, d(1,3), is 1. The process of calculating the distance for the second data point to the last data point with the centroids of clusters 1 to 3 uses the same process as the calculation of d(1,1) to d(1,3). The results of the overall distance calculation produce clusters in the first iteration.

The data is then grouped into clusters based on the closest distance to the centroid. These steps are repeated in the next iteration, updating the centroid of each cluster, until there is no change in the centroid. After that, if there is no change in the centroid, the next step is to calculate the Silhouette Score for each k value. The Silhouette Score measures how well the data is grouped into clusters, and the formula is used to determine the total silhouette in the cluster. The silhouette formula is shown in equation (4).

$$S(i) = b(i) - a(i) / max(a(i), b(i)) \quad (4)$$

Where a(i) is the average distance between data point i and all other points in the same cluster, and b(i) is the average distance between data point i and all points in the nearest different cluster. The Silhouette value ranges from -1 to 1, where a higher value indicates that the data point is better grouped in the correct cluster. This formula is used to assess the quality of clustering, helping to determine the optimal number of clusters k. According to Rousseeuw (1987), who first introduced this method, a high Silhouette Score indicates that the objects in the cluster have a high similarity to each other and a significant difference from objects in other clusters. Further research by Kaufman and Rousseeuw (2005) confirmed that the Silhouette Score is an effective tool for assessing and validating cluster results [23], providing useful insights into selecting the appropriate clustering model [24]. Thus, the use of this formula is very important to ensure that customer segmentation is not only based on algorithms, but also measurable in terms of quality and accuracy in grouping data. To implement equation 3 above, for example, the distance to determine the silhouette value of data point 3 is first calculated as the average distance of data point 3 to all other data points in the same cluster (cluster 1). Next, the average distance of data point 3 to data points in the nearest neighboring cluster is calculated using the formula S(i)= max(a(i),b(i))b(i)−a(i). The calculation results will provide an overview of the variation between the attribute values analyzed in the two data points. Finally, the average Silhouette Score is calculated to assess the quality of data clustering. The overall

research framework flowchart can be seen in Fig. 1, while the detailed structural steps of the K-Means clustering algorithm are illustrated in Fig. 2. Furthermore, the evaluation process using the Silhouette coefficient method is presented in Fig. 3.

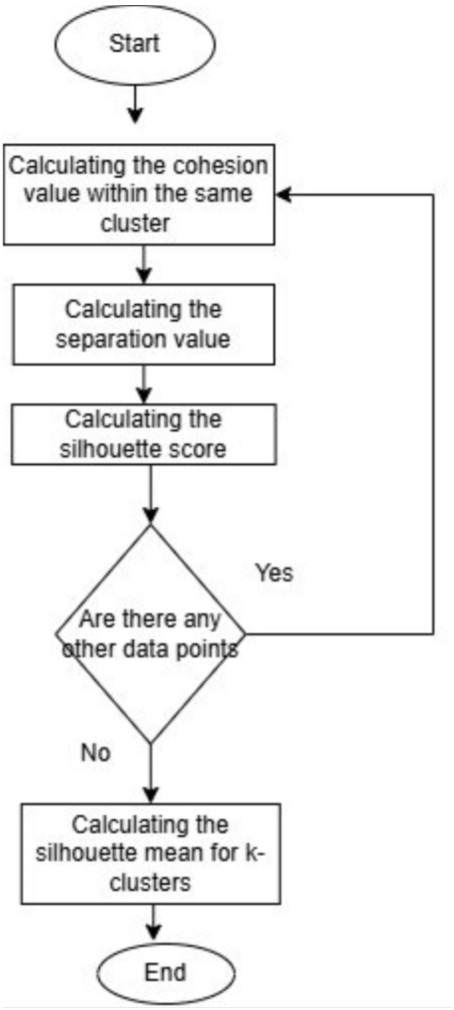

Fig. 3. Flowchart Method Silhouette

The selection of k = 3 as the optimal number of clusters was determined based on an empirical evaluation of multiple cluster configurations ranging from k = 2 to k = 5. The selection of k = 3 as the optimal number of clusters was determined based on an empirical evaluation of multiple cluster configurations ranging from k = 2 to k = 5. Through systematic validation using the Silhouette coefficient, k = 3 yielded the most balanced and interpretable configuration for the business context, effectively separating customers into distinct behavioral archetypes (New, Active, and Loyal) without over-segmenting the dataset. Through systematic validation using the Silhouette coefficient, k = 3 yielded the most balanced and interpretable configuration for the business context, effectively separating customers into distinct behavioral archetypes (New, Active, and Loyal) without over-segmenting the dataset. The Silhouette Coefficient evaluates the clustering quality by combining two metrics: cohesion and separation. Calculating the cohesion value a(i) is done by computing the average distance between data object i and all other data objects within the same cluster, as presented in Equation (5). Meanwhile, the separation value b(i) is determined by calculating the average distance between data object i and all data objects in the nearest neighboring cluster.

$$a_i = avr(D1 + D2 + D3 + \ldots + Dn) \qquad (5)$$

Where a_i is the cohesion value of each data point in the same cluster. D_1 is the distance between data point i and its first neighbor in the same cluster. And so on until D_n, which is the distance between data point i and the last data point in the same cluster. Next, calculate the separation value b_i by calculating the average distance between data point i in one cluster to all data points in other clusters.

This formula is very important in the clustering process because it provides an overview of how well the objects in a cluster are related to each other. According to Jain and Dubes (1988), cohesion is a key indicator of cluster quality because it measures the degree of closeness between data points in the same cluster. Using this formula helps determine how well the clustering model groups data. In addition, Bock et al. (2006) emphasize that cohesion measurement allows researchers to evaluate homogeneity within a cluster and can serve as a basis for comparison between different clusters [25]. Thus, a clear understanding of cohesion is essential in ensuring the effectiveness and accuracy of data grouping in cluster analysis.

Next, the separation value ($b(i)$) is obtained by calculating the average distance between data object i and all data objects in the nearest different cluster. To do this, calculate the average distance of object i to all neighboring objects in other clusters. From the average distance values that have been calculated to all neighboring clusters, select the smallest value. This smallest value is called $b(i)$ and reflects how close object i is to the nearest neighboring cluster, other than its own cluster. The formula for $b(i)$ can be shown in equation (6).

$$b_i = min(A1, A2, A3, \ldots., An) \quad (6)$$

Suppose there are four clusters, namely $A1$, $A2$ dan $A3$.. The value of $A1$is the average distance of the data point to all data points within cluster $A1$. The same applies to , $A2$ dan $A3$. To obtain the value of bi, the smallest value among $A1$, $A2$ dan $A3$is calculated. In other words, bi is the average distance between the i data point and all data points within the nearest cluster. After obtaining the cohesion and separation values, the silhouette score coefficient for data point i is calculated. The formula s(i) is shown in equation (7).

$$S(i) = b(i) - a(i) / max(a(i), b(i)) \quad (7)$$

Where S_i is the silhouette coefficient value of data point -i. This value is obtained by subtracting the separation value of data point i i (b_i) from the cohesion value of data point i (a_i). Then, it is divided by the greater value (max) between the values a_i dan b_i.. If the silhouette coefficient value is the cluster evaluation value for each data point in a dataset with k clusters, then the silhouette score is the total cluster evaluation value that reflects the quality of the clusters in the dataset with k clusters.

The subjective criteria for measuring clustering based on the Silhouette Coefficient according to Kaufman and Rousseeuw can be seen in Table 2.

Table 2. Silhouette coefficient values [19]

No	Silhouette Coefficient (SC) Value	Description
1	0,7 < SC ≤ 1	Strong Structure
2	0,5 < SC ≤ 0,7	Medium Structure
3	0,25 < SC ≤ 0,5	Weak Structure
4	SC ≤ 0,25	No Structure

Table 2 indicates how strong the dataset grouping structure is for each k value or each different cluster composition of the dataset. A Silhouette Score (SS) > 0.7 indicates a strong cluster, 0.5-0.7indicates a medium cluster, 0.25-0.5 indicates a weak cluster, and SC ≤ 0.25 indicates that there is no clear cluster structure.

2.5. Clustering Model Analysis

The RFM (Recency, Frequency, Monetary) model is one of the most widely used methods in customer segmentation to analyze and measure customer behavior based on three key attributes: Recency (R), which indicates how recent a customer's interaction with the company is; Frequency (F), which measures how often customers make purchases or interact; and Monetary (M), which measures the amount of money spent by customers in a given period. According to Stone and Jacob, RFM is an effective tool for identifying high-potential customers by utilizing existing transaction data, as it can predict future purchases based on previous transaction patterns. Kotler et al. also state that the RFM model provides important insights into customer segmentation and supports decisions in designing more focused marketing strategies [26], [27]. By using RFM, companies can identify

groups of customers with similar characteristics and design more personalized approaches according to their shopping habits.

In analyzing clustering models using the K-Means algorithm on RFM (Recency, Frequency, Monetary) data, cluster naming is usually based on the main characteristics of each cluster. The RFM model is an approach used to measure customer value based on three main attributes: Recency (R), which is how recent the customer's last transaction or interaction with the business was; Frequency (F), which is how often the customer makes transactions within a certain period; and Monetary (M), which is the amount of money spent by the customer during a certain period. This model is often used in customer behavior analysis because it provides deep insights into transaction patterns and customer value to the business. In the Recency attribute, customers can be categorized as New if their Recency value is low (customers have just made a purchase), Almost Lost if their Recency value is high (customers have not interacted for a long time), or Retained if their Recency value is moderate (customers are fairly consistent in interacting). In the Frequency attribute, customers can be classified as Loyal if their Frequency value is high (they make frequent purchases), Occasional if their Frequency is medium (they buy periodically), or Rare if their Frequency is low (they only make occasional purchases). Meanwhile, in the Monetary attribute, customers can be grouped as High Spenders if their Monetary value is large, Medium Spenders if their value is moderate, or Low Spenders if their value is small. By adopting the RFM approach in clustering, the analysis results not only produce more structured customer segmentation but also provide insights that can be used to tailor marketing strategies based on the value and behavior of each cluster. This approach helps businesses understand which customer priorities need to be maintained or reactivated to increase overall profitability and customer loyalty.

3. RESULTS AND DISCUSSION

3.1. Research Data

The data excerpt used in the study can be seen in Table 3, which shows the dataset before normalization.

Table 3. Dataset before normalization

No	CustomerID	Recency	Frequency	Monetary
1	16446	0	2	168472,5
2	12901	8	28	17654,54
3	14609	71	4	601,56
4	15749	234	3	44534,3
...	....	....	.....	.....
100	17315	1	3	11482,86
101	16525	1	26	13027,45

Table 3 shows the row dataset used in this study, which consists of customer segmentation. This dataset was taken from the Kaggle repository and includes three main attributes for each customer, namely Recency, Frequency, and Monetary. Each row represents customer transaction data, with the Recency value indicating the last time the customer interacted, Frequency describing the number of interactions or purchases, and Monetary recording the customer's total expenditure.

3.2. Pre-processing Data Results

The pre-processing analysis stage, which includes handling missing values and data normalization, can be seen in Table 4, which shows the dataset after normalization.

Table 4. Dataset after normalization

CustomerID	Recency	Frequency	Monetary
16446	0	0.010	0.601
12901	0.020	0.136	0.063

14609	0.192	0.019	0.002
15749	0.631	0.015	0.159
...	....	.....	.....
17315	0.003	0.015	0.041
16525	0.003	0.126	0.046

Table 4 shows the dataset that has been normalized using the min-max scaling method. Each value in the Recency, Frequency, and Monetary variables is converted to a range of 0 to 1. This normalization process aims to equalize the scale between variables, thereby enabling more effective and accurate statistical analysis and machine learning algorithm application.

3.3. Clustering Model Creation

The number of clusters (k) is determined by dividing the data into three clusters. Cluster 1 (k1) is referred to as New Customers. Cluster 2 (k2) is Active Customers. Cluster 3 (k3) is Loyal Customers. After determining the number of clusters, the next step is to randomly determine the initial centroid points. This stage is the first iteration where the initial values of the cluster centers are identified and updated during the algorithm iteration process until the final cluster configuration is reached. Table 5 presents the cluster centers or initial centroids used in the clustering process.

Table 5. Initial centroids

Customer ID	Recency	Frequency	Monetary
17553	0,346	0,005	0,005
17511	0,004	0,150	0,325
18079	0,119	0,019	0,013

Table 5 shows customer data that already has three initial centroids selected for the clustering process. These centroids are taken based on customers with Customer IDs 17553, 17511, and 18079, which have Recency, Frequency, and Monetary values that represent the characteristics of each cluster. The selection of these initial centroids is important to start the K-Means clustering process. After the clustering process, the results show that the algorithm underwent 8 iterations until it reached a stable final configuration, where the cluster centers or centroids no longer changed. This iteration process illustrates how the K-Means algorithm gradually groups data until it finds a consistent pattern in customer segmentation. The K-Means Clustering calculation ended on the eighth iteration with the cluster results presented in Table 6.

Table 6. Results of iteration 8 centroid

Customer ID	Recency	Frequency	Monetary	Jumlah Data
17553	0.346	0.005	0.005	17
17511	0.004	0.150	0.325	11
18079	0.119	0.019	0.013	73

Table 6 shows the centroid results after 8 iterations in the clustering process. Each selected Customer ID, namely 17553, 17511, and 18079, represents the center point of the cluster identified based on the standardized Recency, Frequency, and Monetary attributes. It can be seen that cluster 2 has the most members, around 73 members, indicating that the majority of customers are in this cluster. Meanwhile, cluster 0 has 17 members and cluster 1 has the fewest members, namely 11 members. These centroid values indicate the center position of each cluster that has been formed after going through a repeated iteration process, in which the K-Means algorithm successfully grouped the data optimally. In other words, these center points describe the general characteristics of customers in each cluster, which can be used to design more focused and efficient marketing strategies. The following is a graph of the number of members per cluster.

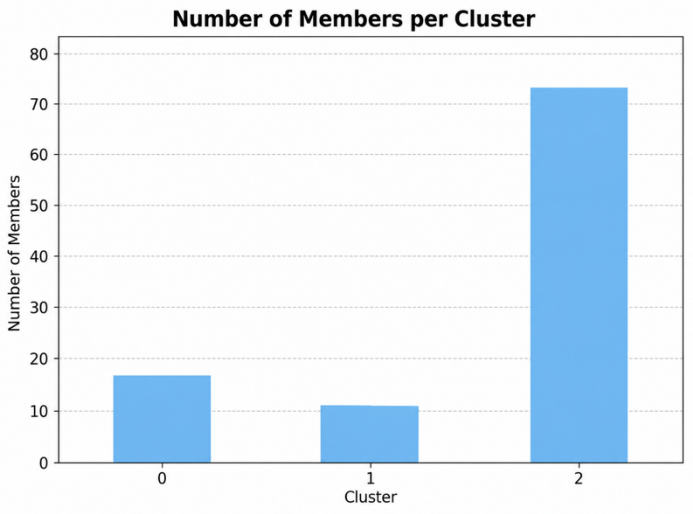

Fig. 4. Number of members per cluster

Fig. 4 shows the distribution of members in each cluster resulting from the clustering process. It can be seen that cluster 2 has the most members, around 73 members, indicating that the majority of customers are in this cluster. Meanwhile, cluster 0 has 17 members and cluster 1 has the fewest members, namely 11 members. This distribution reflects the variation in customer characteristics based on RFM (Recency, Frequency, and Monetary) attributes, where cluster 2 dominates with a much larger number of members than the other two clusters.

After the K-Means clustering process was completed, several evaluations were carried out to ensure the quality and accuracy of the resulting grouping. In addition to visualization using Principal Component Analysis (PCA) to view the distribution of data in two dimensions, another evaluation was conducted using the Silhouette Score to measure how well each data point fits its cluster compared to other clusters. In addition, scatter plots were also used to provide a clearer picture of the distribution of data in the RFM (Recency, Frequency, and Monetary) model. All of these visualizations provide deeper insights into the effectiveness of clustering and how well the RFM model can describe customer segmentation. The following presents the visualization results, which include PCA, Silhouette Score, and scatter plots to further explore the clustering results that have been carried out. Fig. 4 below shows the PCA visualization results of three clusters, with each cluster represented by a different color. The centroid of each cluster is marked with a black cross, providing a clear picture of the data distribution in two dimensions (PC1 and PC2).

Fig. 5 displays the visualization results using Principal Component Analysis (PCA) to illustrate the distribution of the three formed clusters. Each point represents a customer, grouped and distinguished by color: Green for Cluster 0 (17 customers), Blue for Cluster 1 (11 customers), and Orange for Cluster 2 (73 customers).

Fig. 6 shows the results of the Silhouette Plot visualization used to evaluate the quality of separation between clusters. The obtained Silhouette Score of 0.2686 indicates that the clustering structure is relatively weak, suggesting that the separation between clusters is not well-defined. According to unsupervised learning thresholds, a Silhouette Score of 0.2686 indicates a moderate or relatively weak cluster structure, implying a potential overlap between customer segments. In the context of e-commerce behavioral analytics, this structural overlap is expected and acceptable because transactional boundaries between customer groups (such as the transition from an active customer to a loyal one) are naturally fluid and dynamic rather than strictly rigid. This may be caused by overlapping customer characteristics, particularly between active and loyal customer segments, which share similar purchasing behaviors. Therefore, although the clustering results can still provide useful insights, further improvement is needed. Future studies are recommended to explore alternative clustering methods such as DBSCAN or Hierarchical Clustering to achieve better segmentation performance.

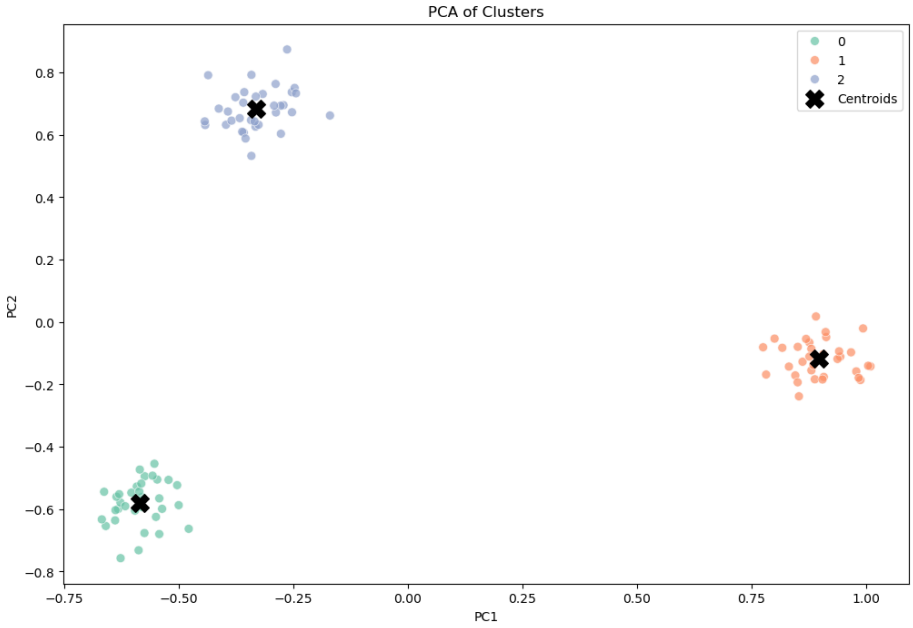

Fig. 5. PCA of Clusters

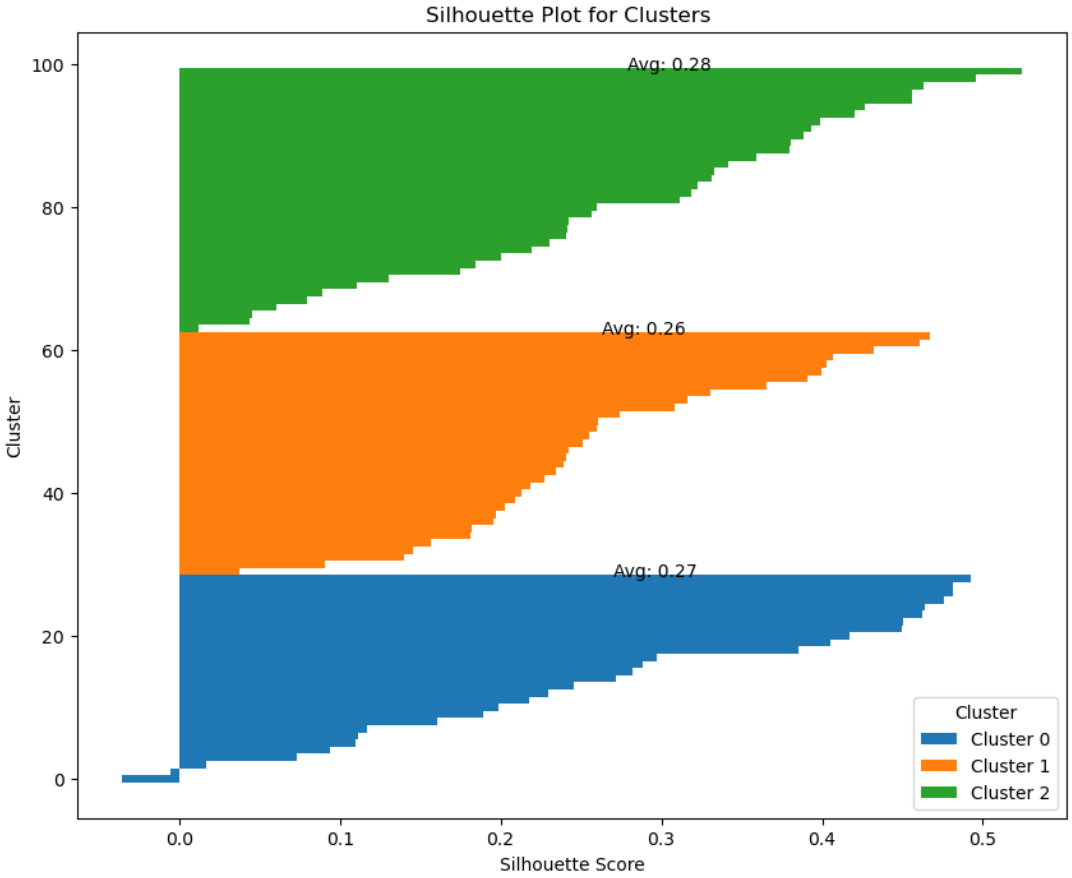

Fig. 6. Silhouette score

In detail, Cluster 1 has a score of 0.26, indicating that customers in this cluster are quite suitable for their cluster even though there are some customers who may be more appropriate in other clusters. Cluster 0 gets a score of 0.27, slightly higher, indicating fairly good separation even though there is still potential overlap. Cluster 2, with the highest score of 0.28, shows the clearest separation among the three. This result is in line with the previous RFM attribute analysis, where Cluster 1 consists of customers with low Recency and Frequency, which are difficult to distinguish from other clusters,

while Cluster 3, with moderate Recency and low Frequency, shows better separation. Overall, although cluster separation is quite effective, there is still room for improvement in some clusters. The following is a Scatter Plot visualization that illustrates the distribution of customers based on RFM attributes (Recency, Frequency, and Monetary). This image provides a clear picture of the relationship between these three attributes in each cluster formed. Each point represents a customer, with the position on the X-axis showing the Recency value, the Y-axis showing the Frequency value, and the size of the point representing the Monetary value. This visualization helps clarify the separation of clusters based on the RFM characteristics that have been analyzed previously, as well as providing a deeper insight into the characteristics of customers in each cluster. The following shows the results of the scatter plot visualization of the clustering results from the RFM (Recency, Frequency, and Monetary) model, where each model is represented by a different color.

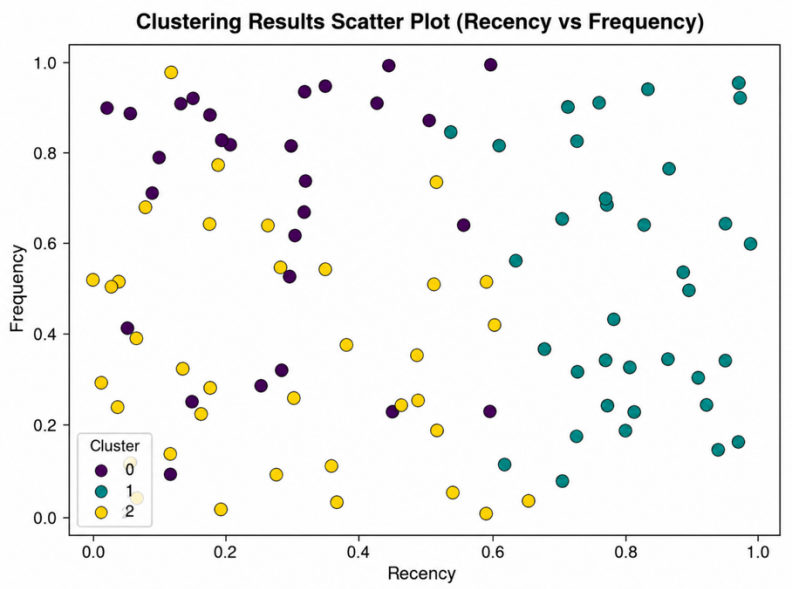

Fig. 7. Scatter plot RFM

Fig. 7 scatter plot of clustering results illustrating the relationship between Recency and Frequency shows that there are three visually distinct clusters. Each cluster is represented by a different color: purple (Cluster 0), green (Cluster 1), and yellow (Cluster 2). Clusters with high Recency and low Frequency values (as seen in the purple dots in the lower right part of the graph) indicate less active customers, which is consistent with the characteristics of Cluster 1, where Recency reaches 216 days and Frequency is low (4 transactions). These customers tend to make transactions infrequently, and a reactivation strategy is needed to regain their interest. On the other hand, clusters with yellow dots, which are more scattered in the upper left part of the graph, have higher Frequency and lower Recency. This matches the description of Cluster 2, where customers are very active with an average Recency of 10 days and fairly high Frequency (69 transactions). Strategies to maintain the loyalty of these customers can be done by providing special rewards. As for Cluster 3 (green), located in the middle of the scatter plot, it describes customers with moderate Recency (29 days) and not very high Frequency (14 transactions). These customers are quite active, but rarely make purchases with small total values, so upselling campaigns can be used to increase their transaction frequency and value. Overall, this scatter plot helps clarify the differences in characteristics between clusters and supports strategies that can be applied according to customer behavior in each cluster.

3.4. Clustering Model Analysis

A more detailed analysis of the characteristics of the RFM attributes (Recency, Frequency, and Monetary) was performed on each cluster based on the last iteration centroid. There were 3 clusters: Cluster 1 with 17 customers, Cluster 2 with 11 customers, and Cluster 3 with 73 customers. The original scale for RFM attributes shows the maximum values in the original data as follows: for the Recency attribute, the maximum value is 371 days, obtained through the min-max scaling process; for the Frequency attribute, the highest frequency recorded is 206 transactions; while for the Monetary attribute, the highest value recorded is $280,206.02. After conversion to the original scale, the centroid of each cluster reflects different characteristics. Cluster 1, consisting of 17 customers,

has different centroid values for each RFM attribute compared to other clusters, as do cluster 2, consisting of 11 customers, and cluster 3, which has the largest number of customers, namely 73 customers. Further analysis of these centroid values provides a deeper understanding of the differences in customer behavior and characteristics within each cluster. To return the centroid values to their original scale, the following calculations were performed: In Cluster 1, the original Recency value is calculated as 0.5828 × 371 = 216 days, the original Frequency value becomes 0.0215 × 206 = 4 transactions, and the original Monetary value is obtained from 0.0169 × 280,206.02 = $4,733. In Cluster 2, the original Recency value is calculated as 0.0258 × 371 = 10 days, the original Frequency value becomes 0.3370 × 206 = 69 transactions, and the original Monetary value is obtained from 0.5190 × 280,206.02 = $145,027. Meanwhile, in Cluster 3, the original Recency value is calculated as 0.0786 × 371 = 29 days, the original Frequency value becomes 0.0684 × 206 = 14 transactions, and the original Monetary value is obtained from 0.0497 × 280,206.02 = $13,922. These calculations provide a clear picture of the RFM attribute values on the original scale for each cluster, which can be used for further analysis. Analysis of RFM attributes per cluster on the original scale shows significant differences in characteristics between clusters. Cluster 1 consists of customers with low Recency (216 days) and Frequency (4 transactions), as well as low Monetary ($4,733). The high Recency value indicates customer inactivity, as they have not made a transaction in more than 7 months, while the low transaction frequency indicates that they rarely make purchases with a small total value. For this cluster, the appropriate strategy is a reactivation program, such as special offers or discounts, to attract customers back to make transactions. Cluster 2, on the other hand, consists of highly active customers with very low Recency (10 days), high Frequency (69 transactions), and very high Monetary ($145,027). Customers in this cluster demonstrate strong loyalty with frequent transactions and significant total purchase values. The appropriate strategy is to maintain their loyalty through reward programs or exclusive offers. Cluster 3 has moderate Recency (29 days), fairly low Frequency (14 transactions), and low Monetary value ($13,922). Although customers in this cluster are fairly active, they do not transact often, and their total purchase value is relatively small. For this cluster, the appropriate strategy is promotions or upselling campaigns to encourage them to transact more often and increase their purchase value.

4. CONCLUSION

The conclusion of the strategy based on cluster characteristics shows a tailored approach for each customer group. For Cluster 1, which consists of customers who rarely make transactions and have low value, the right strategy is to focus on reactivation through special offers or discounts to attract them back to make transactions. Cluster 2, which consists of active and high-value customers, requires loyalty programs, exclusive offers, or gifts to maintain good relationships and increase their engagement. As for Cluster 3, which consists of fairly active but low-value customers, the recommended strategy is promotions or upselling campaigns to encourage them to make more frequent purchases and increase their transaction value. With the right approach for each cluster, companies can strengthen customer relationships and maximize sales potential. A limitation of this study is the relatively compact size of the evaluated sample dataset. Future research could enhance the robustness and generalizability of the clustering model by deploying larger, real-time transaction datasets and exploring deep learning architectures or dynamic clustering techniques to capture continuous shifts in customer purchasing behavior over longer periods.

Acknowledgment

We would like to express our gratitude to the Faculty of Science and Technology, Informatics Study Program, Universitas Islam Mulia Yogyakarta for their invaluable support in this research. We greatly appreciate the collaboration and support from all parties for the smooth running of this research. We hope that the results of this research can make a positive contribution to the advancement of science and public health.

REFERENCES

[1] A.-Y. Al-Yasir, M. Afdal, Z. Zarnelly, and A. Marsal, “Analisis Loyalitas Pelanggan Business To Business Berdasarkan Model RFM Menggunakan Algoritma Fuzzy C-Means: Business to Business

Customer Loyalty Analysis Based on RFM Model Using Fuzzy C-Means Algorithm," *MALCOM Indones. J. Mach. Learn. Comput. Sci.*, vol. 4, no. 1, pp. 359–365, Jan. 2024, doi: 10.57152/malcom.v4i1.1163.

[2] M. Skare, B. Gavurova, and M. Rigelsky, "Innovation activity and the outcomes of B2C, B2B, and B2G E-Commerce in EU countries," *J. Bus. Res.*, vol. 163, p. 113874, Aug. 2023, doi: 10.1016/j.jbusres.2023.113874.

[3] E. Yıldız, C. Güngör Şen, and E. E. Işık, "A Hyper-Personalized Product Recommendation System Focused on Customer Segmentation: An Application in the Fashion Retail Industry," *J. Theor. Appl. Electron. Commer. Res.*, vol. 18, no. 1, pp. 571–596, Mar. 2023, doi: 10.3390/jtaer18010029.

[4] T. T. A. Ngo, H. L. T. Nguyen, H. P. Nguyen, H. T. A. Mai, T. H. T. Mai, and P. L. Hoang, "A comprehensive study on factors influencing online impulse buying behavior: Evidence from Shopee video platform," *Heliyon*, vol. 10, no. 15, p. e35743, Aug. 2024, doi: 10.1016/j.heliyon.2024.e35743.

[5] P. Anitha and M. M. Patil, "RFM model for customer purchase behavior using K-Means algorithm," *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 34, no. 5, pp. 1785–1792, May 2022, doi: 10.1016/j.jksuci.2019.12.011.

[6] B. Sohrabi and A. Khanlari, "Targeting Customers : a Fuzzy Classification Approach," *Int. J. Eng.*, vol. 23, no. 3, pp. 323–336, 2010.

[7] A. J. Christy, A. Umamakeswari, L. Priyatharsini, and A. Neyaa, "RFM ranking – An effective approach to customer segmentation," *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 33, no. 10, pp. 1251–1257, Dec. 2021, doi: 10.1016/j.jksuci.2018.09.004.

[8] F. Barrera, M. Segura, and C. Maroto, "Multiple criteria decision support system for customer segmentation using a sorting outranking method," *Expert Syst. Appl.*, vol. 238, p. 122310, Mar. 2024, doi: 10.1016/j.eswa.2023.122310.

[9] Nofrizal, U. Juju, Sucherly, A. N, I. Waldelmi, and Aznuriyandi, "Changes and determinants of consumer shopping behavior in E-commerce and social media product Muslimah," *J. Retail. Consum. Serv.*, vol. 70, p. 103146, Jan. 2023, doi: 10.1016/j.jretconser.2022.103146.

[10] C. Rungruang, P. Riyapan, A. Intarasit, K. Chuarkham, and J. Muangprathub, "RFM model customer segmentation based on hierarchical approach using FCA," *Expert Syst. Appl.*, vol. 237, p. 121449, Mar. 2024, doi: 10.1016/j.eswa.2023.121449.

[11] B. N. Yulisasih, R. A. Surya, P. Widiandana, M. I. Aulia, and S. Hartina, "Pemberdayaan Pemuda Sebagai Pendamping Teknologi Dalam Digitalisasi Layanan Posyandu," *J. Pengabdi. Abhinaya*, vol. 1, no. 2, pp. 83–89, Oct. 2025.

[12] F. Zhao, G. Jiang, Y. Zhang, and S. Sayed Merajuddin, "Online sales and corporate innovation preference: The impact of e-commerce emergence on corporate innovation behavior," *Finance Res. Lett.*, vol. 64, p. 105447, Jun. 2024, doi: 10.1016/j.frl.2024.105447.

[13] S. H. Shihab, S. Afroge, and S. Z. Mishu, "RFM Based Market Segmentation Approach Using Advanced K-means and Agglomerative Clustering: A Comparative Study," in *2019 International Conference on Electrical, Computer and Communication Engineering (ECCE)*, Cox'sBazar, Bangladesh: IEEE, Feb. 2019, pp. 1–4. doi: 10.1109/ECACE.2019.8679376.

[14] X. Pu, C. Song, and J. Huang, "Research on Optimization of Customer Value Segmentation Based on Improved K-Means Clustering Algorithm," in *2020 IEEE 3rd International Conference on Information Systems and Computer Aided Education (ICISCAE)*, Dalian, China: IEEE, Sep. 2020, pp. 538–542. doi: 10.1109/ICISCAE51034.2020.9236867.

[15] S. Guney, S. Peker, and C. Turhan, "A Combined Approach for Customer Profiling in Video on Demand Services Using Clustering and Association Rule Mining," *IEEE Access*, vol. 8, pp. 84326–84335, 2020, doi: 10.1109/ACCESS.2020.2992064.

[16] H.-H. Zhao, X.-C. Luo, R. Ma, and X. Lu, "An Extended Regularized K-Means Clustering Approach for High-Dimensional Customer Segmentation With Correlated Variables," *IEEE Access*, vol. 9, pp. 48405–48412, 2021, doi: 10.1109/ACCESS.2021.3067499.

[17] K. Tabianan, S. Velu, and V. Ravi, "K-Means Clustering Approach for Intelligent Customer Segmentation Using Customer Purchase Behavior Data," *Sustainability*, vol. 14, no. 12, p. 7243, Jun. 2022, doi: 10.3390/su14127243.

[18] H. (Hojatollah) Hamidi and B. Haghi, "An approach based on data mining and genetic algorithm to optimizing time series clustering for efficient segmentation of customer behavior," *Comput. Hum. Behav. Rep.*, vol. 16, p. 100520, Dec. 2024, doi: 10.1016/j.chbr.2024.100520.

[19] W. A. Prastyabudi, A. N. Alifah, and A. Nurdin, "Segmenting the Higher Education Market: An Analysis of Admissions Data Using K-Means Clustering," *Procedia Comput. Sci.*, vol. 234, pp. 96–105, 2024, doi: 10.1016/j.procs.2024.02.156.

[20] C. Shi, B. Wei, S. Wei, W. Wang, H. Liu, and J. Liu, "A quantitative discriminant method of elbow point for the optimal number of clusters in clustering algorithm," *EURASIP J. Wirel. Commun. Netw.*, vol. 2021, no. 1, p. 31, Dec. 2021, doi: 10.1186/s13638-021-01910-w.

[21] A. Abernathy and M. E. Celebi, "The incremental online k-means clustering algorithm and its application to color quantization," *Expert Syst. Appl.*, vol. 207, p. 117927, Nov. 2022, doi: 10.1016/j.eswa.2022.117927.

[22] M. Kanwal, N. A. Khan, and A. A. Khan, "A Machine Learning Approach to User Profiling for Data Annotation of Online Behavior," *Comput. Mater. Contin.*, vol. 78, no. 2, pp. 2419–2440, 2024, doi: 10.32604/cmc.2024.047223.

[23] Sulistiani, Rakyatol Hasanah, Anggun Sindiana, Ahmad Rizky Nusantara Habibi, Minhajul Abidin, and Asno Azzawagama Firdaus, "Peningkatan Literasi Gizi Melalui Kegiatan Edukasi dan Pendokumentasian Makan Bergizi Gratis," *J. Pengabdi. Abhinaya*, vol. 1, no. 2, pp. 49–55, Oct. 2025, doi: 10.64021/jpa.1.2.49-55.2025.

[24] A. Wasilewski, K. Juszczyszyn, and V. Suryani, "Multi-factor evaluation of clustering methods for e-commerce application," *Egypt. Inform. J.*, vol. 28, p. 100562, Dec. 2024, doi: 10.1016/j.eij.2024.100562.

[25] J. Zhou, L. Zhai, and A. A. Pantelous, "Market segmentation using high-dimensional sparse consumers data," *Expert Syst. Appl.*, vol. 145, p. 113136, May 2020, doi: 10.1016/j.eswa.2019.113136.

[26] J.-J. Jonker, N. Piersma, and D. Van Den Poel, "Joint optimization of customer segmentation and marketing policy to maximize long-term profitability," *Expert Syst. Appl.*, vol. 27, no. 2, pp. 159–168, Aug. 2004, doi: 10.1016/j.eswa.2004.01.010.

[27] L. Wang, T. R. A. L. Pertheban, T. Li, and L. Zhao, "Application of business intelligence based on big data in E-commerce data evaluation," *Heliyon*, vol. 10, no. 21, p. e38768, Nov. 2024, doi: 10.1016/j.heliyon.2024.e38768.

AUTHOR BIOGRAPHY

Baiq Nikum Yulisasih is an academic and researcher with a strong interest in Machine Learning and data analysis. She is actively involved in developing machine learning models for various applications, including data classification and predictive analytics. Her research focuses on applying machine learning techniques to improve the efficiency of data-driven systems. She is affiliated with the Informatics Study Program, Faculty of Science, Technology, and Health, Universitas Islam Mulia Yogyakarta.

Panggah Widiandana is a practitioner and researcher in the fields of Artificial Intelligence (AI) and Cybersecurity. He has a strong interest in developing intelligent systems and securing digital infrastructures. His research focuses on integrating AI technologies to enhance cybersecurity systems, including threat detection and intelligent defense mechanisms. He is affiliated with the Informatics Study Program, Faculty of Science, Technology, and Health, Universitas Islam Mulia Yogyakarta.

Muhammad Immawan Aulia is a researcher specializing in Cybersecurity and Software Engineering. He has experience in developing secure and efficient software systems as well as analyzing system vulnerabilities. His research interests include secure application development, software testing, and the application of software engineering principles in modern systems. He is affiliated with the Informatics Study Program, Faculty of Science, Technology, and Health, Universitas Islam Mulia Yogyakarta.

Rizky Andhika has expertise in Cybersecurity and Digital Image Forensics. He is particularly interested in analyzing the authenticity of
Digital images and detecting image manipulation. His research involves applying digital forensic techniques to identify alterations in images and developing security systems based on image analysis. He is affiliated with the Informatics Study Program, Faculty of Science, Technology, and Health, Universitas Islam Mulia Yogyakarta.

Siti Hartinah is a systems analyst specializing in the design and development of information systems. She has experience in analyzing user requirements and designing effective and efficient system solutions. Her interests include systems analysis, application development, and optimizing business processes through information technology. She is affiliated with the Informatics Study Program, Faculty of Science, Technology, and Health, Universitas Islam Mulia Yogyakarta.