ILKOM Jurnal Ilmiah Vol. 16, No. 3, December 2024, pp.330-342
Accredited 2nd by KEMDIKTISAINTEK No. 10/C/C3/DT.05.00/2025; E-ISSN 2548-7779 | P-ISSN 2087-1716

Evaluation of K-Means Clustering Using Silhouette Score Method on Customer Segmentation

Baiq Nikum Yulisasih [a,1], Herman [a,2*], Sunardi [a,3], Herman Yuliansyah [a,4]
[a] *Ahmad Dahlan University, Jl. Prof. DR. Soepomo Sh, Umbulharjo, Yogyakarta 55166, Indonesia*
[1] *2307048005@webmail.uad.ac.id;* [2] *hermankaha@mti.uad.ac.id;* [3] *sunardi@mti.uad.ac.id;* [4] *herman.yuliansyah@tif.uad.ac.id*
** Corresponding author: hermankaha@mti.uad.ac.id*

Article history: Received September 20, 2024; Revised October 10, 2024; Accepted December 29, 2024; Available *online* December 31 2024.

Abstract

Customer segmentation is a critical process in businesses to understand and meet the diverse needs of customer. This study focused on the challenges of managing large and complex volumes of customer data and identifying the right segments to personalize marketing strategieshow about if I . K-Means Clustering has been widely utilized for its ability to group multidimensional data, but this method often generated broad clusters that lack detailed insights. Therefore, cluster evaluation with the Silhouette Score method became essential to ensure the optimality and validity of the generated groupings. The purpose of this study was to evaluate the quality of K-Means Clustering using the Silhouette Score method on customer segmentation. This research began with the acquisition of a dataset comprising 2,000 data points characterized with 7 attributes: sex, marital status, age, education, income, occupation, and settlement size. The data then underwent pre-processing by checking missing values and normalizing data. K-Means Clustering was then applied to group data into several clusters based on their proximity to the cluster center (centroid). The results of the clusters were assessed using the Silhouette Score method to determine the most optimal number of clusters. The results of this study consisted of manual calculations using Microsoft Excel on 27 data points to facilitate understanding of the logic, steps, methods and practical foundations before implementation on the complete dataset. Furthermore, the results of the Python calculation in 2000 data points showed that the optimal number of clusters (close to the value of 1) between k = 2 to k = 7 was the k = 4 cluster with a Silhouette Score value of 0.43, categorized as a weak structure. Although this value indicated a weak cluster structure, it was the highest value in the test, indicating that the division of data into four clusters (k = 4) was better than the number of other clusters. However, the quality of this cluster indicates the need for futher improvement. Future work should review the used attributes, data normalization methods, or consider other clustering algorithms to achieve a more robust structure and more meaningful interpretation.

Keywords: Cluster Evaluation; Customer Segmentation; K-Means Clustering; Silhouette Score.

Introduction

Business competition has been intensifying and increasingly complex, a comprehensive understanding of customers is one of the strategic factors that can help companies achieve competitive advantage [1]. In the midst of increasing volumes of customer data and changing consumer behavior, companies need to implement a more targeted and effective marketing approach. One of the approaches used is customer segmentation which helps companies understand customer characteristics and preferences more effectively [2]. Customer segmentation is the practice of dividing a company's customers into groups based on the similar characteristics. The main goal of customer segmentation is to divide a broad and heterogeneous market into smaller, homogeneous groups, so that companies can design more effective and efficient marketing strategies. This can improve the effectiveness of marketing strategies, operational efficiency, and customer experience. Customer segmentation has been becoming increasingly important due to the ever-expanding diversity and complexity of customer preferences [3]. With so much data available, companies are facing the challenge to understand the increasingly diverse customer needs and behaviors.

Customer segmentation is an important strategy in modern businesses to understand and meet the needs of diverse consumers [4], [5]. However, challenges arise when companies must manage large and diverse volumes of data, identify the right customer segments, and effectively personalize marketing strategies. Without proper segmentation, companies may struggle to tailor products or services to specific consumer preferences, which can reduce the potential

for purchase. With accurate segmentation, companies can offer products that better suit the needs of the target market, improve the effectiveness of marketing strategies, and ultimately improve consumer purchasing decisions [6]. The current trend shows that more and more consumers shift from offline to online transaction using various e-commerce platforms [7]. This move is due to several advantages offered by online transactions such as convenience, availability and variety of items, the number of discounts, and promotional packages. The shifting is also driven by the rapid development of technology, especially the expansion of internet networks, smartphone penetration, the integration of big data analysis, and the wider application of machine learning [8]. Machine learning modeling has been used in several studies to create customer segmentation based on customer preference factors as one of the inputs for e-commerce marketing strategies. Previous research shows that K-Means Clustering as one of the clustering techniques in machine learning has been widely used to segment customers [9]. For example, some studies have successfully divided customers into clusters based on purchase frequency, transaction patterns, or credit card usage [10]. A theoretical study of customer segmentation shows that K-Means Clustering is a fast, simple, and effective method of grouping data into homogeneous groups.

Previous studies have demonstrated the effectiveness of this method in dividing customers based on various attributes such as transactions online, credit card usage, and customer loyalty [11]. Another study that applied customer segmentation based on loyalty found four different groups: premium loyalty, inertial loyalty, latent loyalty, and disloyal [12]**.** The research of [13] shows that K-Means produces faster and better grouping than agglomerative grouping. [14] use K-Means to identify customers in five categories that are useful for customer management and marketing strategies. The research of [13]. has successfully managed to identify four main customer groups using the proposed approach showing an effective approach in forming customer segments and association rules that are useful for IPTV service providers. Research by [15] found that the clustering method produced smaller grouping errors compared to standard K-Means. The results of the study helped e-commerce platforms to compare multidimensional consumer purchasing patterns with interconnected variables. The study by Kayalvily et al. compared purchasing patterns and the identification of inequality of advantageous segments in small regions and emphasized the need for further research such as deep learning. customer segmentation requires K-Means Clustering because it can handle the diversity of complex customer data efficiently and systematically. This method can group multidimensional data, such as demographics, purchasing behavior, and customer preferences, into more homogeneous segments based on certain similarities. However, segmentation with this method tends to only result in large groups without providing deep insights for more personalized marketing strategies. K-Means Clustering requires cluster evaluation to ensure that the clustering results reflect the data structure accurately. Without evaluation, it is difficult to determine whether the number of clusters selected is appropriate or whether the resulting clusters can separate the data based on relevant characteristics.

Cluster evaluation plays an important role in assessing cluster quality, ensuring that data in one cluster has a high similarity compared to data in other clusters, and helping to determine the optimal number of clusters. Some commonly adopted evaluation methods include Silhouette Score, Elbow Method, Davies-Bouldin Index, and Dunn Index [16], [17]. Among these, Silhouette Score often becomes the first choice in customer segmentation because it measures both density within clusters and segregation between clusters. With values ranging from -1 to 1, this method provides an intuitive evaluation, where values closer to 1 indicates high cluster quality. In addition, Silhouette Score is independent of specific functions and suitable for a variety of distance metrics, making them more flexible than other methods. Related studies show that the evaluation of cluster quality can be improved by the Silhouette Score method, which provides a measure of how well objects in a cluster are grouped. The study compared the Elbow method and Silhouette Score and found that the Silhouette Score method tended to produce more accurate and higher-quality clusters [18], [19]. Additional studies have explored clustering evaluations using Davies-Bouldin Index and the Elbow method as benchmarks. For instace, clustering testing using the Davies Bouldin Index (DBI) method resulted in a DBI value of 1.10 whereas the result of 7 clustering using the Silhouette coefficient indicated a DBI value of 1.06. This shows that the results of K-Medoid clustering with Silhouette coefficient produce superior cluster quality because it has a lower DBI value than K-Medoid clustering with the Elbow method [20], [21].

A comparative analysis between previous and current studies shows significant differences in research approaches and areas of emphasis. Previous research has focused more on comparing cluster evaluation methods, such as Elbow and Silhouette Score, in the context of clustering algorithms such as K-Means and K-Medoids. Meanwhile, the current research focused on the application of the K-Means Clustering method to customer segmentation by evaluating using the Silhouette Score method [22], [23]. This research offers novelty by applying the K-Means Clustering method for customer segmentation which has not been widely explored in the existing literature. This approach both allowed companies with deeper insights into customer needs and preferences and supported more effective and targeted decision-making in marketing strategies.

The Silhouette Score was selected as the primary evaluation method because it provided a balance between intra-cluster cohesion and inter-cluster separation which were crucial in customer analysis [16]**.** This study emphasized the importance of robust cluster evaluation to ensure optimal segmentation results, enabling companies to identify customer needs more accurately and design more effective marketing strategies. The purpose of this study was to evaluate the quality of clusters produced by the K-Means Clustering algorithm in the customer segmentation process by using Silhouette Score as an evaluation method. The combination of K-Means and Silhouette Score in generating optimal customer segmentation can help companies deeply understand customer characteristics, improve service personalization, and optimize marketing strategies to increase purchase potential. The anticipated contribution of this research was to make a real contribution to data-driven customer management. These results can support companies in developing more effective and efficient strategies, increasing customer satisfaction, and driving business growth.

Method

A. Research Object

This study utalized a customer dataset from the Kaggle site provided by Dev Sharma as the research object. The dataset, consisting of 2000 data points, is presented in **Table 1**.

Table 1. Customer segmentation dataset details

Attribute	Description
Sex	0= Male, 1 = Female
Marital Status	0= Single, 1=Married
Age	minimum= 18, maximum= 78
Education	0 = Other/Unknown, 1 = High School, 2 = University, 3 = Graduate
Income	Minimum=$35832, Maximum= $309364 Annual Revenue (US$)
Occupation	0= Unemployed/Unskilled, 1= Skilled Employees/Officials, 2= High quality Management/Self-Employed/Employees/Officials
Settlement Size	0= Small city, 1= Medium city, 2= Large city

Table 1 provides a detail overview of the characteristics of the customer dataset. Each data point has 7 attributes. The first attribute is sex with two value options, 0 for men and 1 for women. The marital status attribute indicates the marital status of the customer, with two categories: 0 for single and 1 for non-single (including divorced, separated, married, or widowed). The age attribute indicates the age of the customer in years. The minimum age observed was 18 years and the maximum age was 76 years. The education attribute shows the level of education of customers in four categories: 0 for other/unknown, 1 for high school, 2 for university, and 3 for graduate school. The revenue attribute reflects the customer's self-reported annual revenue in U.S. dollars. Occupation attributes are customer occupations grouped into three categories, 0 for unemployed/unskilled workers, 1 for skilled employees/employees, and 2 for management/self-employed/highly qualified employees/officials. Finally, the settlement size indicates the city where the customer lives, with three categories, 0 for small cities, 1 for medium-sized cities, and 2 for large cities.

B. Data Pre-processing

After acquiring the data, the next stage was Pre-processing data. It was conducted before the implementation of K-Means Clustering. During this phase, the dateset underwent two crucial steps: missing value and data normalization. Missing value analysis was the process of identifying and handling blank and incomplete entries, while data normalization was a step to change the scale of an attribute or variable value to achieve uniformity [24]. These two steps were important in the implementation of K-Means Clustering because of the missing value step ensuring the integrity of the data before the analysis, while normalization assisted to overcome scale issues that can affect cluster results. The min-max normalization was a method to scale data so that it was between 0 and 1. The formula used to to implement min-max normalization to each attribute in the dataset is presented in Equations 1 [25].

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}} \quad (1)$$

Let X' represent the normalized data value, X denotes the original data value prior to normalization, X_{min} signifies the minimum value in the dataset, and X_{max} indicates the maximum value in the dataset.

B. Clustering Model Design

The next process after pre-processing was the application of K-Means Clustering with the Silhouette Score method. K-Means Clustering is the most well-known algorithm in the grouping technique for dividing data into groups. K-

Means was first introduced in 1967 by Macqueen [4]. K-Means is also one of the most prominent grouping techniques in science and technology [26] . Furthermore, the application of the Silhouette Score method was to evaluate the quality of the results of each cluster in K-Means Clustering. Silhouette Score is an effective method to determine the optimal number of clusters in the K-Means Clustering process. The Silhouette Score value ranges from -1 to 1. Where a value closer to 1 is the most optimal cluster and vice versa. Using the Silhouette Score, the evaluation of the cluster can ensure that the clusters formed are not only well separated, but also homogeneous within them. This is crucial for a more effective marketing strategy, as each cluster must represent a segment of customers who have similar characteristics and preferences. The application of the combination of K-Means Clustering with the Silhouette Score method is shown in the form of a flowchart in **Figure 1**.

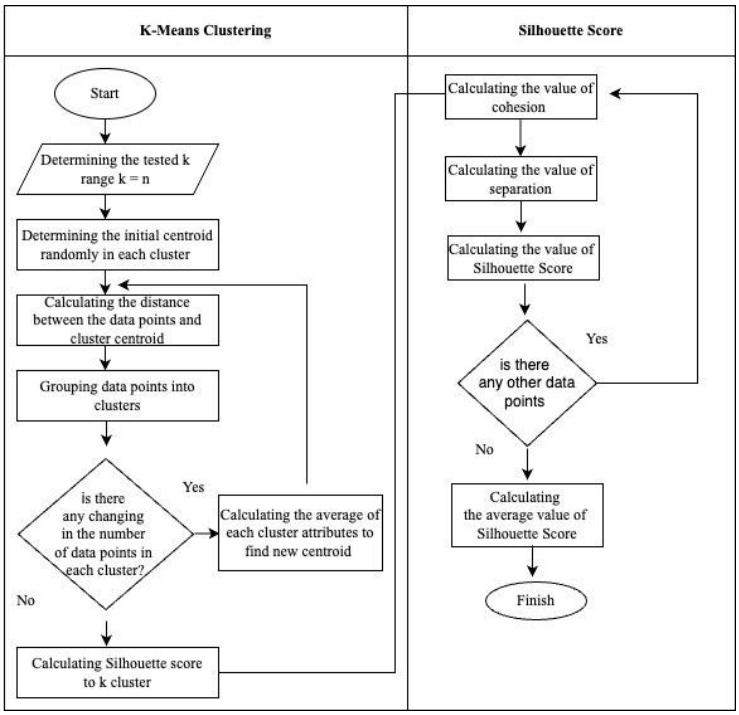

Figure 1. Flowchart of the application of K-Means Clustering with the Silhouette Score method

Figure 1 explains that the K-Means Clustering algorithm with the Silhouette Score method started by determining the number of clusters (k) to be tested. The first step of the algorithm was the selection of centroid random start in each cluster which is the initial process of iteration 1. The next step was to calculate the distance between the data point and centroid cluster using the Euclidean Distance formula. The formula was used to determine the grouping of each cluster based on the results of the distance. The formula is shown in Equation 2 [27], [28].

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

Where $d\,(x,y)$ is the distance between data point x and centroid y, y_i donates the value of the i-th attribute value of centroid y, and x_i represents the i-th attribute value of the data point x. Here, i refers to the index of the attribute, and n indicates the total number of the attribute for each data point.

After calculating the distance, each data point was then grouped based on the closest distance between the first data point with three centroid clusters. Then, the data points were grouped into clusters based on the closest distance to the centroid. Furthermore, the results of managing each cluster in iteration 1 would be compared to each cluster in iteration 2. However, before the grouping of iteration 2, a new centroid would be searched by calculating the average value of all data points from each cluster. Once the new centroid value of each cluster was obtained, then the calculation of the euclidean distance was conducted just like the previous process. Then the process continued with the grouping of iterations 2. The results of this grouping were compared to the results of the grouping of iteration 1. If there was a change in the composition of the number of data points and centroid values, then the next iteration would be continued. On the other hand, if there was no change in the composition of the data points of each cluster or the centroid value, the iteration process ended or the k-count. If there was a change in the composition of the number of data points and centroid values, the next iteration would be continued. On the other hand, if there was no change in

the composition of the data points of each cluster or the centroid value, the iteration process was over, or the calculation of K-Means Clustering has ended.

The cluster quality formed from dataset using K-Means Clustering was evaluated using the Silhouette Score method. The first step was to calculate the cohesion value by calculating the average distance between the i-th data point and all other data points in the same cluster. The formula for finding the cohesion value is shown in equation (3) [29].

$$a_i = avr(D_1 + D_2 + D_3 + \cdots + D_n) \quad (3)$$

Where a_i represents the cohesion value of each data point within the same cluster. D_1 is the distance between the i-th data point and the first neighbor data point within the same cluster. This continues until D_n, which the the distance between the i-th data point and the last data point within the same cluster. This formula is very important in the clustering process because it gives an idea of how strongly the objects in a cluster are related to each other. Cohesion is a key indicator of cluster quality, as it measures the degree of closeness between data points in the same cluster [20]. The calculation of this cohesion value is very essential in ensuring the effectiveness and accuracy of data grouping in clustering analysis. Next, the separation value b_i is calculated by finding the average distance between the i-th data points in one cluster and all data points in other clusters. The separation formula is presented in Equation 4 [30].

$$b_i = min(A_1, A_2, A_3, \ldots, A_n) \quad (4)$$

Suppose there are four clusters, A_1, A_2, A_3 and A_4. The value of A_1 represents the average distance of a data point to all other data points within cluster A_1. Similarly, this applies to A_2 and A_3. To obtain the value of b_i, the smallest value among A_1, A_2, A_3 is selected. In other words, b_i represents the average distance between the i-th data point and all data points in the nearest cluster. The separation value is as important as the cohesion value in the clustering evaluation process for assessing the quality of the formed clusters. Cohesion measures the proximity of objects within a cluster to the cluster's centroid, whereas separation measures the distance between data points across different clusters. After obtaining the cohesion and separation values, the Silhouette Score coefficient for the i-th data point is calculated, as shown in Equation 5 [9].

$$S_i = \frac{b_i - a_i}{max(a_i, b_i)} \quad (5)$$

Where S_i is the Silhouette coefficient value for the i-th data point. This value is obtained by subtracting the value of the separation value (b_i) of the i-th data point from its cohesion value of the data point (a_i). Then, it is divided by a larger value (max) between the value of a_i and b_i. If the Silhouette coefficient value is the clustering evaluation value on each data point in a dataset with k clusters, the Silhouette Score value is the total cluster evaluation value that reflects the quality of the cluster in dataset with k clusters. The Silhouette Score value is obtained by calculating the average Silhouette coefficient value of all data points from k clusters in the dataset. The Silhouette Score ranges from -1 and 1 [4]. A dataset composition with k clusters with a Silhouette Score value close to 1 indicates as the strongest structure. This means that the dataset with k clusters has been optimally grouped and reflects a strong cluster structure. The Silhouette Score calculation is repeated for all dataset compositions, starting from k = 1 to k = n. From the calculation, the dataset composition with the best k value or number of clusters is selected, that is composition with the Silhouette Score value close to 1. **Table 2** shows the assessment of cluster quality using the Silhouette Score according to Kaufman and Rousseeuw [31].

Table 2. Silhouette Score range

It	Silhouette Score (SS) Value Range	Information
1	0.7 < S S ≤ 1	Strong Structure
2	0.5 < S S ≤ 0.7	Moderate Structure
3	0.25 < S S ≤ 0.5	Weak Structure
4	S S ≤ 0.25	No Structure

This assessment indicates the strength of the clustering structure of the dataset for each k-value or each composition of the different cluster numbers from in the dataset. A Silhouette Score (SS) > 0.7 indicates a strong, 0.5-0.7 is moderate, 0.25-0.5 is weak, and SC ≤ 0.25 means there is no clear cluster structure.

Results and Discussion

A. Research Dataset

The application of K-Means Clustering algorithm combination with Silhouette Score on this research dataset used Microsoft Excel and Python tools. Microsoft Excel was used for manual initial calculations on a portion of the dataset whereas Python for the entire dataset. Manual calculations were part of the preliminary research by taking a part of randomly selected data points to simplify the calculation without reducing the representative value of the analysis. The selection of the number of data points aimed both to provide a practical overview of the application of the segmentation method with algorithms used on a small scale and to understand the calculation steps in detail before proceeding to automated analysis with Python on the entire dataset. The results of this manual calculation were to make it easier to understand the logic of the method used before applying the technique on a larger data scale. A snippet of the data points for the manual calculation used can be seen in **Table 3**.

Table 3. Datasets before normalization

It	ID	Sex	Marital status	Age	Education	Income	Occupation	Settlement size
1	100000003	0	0	49	1	89210	0	0
2	100000006	0	0	35	1	144848	0	0
3	100000009	0	1	61	2	151591	0	0
4	……	……	……	……	……	……	……	……
26	100000028	1	1	42	2	163025	1	1
27	100000035	1	1	33	1	155569	2	1

Table 3 shows a dataset of 27 data points that would be calculated manually using Microsoft Excel. Furthermore, this study displayed calculations on the actual dataset with up to 2000 data points using Python. Managing large datasets was more efficient using Python because it included libraries such as Pandas and Scikit-learn designed to handle large and complex amounts of data. The use of Python allowed for fast and accurate data processing, reducing the risk of manual errors. Additionally, Python offered flexibility in data analysis and visualization, allowing for the design of more complex models and deeper interpretations.

B. Pre-processing Data Results

The pre-processing data analysis stage included handling missing values and normalizing data. After checking that no missing values were found on this dateset, the dataset was normalized with the results shown in **Table 4**.

Table 4. Datasets after normalization

ID	Sex	Marital status	Age	Education	Income	Occupation	Settlement size
100000003	0	0	0,645	0,333	0,288	0	0
100000006	0	0	0,461	0,333	0,468	0	0
100000009	0	1	0,801	0,667	0,490	0	0
…..	…	……	……	……	……	……	……
100000028	1	1	0,551	0,667	0,527	0,5	0,5
100000035	1	1	0,434	0,333	0,501	1	0,5

Table 4 presents individual data of customers with various demographic and socioeconomic characteristics. This data has been normalized using the min-max formula as in equation 1, meaning that each dataset attribute was scaled to a range of 0 to 1. This normalization was very useful for ensuring that all attributes were at the same scale so that algorithm analysis can be performed more effectively and accurately.

C. Clustering Model Creation

1. *Application of K-Means Clustering*

The first stage of applying the model to K-Means Clustering was the determination of the number of clusters (k). In the study, the determination of the optimal number of clusters was carried out by testing various k values, ranging from k = 2 to k = 7 using Python. Each k value was tested using the Silhouette Score evaluation method to assess the quality of the clusters formed.

However, for the example of applying manual calculations with Microsoft Excel, only *the* k = 3 value was tested. Once the number of k has been determined, the next step was to initialize the centroid position for each cluster. This centroid can be initialized randomly, by selecting a data point as a centroid.

This stage was the first iteration where the initial values of the cluster center were identified and updated during the algorithm iteration until the final configuration of the cluster was reached. The cluster center or initial centroid is presented in **Table 5**.

Table 5. *Initial Centroid*

ID	Sex	Marital status	Age	Education	Income	Occupation	Settlement size
100000003	0	0	0,645	0,333	0,288	0	0
100000007	0	0	0,697	0,333	0,506	0,5	0,5
100000011	1	1	0,329	0,333	0,351	0	0

In **Table 5**, the initial three centroids of the normalized customer data were selected. The selected centroids were customers with 100000003, 100000007, and 100000011 IDs. After that, calculating the distance of the data point to the centroid of each cluster was done using equation (2) on the first data point.

The calculation of the distance between the first data point and the first centroid in cluster 1 was d (1,1), the distance between the first data point and the second centroid in cluster 2 d(1,2), and the distance between the first data point and the third centroid in cluster 3 d(1,3). The following is the process of calculating the distance between the first data point and the three centroids of the cluster.

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

$$d(1,1) = \sqrt{(0-0)^2 + (0-0)^2 + (0.88-0.65)^2 + (0.67-0.33)^2 + (0.40-0.29)^2 + (0.5-0)^2 + (1-0)^2} =$$

$$d(1,1) = 1.25$$

$$d(1,2) = \sqrt{(0-0)^2 + (0-0)^2 + (0.88-0.70)^2 + (0.67-0.33)^2 + (0.40-0.51)^2 + (0.5-0{,}5)^2 + (1-0{,}5)^2} =$$

$$d(1,2) = 0.54$$

$$d(1,3) = \sqrt{(0-1)^2 + (0-1)^2 + (0.88-0.33)^2 + (0.67-0.33)^2 + (0.40-0.35)^2 + (0.5-0)^2 + (1-0)^2} =$$

$$d(1,3) = 1.89$$

In this example, the results of the calculation of the distance of the first data point d (1.1), d (1.2), d (1.3) show that d (1.2) which was the distance of the first data point to the centroid of cluster 2 was the smallest of 0.54. Meanwhile, the distance between the first data point and *the* centroid of cluster one d (1.1) was larger, namely 1.25, and the distance between the first data point and the *centroid* of the third cluster d (1.3) was larger, namely 1.89. The results of the calculation of the distance d (1.1) to d (1.3) were then grouped based on the closest distance between the first data point and the three centroid clusters. This first data point was closer to the centroid of cluster 2, which was 0.54 compared to the distance to the centroid of cluster 1 d (1.1) or to the centroid of cluster 3 d (1.3) so that this data point was grouped in cluster 2.

Table 6. Iteration 1 grouping results

ID	Sex	Marital status	Age	Education	Income	Occupation	Settlement size	Cluster
100000003	0	0	0,645	0,333	0,288	0	0	1
……	….	……	……	……	……	……	……	……
100000036	0	0	0	0,329	0	0,338	0	1
100000001	0	0	0,882	0,667	0,403	0,5	1	2
…..	….		…..	…..	…..	…..	…..	…..
100000020	0	0	0,408	0	0,468	0,5	0,5	2
100000002	1	1	0,289	0,333	0,487	0,5	1	3
…..	….		…..	…..	…..	…..	…..	…..
100000035	1	1	0,434	0,333	0,503	1	0,5	3

The process of calculating the distance for the second data point to the last data point with three centroid clusters went through the same process. The results of grouping each data point into three clusters in the first iteration were shown in **Table 6**. **Table 6** shows the results of the three clusters grouping the first data point in the first cluster was the data point with ID 100000003. This cluster had 9 data points, up to data points ID 100000036. Furthermore, cluster 2 consisted of ID100000001 to 100000020, 10 data points. Cluster 3 consisted of IDs 100000002 to 100000035, 8 data points. The results of the grouping in iteration 1 were analyzed to determine whether any changes occurred in the composition of the data points within each cluster when compared to the grouping from iteration 2. If there was no change, the calculation of K-Means ended in iteration 2, on the other hand, if there was a change, the calculation continued in iteration 3 with the same process. Furthermore, for the grouping of iteration 2, the average of each attribute of the three clusters' data points was first calculated to get a new centroid. The results of the new centroid are presented in **Table 7**.

Table 7. *Centroid* for iteration 2

Cluster	Sex	Marital status	Age	Education	Income	Occupation	Settlement size
1	0	0,222	0,599	0,407	0,322	0	0
2	0,1	0,2	0,571	0,333	0,536	0,65	0,55
3	1	1	0,375	0,375	0,406	0,5	0,437

Table 7 shows the new centroid values of each cluster attribute (k) in the 2nd iteration. The results of this centroid were then used to group the data points of each cluster with the same distance calculation process as in the previous example using Equation 2. Furthermore, after the distance value of each data point was obtained, the data points were grouped into clusters displayed in **Table 8**.

Table 8. Iteration 2 grouping results

ID	Sex	Marital status	Age	Education	Income	Occupation	Settlement size	Cluster
100000003	0	0	0,645	0,333	0,288	0	0	1
……	….	……	……	……	……	……	……	……
100000017	0	0	0,697	0,333	0,405	0,5	0	1
100000001	0	0	0,882	0,667	0,403	0,5	1	2
…..	….		…..	…..	…..	…..	…..	…..
100000020	0	0	0,408	0	0,468	0,5	0,5	2
100000019	1	1	0,578	0,667	0,846	1	1	3
…..	….		…..	…..	…..	…..	…..	…..
100000035	1	1	0,434	0,333	0,503	1	0,5	3

Table 8 shows the grouping results of the three cluster in iteration 2, which had different results from iteration 1. The result of iteration 1 in cluster 1 consisted of 9 data points, while iteration 2 in cluster 1 consisted of 10 data points. This was due to the shift in the composition of the data point with ID 100000017 from cluster 2 in iteration 1 to cluster 1 in iteration 2. Similarly, the results of iteration 1 cluster 2 consisted of 10 data points, while iteration 2 cluster 2 consisted of 8 data points, due to the transfer of the composition of the data point with ID 100000019 from cluster 2 of the first iteration to cluster 3 of iteration 2. Likewise, iteration 1 cluster 3 consisted of 8 data points, while iteration 2 cluster 3 consists of 9 data points, due to the shift in the composition of data points with ID 100000019 from cluster 2 iteration 1 to cluster 3 iteration 2.

Table 9. Iteration 7 grouping results

ID	Sex	Marital status	Age	Education	Income	Occupation	Settlement size	Cluster
100000003	0	0	0,645	0,333	0,288	0	0	1
……	….	……	……	……	……	……	……	……
100000017	0	0	0	0,329	0	0,338	0	1
100000001	0	0	0,882	0,667	0,403	0,5	1	2
…..	….		…..	…..	…..	…..	…..	…..
100000020	0	0	0,408	0	0,468	0,5	0,5	2
100000019	1	1	0,289	0,333	0,487	0,5	1	3
…..	….		…..	…..	…..	…..	…..	…..
100000028	1	1	0,552	0,667	0,526	0,5	0,5	3

Based on these results, it can be said that the grouping of data points in iteration 2 had not reached convergence, so it required to be continued to the next iteration. The calculation continued in iteration 3 until the end of iteration 7 because there was no change in the composition of the data points of each cluster so that it could be said that convergence had been achieved. The results of the grouping of iteration 7 are shown in **Table 9**. **Table 9** shows the grouping results in iteration 7 for each cluster. The results of the data point grouping of iteration 7, then interpreted with the results of the analysis showing three different customer groups based on their characteristics. Cluster 1 consists of 10 customers who were mostly male, single, older, education at the high school level, with relatively low annual incomes, and tended to work in unskilled or unemployed fields. This group mostly lived in small towns. Cluster 2, which consisted of 8 customers, was dominated by men who were mostly single, middle-aged, university graduates, and work in high-quality management or professional positions. Customers in this cluster had relatively high annual income and tended to live in big cities. Cluster 3 included 9 customers, majority of whom were women, married, with young to middle ages. This group had university education, a moderate annual income, worked in a professional field, and lived in a medium or large city. These three clusters show significant differences in demographic, economic, and geographic characteristics, which can be used as a basis for more specific and effective customer segmentation strategies.

1. Cluster Evaluation Using Silhouette Score

The results of K-Means Clustering algorithm on customer datasets, then evaluated using the Silhouette Score method to assess the quality of the clusters formed. The Silhouette Score measured how well each data in a cluster related to other data in the same cluster (cohesion) and then compared to other clusters (segregation). The Silhouette coefficient value was calculated for each data point, which then provided an idea of the internal consistency of the cluster formed. Furthermore, this study calculated the Silhouette Score value for the composition of the dataset with 3 clusters (k = 3). This process involved the following steps:

The first step was to calculate the cohesion value a_i for each data point within the cluster. Cohesion referes to how close or homogeneous the data points were in the same cluster. In the Silhouette Score calculation, the cohesion value was calculated as the average distance between a given data point and all other data points in the same cluster. The value of a_i was the cohesion value of the first data point a_1 to the last data point (a_{27}). In the calculation for the composition of the dataset with 3 clusters (k = 3), cohesion was calculated for 10 data points grouped in cluster 1, 8 data points in cluster 2 and 9 data points in cluster 3. The calculation of the cohesion value (a_i) in cluster 1 was shown in **Table 10**.

Table 10. Cluster 1 cohesion value

D_n	$data\ point_1$	$data\ point_2$	$data\ point_3$	$data\ point_4$	$data\ point_5$	$data\ point_6$	$data\ point_7$	$data\ point_8$	$data\ point_9$	$data\ point_{10}$
D_1	0	1,229	1,040	0,136	0,029	1,001	1,171	0,624	0,501	0,964
D_2	0,066	0	1,180	0,214	1,019	0,190	1,159	0,379	0,213	0,778
….	……	……	……	……	……	……	……	……	……	……
D_9	0,267	0,067	1,229	0	0,137	0,029	1,001	1,171	0,624	1,010
a_i	0,337	0,380	1,144	0,445	0,335	0,364	1,023	0,432	0,593	1,157

In cluster 1, the first data point was the customer with ID 100000003, the cohesion value for this first data point was 0.337 (a_1), the second point data obtained a cohesion value of 0.380 (a_2). Furthermore, the results of the cohesion calculation were displayed as an example in cluster 1. The same calculation process was also carried out to the data points in cluster 2 and cluster 3.

The second step was to calculate the separation value. Separation measures the distance between a data in a particular cluster and the datae points in other clusters. The value of A_1 was the average distance of the first data point in cluster 1 with all cluster 2 data points, while A_2 was the average distance of the first data point in cluster 1 with all cluster 3 data points. The separation value was calculated as the minimum or smallest value of the average distance between the data point in cluster 1 and its neighboring data points in cluster 2 (A_1) and cluster 3 (A_2).

The separation calculation at k = 3 had 10 data points grouped in cluster 1, then in cluster 2 there were 8 data points, and in cluster 3 there were 9 data points. The results of the calculation of the distance of the data point of cluster 1 to its neighboring clusters (cluster 2 and cluster 3) to determine the separation value (b_i) can be seen in **Table 11**.

Table 11 is the result of the value calculation of b_i for all data points in cluster 1. The calculation of cohesion and separation values is carried out with the same calculation process in clusters 2 and 3.

Table 11. Cluster 1 separation value

Cluster	$data\ point_1$	$data\ poi$	$data\ point_3$	$data\ point_4$	$data\ point_5$	$data\ point_6$	$data\ point_7$	$data\ point_8$	$data\ point_9$	$data\ point_{10}$
A_1	1,106	1,058	1,998	1,292	1,103	1,133	1,885	1,278	1,212	0,895
A_2	2,955	2,870	2,998	3,109	2,948	2,921	1,934	3,081	3,045	2,581
b_i	1,106	1,058	1,998	1,292	1,103	1,133	1,885	1,278	1,212	0,895

The last step was to calculate the Silhouette Score value which was the average of the Silhouette coefficient (s_i) for all data point in the tested cluster (k = 3). The value of s_i was calculated using equation 5, namely the separation value (b_i) minus the cohesion value (a_i). The result of the reduction was divided by the largest (maximum) value between the two values (a_i, b_i). In the number of clusters k = 3, there were 27 data points whose Silhouette coefficient values had been calculated (s_1 to s_{27}). The Silhouette Score was then obtained as the average of all Silhouette coefficient for the data points. The results of the Silhouette Score calculation is presented in **Table 12**.

Table 12. Silhouette score score k=3

s_i	Silhouette Coefficient	Silhouette Score
s_1	0.695	0.513
s_2	0.641	
s_3	0.427	
…	…	
s_{27}	0.698	

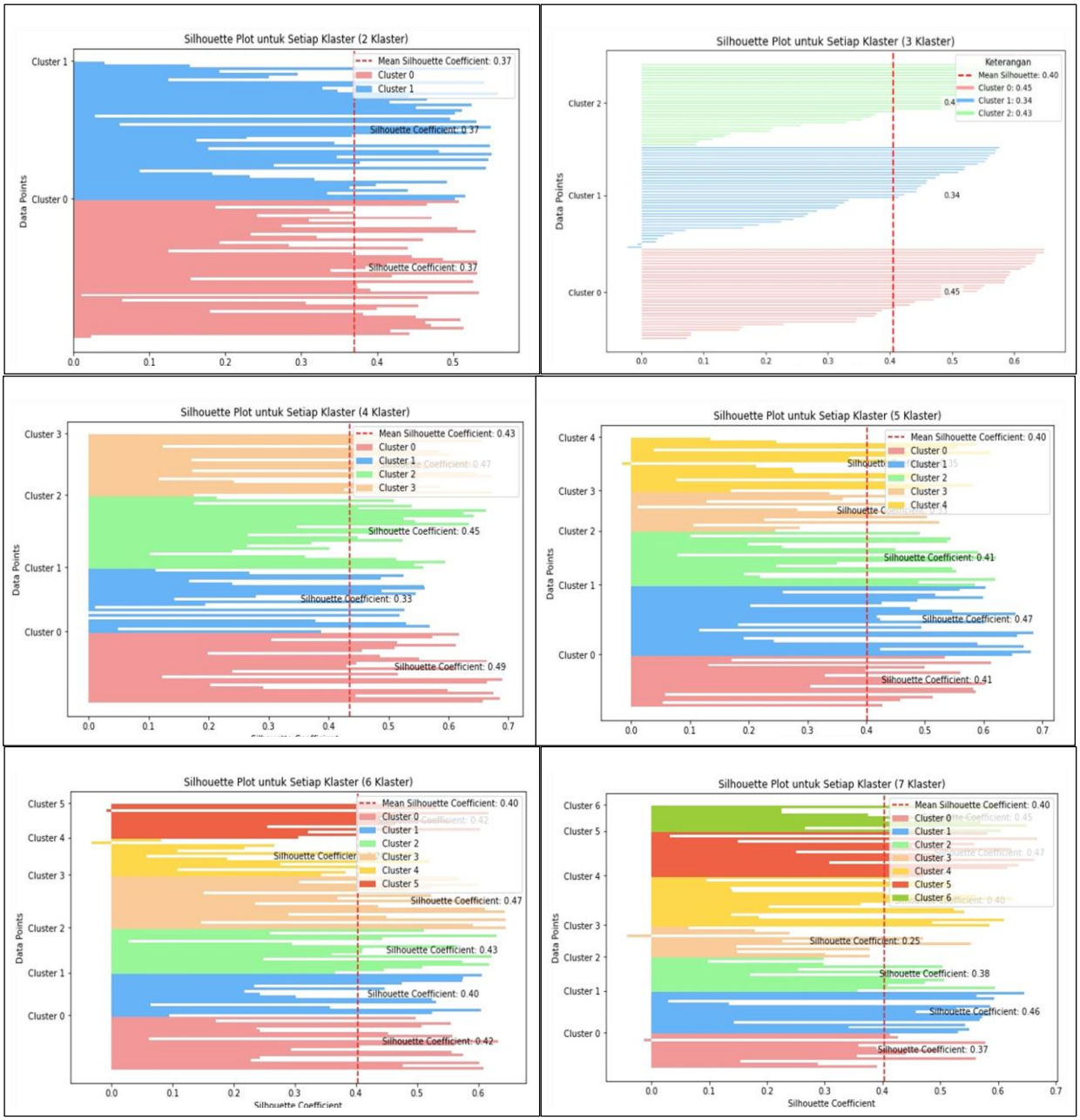

Figure 2. Cluster testing k=2 to k=7 with sillhouette score

Table 12 shows that the Silhouette Score value for all 3 clusters was 0.513. From these results, it can beconcluded that the Silhouette Score value of k = 3 can be categorized in structures with moderate strength. The calculation of the Silhouette Score value was carried out with the same calculation process on different clusters being tested (k = n). Furthermore, the application of K-Means grouping and evaluation of Silhouette Score clusters were applied for cluster numbers ranging from k = 2 to k = 7 using Python for 2000 data points. The dataset had gone through the same process as the previous dataset, namely pre-processing (checking missing values and normalizing data), clustering K-Means, then the cluster results were evaluated using Silhouette Scores to find the optimal number of clusters. The following results of the Silhouette Score test with Pyhtons from k = 2 to k = 7 were presented in **Figure 2**

The clustering test results shown in **Figure 2** consisted of cluster 1 encoded with cluster 0, cluster 2 encoded with cluster 1 and so on. The results of the Silhouette Score were seen from the average value (mean) of Silhouette coefficient. The value closest to 1 was the k = 4 cluster, with a value of 0.43 categorized as a weak structure category. Furthermore, the cluster values k = 3, k = 5 to k = 7 had a smaller value of 0.40 and the k = 2 cluster had the smallest value, 0.37. All of them were in the category of weak structure.

Based on the test, it can be concluded that the most optimal number of clusters was four clusters (k = 4). Although referring to **Table 2**, a value of 0.43 indicated as a weak cluster structure, this was the highest value compared to the other cluster numbers in the test. This indicated that in the context of the dataset and attributes used, dividing the data into four clusters provided better results compared to other cluster numbers, although the quality of the clusters still needed improvement. Higher Silhouette coefficient values generally indicated better clusters. Therefore, it was advisable to consider other methods or aspects that can improve the quality of the cluster, such as reviewing the attributes used, data normalization methods, or even clustering algorithms that are likely to be more suitable.

Conclusion

This research has applied the K-Means Clustering algorithm and the Silhouette Score method to customer segmentation using manual calculations in Microsoft Excel and using Python for the automatic calculation. In manual calculations, k = 3 was tested with 27 data points, while calculations using Python, k = 2 to k = 7 were tested with 2000 data points. The results of this manual calculation were expected to make it easier to understand the logic of the method used and to provide a practical basis before applying the technique on a larger data scale. Meanwhile, the results using Python with a large data scale showed that the optimal number of clusters was four clusters (k = 4) and the Silhouette Score value was 0.43, categorized as weak cluster structure. Thus, although four clusters (k = 4) were considered as the optimal number of clusters in this test, the quality requires further improvement. Further research is suggested to explore other clustering algorithms, such as the Gaussian Mixture Model (GMM) or DBSCAN, as well as consider dynamic data analysis to improve segmentation accuracy. The contribution of this research to the public is to provide a framework that can be used by various companies to understand their customer behavior, thereby improving customer satisfaction and competitiveness in the market.

Acknowledgments

Thank you to the UAD Institute for Research and Community Service (LPPM) for the approval of funding for this research through the DRTPM 2024 Research Program with the main contract number 107/E5/PG.02.00.PL/2024 and the derivative contract number 0609.12/LL5-INT/AL.04; 059/PTM/LPPM-UAD/VI/2024, as well as to Kaggle for providing a very valuable dataset and to the Master of Informatics Engineering Study Program, Ahmad Dahlan University for their very important support in this research. We highly appreciate the collaboration and support from all parties. Hopefully the results of this research can make a positive contribution to the progress of science and public health.

Reference

[1] S. Magatef, M. Al-Okaily, L. Ashour, and T. Abuhussein, "The impact of electronic customer relationship management strategies on customer loyalty: A mediated model," *J. Open Innov. Technol. Mark. Complex.*, vol. 9, no. 4, p. 100149, 2023, doi: 10.1016/j.joitmc.2023.100149.

[2] H. (Hojatollah) Hamidi and B. Haghi, "An approach based on data mining and genetic algorithm to optimize time series clustering for efficient segmentation of customer behavior," *Comput. Hum. Behav. Reports*, vol. 16, no. November, p. 100520, 2024, doi: 10.1016/j.chbr.2024.100520.

[3] X. Ma and X. Gu, "New marketing strategy model of E-commerce enterprises in the era of digital economy," *Heliyon*, vol. 10, no. 8, p. e29038, 2024, doi: 10.1016/j.heliyon.2024.e29038.

[4] J. M. John, O. Shobayo, and B. Ogunleye, "An Exploration of Clustering Algorithms for Customer Segmentation in the UK Retail Market," *Analytics*, vol. 2, no. 4, pp. 809–823, 2023, doi: 10.3390/analytics2040042.

[5] S. Abdul-Rahman, N. F. K. Arifin, M. Hanafiah, and S. Mutalib, "Customer Segmentation and Profiling for Life Insurance using K-Modes Clustering and Decision Tree Classifier," *Int. J. Adv. Comput. Sci. Appl.*, vol. 12, no. 9, pp. 434–444, 2021, doi: 10.14569/IJACSA.2021.0120950.

[6] S. J, C. Gangadhar, R. K. Arora, P. N. Renjith, J. Bamini, and Y. devidas Chincholkar, "E-commerce customer churn prevention using machine learning-based business intelligence strategy," *Meas. Sensors*, vol. 27, no. February, p. 100728, 2023, doi: 10.1016/j.measen.2023.100728.

[7] L. Li, L. Yuan, and J. Tian, "Influence of online E-commerce interaction on consumer satisfaction based on big data algorithm," *Heliyon*, vol. 9, no. 8, 2023, doi: 10.1016/j.heliyon.2023.e18322.

[8] M. S. E. Kasem, M. Hamada, and I. Taj-Eddin, "Customer profiling, segmentation, and sales prediction using AI in direct marketing," *Neural Comput. Appl.*, vol. 36, no. 9, pp. 4995–5005, 2024, doi: 10.1007/s00521-023-09339-6.

[9] 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.

[10] 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, no. April, p. 113874, 2023, doi: 10.1016/j.jbusres.2023.113874.

[11] M. Hänninen, L. Mitronen, and S. K. Kwan, "Multi-sided marketplaces and the transformation of retail: A service systems perspective," *J. Retail. Consum. Serv.*, vol. 49, no. April, pp. 380–388, 2019, doi: 10.1016/j.jretconser.2019.04.015.

[12] A. Griva, E. Zampou, V. Stavrou, D. Papakiriakopoulos, and G. Doukidis, "A two-stage business analytics approach to perform behavioural and geographic customer segmentation using e-commerce delivery data," *J. Decis. Syst.*, Vol. 33, No. 1, pp. 1–29, 2024, doi: 10.1080/12460125.2022.2151071.

[13] 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.

[14] 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, 2004, doi: 10.1016/j.eswa.2004.01.010.

[15] K. Tabianan, S. Velu, and V. Ravi, "K-Means Clustering Approach for Intelligent Customer Segmentation Using Customer Purchase Behavior Data," *Sustain.*, vol. 14, no. 12, pp. 1–15, 2022, period: 10.3390/su14127243.

[16] 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, 2021, doi: 10.1186/s13638-021-01910-w.

[17] R. Passarella, T. Marsyah, O. Arsalan, and M. Shahriman, "Anomaly detection in commercial aircraft landing at SSK II airport using clustering method," *Aerosp. Traffic Saf.*, no. July, pp. 0–1, 2024, doi: 10.1016/j.aets.2024.12.004.

[18] 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, 2024, doi: 10.1016/j.heliyon.2024.e38768.

[19] J. Meng *et al.*, "Nano-integrating green and low-carbon concepts into ideological and political education in higher education institutions through K-Means Clustering," *Heliyon*, vol. 10, no. 10, p. e31244, 2024, doi: 10.1016/j.heliyon.2024.e31244.

[20] M. A. I. Gazi, A. Al Mamun, A. Al Masud, A. R. bin S. Senathirajah, and T. Rahman, "The relationship between CRM, knowledge management, organization commitment, customer profitability and customer loyalty in telecommunication industry: The mediating role of customer satisfaction and the moderating role of brand image," *J. Open Innov. Technol. Mark. Complex.*, vol. 10, no. 1, p. 100227, 2024, doi: 10.1016/j.joitmc.2024.100227.

[21] W. Zhang, L. Wu, and S. Zhang, "Clinical phenotype of ARDS based on K-Means cluster analysis: A study from the eICU database," *Heliyon*, vol. 10, no. 20, p. e39198, 2024, doi: 10.1016/j.heliyon.2024.e39198.

[22] Y. Li, X. Chu, D. Tian, J. Feng, and W. Mu, "Customer segmentation using K-Means Clustering and the adaptive particle swarm optimization algorithm," *Appl. Soft Comput.*, vol. 113, p. 107924, 2021, doi:

10.1016/j.asoc.2021.107924.

[23] 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, no. October 2023, 2024, doi: 10.1016/j.eswa.2023.122310.

[24] K. ŞAHİNBAŞ, "Performance Comparison of K-Means and DBSCAN Methods for Airline Customer Segmentation," *Black Sea J. Eng. Sci.*, vol. 5, no. 4, pp. 158–165, 2022, doi: 10.34248/bsengineering.1170943.

[25] M. Faisal, E. M. Zamzami, and Sutarman, "Comparative Analysis of Inter-Centroid K-Means Performance using Euclidean Distance, Canberra Distance and Manhattan Distance," *J. Phys. Conf. Ser.*, vol. 1566, no. 1, 2020, doi: 10.1088/1742-6596/1566/1/012112.

[26] A. Al Mamun, P. P. Em, M. J. Hossen, B. Jahan, and A. Tahabilder, "A deep learning approach for lane marking detection applying encode-decode instant segmentation network," *Heliyon*, vol. 9, no. 3, p. e14212, 2023, doi: 10.1016/j.heliyon.2023.e14212.

[27] K. Sya, H. Yuliansyah, and I. Arfiani, "Clustering Student Data Based On K - Means Algorithms," vol. 8, no. 08, pp. 1014–1018, 2019.

[28] F. Grandoni, R. Ostrovsky, Y. Rabani, L. J. Schulman, and R. Venkat, "A refined approximation for Euclidean K-Means," *Inf. Process. Lett.*, vol. 176, pp. 1–9, 2022, doi: 10.1016/j.ipl.2022.106251.

[29] A. Rachwał *et al.*, "Determining the Quality of a Dataset in Clustering Terms," *Appl. Sci.*, vol. 13, no. 5, pp. 1–20, 2023, doi: 10.3390/app13052942.

[30] 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, 2022, doi: 10.1016/j.jksuci.2019.12.011.

[31] G. Liu, "A New Index for Clustering Evaluation Based on Density Estimation," 2022.