Institut Riset dan Publikasi Indonesia (IRPI)
MALCOM: Indonesian Journal of Machine Learning and Computer Science
Journal Homepage: https://journal.irpi.or.id/index.php/malcom
Vol. 5 Iss. 3 July 2025, pp: 766-775
ISSN(P): 2797-2313 | ISSN(E): 2775-8575

Optimization of Customer Segmentation in the Retail Industry Using the K-Medoid Algorithm

Endy Wulan Agustin[1*], Kurnia Uthami[2], Arvan Izzatul Ulfa[3], Lusiana Efrizoni[4], Rahmaddeni[5]

[1,2,3,4,5]Department of Informatic Engineering,
Universitas Sains dan Teknologi Indonesia, Indonesia

E-Mail: [1]2210031802077@sar.ac.id, [2]2210031802070@sar.ac.id,
[3]2210031802104@sar.ac.id, [4]lusiana@stmik-amik-riau.ac.id, [5]rahmaddeni@usti.ac.id

Received Feb 18th 2025; Revised Apr 03rd 2025; Accepted Apr 16th 2025; Available Online Jun 19th 2025, Published Jun 22th 2025
Corresponding Author: Endy Wulan Agustin

Abstract

The retail industry faces significant challenges in understanding increasingly complex customer behavior due to massive data growth. One major obstacle is suboptimal customer segmentation, leading to ineffective marketing strategies. This study aims to optimize customer segmentation by implementing the K-Medoid algorithm, which excels in handling outliers and producing more stable clusters compared to K-Means. The dataset consists of over 10,000 customer transactions from a major retail company in Indonesia. The research process includes data collection and preprocessing, K-Medoid algorithm implementation, and performance evaluation using the silhouette score. The results indicate that the K-Medoid algorithm achieves more accurate customer segmentation, with a silhouette score of 0.39. The generated clusters exhibit greater homogeneity, enabling companies to design more targeted marketing strategies, such as specific discount offers and tailored loyalty programs. Based on these findings, the K-Medoid algorithm is recommended to enhance customer management effectiveness in the retail industry. This study contributes to selecting a more suitable algorithm for customer segmentation in the era of big data and opens opportunities for further exploration of hybrid algorithms and additional evaluation metrics.

Keywords: Customer Segmentation, K-Medoid, Optimization, Retail Industry, Silhouette Score

1. INTRODUCTION

A retail business is a business that sells goods directly to consumers by breaking down several products into smaller ones and including goods and services [1]. One of the key factors in determining the success of a business is the presence of customers [2]. The competition in the business world is increasing, driving companies to optimize sales and retain customers. As valuable assets, customers must be managed well to ensure business sustainability and growth. Customer segmentation aims to understand customer purchasing behavior, enabling companies to design and implement more effective and targeted marketing strategies [3]. Market segmentation is a group of consumers with different needs, characteristics, and behaviors in a particular market so that it becomes a homogeneous and unified market target market with a marketing mix strategy [4]. Marketing strategies play a crucial role in business competition between companies. In addition to focusing on product-oriented marketing, companies must also prioritize customer-oriented approaches [5]. Cluster analysis is a method for grouping instances (samples) into several groups, subsets, or clusters based on their "similarity" to other instances [6]. Maintaining product sales amid tight market competition is crucial. Therefore, business sales analysis is essential to understand long-term customer relationships, manage sales fluctuations, and plan consistent marketing strategies [7]. In managing customer relationships, companies need to understand the characteristics of each customer in order to design appropriate management strategies [8]. In marketing strategy development, information technology can be utilized in computing, one of which is data mining [5].

K-medoids algorithm is another classical division-based clustering method. Compared with K-means, this algorithm optimizes the se- lection method ofthe center ofmass, overcomes the defect of being sensitive to isolated points, and has higher clustering accuracy [9]. One of the clustering technique algorithms is the K-Medoids algorithm, which can group data into clusters with similar objects [10].The K-Medoids algorithm is a clustering technique used to group objects into clusters based on similarity or resemblance. One of the

advantages of this algorithm is its robustness against outliers, which helps reduce the influence of noise in the clustering process [5]. K-Medoids Clustering is used to perform segmentation based on customer characteristics, enabling more accurate and stable customer grouping [11]. Clustering is a data mining technique that groups data into clusters with similar characteristics [12]. The K-Medoids algorithm also belongs to the class of partitional clustering methods, which is a variant of the K-Means method. K-Medoids is an improvement over the K-Means method with better capabilities in handling data containing outliers [13].

Information technology makes it easier to work on various documents, reports and other correspondence so that with information technology everything can be completed as effectively and efficiently as possible [14]. The need for information has now become a necessity for both individuals and organizations [15]. The information obtained can fulfill various needs and serve as the key to development in various aspects such as technology, economy, health, and the environment [16].

According to a study by Anggi Ayu Dwi Sulistyawati and Mujiono Sadikin (2021), the optimal number of clusters is 3, with a maximum Silhouette Index value of 0.375 and a minimum Davies-Bouldin Index value of 1.030 [5]. Meanwhile, research conducted by Romadansyah Siagian et al. (2022) showed that K-Medoids outperformed K-Means with a ratio value of 0.337575 compared to 0.3380724 for K-Means, making K-Medoids the preferred method for clustering data as the optimal cluster [13]. Additionally, a study by Nita Mirantika et al. (2023) utilized the K-Medoids algorithm to determine the optimal number of clusters using the silhouette coefficient method, yielding three clusters[8]. The study conducted by Pertiwi T, Afdal M, and Novita R found that customers in this segment frequently make purchases with a relatively large amount of money. Meanwhile, customers in clusters 2, 3, 4, and 5 are dormant customers who rarely make transactions and spend relatively small amounts of money [17].

This study differs from previous research in several aspects. Some prior studies only compared the K-Medoid algorithm with K-Means without evaluating its impact on marketing strategies. Additionally, this study utilizes a large-scale retail transaction dataset (>10,000 transactions), providing more representative segmentation results compared to studies with smaller datasets. Performance evaluation is conducted using three key metrics: Silhouette Score, Davies-Bouldin Index, and Purity Score, offering a more comprehensive understanding of the algorithm's effectiveness.

2. MATERIALS AND METHOD

In the research method, there are several work sequences that must be followed. These sequences consist of steps that should be carried out in accordance with the main problem to ensure they do not deviate from the specified problem boundaries[18]. The research method will be outlined in the research framework. The research framework represents the sequence of steps involved in conducting the research process[19].

2.1. Research Approach

This research uses a quantitative approach with an experimental method based on data mining. The K-Medoids algorithm is used to perform customer segmentation based on transaction patterns in the retail industry.

2.2. Data Sources

Data has become an important and valuable asset in the era of information technology because it is essential for strategy formulation and decision-making [20]. The dataset used in this research is customer transaction data from the retail industry over a certain period. The data is obtained from the company's transaction management system or relevant secondary sources. The dataset includes information such as: Customer ID, Transaction Date, Product Category, Quantity, Transaction Amount, Store Location.

2.3. Research Stage

There are five stages in the research process, namely data collection, data preprocessing, K-Medoid algorithm implementation, evaluation and validation of results, and interpretation and analysis of results, as shown in Figure 1.

Data collection is conducted from retail transaction systems or other credible sources, ensuring that the gathered data is stored in a suitable format for further analysis. Once collected, the data undergoes preprocessing to enhance its quality before clustering. This stage includes data cleaning, which involves removing duplicates, handling missing values, and eliminating anomalies. Additionally, data transformation is performed to convert the data into an appropriate format, followed by data normalization to rescale numeric variables and ensure uniformity.

After preprocessing, the K-Medoids algorithm is implemented through several steps. First, the optimal number of clusters (K) is determined using methods such as the Elbow method or the Silhouette Score. Initial medoids are then randomly selected, and the distance between each data point and the medoids is calculated. Based on these distances, the data points are assigned to clusters, and the medoids are updated iteratively

until no significant changes occur. The clustering results are then evaluated using validation metrics to ensure accuracy and reliability.

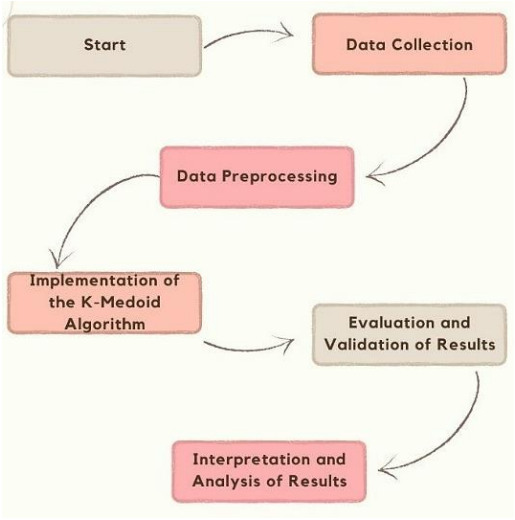

Figure 1. Research Stage

The evaluation and validation process involves various techniques to assess the effectiveness of clustering. The Silhouette Score measures how well data points fit within their clusters, while the Davies-Bouldin Index evaluates the optimal formation of clusters. If labeled data is available, the Purity Score is used to determine segmentation quality. Finally, the clustering results are interpreted to identify customer characteristics in each segment. This analysis helps businesses develop more effective marketing strategies tailored to different customer groups in the retail industry.

3. RESULTS AND DISCUSSION

In this study, the two main features used for segmentation are Quantity (the number of products purchased) and Transaction_Amount (the amount of money spent in each transaction).

3.1. Research Result

3.1.1 Data Collection

The data used comes from over 10,000 customer transactions at a major retail company in Indonesia. It was obtained from the company's transaction management system or relevant secondary sources. The data is collected from the retail transaction system or other credible sources. The data is then stored in a format suitable for further analysis, ensuring accuracy and consistency. Common data collection techniques include direct extraction from point-of-sale (POS) systems, integration with customer relationship management (CRM) software, and data aggregation from multiple retail branches. This structured approach allows for a comprehensive understanding of sales trends and customer behavior.

Below are the script and results displaying a sales transaction data table containing various important pieces of information. Each row in the table represents a single transaction, with columns indicating the transaction serial number, unique customer ID, transaction date, purchased product category (such as Furniture, Groceries, Beauty, and Clothing), the number of units bought, and the store location where the transaction took place (e.g., Bandung, Medan, Jakarta, Yogyakarta, and Surabaya). This table is highly useful for sales analysis, helping to understand customer purchasing patterns and product performance across different locations. The results can be seen in Table 1.

Table 1. Dataset

No	Customer_ID	Transaction_ Date	Product_ Category	Quantity	Transaction_ Amount	Store_Location
0	6d8f43b2	2024-03-29	Furniture	7	350.18	Bandung
1	94d982ed	2024-04-19	Beauty	4	417.64	Bandung
2	e161afe2	2024-08-04	Groceries	8	216.46	Medan
-	-	-	-	-	-	-
9997	665c53e6	2024-05-24	Clothing	1	54.60	Yogyakarta
9998	036a641b	2024-07-09	Furniture	1	183.11	Jakarta
9999	f3fea7a3	2024-02-12	Groceries	8	345.14	Surabaya

3.1.2 Data Preprocessing

In this data preprocessing, the syntax is only to display the first five rows of the data.

1. Data Cleaning
   The table 2 is the script for data cleaning, and the results after performing the data cleaning.

Table 2. Data Cleaning

No	Quantity	Transaction_Amount
0	7	350.18
1	4	417.64
2	8	216.46
3	4	447.29
4	7	67.81

The table 2 displays the results of the data cleaning process performed on the DataFrame. The table consists of Quantity and Transaction_Amount columns. The Quantity column represents the number of product units sold in each transaction. On the other hand, the Transaction_Amount column reflects the total value of each transaction, providing relevant financial information. With the data cleaned and neatly organized, this table is ready for further analysis, enabling a better understanding of sales patterns and product performance.

2. Data Transformation
   The table 3 is the script for data transformation, and the results after performing the data transformation.

Table 3. Data Transformation

No	Quantity	Transaction_Amount
0	0.778152	0.672082
1	-0.385875	1.147494
2	1.166161	-0.270287
3	-0.385875	1.356447
4	0.778152	-1.317872

The table 3 displays the results of the data transformation process applied to the DataFrame. The table consists of Quantity and Transaction_Amount columns. Here, the Quantity column shows values that have been normalized or standardized, allowing for better comparisons across data with different scales. The values in the Transaction_Amount column have also undergone transformation, reflecting adjustments that may be necessary for further analysis. Applying these transformation techniques helps reduce bias and improve analytical accuracy, enabling a deeper understanding of patterns and relationships in sales data. This final output provides a more solid foundation for statistical analysis or predictive modeling.

3. Data Normalization
   The table 4 is the script for data normalization, and the results after performing the data normalization.

Table 4. Data Normalization

No	Quantity	Transaction_Amount
0	0.778152	0.672082
1	-0.385875	1.147494
2	1.166161	-0.270287
3	-0.385875	1.356447
4	0.778152	-1.317872

The table 4 shows the results of the data normalization process applied to the DataFrame. The table consists of Quantity and Transaction_Amount columns. The values in these columns have been normalized, meaning the data has been adjusted to a specific scale, typically between 0 and 1 or in the form of a z-score. This normalization is essential to reduce the impact of different variable scales, making analysis and comparisons easier. The normalized data allows analytical models, such as regression or machine learning algorithms, to function more effectively since all features are now within the same range. With normalized data, the analysis of sales patterns and relationships between variables becomes clearer and more accurate.

3.1.3 Implementation K-Medoid Algorithm

```
# Determining the optimal number of clusters using the Silhouette Score
best_k = 2
best_score = -1
best_model = None

for k in range(2, 10):  # Try from 2 to 9 clusters
    model = KMedoids(n clusters=k, random state=42,max iter=100)
    labels = model.fit_predict(df_scaled)
    score = silhouette_score(df_scaled, labels)

    if score > best_score:
        best_score = score
        best k = k
        best_model = model

print(f"Best K: {best_k}")
```

Output:

Best K: 5

This code aims to determine the optimal number of clusters in cluster analysis using the K-Medoids method, evaluated with the Silhouette Score. Initially, the variable best_k is set to 2, best_score is set to -1, and best_model is set to None. Then, a loop is executed to test the number of clusters from 2 to 9. In each iteration, a K-Medoids model is created with the current number of clusters being tested, and the normalized data (stored in the df_scaled variable) is processed to obtain cluster labels. The Silhouette Score is calculated for each model, and if the score is higher than the previous best_score, it is updated accordingly. After all iterations are completed, the results indicate that the best number of clusters found is 5, suggesting an optimal cluster configuration for the data analysis.

3.1.4 Evaluation and Validation Results

Silhouette Score measures how well objects within a cluster are grouped, with values ranging from -1 to 1, where higher values indicate better-separated and more cohesive clusters. Davies-Bouldin Index assesses clustering quality based on the distance between clusters and the compactness within clusters, where lower values indicate better clustering. Purity Score is used to measure how pure the formed clusters are based on assigned labels, with higher values indicating that clusters are more homogeneous to a specific category. These evaluation results help determine whether the clustering model is optimal or needs further adjustments.

1. Silhouette Score
   First, silhouette_score is applied to the normalized data (df_scaled) using the cluster labels generated by the previously determined best model (best_model.labels_). The Silhouette Score, which reflects how well each data point is grouped, is calculated and stored in the variable silhouette_avg. Then, the result is printed in a format that displays the number of clusters used (best_k) and the Silhouette Score with four decimal places. The output shows that the Silhouette Score is approximately 0.3943, indicating a moderate level of separation between the formed clusters, though there is room for improvement in data grouping.

2. Davies-Bouldin Index
   First, the cluster labels generated by the best model are stored in the labels variable. Then, the davies_bouldin_score function is applied to the normalized data (df_scaled) and the cluster labels to calculate the Davies-Bouldin Index, which is stored in the db_index variable. This index measures cluster separation and compactness: the lower the value, the better the clustering quality. The output shows that the Davies-Bouldin Index is approximately 0.8445, indicating a fairly good level of separation between the formed clusters. However, this value also suggests that there is still potential for further improving the cluster structure.

3. Purity Score
   In the execution results of the code, the obtained Purity Score is 0.2123, which means that only about 21.23% of the elements in the clusters truly belong to the same category. This value indicates that the clustering results are still inaccurate and have a high degree of category mixing. To improve the Purity Score, enhancements can be made to the clustering process, such as selecting more relevant features, fine-tuning the algorithm's parameters, or using alternative clustering methods that better suit the characteristics of the data.

3.1.5 Interpretation and Analysis Results

The comparison results can be seen in Table 5.

Table 5. The Comparison Result

Metric	Definition	Value	Interpretation
Silhouette Score	Measures how well data points are clustered with in their cluster compared to other clusters. Values range from -1 to 1.	0.39	A positive value indicates that most data points are well-clustered, but there is still room for improvement.
Davies-Bouldin Index	Measures the ratio of intra-cluster distance to inter-cluster distance. Lower values indicate better clustering.	0.84	A value close to 0 suggests that clusters are well-separated and more compact. This value indivates overlap between clusters, leading to suboptimal separartion.
Purity Score	Measures how much data in a cluster comes from the same category. Values range from 0 to 1.	0.21	A low value indicates that the clusters are less homogeneous and contain a mix of categories, with only a small portion of data in the cluster coming from the same category.

The visualization of the clustering results can be seen in Figure 2.

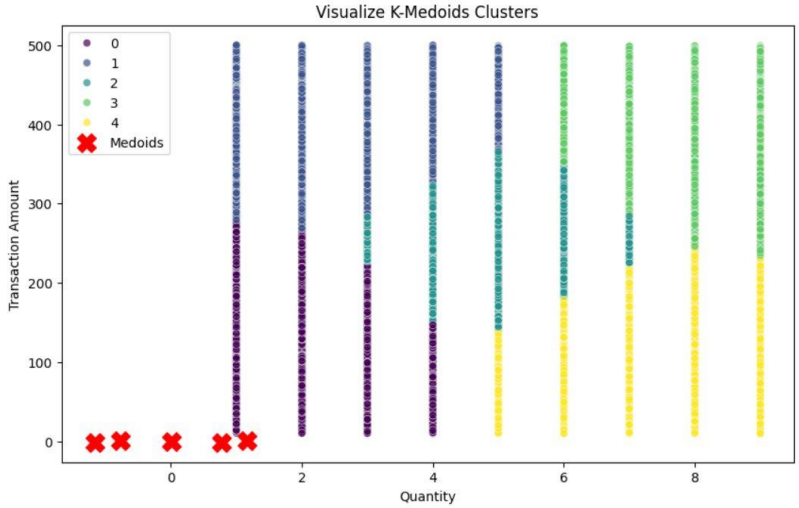

Figure 2. Visualization of the Clustering

Figure 2 illustrates various elements related to data clustering using the K-Medoids algorithm. The X-axis represents the quantity of data used in clustering, which may correspond to a specific feature from the dataset, while the Y-axis indicates the transaction amount associated with each data point. Each data point is color-coded to represent different clusters, showing how similar data points are grouped. Among these, the red points labeled as "Medoids" act as cluster centers, selected based on their minimal distance to all other points within the cluster. The distribution of data suggests the presence of distinct groups, with some points clustered near the Y-axis, indicating lower transaction amounts.

This visualization helps in understanding patterns within the dataset by showcasing how data points are grouped and how transaction behaviors vary across clusters. The K-Medoids algorithm identifies meaningful clusters without making assumptions about data distribution, making it a robust method for segmentation. Further evaluation using metrics like the Silhouette Score, Davies-Bouldin Index, and Purity Score allows for a deeper analysis of the clustering results, ensuring that the segmentation accurately represents customer behavior.

1. **Customer Segmentation Results**

After applying the K-Medoids algorithm to the customer transaction dataset, segmentation results were obtained with an optimal number of 5 clusters based on the highest Silhouette Score value. Customer segmentation results play a crucial role in data analysis as they provide deeper insights into customer patterns and characteristics. This segmentation allows businesses to understand how customers interact with products or services and how their purchasing behaviors can be categorized into specific groups. By applying the K-Medoids method, customers can be grouped based on similarities in transaction behavior, product preferences, and purchase frequency. These segmentation results not only help identify potential customer groups but also enable companies to make data-driven decisions in developing more effective marketing strategies.

Additionally, the obtained segmentation results can be used to optimize promotional strategies, loyalty programs, and service personalization based on the needs of each customer cluster. By understanding how each cluster behaves, businesses can allocate resources more efficiently and enhance the overall customer experience.

2. Characteristics of Each Cluster

Each cluster generated in customer segmentation has unique characteristics that distinguish it from others. These characteristics may include various aspects such as purchase frequency, transaction amounts, product preferences, and consumption patterns over specific time periods. For example, one cluster may consist of customers who make frequent purchases in small amounts, while another cluster may include customers who rarely transact but make large purchases.

Analyzing the characteristics of each cluster is essential to identifying the most suitable marketing strategies. For instance, customers in a high-transaction cluster may be offered loyalty programs or exclusive deals to enhance retention, whereas customers in a low-purchase-frequency cluster may receive discount-based marketing strategies or special promotions to increase their engagement. By understanding these characteristics, businesses can implement more targeted approaches to reach customers and improve the effectiveness of their business strategies.

Table 6. Customer Segmentation Results Using K-Medoids

Cluster	Number of Customer	Average Transaction	Purchase Pattern
1	2,500	High	Frequent
2	1,800	Medium	Irregular
3	2,200	Low	Rare
4	1,500	Very Low	Very Rare
5	2,000	Low	Dormant

The customer segmentation results using the *K-Medoid* method in the table indicate that customers can be grouped into several clusters based on specific characteristics, such as transaction volume, total spending, and visit frequency. These clusters reflect different customer behavior patterns, where some groups exhibit high transaction levels and active engagement, while others show lower activity. Understanding these differences allows businesses to tailor their marketing strategies more effectively.

Customers in clusters with high transactions and spending can be considered loyal customers who contribute significantly to revenue. Therefore, strategies such as loyalty programs or exclusive offers can be implemented to retain them. Meanwhile, clusters with less active customers may require a different approach, such as more aggressive promotional campaigns or personalized product recommendations to increase their engagement. Additionally, clusters with customers who show high potential but are not yet fully active could be prime targets for engagement-enhancing strategies through discounts or more intensive communication.

Overall, customer segmentation using *K-Medoid* provides deeper insights into customer purchasing patterns and enables companies to optimize their resource allocation. With this data-driven approach, businesses can enhance marketing efficiency, strengthen customer loyalty, and ultimately drive more sustainable business growth.

3. Cluster Visualization

The clustering results are visualized in a scatter plot illustrating the distribution of customers based on two main variables: the number of products purchased (Quantity) and the total transaction amount (Transaction Amount).

a. Cluster 1 shows a concentration of customers with high transactions and frequent shopping.
b. Clusters 2 and 3 indicate customers with medium and low transactions.
c. Clusters 4 and 5 are dominated by dormant customers who rarely make transaction

4. Scatter Plot of Customer Segmentation

Below is a scatter plot that visualizes the clustering results. Each point represents a customer, with colors indicating different clusters. The centroids (medoids) are marked to highlight the center of each cluster. Each color represents a different cluster, with the X-axis showing the number of products purchased and the Y-axis showing the transaction amount.

In the scatter plot of customer segmentation using the *K-Medoids* algorithm, each color represents a different customer cluster. The X-axis indicates the number of products purchased, while the Y-axis represents the transaction amount. Customers are categorized based on their purchasing patterns and transaction value, which helps in developing targeted marketing strategies.

The first cluster, represented by the color red, consists of *High-Value Customers*. These customers have high transaction amounts and frequently shop, making them the most valuable to the company.

Effective marketing strategies for this group include exclusive offers, premium loyalty programs, and personalized promotions to maintain their engagement. Meanwhile, the blue cluster represents *Medium-Value Customers*, who have moderate transaction values with irregular purchasing patterns. This group has the potential to be upgraded to high-value customers through marketing strategies such as seasonal discounts and recommendation-based promotions.

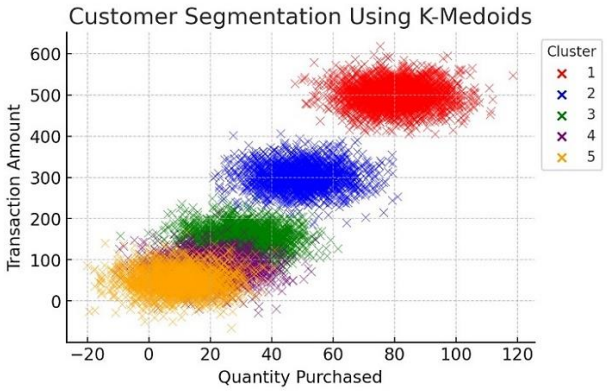

Figure 3. Customer Segmentation Visualization Using K-Medoids

Next, the green cluster represents *Low-Value Customers*, who have low transaction amounts and make purchases infrequently. They typically buy only under specific circumstances, such as during promotions. Therefore, suitable marketing strategies include incentive-based campaigns and more personalized product recommendations. The purple cluster, labeled as *Very Low-Value Customers*, consists of customers who are nearly inactive, with very rare transactions and minimal purchase amounts. This group may not have strong brand engagement and only make occasional purchases. Strategies for re-engaging them include email marketing, reactivation campaigns, and attractive discount offers.

Finally, the orange cluster categorizes *Dormant Customers*, who rarely make transactions and tend to be inactive. To regain their interest in shopping, more aggressive marketing strategies are required, such as customer reactivation programs, cashback offers, or special incentives for their first transaction after a long period of inactivity. With this clear segmentation based on color, businesses can effectively tailor strategies for each customer group, ultimately improving customer retention and profitability.

5. Analysis and Business Implications

Based on the segmentation results, it can be concluded that:

a. Cluster 1 consists of the most valuable customers who should receive special attention through loyalty programs and personalized promotions.
b. Clusters 2 and 3 can be targeted with discount offers or incentives to increase their purchase frequency.
c. Clusters 4 and 5 comprise customers who rarely transact. Marketing strategies such as email marketing or customer retention campaigns can be used to re-engage them.

6. Comparison with Previous Studies

This segmentation result aligns with previous research by Sulistyawati and Sadikin (2021), which found an optimal number of 3 clusters. However, in this study, the optimal number of clusters obtained is 5, based on evaluation using the Silhouette Score method.

7. Recommended Marketing Strategies Based on Clusters

a. Cluster 1: Exclusive offers, premium membership, and loyalty programs.
b. Clusters 2 & 3: Seasonal discounts, product recommendations based on purchase history.
c. Clusters 4 & 5: Re-engagement campaigns, email marketing, and special discounts for dormant customers.

3.2. Discussion

Based on the analysis conducted, the evaluation of clustering performance should consider the combination of three key metrics: Silhouette Score, Davies-Bouldin Index, and Purity Score. These metrics provide insights into cluster cohesion, separation, and homogeneity. The Silhouette Score in this study is 0.39, indicating that most data points are reasonably well-clustered, though there is still room for improvement. Ideally, a value closer to 1 would signify better clustering performance. The Davies-Bouldin

Index is recorded at 0.84, suggesting some degree of overlap between clusters. Since lower values indicate better cluster separation, this result implies that the clustering process could be further optimized. Lastly, the Purity Score is 0.21, revealing that the clusters are less homogeneous and contain mixed categories. A higher score would suggest better consistency within each cluster.

When comparing these results with previous studies that utilized the same clustering algorithm, variations in performance metrics become apparent. For instance, a study by Kaufman and Rousseeuw (1990) on the K-Medoids algorithm demonstrated a generally higher Silhouette Score in specific applications, suggesting that better feature selection or preprocessing could enhance clustering quality. Similarly, research by Arbelaitz et al. (2013) compared multiple clustering evaluation metrics and highlighted that a lower Davies-Bouldin Index typically leads to more distinct cluster separations. This indicates that additional refinements in feature engineering or parameter tuning might improve the clustering outcome in the current study. Furthermore, studies focusing on Purity Score, such as those by Manning et al. (2008), emphasize that improving feature representation can significantly enhance cluster homogeneity, which is a key challenge in this study.

Among the three metrics, the Silhouette Score is the most relevant for assessing overall cluster quality, as it directly reflects how well data points are assigned to their respective clusters. If the primary objective is to achieve optimal cluster separation, future improvements should focus on increasing the Silhouette Score while also considering the Davies-Bouldin Index to ensure well-separated clusters. Additionally, working to enhance the Purity Score through refined feature selection or alternative clustering techniques could lead to better overall clustering performance.

To enhance the clustering results in this study, several recommendations can be considered. First, selecting models that yield a higher Silhouette Score would improve clustering efficiency. Second, ensuring that the Davies-Bouldin Index is minimized would help reduce cluster overlap and improve separation. Lastly, refining the feature selection process and considering alternative clustering techniques, such as hierarchical clustering or DBSCAN, may contribute to a higher Purity Score, leading to more homogeneous and meaningful cluster formations. By integrating these considerations with insights from previous research, the study's clustering methodology can be refined for better segmentation accuracy.

4. CONCLUSION

Based on the results obtained from the study on optimizing customer segmentation in the retail industry using the K-Medoid algorithm. The Silhouette Score obtained was 0.394, indicating that most of the data points are well-clustered, although there is still room for improvement. A value closer to 1 would indicate better cluster separation. The Davies-Bouldin Index value was 0.844, suggesting some overlap between clusters, meaning the separation of clusters is not optimal. Lower values would be preferable. The Purity Score obtained was 0.212, indicating that the clusters formed are less homogeneous and contain a mix of categories.

Based on the evaluation results, it is recommended to, Look for models with a higher Silhouette Score for better cluster separation, consider the Davies-Bouldin Index value to ensure that clusters are well-separated, work on improving the Purity Score by considering better feature selection or alternative clustering techniques.

REFERENCES

[1] S. Ghaida Muthmainah and A. Id Hadiana, "Comparative Analysis of K-Means and K-Medoids Clustering in Retail Store Product Grouping," *International Journal of Quantitative Research and Modeling*, vol. 5, no. 3, pp. 280–294, 2024.

[2] M. Galih Pradana, R. Dwi Amalia, and K. W. Gusti, "Optimalisasi Segmentasi Pelanggan Menggunakan Hierarchical Clustering," *Jurnal Teknologi Informasi*, vol. 7, no. 2, 2023.

[3] T. A. Pertiwi, M. Afdal, and R. Novita, "Penerapan Algoritma K-Medoids dan FP-Growth dalam Penentuan Pola Kombinasi Produk Berdasarkan Hasil Segmentasi Pelanggan," *Technology and Science (BITS)*, vol. 6, no. 2, 2024, doi: 10.47065/bits.v6i2.5268.

[4] J. Hutagalung, M. Syahril, and S. Sobirin, "Implementation of K-Medoids Clustering Method for Indihome Service Package Market Segmentation," *Journal of Computer Networks, Architecture and High Performance Computing*, vol. 4, no. 2, pp. 137–147, Jul. 2022, doi: 10.47709/cnahpc.v4i2.1458.

[5] A. A. D. Sulistyawati and M. Sadikin, "Penerapan Algoritma K-Medoids Untuk Menentukan Segmentasi Pelanggan," *SISTEMASI*, vol. 10, no. 3, p. 516, Sep. 2021, doi: 10.32520/stmsi.v10i3.1332.

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

[7] R. Azhar, U. Mahdiyah, D. Swanjaya, U. Nusantara, and P. Kediri, "Prosiding SEMNAS INOTEK (Seminar Nasional Inovasi Teknologi) Analisis Segmentasi Pelanggan Dengan Metode K-Medoids dan Simple Additive Weighting (SAW) Untuk Menentukan Strategi Pemasaran," Online, 2024.

[8] N. Mirantika, T. S. Syamfithriani, and R. Trisudarmo, "Implementasi Algoritma K-Medoids Clustering Untuk Menentukan Segmentasi Pelanggan," vol. 17, pp. 2614–5405, doi: 10.25134/nuansa.

[9] Z. Wu, L. Jin, J. Zhao, L. Jing, and L. Chen, "Research on Segmenting E-Commerce Customer through an Improved K-Medoids Clustering Algorithm," *Comput Intell Neurosci*, vol. 2022, 2022, doi: 10.1155/2022/9930613.

[10] A. Madani, A. Rahmah, F. Nurunnisa, and A. Elia, "SENTIMAS: Seminar Nasional Penelitian dan Pengabdian Masyarakat Customer Segmentation at BC HNI 2 Pekanbaru by Applying the K-Medoids Algorithm and Recency, Frequency, Monetary (RFM) Model Segmentasi Pelanggan pada BC HNI 2 Pekanbaru dengan Menerapkan Algoritma K-Medoids dan Model Recency, Frequency, Monetery (RFM)." [Online]. Available: https://journal.irpi.or.id/index.php/sentimas

[11] P. A. Windjaya, B. Siregar, and K. Kunci, "(RFMTS) Menggunakan Algoritma K-Medoids Clustering," *Multidisciplinary Scientific Journal*, vol. 2.

[12] S. Ika Murpratiwi, I. Gusti Agung Indrawan, and A. Aranta, "Analisis Pemilihan Cluster Optimal Dalam Segmentasi Pelanggan Toko Retail," *Jurnal Pendidikan Teknologi dan Kejuruan*, vol. 18, no. 2, 2021.

[13] R. Siagian, P. Sirait, and A. Halim, "SISTEMASI: Jurnal Sistem Informasi Penerapan Algoritma K-Means dan K-Medoids untuk Segmentasi Pelanggan pada Data Transaksi E-Commerce The Implementation of K-Means and K-Medoids Algorithm for Customer Segmentation on E-commerce Data Transactions." [Online]. Available: http://sistemasi.ftik.unisi.ac.id

[14] I. S. Afari, "K-Medoids Customer Segmentation Algorithm by Utilizing Customer Relationship Management," *Journal of Computer Scine and Information Technology*, pp. 89–93, Apr. 2023, doi: 10.35134/jcsitech.v9i2.69.

[15] D. Ispandi, "Membangun Sistem Informasi Manajemen Laboratorium Komputer (SILABKOM) STMIK-AMIK Riau."

[16] R. Rahmaddeni, M. K. Anam, Y. Irawan, S. Susanti, and M. Jamaris, "Comparison of Support Vector Machine and XGBSVM in Analyzing Public Opinion on Covid-19 Vaccination," *ILKOM Jurnal Ilmiah*, vol. 14, no. 1, pp. 32–38, Apr. 2022, doi: 10.33096/ilkom.v14i1.1090.32-38.

[17] T. A. Pertiwi, M. Afdal, and R. Novita, "Penerapan Algoritma K-Medoids dan FP-Growth dalam Penentuan Pola Kombinasi Produk Berdasarkan Hasil Segmentasi Pelanggan," *Technology and Science (BITS)*, vol. 6, no. 2, 2024, doi: 10.47065/bits.v6i2.5268.

[18] E. Hermika and S. Zuhri Harahap, "Application Of Data Mining In Selecting Superior Products Using The K-Means And K-Medoids Algorithm Methods".

[19] Y. Diana *et al.*, "Analisa Penjualan Menggunakan Algoritma K-Medoids Untuk Mengoptimalkan Penjualan Barang," *JOISIE Journal Of Information System And Informatics Engineering*, vol. 7, no. 1, pp. 97–103, 2023.

[20] S. Ika Murpratiwi, I. Gusti Agung Indrawan, and A. Aranta, "Analisis Pemilihan Cluster Optimal Dalam Segmentasi Pelanggan Toko Retail," *Jurnal Pendidikan Teknologi dan Kejuruan*, vol. 18, no. 2, 2021.