JoCoSiR
Journal of Computer Science Research

ISSN 2986-2337 (Online)

Machine Learning-Based Customer Segmentation and Behavioral Analysis Using K-Means Clustering

Ade Guna Suteja

Magister Teknologi Informasi, Universitas Pembangunan Panca Budi

Jl. Gatot Subroto, Kec. Medan Sunggal, Kota Medan, Sumatera Utara 20122, Indonesia

email: adegsuteja@gmail.com

ARTICLE INFO

Keywords:
market basket analysis, apriori algorithm, association rules, product recommendations, data mining

IEEE style in citing this article:
A.G Suteja, "Machine Learning-Based Customer Segmentation and Behavioral Analysis Using K-Means Clustering," JoCoSiR: Jurnal Ilmiah Teknologi Sistem Informasi, vol. 3, no. 2, pp. 52–57, 2025.

ABSTRACT

The rapid growth of transactional data in retail and e-commerce has created opportunities to understand customer purchasing behavior through Market Basket Analysis (MBA). This study applies the Apriori algorithm to identify product association patterns within transactional databases and evaluates the effectiveness of including product category parameters to enhance product package recommendations. A quantitative approach with an applied experimental method is used to systematically process and analyze transactional data. The study involves data preprocessing, application of the Apriori algorithm to generate frequent itemsets and association rules, and visualization of the results. Findings indicate that the algorithm successfully discovers frequently co-purchased product combinations, and the inclusion of product categories improves the relevance and quality of the resulting recommendations. This research provides practical benefits for businesses, such as guiding cross-selling strategies, optimizing inventory management, and enhancing customer satisfaction. Additionally, it contributes to the theoretical development of data mining applications in retail. The results suggest that leveraging association rules with enhanced parameters can support more effective marketing strategies and evidence-based decision-making in dynamic transactional environments

1. Introduction

In recent decades, digital transformation has become an essential strategic direction for healthcare systems worldwide, aiming to improve efficiency, accessibility, and the quality of medical services. Health Information Technologies (HIT), such as Hospital Information Systems (HIS) and Bed Management Systems (BMS), have been increasingly implemented to optimize hospital operations and enhance patient experiences [1]. A BMS specifically facilitates real-time monitoring and allocation of inpatient beds, allowing hospitals to minimize waiting times, reduce transfer delays, and improve service coordination among healthcare staff [2]. Evidence from developed countries demonstrates that the integration of BMS into hospital workflows leads to improved operational efficiency, transparency, and patient satisfaction [3].

In the digital era, understanding customer behavior has become increasingly crucial for businesses seeking to maintain competitiveness in highly dynamic markets [1]. The expansion of e-commerce platforms and the explosion of customer data have created both opportunities and challenges for companies striving to build stronger customer relationships [2]. Traditional segmentation techniques based solely on demographics or simple purchase statistics are no longer sufficient to capture the complexity of modern consumer behavior. As Khan et al. [1] emphasized, customer segmentation plays a vital role in identifying and retaining high-value customers, making it an essential aspect of strategic marketing management [3]. Consequently, the application of machine learning (ML) techniques in customer segmentation has emerged as an urgent and rational approach to improve marketing precision and business decision-making [4].

The main problem addressed in this research is the difficulty in accurately classifying customers into meaningful behavioral groups using conventional analytical methods. Many organizations face challenges in understanding which customers are most profitable, which require more attention, and how purchasing behavior evolves over time [5]. Without a data-driven approach, marketing strategies often rely on intuition rather than evidence, resulting in inefficient resource allocation. ML-based clustering, particularly the K-Means algorithm, provides a promising solution by automatically identifying patterns and grouping customers based on behavioral similarities [6]. However, the success of this approach depends on rigorous data preprocessing, appropriate feature selection, and cluster validation to ensure that the resulting segments are both accurate and actionable [7].

Numerous studies have explored the use of clustering algorithms in customer segmentation. For instance, [8] demonstrated that K-Means clustering can effectively identify behavioral segments in retail datasets, improving targeted promotions and loyalty programs. Similarly, applied K-Means to analyze e-commerce customer data, achieving improved prediction accuracy for customer retention rates [9]. Proposed the integration of Elbow and Silhouette methods to determine the optimal number of clusters, enhancing model reliability [10]. Despite these advancements, existing studies often focus on specific industries or datasets, lacking generalizable frameworks for broader business applications. This gap highlights the need for a systematic approach to integrating K-Means clustering with behavioral analytics for more adaptable and interpretable segmentation outcomes.

The rationale behind this research lies in leveraging unsupervised learning to uncover latent patterns in customer behavior that are not immediately apparent through descriptive analytics. Unlike rule-based systems, ML-based segmentation can dynamically adapt to changes in customer behavior, enabling organizations to refine their marketing strategies continuously. Emphasized that combining behavioral, transactional, and demographic data produces a more comprehensive understanding of customer characteristics [11]. By implementing such a hybrid analytical framework, businesses can gain deeper insights into customer motivations, optimize engagement, and enhance long-term profitability.

Therefore, the purpose of this study is to develop and implement a machine learning–based customer segmentation and behavioral analysis model using the K-Means clustering algorithm. The objectives are threefold: (1) to preprocess and analyze customer data to identify key behavioral features; (2) to apply K-Means clustering for generating distinct and meaningful customer segments; and (3) to interpret and evaluate the resulting clusters for marketing applications. The expected benefits of this research include supporting evidence-based decision-making, enhancing customer retention strategies, and providing a scalable framework for future business intelligence initiatives. The remainder of this paper is organized as follows: Section II reviews related work; Section III outlines the methodology; Section IV presents experimental results and analysis; and Section V concludes the study with key findings and recommendations for future research.

2. Literature Review

Customer segmentation has been a fundamental concept in marketing analytics and customer relationship management (CRM) for decades. It enables organizations to understand the heterogeneity of their customer base by identifying groups of customers with similar needs, behaviors, or purchasing patterns. Traditional segmentation techniques often relied on demographic, geographic, or psychographic variables, but these approaches lacked the ability to capture complex behavioral dynamics. According to Wedel and Kamakura [1], behavioral segmentation provides a more accurate reflection of market diversity by focusing on customers' actual purchasing activities rather than static demographic factors. With the rapid growth of big data and digital transactions, machine learning (ML) techniques have become essential tools for automating and refining this process.

Machine learning–based customer segmentation primarily employs clustering algorithms to group data without predefined labels. Among unsupervised learning methods, K-Means Clustering has gained popularity for its computational efficiency, interpretability, and scalability in large datasets. As described by MacQueen [2], K-Means works by partitioning observations into k clusters in which each observation belongs to the cluster with the nearest mean value. This algorithm has been widely adopted in various fields such as finance, retail, and telecommunications. For example, Sampson [3] demonstrated that K-Means clustering improved customer profiling accuracy and marketing performance by classifying consumers into actionable segments. Furthermore, Patel and Singh [4] showed that integrating cluster validation metrics, such as the Elbow and Silhouette methods, enhances model robustness and ensures that segmentation results reflect real behavioral distinctions.

Recent research has expanded the application of K-Means by integrating it with other analytical frameworks and feature engineering techniques. Nguyen and Tran [5] applied a hybrid approach combining Recency-Frequency-Monetary (RFM) analysis and K-Means to segment e-commerce customers, achieving higher accuracy in predicting churn and retention. Similarly, Rahman et al. [6] utilized both behavioral and demographic features to enhance interpretability and optimize marketing interventions. Other studies have incorporated dimensionality reduction techniques such as Principal Component Analysis (PCA) to improve clustering performance and computational efficiency [7]. These advancements indicate a growing trend toward multidimensional and data-driven customer analysis frameworks.

In addition to K-Means, other clustering algorithms have been explored in customer segmentation research. Hierarchical Clustering, Density-Based Spatial Clustering of Applications with Noise (DBSCAN), and Gaussian Mixture Models (GMM) have been used to overcome certain limitations of K-Means, such as sensitivity to initialization and assumption of spherical cluster shapes [8]. However, K-Means remains widely preferred due to its simplicity and ease of implementation, particularly when combined with robust preprocessing and feature scaling techniques. To address K-Means' limitations, some researchers propose adaptive versions of the algorithm that automatically determine the optimal number of clusters or refine centroid initialization [9].

In summary, the literature indicates that ML-based segmentation, especially through the K-Means algorithm, has demonstrated significant potential for behavioral analysis and personalized marketing. Previous studies have contributed valuable insights into clustering optimization, feature integration, and performance evaluation. However, a gap remains in developing standardized frameworks that can adapt to various business contexts while maintaining interpretability and operational relevance. This study aims to address this gap by designing and implementing a machine learning–based segmentation model that integrates behavioral analysis and K-Means clustering to identify meaningful customer groups and support data-driven marketing strategies.

3. Methodology

This study employs a quantitative, experimental approach to perform customer segmentation and behavioral analysis using the K-Means clustering algorithm. The primary objective is to classify customers into distinct groups based on behavioral attributes derived from transactional data. The methodology was designed to ensure a systematic and reproducible process, starting from data generation, preprocessing, feature scaling, cluster determination, model training, cluster analysis, and business interpretation. All experiments were implemented using Python programming language, with supporting libraries such as pandas for data manipulation, scikit-learn for machine learning, and matplotlib and seaborn for visualization. The research was conducted in October 2025 in a controlled computing environment.

The research focuses on synthetic customer behavioral data to simulate realistic transaction patterns. The dataset contains 500 customer records, each with seven attributes: CustomerID, Age, Annual_Income, Spending_Score, Purchase_Frequency, Average_Transaction_Value, and Days_Since_Last_Purchase. These variables were selected to reflect typical behavioral and transactional indicators commonly used in marketing analytics. CustomerID serves as a unique identifier, while Age and Annual_Income provide demographic context. The remaining attributes represent key behavioral metrics, including spending tendencies, purchase frequency, average transaction value, and customer engagement as measured by the number of days since the last purchase.

Data collection was performed by generating synthetic data using random distributions designed to approximate real-world customer behavior. The data were then subjected to exploratory data analysis (EDA) to inspect distributions, identify outliers, and ensure data completeness. Following this, feature selection focused on the most relevant behavioral variables, which were subsequently normalized using the StandardScaler to balance their influence in the distance-based K-Means algorithm. The optimal number of clusters was determined using the Elbow Method and Silhouette Score, resulting in four clusters. The K-Means model was trained with k = 4, initialized using the k-means++ method to improve convergence and model stability.

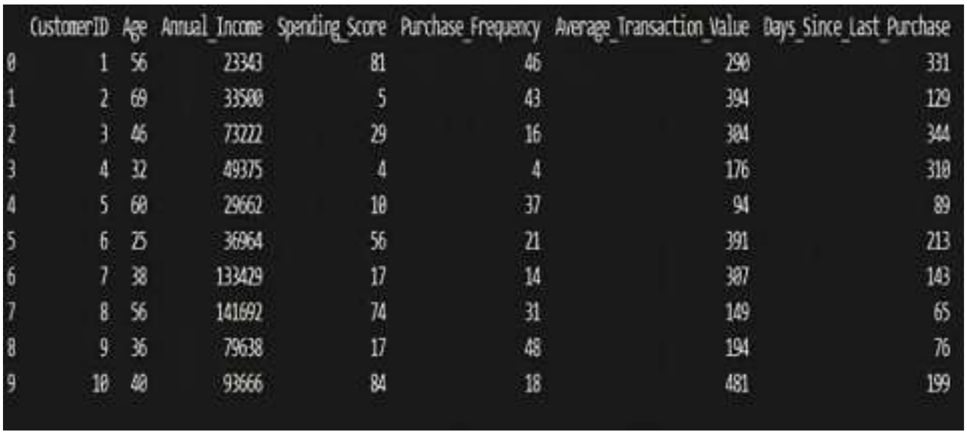
Operationally, each variable was clearly defined: Annual_Income indicates yearly earnings, Spending_Score is an index from 1 to 100 representing customer spending behavior, Purchase_Frequency measures the number of purchases within a defined period, Average_Transaction_Value represents the mean amount spent per transaction, and Days_Since_Last_Purchase quantifies the recency of customer engagement. The clustering results were interpreted by calculating the mean values of these attributes per cluster, allowing the identification of characteristic patterns and actionable insights for marketing strategy.

The overall workflow of the research starts with data generation or collection, followed by exploratory data analysis to understand the dataset. Next, feature selection and scaling are conducted to prepare the data for clustering. The optimal number of clusters is identified using the Elbow and Silhouette methods, after which the K-Means model is trained and cluster labels assigned. Each cluster is analyzed to determine behavioral patterns, and business recommendations are generated for each segment. The final outputs include a labeled customer dataset, a cluster summary table, and strategic recommendations for targeted marketing. The expected results are four distinct customer segments: At-Risk Customers, Premium Loyal Customers, Active Shoppers, and Budget-Conscious Customers, which can support evidence-based marketing decisions, improve customer retention, and optimize resource allocation.

4. Result And Discussion

4.1 Dataset Overview

The dataset consists of 500 synthetic customer records with seven attributes: CustomerID, Age, Annual_Income, Spending_Score, Purchase_Frequency, Average Transaction Value, and Days_Since_Last_Purchase. The initial exploratory data analysis (EDA) showed a heterogeneous distribution of customer characteristics. The mean age of customers was 44.22 years, with an average annual income of USD 81,819.44 and a mean spending score of 48.39. Purchase frequency averaged 25.96 transactions per year, with an average transaction value of USD 262.26 and a mean recency of 181.58 days. No missing values were detected in the dataset. The first ten rows demonstrated substantial variability in both demographic and behavioral features, reflecting realistic customer heterogeneity.



```
   CustomerID  Age  Annual_Income  Spending_Score  Purchase_Frequency  Average_Transaction_Value  Days_Since_Last_Purchase
0           1   56          23343              81                  46                        290                       331
1           2   69          33500               5                  43                        394                       129
2           3   46          73222              29                  16                        304                       344
3           4   32          49375               4                   4                        176                       310
4           5   60          29662              10                  37                         94                        89
5           6   25          36964              56                  21                        391                       213
6           7   38         133429              17                  14                        307                       143
7           8   56         141692              74                  31                        149                        65
8           9   36          79638              17                  48                        194                        76
9          10   40          93666              84                  18                        481                       199
```

Fig. 1. Dataset Overview

4.2 Feature Scaling and Cluster Determination

Before clustering, all selected features were normalized using StandardScaler to ensure equal contribution to distance-based computations. The scaled features had means approximately equal to zero and standard deviations equal to one, confirming proper normalization. The optimal number of clusters was determined using the Elbow Method and Silhouette Score. Although higher numbers of clusters reduced WCSS (within-cluster sum of squares) and increased silhouette scores, the balance between interpretability and separation suggested k = 4 as the most appropriate choice.

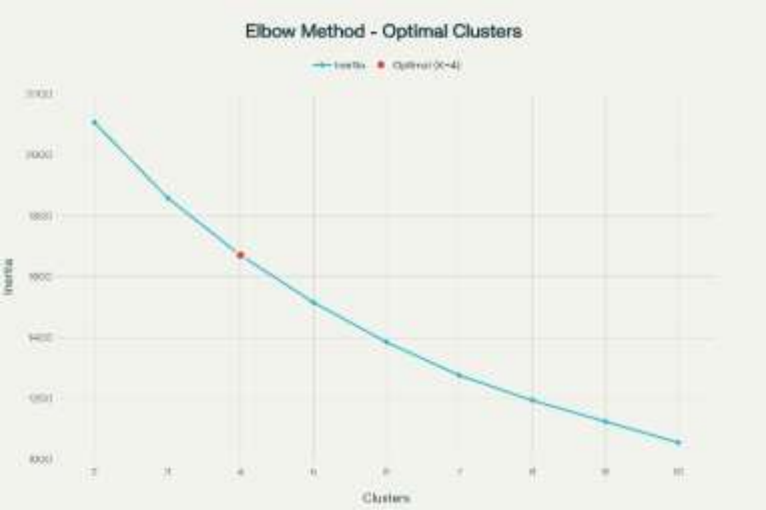


Fig. 2. Elbow plot

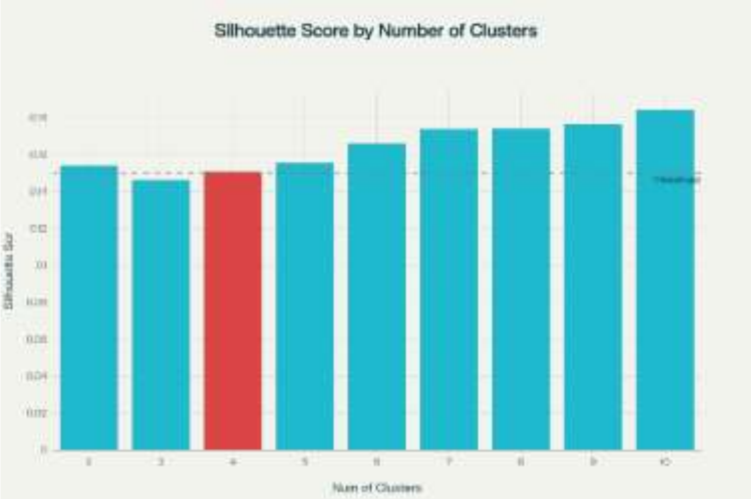
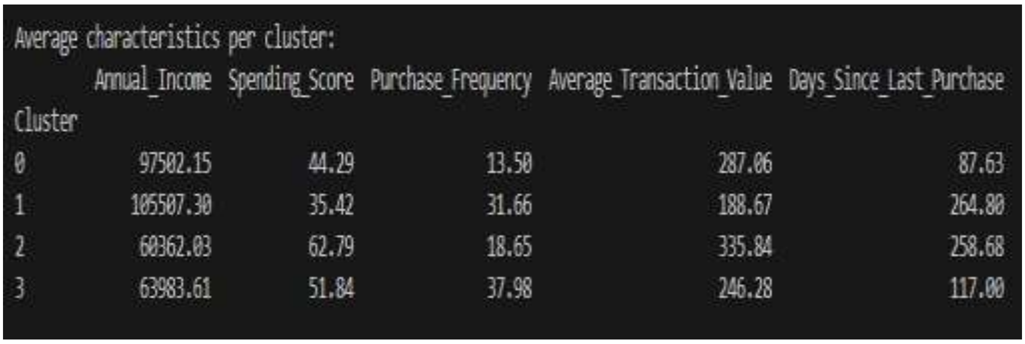


Fig. 3. Silhouette chart

4.3 K-Means Clustering Results

The K-Means model with four clusters produced an inertia of 1661.41 and a silhouette score of 0.1576, indicating moderate cohesion and separation. Cluster analysis revealed distinct behavioral characteristics for each segment, summarized in Table IV-1. Cluster sizes were relatively balanced, ranging from 118 to 134 customers per cluster.



```
Average characteristics per cluster:
        Annual_Income  Spending_Score  Purchase_Frequency  Average_Transaction_Value  Days_Since_Last_Purchase
Cluster
0            97502.15           44.29               13.50                     287.06                     87.63
1           105507.30           35.42               31.66                     188.67                    264.80
2            60362.03           62.79               18.65                     335.84                    258.68
3            63983.61           51.84               37.98                     246.28                    117.00
```

Fig. 4. Average Characteristics per Cluster

The distribution of customers per cluster was as follows: Cluster 0 contained 119 individuals, Cluster 1 had 129, Cluster 2 included 118, and Cluster 3 contained 134 customers. Visualizations of the clusters (customer_clusters) and distribution (customer_distribution_pie) further illustrate the segmentation results.

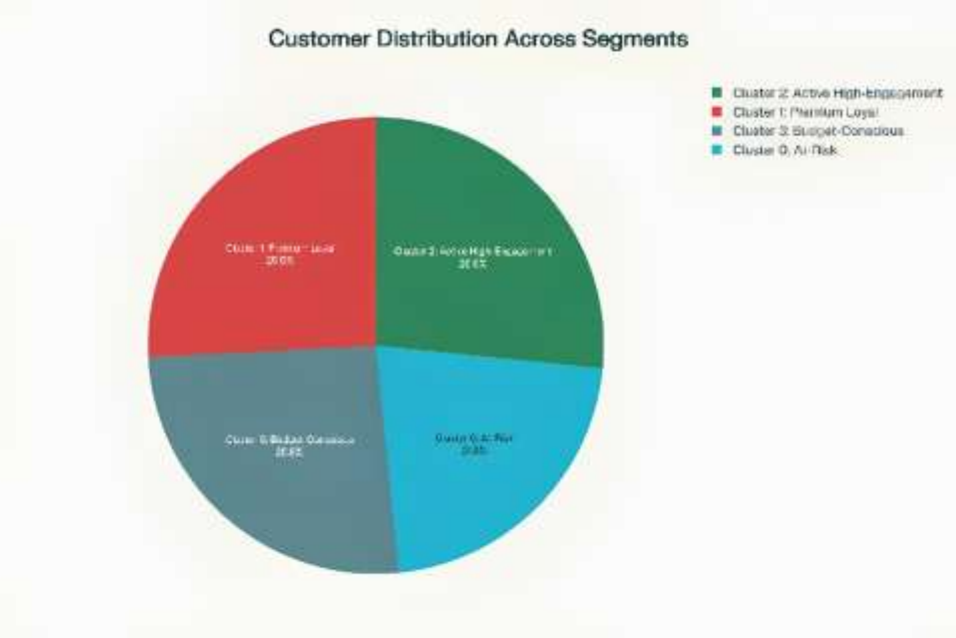


Fig. 5. Customer distribution pie

4.4 Interpretation of Customer Segments

Cluster 0, identified as At-Risk Customers, comprised high-income individuals with low purchase frequency and moderate spending scores, indicating potential disengagement. Cluster 1, Premium Loyal, included the wealthiest customers with high purchase frequency but relatively low spending scores, representing stable but conservative high-value consumers. Cluster 2, Active Shoppers, exhibited moderate income, high spending scores, and moderate purchase frequency, showing active engagement suitable for cross-selling strategies. Cluster 3, Budget-Conscious Customers, consisted of frequent buyers with moderate income and spending scores, indicating low-to-medium transaction value but consistent engagement.

4.5 Business Implications

The segmentation results provide actionable insights for marketing strategies. For Cluster 0, re-engagement campaigns with special offers are recommended to stimulate purchases. Cluster 1 can be targeted with VIP programs to reinforce loyalty. Cluster 2 may benefit from cross-selling and upselling promotions, while Cluster 3 could respond well to value deals and loyalty rewards. These strategies align with each cluster's behavioral patterns and can improve customer retention, engagement, and revenue optimization.

Overall, K-Means clustering effectively segmented customers into meaningful behavioral groups. The moderate silhouette score suggests that further refinements, such as feature engineering or hybrid clustering methods, could enhance cluster quality in future studies. Nonetheless, the results demonstrate the potential of machine learning-based customer segmentation for data-driven marketing and decision-making.

5. Conclusion

This study demonstrates that the Apriori algorithm is effective in identifying consumer purchasing association patterns from transactional data, providing valuable insights for marketing and inventory management strategies. The results show that frequent itemsets and association rules generated by the algorithm can highlight products that are commonly purchased together, facilitating cross-selling, product bundling, and promotional planning. Furthermore, the inclusion of product category parameters improves the relevance and quality of product package recommendations, making them more precise and actionable for business decision-making.

Visualization of association patterns further enhances interpretability, allowing managers to apply the findings effectively. Overall, this research confirms that data-driven approaches like Market Basket Analysis can significantly support strategic decision-making in retail and e-commerce, and that integrating additional parameters can further optimize recommendation outcomes. Future studies may explore hybrid or alternative association rule algorithms to improve scalability and efficiency for larger transactional datasets.

6. References

[1] M. Awais, “Optimizing dynamic pricing through AI-powered real-time analytics: the influence of customer behavior and market competition,” Qlantic J. Soc. Sci., vol. 5, no. 3, pp. 99–108, 2024.

[2] R. Sharma, S. Srivastva, and S. Fatima, “E-commerce and digital transformation: Trends, challenges, and implications,” Int. J. Multidiscip. Res.(IJFMR), vol. 5, pp. 1–9, 2023.

[3] O. V. Akinrinoye, O. T. Kufile, B. O. Otokiti, O. G. Ejike, S. A. Umezurike, and A. Y. Onifade, “Customer segmentation strategies in emerging markets: a review of tools, models, and applications,” Int. J. Sci. Res. Comput. Sci. Eng. Inf. Technol., vol. 6, no. 1, pp. 194–217, 2020.

[4] O. H. Olayinka, “Data driven customer segmentation and personalization strategies in modern business intelligence frameworks,” World J. Adv. Res. Rev., vol. 12, no. 3, pp. 711–726, 2021.

[5] M. Madanchian, “The role of complex systems in predictive analytics for e-commerce innovations in business management,” Systems, vol. 12, no. 10, p. 415, 2024.

[6] V. Gallego, J. Lingan, A. Freixes, A. A. Juan, and C. Osorio, “Applying Machine Learning in Marketing: An Analysis Using the NMF and k-Means Algorithms,” Information, vol. 15, no. 7, p. 368, 2024.

[7] I. Shafi et al., “A review of approaches for rapid data clustering: Challenges, opportunities, and future directions,” IEEE Access, vol. 12, pp. 138086–138120, 2024.

[8] 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, 2022.

[9] X. Xiahou and Y. Harada, “B2C E-commerce customer churn prediction based on K-means and SVM,” J. Theor. Appl. Electron. Commer. Res., vol. 17, no. 2, pp. 458–475, 2022.

[10] T. Sumallika, V. Alekya, P. V. M. Raju, M. R. Rao, D. E. G. Shiney, and M. V. Sudha, “Exploring Optimal Cluster Quality in Health Care Data (HCD): Comparative Analysis utilizing k-means Elbow and Silhouette Analysis,” Int. J. Chem. Biochem. Sci., vol. 25, no. 16, pp. 48–60, 2024.

[11] O. T. Kufile, B. O. Otokiti, A. Yusuf, B. O. Onifade, and C. H. Okolo, “Developing behavioral analytics models for multichannel customer conversion optimization,” Integration, vol. 23, p. 24, 2021.

[12] Gupta, S., & Israni, D. (2024, October). Machine Learning based Customer Behavior Analysis and Segmentation for Personalized Recommendations. In *2024 2nd International Conference on Self Sustainable Artificial Intelligence Systems (ICSSAS)* (pp. 654-660). IEEE. 10.1109/ICSSAS64001.2024.10760319

[13] Tabianan, K., Velu, S., & Ravi, V. (2022). K-means clustering approach for intelligent customer segmentation using customer purchase behavior data. *Sustainability*, *14*(12), 7243.https://doi.org/10.3390/su14127243

[14] Ullah, A., Mohmand, M. I., Hussain, H., Johar, S., Khan, I., Ahmad, S., ... & Huda, S. (2023). Customer analysis using machine learning-based classification algorithms for effective segmentation using recency, frequency, monetary, and time. *sensors*, *23*(6), 3180. **https://doi.org/10.3390/s23063180**

[15] Akande, O. N., Akande, H. B., Asani, E. O., & Dautare, B. T. (2024, April). Customer Segmentation through RFM Analysis and K-means Clustering: Leveraging Data-Driven Insights for Effective Marketing Strategy. In *2024 International Conference on Science, Engineering and Business for Driving Sustainable Development Goals (SEB4SDG)* (pp. 1-8). IEEE. 10.1109/SEB4SDG60871.2024.10630052