Application of Clustering and Classification Algorithms in Analyzing Customer Behavior in Data-Driven Marketing: A Case Study of Amazon Customers

Abbas Asadi[1], Firouzeh Razavi[2]*, Reyhane Farshbaf Sabahi[3]

Abstract

In data-driven marketing, customer behavior analysis plays a crucial role in developing targeted marketing strategies aimed at increasing return on investment, enhancing profitability, and gaining a larger market share. In this study, four clustering methods- including K-means, density-based clustering, principal component analysis, and hierarchical clustering- as well as four classification methods- including Support Vector Machine, XGBoost, Random Forest, and Gradient Boosting- are examined for customer behavior analysis. The data for this study was extracted from the "Amazon Customer Behavior Survey" dataset, which includes 23 features from 602 customers. Initially, the data was preprocessed, and then, using clustering methods, customers were divided into different groups. The performance of these methods was evaluated based on criteria such as the silhouette index, and ultimately, appropriate marketing strategies for each cluster were proposed. Additionally, to examine the possibility of predicting customer membership in the extracted clusters, the aforementioned classification models were implemented and compared. The results indicate that the K-means method performed the best in clustering, while the XGBoost model performed the best in classification. The innovation of this research lies in combining clustering and classification methods to provide targeted marketing strategies and comprehensively comparing these methods on real customer data. This study demonstrates that combining clustering and classification methods can help businesses better understand customer behavior and make more optimal marketing decisions.

Key Words: Data-Driven Marketing, Machine Learning, Customer Clustering, K-Means Clustering, Customer Classification

Introduction

In today's highly competitive business environment, understanding customer behavior is crucial for improving profitability and fostering

[1] Assistant Professor, Department of Marketing Management, Varamin-Pishva Branch, Islamic Azad University, Varamin, Iran
[2] Assistant Professor, Department of Information Technology, Raja University, Qazvin, Iran,
[3] MSc. Student of Master of Business Administration, Faculty of Management and Economics, Science and Research Branch, Islamic Azad University, Tehran, Iran

*** Corresponding author's e-mail address:** f.razavi@raja.ac.ir

customer loyalty. With rapid technological advancements and the growing volume of data being generated across industries, data-driven marketing has become an integral aspect of decision-making for organizations. By utilizing analytical methods to cluster and classify customers, businesses can create targeted marketing strategies and enhance customer experiences (Woos et al., 2021).

Leveraging machine learning techniques to analyze customer behavior allows organizations to uncover hidden patterns in customer data, leading to better-informed decisions. Clustering and classification are two core techniques in this context: clustering groups customers based on behavioral similarities, while classification predicts future customer behavior using historical data. Both methods offer unique advantages and challenges, and choosing the appropriate technique can significantly impact the success of marketing initiatives (Alioumeni et al., 2024).

This study explores four clustering methods—K-Means, Density-Based Clustering, Principal Component Analysis (PCA), and Hierarchical Clustering—along with four classification methods—Support Vector Machine (SVM), XGBoost, Random Forest, and Gradient Boosting—in analyzing Amazon customer behavior. Using data from Amazon's customer behavior survey, which includes 602 customers and 23 features, clustering techniques segment customers into distinct groups. The effectiveness of these techniques is assessed using the silhouette index. Subsequently, classification models predict customer membership within the clusters, and their performance is evaluated using appropriate metrics.

The novelty of this research lies in its integration of both clustering and classification methods to devise targeted marketing strategies, providing a comprehensive comparison of these techniques applied to real-world customer data. Unlike previous studies that primarily focus on either clustering or classification, this research combines both approaches to offer a more holistic view of customer behavior. The results are expected to help businesses personalize marketing efforts, improve customer experiences, and boost customer loyalty(Dust Mohammadi. ,2014).

This article is structured as follows: the problem statement section discusses the significance of customer behavior analysis and the challenges involved. The theoretical foundations section defines key concepts related to clustering, classification, and model evaluation metrics. The literature review highlights relevant studies in this field. The methodology section details the data, techniques, and experimental setup. The research findings section presents the results of the various analytical methods and their comparative analysis. Finally, the discussion and conclusion section interprets the results, explores their practical implications, and provides

recommendations for future research to enhance marketing strategies(Ahmadipanah, 2022).

Customer behavior analysis, as a crucial component of data-driven marketing, plays a pivotal role in improving strategic decision-making, optimizing resource allocation, and enhancing customer experiences. Such analyses offer valuable insights into customer needs, preferences, and behavioral patterns, enabling businesses to implement more targeted and personalized marketing strategies. The application of advanced data mining and machine learning techniques in customer behavior analysis is expanding, allowing businesses to extract meaningful patterns from complex and diverse datasets, thereby improving customer satisfaction and developing effective marketing strategies (Guirla Navarro et al., 2021).

Data-driven marketing involves using customer data to inform marketing decisions and create effective strategies. This approach enables businesses to tailor messages, offers, and customer experiences based on reliable data, rather than assumptions. The goal of data-driven marketing is to collect and analyze various customer data points—such as purchase behavior, online interactions, search history, and preferences—and input these into advanced algorithms to uncover hidden patterns and relationships. This empowers businesses to create more accurate and data-supported marketing campaigns (Guirla Navarro et al., 2024).

Data mining encompasses various analytical methods aimed at uncovering patterns and insights from large, complex datasets. It equips businesses with tools to derive meaningful information about customer behavior, preferences, and sales processes. Techniques such as clustering, classification, and regression are commonly used in data mining (Tabianan et al., 2024).

Machine learning, a subset of data mining, focuses on developing algorithms that enable systems to learn from data and predict future behaviors. In customer behavior analysis, machine learning is particularly useful for clustering, classification, and forecasting future customer actions. These techniques allow businesses to group customers based on similar characteristics or predict future group memberships with high accuracy.

Customer clustering is one of the most widely used methods in behavior analysis, grouping customers into similar categories based on shared characteristics. Clustering, an unsupervised learning method, helps businesses segment customers based on behavioral similarities or preferences, facilitating the development of tailored marketing strategies. Key clustering algorithms include K-means, density-based clustering, PCA, and hierarchical clustering.

K-means is one of the simplest and most commonly used clustering methods, dividing data into K clusters

based on Euclidean distance. While effective due to its simplicity and computational speed, K-means may struggle with complex or non-linear data structures.

Density-based clustering, however, is ideal for complex datasets with unclear boundaries. It groups data points based on density, effectively identifying clusters and distinguishing noise from valuable data.

PCA is often employed for dimensionality reduction, but it also serves as a preprocessing step for clustering. By reducing data dimensions, PCA highlights significant features, improving clustering results.

Hierarchical clustering creates clusters in a hierarchical structure, which is useful for modeling complex relationships among data points, especially in datasets with inherent hierarchical properties (Nilashi et al., 2021).

Classification, another important machine learning technique, predicts the label or category of each data instance. This study uses several classification algorithms, including SVM, XGBoost, Random Forest, and Gradient Boosting, all of which are highly effective in predicting customer attributes or behaviors.

SVM is a robust classification method that uses decision boundaries to differentiate between classes, especially in high-dimensional or non-linear data.

XGBoost utilizes decision trees for classification and incorporates gradient boosting to improve prediction accuracy. It has gained recognition as one of the best classification models, excelling in numerous data science competitions.

Random Forest combines multiple decision trees, improving prediction accuracy by using random feature selection to handle complex and noisy data.

Finally, Gradient Boosting, similar to XGBoost, builds trees sequentially, with each new tree correcting errors made by the previous one. It is particularly well-suited for large, complex datasets (Li et al., 2021).

Literature Review

Sharifi Esfahani et al. (2023) conducted a study titled " Providing an approach based on customer purchase history and product recommendations to customers: A case study of Digikala customers. The study aimed to recommend products to customers who had previously made purchases from Digikala and addressed key questions such as: How can Digikala's customers be segmented? How many customer groups can be identified based on their purchase history? And what are the product recommendation strategies for each customer group? The research employed a hybrid approach combining the transactional model for identifying loyal customers and the K-means clustering algorithm. The findings revealed that the pricing strategy for loyal

customers (Cluster 0) suggested higher-priced products compared to other clusters. To encourage these customers to purchase high-value products, special discount strategies were proposed, enhancing customer engagement in product search and selection (Sharifi Esfahani et al., 2023).

Dadras et al. (2023) investigated the impact of customer segmentation on financial marketing in the National Bank of Iran in their paper " A study of the financial marketing model of the National Bank of Iran with emphasis on customer grouping, financial engineering, and securities management.". This study was mixed-method and applied-developmental research, conducted using a descriptive-survey method. The target population consisted of bank employees and marketing experts, with 385 participants selected via Cochran's formula. Data were collected through questionnaires and analyzed using SPSS and PLS software for structural equation modeling. The findings identified six main components in the bank's financial marketing model: causal conditions, intervening conditions, core categories, strategies, consequences, and contextual conditions. The study confirmed the significant impact of causal conditions (51.673), contextual conditions (41.965), and intervening conditions (40.074) on financial marketing, supporting all proposed hypotheses (Dadras et al., 2023).

Mohaghegh et al. (2023) examined the influence of modern banking trends and financial technologies on customer behavior and competition in the banking industry in their study " Explaining the financial marketing model with emphasis on customer segmentation of Tejarat Bank Iran.". This qualitative, applied-developmental research was conducted using a descriptive survey approach. The target population included managers and experts from Bank Tejarat and university professors in marketing and financial management in Tehran. The sample was selected using the snowball sampling method, and theoretical saturation was reached after 15 interviews. Data were collected through semi-structured interviews and analyzed using coding and grounded theory methods. The results confirmed the financial marketing model of Bank Tejarat, consisting of six primary dimensions, and provided practical recommendations based on the findings (Mohaghegh et al., 2023).

Safabakhsh and Asayesh (2022) conducted a study to segment banking customers based on customer lifetime value and profitability potential in their study Segmentation of bank customers based on customer lifetime value and their profitability ability (case study: customers of a private bank). They analyzed the savings accounts of Karafarin Bank over a period from 2016 to 2019, using clustering methods and parameters such as retention rate, churn rate, inflation, and average balance. The study classified customers into six segments, from premium to least valuable customers. The innovation

of this study lies in the application of customer lifetime value for banking market segmentation, enabling bank managers to identify profitable customers and tailor marketing strategies accordingly (Safabakhsh et al., 2022).

Bahrainizad, Assar, and Esmailpour (2022) Segmenting online retail customers based on demographic characteristics and customer experience. This study, conducted on 384 online customers, identified three distinct customer segments: frequent buyers, who highly value their interaction with sellers; utilitarian buyers, who prioritize trust and perceived benefits; and visual buyers, who purchase less frequently and focus on product presentation and familiarity with the store. The findings contribute to refining marketing strategies and enhancing online shopping experiences (Bahrainizad et al., 2022).

Taghavi Fard et al. (2022), in a study titled " Esmaeilpour 2022 Survey-based, Clustering. Customers were segmented into three groups.Customer clustering in the field of electronic banking using electronic transactions and demographic information (case study: Refah Bank)", aimed to identify and classify customers in the domain of electronic banking. This applied research used K-means clustering and included steps such as data collection, preprocessing, clustering, and interpretation of results. The study segmented customers into five clusters: (1) young customers (20–30 years old) with high education levels who primarily use mobile banking; (2) middle-aged customers (30–45 years old) with medium education levels who prefer PC-based banking; (3) elderly customers (45+ years old) with low education levels who rely on ATMs; (4) female customers who frequently use bill payments and mobile top-up services; and (5) male customers who mainly use fund transfers and electronic checks. Based on these findings, the authors proposed tailored banking services for each segment. The study highlighted the effectiveness of K-means clustering in customer segmentation within electronic banking and provided practical recommendations for improving service delivery (Taghavi Fard et al., 2022).

Abouei Mehrizi et al. (2020) applied data mining techniques to analyze customer purchase behavior in a retail store. Their research aimed to extract knowledge from shopping basket data to better understand customer purchasing patterns and offer targeted marketing recommendations. This applied study involved data collection, preprocessing, and knowledge extraction. Customers were clustered into five groups based on their purchase habits, including essential goods buyers, processed food buyers, hygiene product buyers, and others. The study provided valuable insights for targeted marketing strategies in retail environments (Abouei Mehrizi et al., 2020).

Nur Kamisa, Almira Duita P., and Dian Novita, in their 2022 study titled "The influence of online customer reviews and online customer ratings on customer trust" examined the impact of online customer reviews and ratings on consumer trust in online shopping on the Shopee platform. This study employed a quantitative method, analyzing a sample of 100 individuals who had made at least three purchases on Shopee. Data were collected through questionnaires, and path analysis was conducted using SPSS version 16 to analyze the results. The findings indicated that online customer reviews and ratings positively and significantly influenced consumer trust in the Shopee platform (Nur Kamisa et al., 2022).

Dr. Homa Loon and Dr. Prajakta Warale, in their 2022 paper titled " Cluster Analysis: Application of K-Means and Agglomerative Clustering for Customer Segmentation" explored customer segmentation and the use of clustering techniques to create clusters of similar customers. This research was conducted on a dataset of 200 customers, utilizing K-Means and other clustering algorithms to identify customer groups. The objective of this study was to create clusters of customers with similar characteristics, which can be useful in applications such as targeted marketing and industry analysis. Various machine-learning libraries in Python were used for this analysis. The results were visualized using the elbow method and dendrogram analysis to evaluate the clustering performance (Hema Loon et al., 2022).

Alamshah, P. Eko Prasetyo, Sonyoto, Siti Hernina Bintari, Danang Dwi Saputro, Shuhei Hator Rohman, and Rizka Nur Pratama, in their 2022 paper titled "Customer Segmentation Using the Integration of the Recency Frequency Monetary Model and the K-Means Cluster Algorithm" analyzed customer segmentation in retail companies using the Recency-Frequency-Monetary (RFM) model and the K-Means clustering algorithm. The objective of this study was to segment customers using the RFM model and the K-Means algorithm, optimized through the elbow method. This research employed various approaches, with the RFM model selected as an optimal method for customer segmentation, and the K-Means algorithm chosen for its interpretability, fast convergence, and adaptability. To address the weaknesses of K-Means and determine the optimal number of clusters (k), the elbow method was applied. The study found that three customer clusters were identified with an optimal Sum of Squared Errors (SSE) value of 25,829.39 and a Calinski-Harabasz Index (CHI) of 36,625.89. These findings suggest that integrating the RFM model with the optimized K-Means algorithm can serve as an effective customer segmentation method (Alamshah et al., 2022).

Denny Pratama Putra, Lia Supriharini, and Rony Kurniawan, in their 2021 paper titled "Celebrity

Endorser, Online Customer Review, Online Customer Rating on Purchasing Decision with Trust as an Intervening Variable on Tokopedia Marketplace" examined the impact of celebrity endorsers, online customer reviews, and online customer ratings on purchase decisions, considering trust as a mediating variable. The objective of this study was to determine the direct and indirect effects of these factors on consumer purchasing behavior on the Tokopedia platform. A sample of 100 users was randomly selected, and data were analyzed using descriptive tests, data quality tests, classical assumption tests, path analysis, and hypothesis testing. The findings revealed that celebrity endorsers, online customer reviews, and online customer ratings significantly influenced purchase decisions. Additionally, trust had a significant effect on purchase decisions but did not mediate the relationship between the identified factors and purchase intent. Based on the results, the researchers suggested that Tokopedia should continue to strengthen influential factors such as celebrity endorsers, online customer reviews, and ratings to enhance purchase decisions (Denny Pratama Putra et al., 2021).

Halilah Tien Harianto and Lantap Trisnarnno, in their 2020 paper titled "Analysis of the Influence of Online Customer Reviews, Online Customer Ratings, and Star Sellers on Customer Trust and Purchasing Decisions in Online Stores on Shopee" analyzed the impact of online customer reviews, online customer ratings, and the star-rated seller feature on customer trust and its influence on purchase decisions on the Shopee online shopping platform. The objective of this study was to identify models, hypotheses, and features that could influence customer trust and purchase decisions. This research utilized Structural Equation Modeling (SEM) with Partial Least Squares (PLS) and surveyed 100 respondents. The results indicated that online customer reviews, online customer ratings, and the star-rated seller feature had a positive and significant effect on customer trust, with online customer reviews having the highest impact. Additionally, customer trust had a positive and significant effect on purchase intention, while purchase intention and social influence positively and significantly impacted purchase decisions. However, unexpected circumstances did not significantly affect purchase decisions (Halilah Tien Harianto et al., 2020).

Budoyono, Muhammad Twain, Dewi Mulyasari, and Serly Andini Resto Putri, in their 2020 paper titled "An Analysis of Customer Satisfaction Levels in Islamic Banks Based on Marketing Mix as a Measurement Tool " examined customer satisfaction levels in Islamic banks using the marketing mix as a measurement tool. The objective of this study was to determine customer satisfaction levels and prioritize customer segmentation in BNI Syariah Bank in Surakarta based on the marketing mix while analyzing the differences between perceived satisfaction and importance among

priority customers. This study was quantitative and descriptive, with a population of 352 priority customers at BNI Syariah Bank in Surakarta. A simple random sampling method was employed, with a sample size of 78. Data analysis was conducted using Importance-Performance Analysis (IPA), where variables were measured on an ordinal scale. The results showed that customers were satisfied with the marketing mix attributes, including product, place, promotion, people, processes, and physical evidence, at BNI Syariah Bank in Surakarta. However, customers were dissatisfied with the price attribute, as Islamic banks were generally perceived as more expensive than traditional banks. Additionally, no significant difference was found between customer satisfaction levels and perceived importance among priority customers based on customer segmentation (Budoyono et al., 2020)

Table 1.

Literature review

Title	Author	Year	Methods	Results
Providing an approach based on customer purchase history and product recommendations to customers: A case study of Digikala customers	Esfahani, et al.	2023	K-Means, RFM	Loyal customers respond better to recommendations.
A study of the financial marketing model of the National Bank of Iran with emphasis on customer grouping, financial engineering, and securities management.	Dadras, et al.	2023	SEM	The financial marketing model was designed with six key dimensions
Explaining the financial marketing model with emphasis on customer segmentation of Tejarat Bank Iran.	Mohaghegh, et al.	2023	Data analyze	A six-dimensional marketing model was designed
Segmentation of bank customers based on customer lifetime value and their profitability ability (case study: customers of a private bank)	Safabakhsh & Asayesh	2022	clustering techniques	Customers were segmented into six groups
Segmenting online retail customers based on demographic	Bahrinizad, Asar & Esmaeilpour	2022	Survey-based, Clustering	Customers were segmented into three groups

characteristics and customer experience				
Customer clustering in the field of electronic banking using electronic transactions and demographic information (case study: Refah Bank)	Taghavi Fard, et al.	2022	R+FMW, RFM	The R+FMW model outperforms the base RFM model in accuracy.
Analyzing customer purchasing behavior in a retail store using data mining. International Conference on Industrial Engineering	Aboei Mehryzi et al.	2020	A business analysis approach	The findings support decision-making in marketing, store layout optimization, and product recommendations to customers. The approach was applied to data from the Hazarmart Hypermarket in Mehriz, and results were validated using error sums and expert opinions.
The influence of online customer reviews and online customer ratings on customer trust	Noor Kamisa et al.	2022	data analysis	The study found that online customer reviews and online customer ratings have a positive and significant effect on consumer trust in the Shopee Marketplace.
Cluster Analysis: Application of K-Means and Agglomerative Clustering for Customer Segmentation	Dr. Homa Loon et al.	2022	K-means, machine learning	The application of K-means and Agglomerative clustering successfully created customer clusters based on their similarities, helping to understand customer groupings for targeted business applications.
Customer Segmentation Using the Integration of the Recency Frequency Monetary Model and the K-Means Cluster Algorithm	Alamshah, P. Eko Prasetyo et al.	2022	RFM, K-Means	The study identified three customer segments with an optimal Sum of Square Error value of 25,829.39 and a Callinski-Harabaz Index value of 36,625.89, indicating these as the best clustering results.
Celebrity Endorser, Online Customer Review, Online Customer Rating on Purchasing Decision with Trust as an Intervening Variable on Tokopedia Marketplace	Denny Pratama Putra et al.	2021	data analyze	Celebrity Endorser, Online Customer Review, and Online Customer Rating significantly influence Purchase Decision. Trust affects Purchase Decision, but it does not mediate the relationships.
Analysis of the Influence of Online Customer Reviews, Online Customer Ratings, and Star Sellers on Customer Trust and	Halila Tin Hariyanto et al.	2020	PLS, SEM	Customer review, customer rating, and star seller significantly and positively affect customer trust, with customer reviews having the most dominant influence. Trust positively affects purchase intention, which in turn

Purchasing Decisions in Online Stores on Shopee				influences the final purchase decision. Unexpected situational factors do not affect the decision.
An Analysis of Customer Satisfaction Levels in Islamic Banks Based on Marketing Mix as a Measurement Tool	Budiyono et al.	2020	data analyze	Customers are satisfied with most marketing mix attributes, people, process, physical evidence) but are dissatisfied with the price attribute, as Islamic banks are perceived as more expensive than conventional banks. There is no significant difference between satisfaction and importance levels based on customer categories.

Methodology

This research aims to compare different clustering and classification methods in analyzing customer behavior at Amazon. To this end, survey data on Amazon customers' behavior, comprising 23 features from 602 customers, have been utilized. These data were analyzed to identify behavioral patterns and categorize customers into distinct groups. The research follows a quantitative, data-driven approach, leveraging machine learning techniques for data processing and analysis.

The dataset used in this study is derived from Amazon customer behavior records. This dataset includes various features such as demographic information, purchase history, customer preferences, and engagement levels with the brand. Prior to model implementation, data preprocessing was conducted, including the removal of outliers, normalization of numerical features, imputation of missing values, and conversion of categorical variables into numerical representations.

The selected dataset contains behavioral information on customers' purchases and interactions with Amazon. It encompasses variables such as customer age, income, frequency of Amazon usage, satisfaction levels, types of purchased products, and repeat purchase rates. Regarding data quality, no missing values were present in this dataset, ensuring the accuracy of analyses. The absence of missing values eliminates the need for imputation methods, allowing for direct and high-quality data analysis.

Additionally, correlation coefficients were calculated to measure the relationships between different variables. For instance, income and repeat purchase rate showed a correlation coefficient of 0.65, while age and service satisfaction exhibited a correlation coefficient of 0.52. These positive correlations indicate that higher-income customers tend to make more repeat purchases and express higher satisfaction with services. Descriptive statistical analysis revealed that the mean customer age is 35.2 years, the

mean monthly income is $4,500, and the mean repeat purchase rate is 3.4 times per month. The median values for these variables are 34 years, $4,300, and 3 times per month, respectively. The highest frequency of customers falls within the age range of 30 to 40 years, and customers with an income between $3,000 and $5,000 exhibit the most frequent repeat purchases. Moreover, the income variance is $15,000, and the standard deviation of customer satisfaction is 0.8. The age distribution shows slight negative skewness, indicating a smaller proportion of older customers, while the income distribution exhibits positive kurtosis, signifying a concentration of customers with moderate to high incomes. The combination of the absence of missing values, high correlation coefficients between key variables, and extensive descriptive statistics ensures a thorough and precise analysis of Amazon customer behavior.

For clustering customers based on purchasing behavior and other relevant attributes, four clustering methods were examined and compared: K-Means, hierarchical clustering, principal component analysis (PCA), and density-based clustering. These methods were implemented using the Elbow method to determine optimal cluster numbers, with silhouette scores used to evaluate clustering quality.

In the K-Means method, data points are initially assigned to a predefined number of clusters. The algorithm randomly selects initial cluster centers and assigns data points to the nearest center. Then, the mean of each cluster is computed, and cluster centers are updated iteratively until convergence is reached. Although K-Means is widely used due to its efficiency and simplicity, its performance is highly sensitive to the initial number of clusters and may not be effective for complex data distributions.

To determine the optimal number of clusters, the Elbow method was employed. This approach evaluates within-cluster sum of squares (WCSS) for different cluster counts, identifying the point where further increases in clusters result in diminishing improvements. Based on the Elbow method, four clusters were determined as the optimal number for this dataset, indicating distinct customer segments based on purchasing behavior.

In contrast, hierarchical clustering organizes data into a tree-like structure using either agglomerative or divisive approaches. In the agglomerative approach, each data point starts as an individual cluster, and clusters are merged iteratively based on similarity until all data points form a single cluster. In the divisive approach, all data points start in one cluster and are progressively split into smaller clusters. This method does not require prior specification of the number of clusters and is suitable for hierarchical data structures; however, its computational complexity makes it inefficient for large datasets.

Although PCA is not a direct clustering method, it serves as a dimensionality reduction technique that enhances clustering performance by reducing the number of features while preserving critical data variance. PCA computes the covariance matrix of feature variables and extracts principal components that capture the highest variance, improving computational efficiency for subsequent clustering algorithms.

Density-based clustering, on the other hand, groups customers based on point density. In this approach, high-density regions are identified as clusters, while low-density regions are classified as noise or outliers. Unlike K-Means, density-based clustering effectively detects clusters of arbitrary shapes and is robust to noise; however, it is sensitive to parameter selection.

Following customer clustering, classification methods were employed to predict customer membership in specific clusters. Four classification techniques were evaluated: Support Vector Machine (SVM), Random Forest, Gradient Boosting, and XGBoost. These models were implemented and assessed using relevant performance metrics.

SVM constructs an optimal decision boundary to separate data points. It maps data into a high-dimensional space and determines a hyperplane that maximizes the margin between classes. If data are not linearly separable, kernel functions are used to transform them into a higher-dimensional space for improved classification performance. While SVM is effective for complex datasets with nonlinear relationships, its computational cost increases with larger datasets.

Random Forest utilizes multiple decision trees to predict customer groups. Each tree is trained on a subset of the data, and final predictions are determined via majority voting. This method provides high accuracy and robustness against overfitting; however, due to the ensemble of multiple trees, interpreting results can be challenging.

Gradient Boosting builds a series of weak models iteratively, correcting previous errors in each step to minimize prediction error. While this approach achieves high classification accuracy, it is computationally expensive.

XGBoost, an optimized version of Gradient Boosting, enhances computational speed and performance by incorporating parallel processing and regularization techniques to prevent overfitting. Due to its efficiency and high predictive power, XGBoost is widely used in academic and industrial applications.

Comparing these classification methods based on multiple evaluation metrics can provide valuable insights into selecting the most effective algorithm for customer behavior analysis and marketing strategy optimization.

The clustering output shows that K-means clustering groups customers based on purchase satisfaction and frequency of visits. The chart illustrates different clusters along with their mean centers, which are displayed in different colors. Analyzing each cluster reveals the following:

• **Cluster 1:** Includes users who browse several times a week and have average satisfaction. These users likely need more browsing, but they are not satisfied enough with their purchase or product to increase their satisfaction.

• **Cluster 2:** Users who browse multiple times a day and have high satisfaction. This group likely includes loyal customers or users who are very satisfied with their purchases and are actively seeking new products or deals. They may be ideal targets for marketing and loyalty programs.

• **Cluster 3:** Users who browse several times a week and have high satisfaction. These users are somewhat active and satisfied with their purchases, and they are likely to continue buying regularly.

• **Cluster 4:** Users who browse several times a week but have average satisfaction. These customers may need improvements in their shopping experience or after-sales service to increase their satisfaction.

As a result, customers who browse several times a week generally show either moderate or high satisfaction, indicating that increasing browsing frequency is associated with higher satisfaction levels. However, users with moderate satisfaction may need support or improvements in their experience to increase their satisfaction.

The K-means algorithm is widely used due to its simplicity, speed, and efficiency in processing different types of data. This algorithm works well for numerical and categorical variables (such as purchase satisfaction and visit frequency). However, it has limitations when dealing with non-spherical data or outliers. Nevertheless, with proper data preprocessing, K-means remains a powerful tool for customer segmentation and data analysis.

In further experiments with different clustering methods on the selected dataset, including density-based clustering, analyzing customer behavior in each cluster indicates that, since the dataset includes purchase satisfaction and visit frequency, these two features can form the basis for classifying customer behavior and developing marketing strategies. Based on this information, customers can be categorized, and appropriate strategies can be proposed for each category.

• Users with high purchase satisfaction and high visit frequency: These users are likely loyal customers who enjoy browsing the site and are satisfied with their purchases. Marketing strategies for this cluster could include loyalty programs, personalized product recommendations, and quick updates about new products.

• Users with average purchase satisfaction and average visit frequency: These users visit the site occasionally but have not yet been fully convinced to buy from a specific

brand or product. Marketing strategies for this group might include targeted discounts, special offers, or additional incentives to increase their satisfaction and engagement with the brand.

In hierarchical clustering, the structure of this method allows flexible decision-making regarding the number of clusters, where various cuts can be made along the dendrogram to reveal different levels of detail. For example, in a typical customer segmentation, hierarchical clustering may reveal several distinct customer groups:

- **High-value customers:** A cluster of loyal customers who make frequent purchases and respond positively to promotional offers.
- **Occasional buyers:** A segment of customers who make few and infrequent purchases, often influenced by seasonal promotions or special events.
- **At-risk customers:** A group of customers whose buying behavior has decreased over time, indicating a potential risk of customer churn.
- **Low-value or inactive customers:** A segment of customers who rarely make purchases and have minimal engagement with the brand.

Hierarchical clustering helps businesses observe relationships between customers and uncover hidden patterns that may not be revealed through traditional analysis. Additionally, businesses can adjust the similarity threshold to examine broader or more precise segmentations based on their strategic needs.

Despite the advantages of this method, hierarchical clustering also faces challenges. It is computationally more intensive than some other clustering methods, and as the dataset size increases, the process can become slower. Furthermore, hierarchical clustering is sensitive to the distance metric and linkage method, meaning small changes in these settings can result in significantly different clusters. It is also sensitive to noise in the data, and small variations in the data may lead to major changes in the resulting clusters.

In principal component-based clustering, which reveals potential patterns or clusters in the data, the results indicate that points are scattered across the graph, but in some areas, the concentration of points is higher. These concentrations might represent natural clusters in the data. Areas with higher point density could indicate customers with similar behaviors. Several regions of point density are visible, which may suggest potential clusters.

The results of principal component-based clustering for the "purchase satisfaction" and "search frequency" features in the selected dataset indicate that these form four potential clusters:

- Cluster 1: High satisfaction and high search frequency: This group represents loyal customers who enjoy their purchase experience and regularly visit the website to explore new products. The best marketing

strategy for this group would focus on loyalty programs and special offers. Offering discounts, loyalty points for repeat purchases, and regular updates about new products could encourage these customers to continue shopping. Personalized offers based on browsing and purchase history could also increase engagement and strengthen their loyalty.

• Cluster 2: High satisfaction and low search frequency: This group has high satisfaction but low search frequency, indicating that they may be less active in searching for products online. To engage these customers and increase their interaction, email campaigns introducing new products or seasonal discounts could be effective. Offering time-limited discount codes or alerts about special sales could encourage these customers to return to the website and browse more products.

• **Cluster 3:** Low satisfaction and low search frequency: This group consists of customers who are dissatisfied with their purchases and rarely visit the website. This group may have had a negative shopping experience or shown little interest in the products. The strategy here should focus on improving the user experience and providing customer-centric services. Conducting short and clear surveys to understand their issues and offering initial discounts or better services could encourage these customers to reconsider and re-engage with the brand.

• **Cluster 4:** High search frequency and low satisfaction: This group browses the site frequently but is dissatisfied with their shopping experience, likely due to high prices, poor product quality, or insufficient customer support. The appropriate strategy for this group could involve improving transparency and providing more detailed information about products. Offering special deals, and guarantees, and showing reviews from other customers could help build trust. Educational content or product guides could assist customers in making better purchasing decisions.

Using principal component-based clustering for a dataset that includes various customer behavior features offers several valuable benefits. Due to the complexity and diversity of the data, features like satisfaction and search frequency may have correlations that make precise analysis challenging. Principal component-based clustering can transform these correlated features into independent principal components, preserving essential information and eliminating unnecessary dimensions and correlations. This dimensionality reduction makes the data more compact and less noisy, ultimately leading to more accurate clustering.

Another advantage of principal component-based clustering in this dataset is a significant improvement in the visual interpretation of customer clusters. By reducing the data to two principal components, the data can be displayed in a two-dimensional space, making it easier to identify different customer groups and their behavioral patterns. This

visual representation of clusters is useful for analysts and marketing teams, as they can quickly identify groups of customers with similar behaviors and better understand the differences between these groups.

Overall, using principal component-based clustering for this dataset helped manage the inherent complexities of customer data and performed clustering more efficiently and interpretably. This clustering also contributes to better marketing strategies and enables targeted services and offers for customers.

In this study, K-means clustering was chosen as the most suitable method for segmenting customer data due to its simplicity, computational efficiency, and ease of interpretation. This method is recognized as an effective clustering technique for large datasets as it allows efficient separation of data into distinct groups based on key features such as customer satisfaction and visit frequency. This method is especially useful when the data is relatively simple, making it easier to interpret, which makes it a valuable tool for analyzing customer behavior.

One of the standout strengths of K-means clustering is its ability to process large datasets quickly, making it an ideal choice for customer segmentation tasks. Using tools like Orange software, this algorithm can be executed quickly to automatically identify different customer segments based on their behavioral patterns. The algorithm works by assigning data points to the nearest cluster center and iteratively adjusting the cluster centers until convergence. This inherent simplicity and computational efficiency make the K-means algorithm a powerful and accessible tool for businesses looking to analyze and segment their customers.

Data analysis revealed that there are four main customer clusters, each with distinct patterns of satisfaction and browsing behavior. These clusters can serve as the foundation for targeted marketing strategies to optimize customer interaction:

- Cluster 1: Customers with moderate satisfaction who browse several times a week. The main goal for this group is to increase overall satisfaction. Offering special discounts, coupons, or exclusive offers could encourage them to interact more and increase their likelihood of purchasing. Additionally, providing feedback channels through surveys or direct communication can help identify factors affecting their satisfaction and improve the overall customer experience.

- Cluster 2: Customers with high satisfaction who browse frequently. Given their high satisfaction and engagement, loyalty programs would be very effective for this group. Rewarding their continuous interaction with personalized offers, special discounts, and other incentives can strengthen their relationship with the brand. Regular updates on new products or services, along with benefits such as early access to sales or special discounts,

can further increase their loyalty and lifetime customer value.

• Cluster 3: Customers who browse frequently and have high satisfaction. Building deeper relationships with this group through exclusive and personalized offers is essential. Since these customers are already highly satisfied, providing time-limited discounts, special offers, or early access to new products can strengthen their loyalty. Additionally, offering educational content—such as guides, tutorials, or product tips—can further increase their engagement. Enhanced post-sale support and follow-up communication to gather feedback can also help strengthen this group.

• Cluster 4: Customers with moderate satisfaction who browse frequently. The main objective for this group is to increase satisfaction and build greater loyalty. Conducting targeted surveys or gathering feedback to understand their concerns is

Table 2.

Siluate score in clustering algorithms

Method	Siluate score
K-means	0.72
DBSCAN	0.65
Hierarchical	0.58
PCA	0.50

Given the superior performance of the K-means algorithm in clustering customer data, this method was chosen as the primary basis for implementing classification techniques. Accordingly, the clusters obtained from this algorithm were used as class labels. The labeling process was performed manually, where each data point received a specific label based on the cluster assigned by the K-means algorithm. This step prepared the data for the application of various classification methods, allowing their effectiveness in predicting customer behavior to be evaluated.

Upon applying the first classification method, the Support Vector Machines (SVM) approach, the results from this model, including the support vectors, confusion matrix, and performance evaluation report, demonstrate the overall model's effectiveness in classifying customer behavior.

The following chart displays the support vectors in the feature space of purchase satisfaction and search frequency. The support vectors, which play a critical role in defining the decision boundaries of the model, are highlighted with circles. The presence of these points indicates that the classification structure.

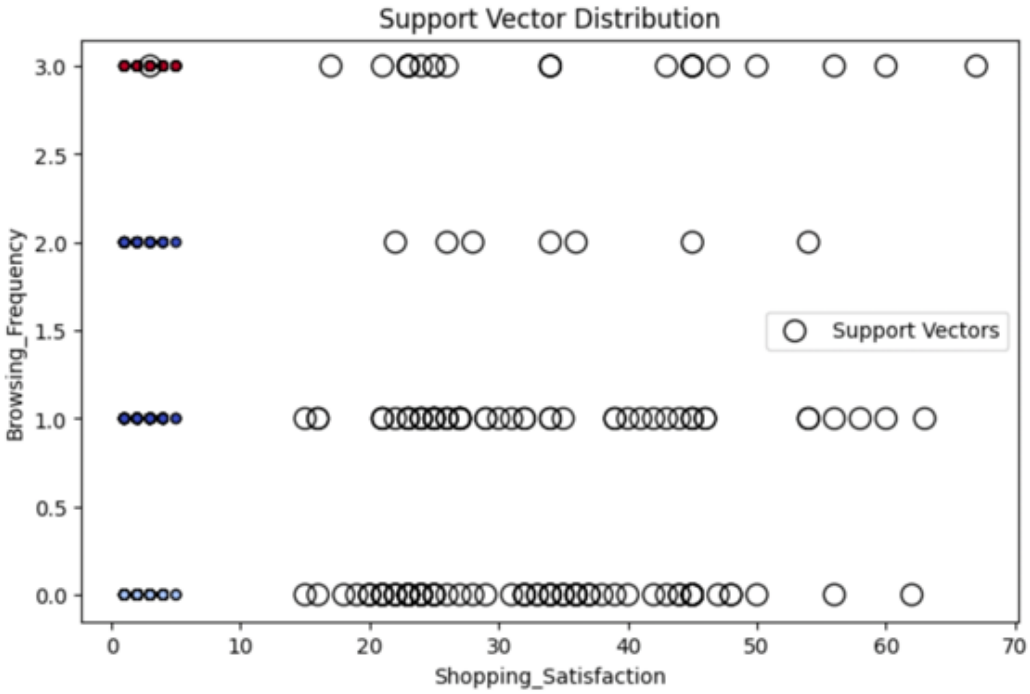

Figure 1. SVM classification

The confusion matrix below illustrates the performance of the Support Vector Machine (SVM) model in predicting the different classes of the dataset. It can be observed that the model has accurately classified most of the samples with high precision. However, a few misclassifications are seen in the third class, which may indicate the model's difficulty in distinguishing this particular class.

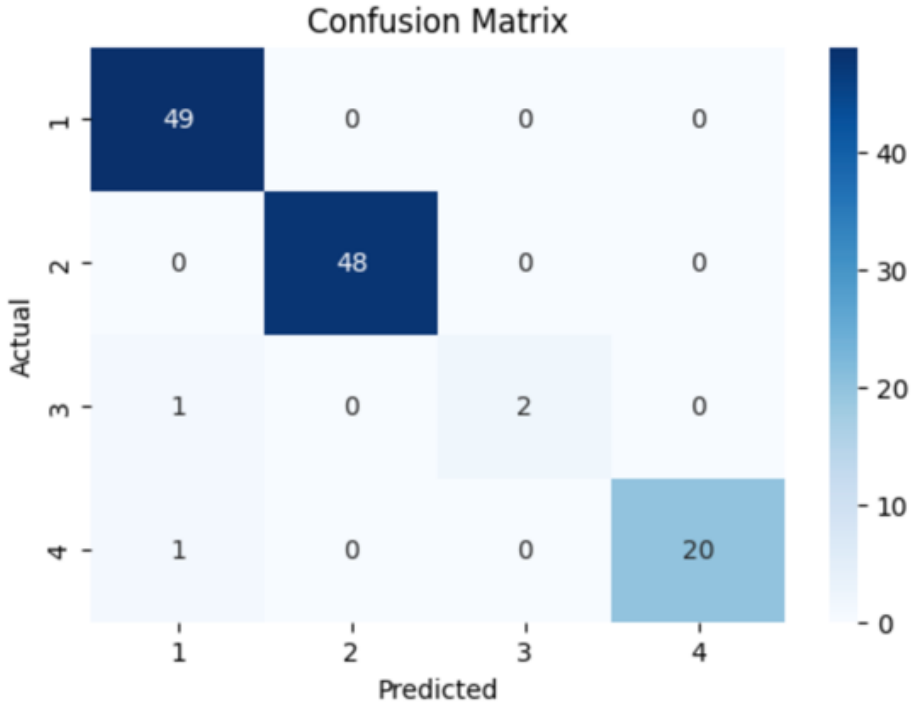

Figure 2. SVM confusion matrix

The model evaluation report indicates that the Support Vector Machine (SVM) algorithm was able to classify 98% of the available data with high accuracy. The precision, recall (sensitivity), and F1 score for each class were calculated, showing that the model performed well in most classes. The decrease in recall (sensitivity) in the third class suggests that some samples from this class were not correctly identified, which could be due to the imbalanced distribution of the data.

Table 3.

SVM classification performance report

Accurcy: 0.98	**precision**	**recall**	**F1-score**	**support**
1	0.96	1.00	0.98	49
2	1.00	1.00	1.00	48
3	1.00	0.67	0.80	3
4	1.00	0.95	0.98	21
Accuracy			0.98	121
Macro avg	0.99	0.90	0.94	121
Weighted avg	0.98	0.98	0.98	121

The Random Forest model, applied to the selected dataset, provides results that require thorough analysis and examination. Initially, the distribution of data in the test set reveals that the different classes have an imbalanced number of samples. Class 1 has the highest number of samples (78 samples), while Classes 3 and 4 have fewer samples (9 and 25 samples, respectively), resulting in an imbalance. This data imbalance can significantly affect the model's performance, particularly for classes with fewer samples, as the model tends to predict the more frequent classes more often.

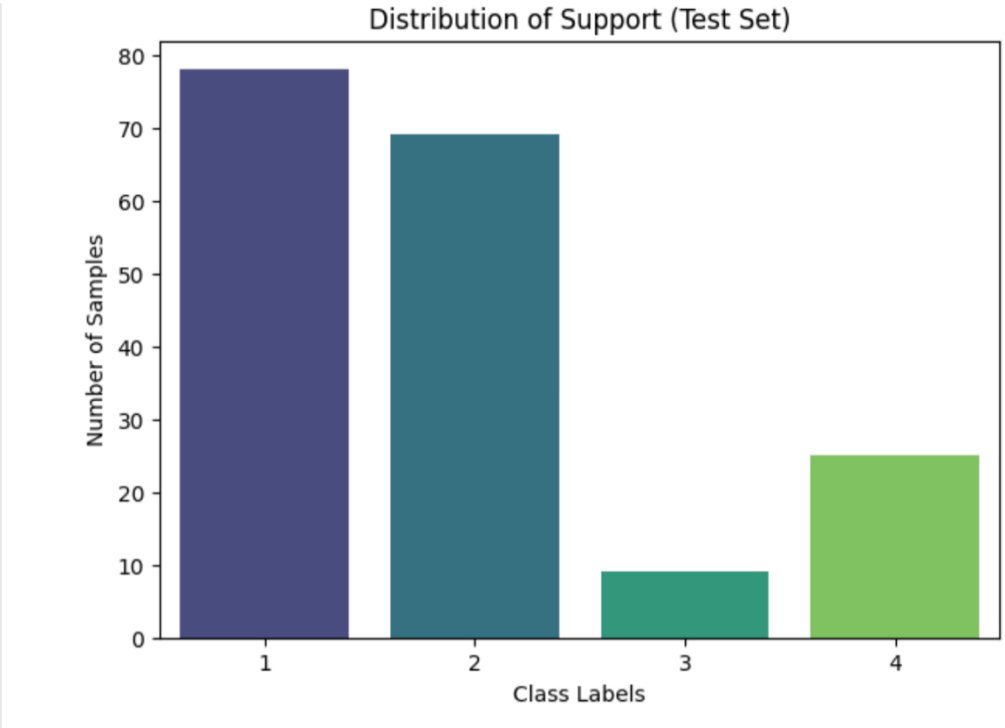

Figure 3. Random forest classification

The confusion matrix of the model shows that Class 1 has the highest correct identification rate; however, some of its samples have still been misclassified into Class 2. On the other hand, Class 2 performs poorly, with many of its samples incorrectly predicted as Class 1. This issue is much more pronounced in Classes 3 and 4, where the model has correctly classified only a few samples, and in most cases, it has assigned them to Class 1. This suggests that the model struggles with classes that have fewer samples and requires further improvement.

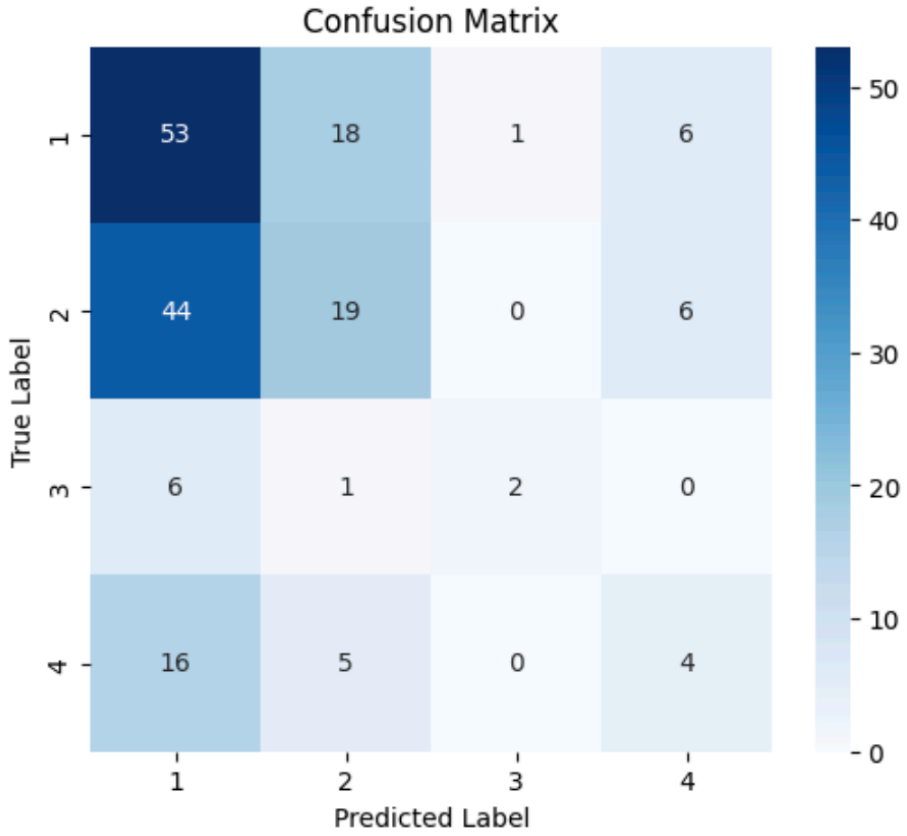

Figure 4. Random forest confusion matrix

performance report of this model further confirms this issue. The model achieves a precision of 0.45 and recall of 0.68 for Class 1, indicating that it correctly identifies most of the samples in this class. However, precision and recall sharply decline in Classes 2, 3, and 4. For instance, the recall for Class 3 is only 0.22, and for Class 4, it is 0.16, suggesting that the model struggles significantly in recognizing these classes. Additionally, the overall precision of the model is 0.43, reflecting a moderate performance. The weighted average F1-score is 0.40, emphasizing that the model faces challenges in achieving a balanced distribution of predictions. The main issue with the model is class imbalance and its tendency to predict more frequent classes.

Table 4.

Random forest classification performance report

Accurcy: 0.43	**precision**	**recall**	**F1-score**	**support**
1	0.45	0.68	0.54	78
2	0.44	0.28	0.34	69
3	0.67	0.22	0.33	9
4	0.25	0.16	0.20	25

Accuracy			0.43	181
Macro avg	0.45	0.33	0.35	181
Weighted avg	0.43	0.43	0.40	181

The results obtained from executing the Gradient Boosting algorithm on the dataset indicate the excellent performance of this model in classifying customer behavior. The support distribution chart shows that the highest number of samples belong to Classes 1 and 2, while Classes 3 and 4 have fewer samples. This distribution indicates that some customer groups are predominant in the reviewed data, while others are in the minority.

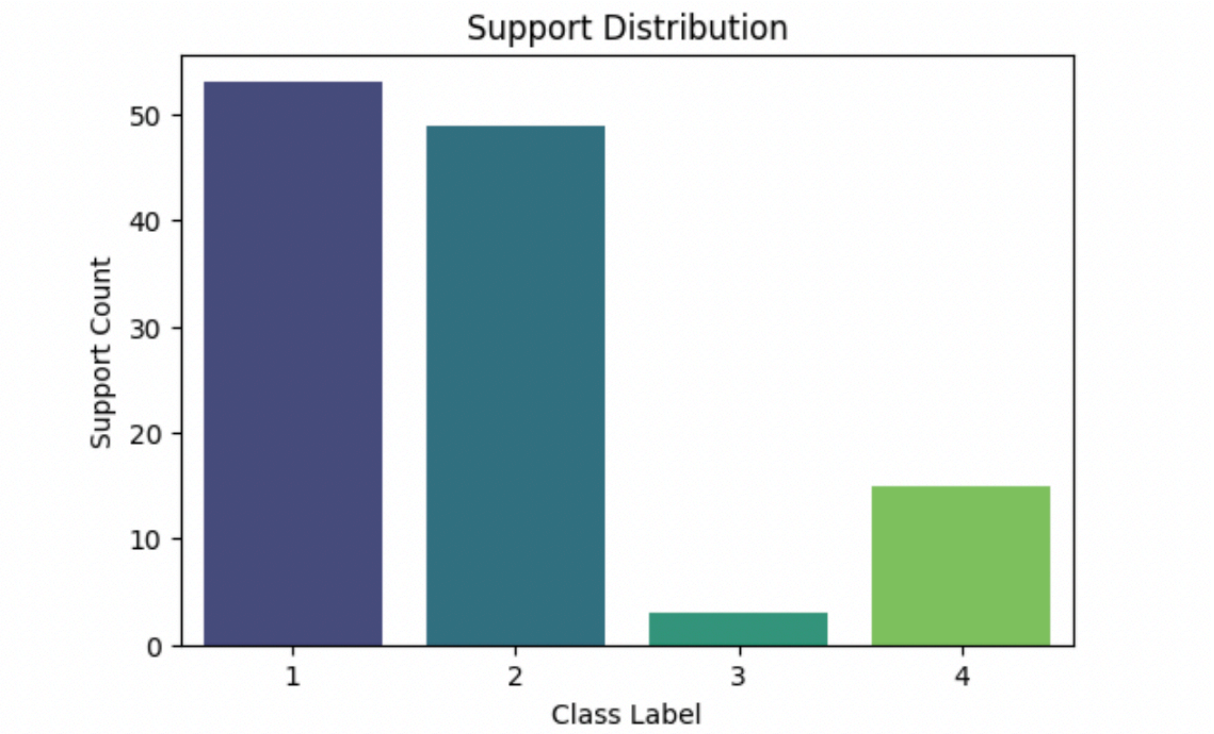

Figure 5. Gradient boosting classification

The results obtained from executing the Gradient Boosting algorithm on the dataset indicate the excellent performance of this model in classifying customer behavior. The support distribution chart shows that the highest number of samples belong to Classes 1 and 2, while Classes 3 and 4 have fewer samples. This distribution indicates that some customer groups are predominant in the reviewed data, while others are in the minority.

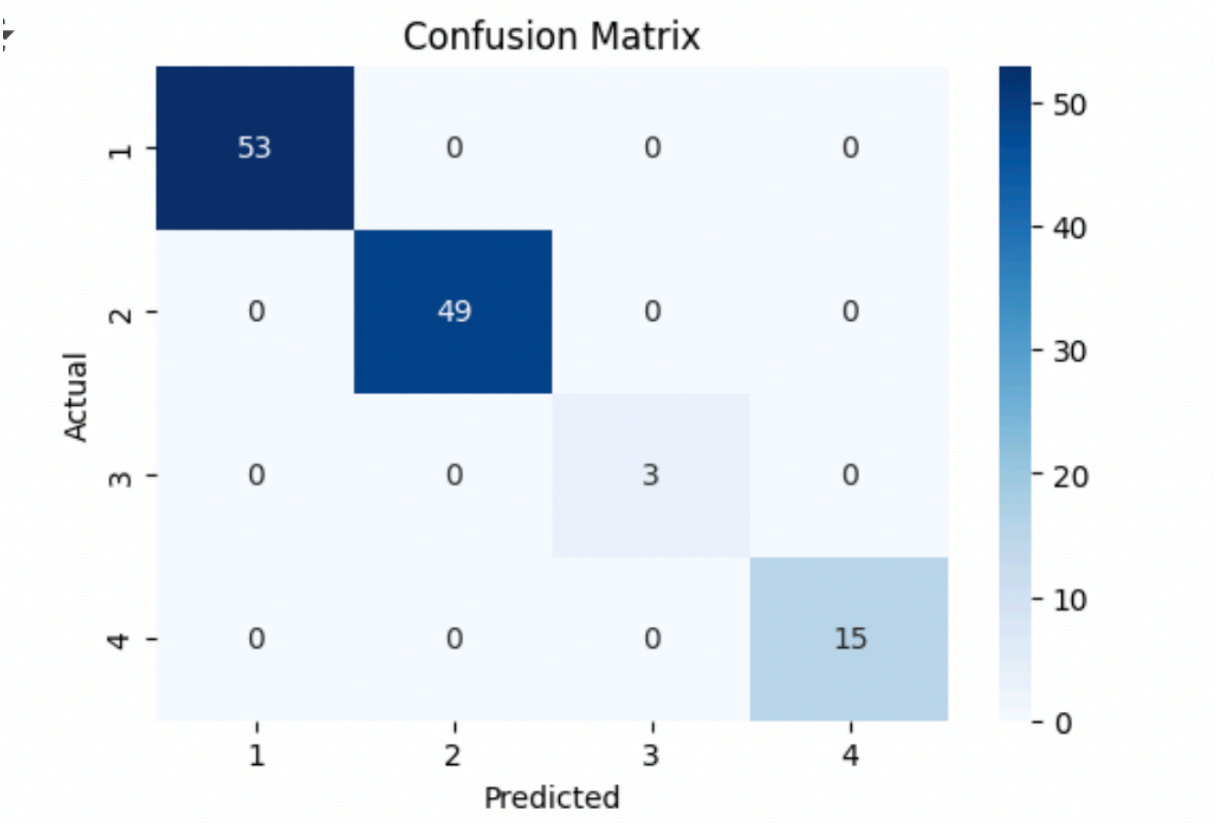

Figure 6. Gradient boosting confusion matrix

The results obtained from executing the Gradient Boosting algorithm on the dataset indicate the excellent performance of this model in classifying customer behavior. The support distribution chart shows that the highest number of samples belong to Classes 1 and 2, while Classes 3 and 4 have fewer samples. This distribution indicates that some customer groups are predominant in the reviewed data, while others are in the minority.

Table 5.

Gradient boosting classification performance report

Accurcy: 1.00	**precision**	**recall**	**F1-score**	**support**
1	1.00	1.00	1.00	53
2	1.00	1.00	1.00	49
3	1.00	1.00	1.00	3
4	1.00	1.00	1.00	15
Accuracy			1.00	120

Macro avg	1.00	1.00	1.00	120
Weighted avg	1.00	1.00	1.00	120

Overall, these results indicate that the Gradient Boosting model, using the selected features, has been able to predict customer behavior with very high accuracy. This high accuracy enables the use of this model in analyzing customer behavior and designing data-driven marketing strategies.

Regarding the performance of the XGBoost model on the Amazon Customer Behavior Survey dataset, the predicted label distribution chart provides valuable insights into the model's behavior. Analyzing the distribution of the predicted labels, the chart shows that the model tends to classify most of the samples into Classes 0 and 1, while the number of predicted samples in Classes 2 and 3 is noticeably lower. This distribution could be due to imbalanced training data or the model's limitations in identifying patterns in underrepresented classes. The density curve further confirms this, as the model's focus is predominantly on Classes 0 and 1, with other classes being less identified.

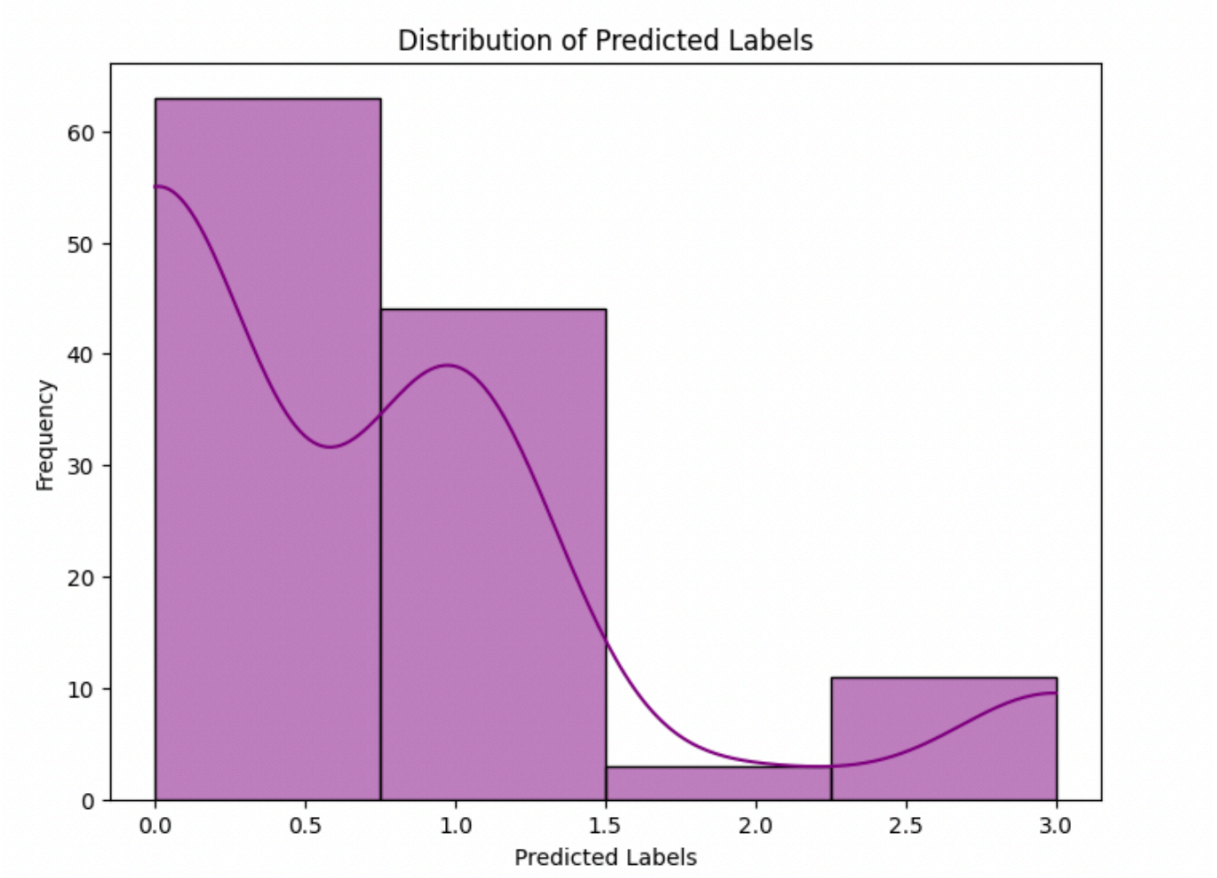

Figure 7. XGBoost classification

The confusion matrix of this model also indicates that the model has varying accuracy in predicting different classes. The highest correct prediction rate is for Class 0, with 30 samples correctly classified, yet 17 samples are mistakenly classified into Class 1. Additionally, Class 1 shows a moderate performance with 20 correctly predicted samples, but the number of incorrectly predicted samples (22 in Class 0) highlights challenges in distinguishing this class.

For Classes 2 and 3, the model performs more poorly. For Class 2, only 1 sample is correctly classified, and 2 samples are assigned to Class 0, indicating imbalanced data or high similarity between class features. Class 3 also faces low accuracy, as 9 of its samples are placed in Class 0, and 7 samples are placed in Class 1. These results suggest that the model struggles to distinguish specific classes and requires further optimizations.

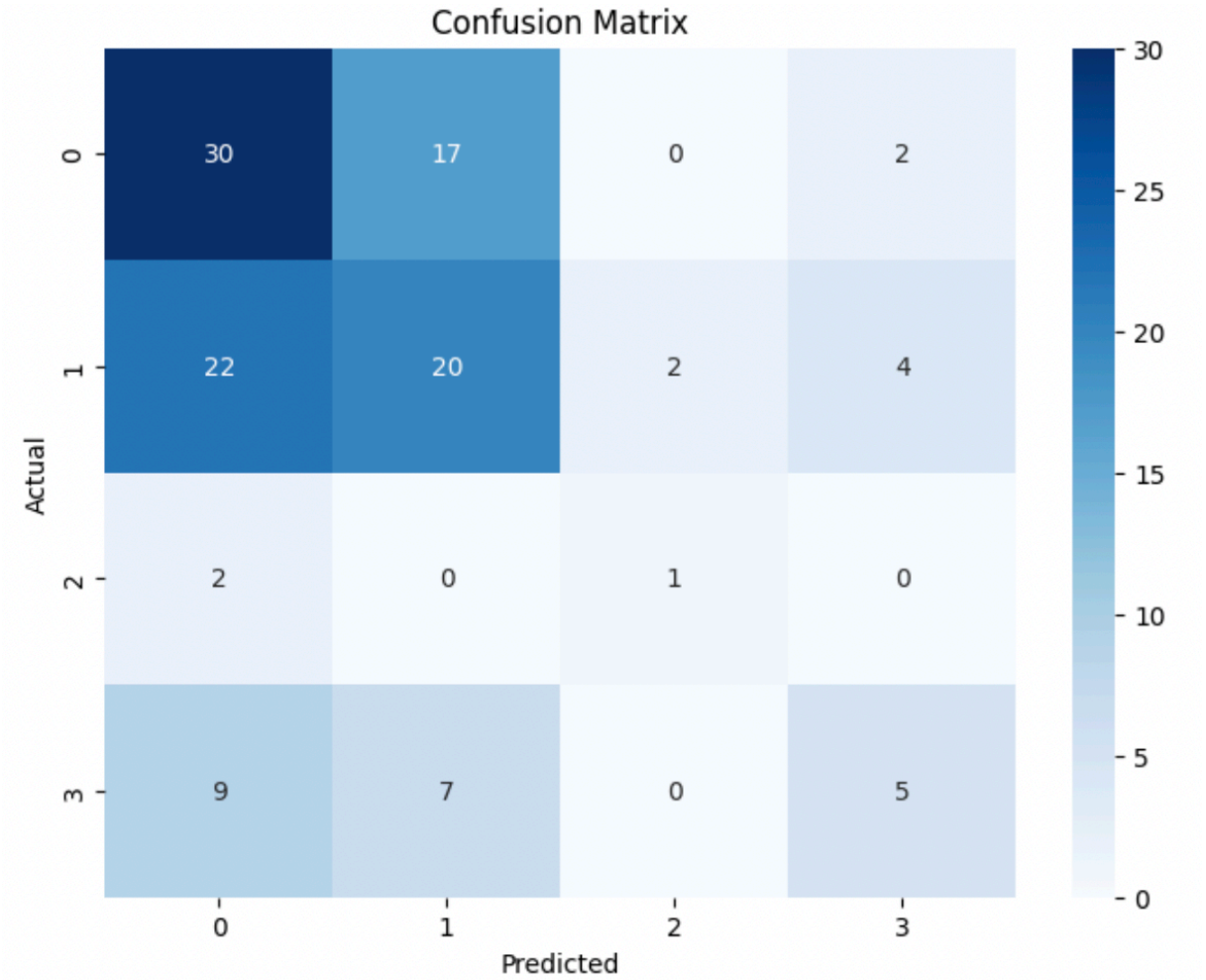

Figure 8. XGBoost confusion matrix

The XGBoost model applied to this dataset has an accuracy of 0.4628, indicating relatively poor performance in predicting the clustering labels. The input data includes features such as age, customer review importance, usage of personalized recommendations,

rating accuracy, and satisfaction with the purchase. The predicted labels suggest that the model tends to favor a specific class, which may indicate data imbalance or weak features in distinguishing the clusters. To improve the model's performance, new features could be added, the data could be balanced, or parameter optimization techniques, such as hyperparameter tuning, could be utilized.

Table 6.

XGBoost classification performance report

Accuracy: 0.46	**age**	**Review_Importance**	**Personalized_Recommendation_Frequency**
0	23	3	3
1	23	1	2
2	34	2	4
3	23	3	3
4	23	3	3

	Rating_Accuracy	**Shopping_Satisfaction**	**Cluster_Label**
0	3	3	3
1	2	2	2
2	4	4	3
3	2	3	3
4	3	3	3

Conclusion

The results above indicate that XGBoost has performed well in classifying certain classes but faces challenges in distinguishing specific classes. To enhance the model's performance, it is suggested to utilize techniques such as data augmentation, class weight adjustment, alternative metrics like the F1-score, or the use of ensemble algorithms to increase prediction accuracy for underrepresented classes.

The results also show that the gradient-boosting model achieved optimal performance with 100% accuracy. However, this value might indicate overfitting, particularly if the test data size is limited. In contrast, the Support Vector Machine (SVM) algorithm achieved 98% accuracy, demonstrating very good performance with results close to optimal without showing signs of overfitting. On the other hand, the Random Forest and XGBoost models showed weaker performance, with accuracies of 40% and 46%, respectively, indicating their inability to generalize well for this dataset.

Based on this analysis, the Support Vector Machine (SVM) algorithm is selected as the best classification method for this dataset, as it not only provides high accuracy but also avoids overfitting and demonstrates better generalizability.

Table 7.

Classification methods accuracy

Method	Accuracy
Gradient Boost	100%
SVM	98%
XGBoost	46%
Random Forest	40%

The results of this study indicate that various clustering and classification methods demonstrated different performances in analyzing customer behavior. In the clustering section, the K-means method, being one of the most widely used, successfully divided customers into meaningful clusters based on their satisfaction with purchases and browsing frequency. The analysis of these clusters reveals that increased customer interaction and page browsing are often associated with higher satisfaction levels. On the other hand, the hierarchical clustering method provided a more flexible analysis of customer clusters and displayed customer relationships in a more structured manner. This method is suitable for strategic decision-making but is computationally more complex when dealing with large datasets. Additionally, methods such as principal component analysis and density-based clustering were also implemented, each showing its own strengths and weaknesses.

In the classification section, four algorithms—Support Vector Machine (SVM), Random Forest, Gradient Boosting, and XGBoost—were examined. The results showed that the SVM algorithm, with an accuracy of 98%, performed better in predicting cluster labels, and accurately classifying customers into appropriate groups. The Gradient Boosting model, with an accuracy of 100%, showed excellent performance; however, this could indicate overfitting, particularly if the test data size is small. In contrast, the Random Forest and XGBoost models, with accuracies of 40% and 46%, respectively, showed weaker generalization performance, indicating their inability to predict customer categories accurately.

Overall, the findings of this study confirm that combining clustering and classification methods can be an effective tool for analyzing customer behavior in data-driven marketing. Among the methods reviewed, the Support Vector Machine (SVM) algorithm is selected as the best classification method for this dataset, as it not only provides high accuracy but also avoids overfitting and demonstrates better generalization.

Based on the results of this study, several recommendations are proposed to improve the accuracy of customer behavior analysis models and enhance the effectiveness of classification methods. First, increasing the volume of data and utilizing data augmentation techniques can help balance the classes and improve the performance of classification models. Imbalanced data may negatively impact the performance of machine learning algorithms, so applying data augmentation methods can enhance the generalization of models.

In addition, optimizing classification models through fine-tuning hyperparameters can contribute to improving prediction accuracy. Optimizing key parameters, such as the learning rate in gradient boosting algorithms, and utilizing comprehensive evaluation methods like the F1-Score, can reduce error rates and improve model performance. Furthermore, employing ensemble models, such as combining Support Vector Machine (SVM) and XGBoost algorithms, can increase overall classification accuracy and mitigate the weaknesses of each individual model.

Finally, improving data quality and applying advanced preprocessing techniques can have a significant impact on enhancing the accuracy of machine learning models. Removing noise, standardizing features, and reducing data dimensions using methods like Principal Component Analysis (PCA) not only reduces computational complexity but also improves model prediction accuracy. Therefore, implementing these solutions can be effective in improving customer behavior analysis and optimizing data-driven marketing strategies.

References

Abu'i Mehrizi, Abbas. (2019). " Analyzing customer purchasing behavior in a retail store using data mining. International Conference on Industrial Engineering " International Conference on Industrial Engineering. DOI: https://doi.org/10.22105/riej.2024.468414.1458

Ahmadipanah, M., Chalaki, K., & Shakeri, R. (2022). "Designing Cell Production Arrangement Scenarios with the Approach of Artificial Neural Networks", *Journal of System Management*, Vol. 8, No. 4, pp. 49–64. Online ISSN: 2322-2301. DOI: 10.30495/JSM.2022.1964485.1673.

Bahrini-Zad, M., Asar, M., & Esmailpour, M. (2022). "Segmenting online retail customers based on demographic characteristics and customer experience ", New Marketing Research, Vol. 44, Issue B, pp. 69-88. DOI: 10.22108/NMRJ.2021.130039.2519

Dezbandi, Afsaneh. (2020). "The Importance of Customer Clustering on Brand Value Creation Based on Aaker's Brand Equity Model: A Case Study of Shahroud Dairy Products - Miami Cheese." 18th International Conference on Management. DOR: IAMS18_002. [In Persian].

Dadres, M., Rahimi Nik, A., & Nematizadeh, S. (2023). " A study of the financial marketing model of the National Bank of Iran with emphasis on customer grouping, financial engineering, and securities management. " Financial Engineering and Securities Management, Winter 2023, Issue 57, A, pp. 65-79. [In Persian]

Dust Mohammadi, V., Albadvi, A., & Teymorpur, B. (2014). "Predicting Customer Churn Using CLV in Insurance Industry", *Shiraz Journal of System Management*, Vol. 2, No. 1, Ser. 5, pp. 39–49.

Sharifi Isfahani, Hamid. (2023). " Providing an approach based on customer purchase history and product recommendations to customers: A case study of Digikala customers" Journal of Industrial Management Perspectives. DOI: 10.48308/JIMP.13.2.99

Safabakhsh, M., & Asayesh, F. (2022). "Segmentation of bank customers based on customer lifetime value and their profitability ability (case study: customers of a private bank)" Islamic Financial Studies and Banking, 8th Year, Issue 19, pp. 53-80. [In Persian]

Mohaghegh, A., Habibnejad Behtash, N., Sheikhzadeh Sani, H., & Karimi, A. (2022). " Explaining the financial marketing model with emphasis on customer segmentation of Tejarat Bank Iran."Proceedings of the 6th International Conference on Interdisciplinary Studies in Management and Engineering, International, pp. 791-811. [In Persian]

Taghavi-Fard, Mohammad Taqi. (2022). " Customer clustering in the field of electronic banking using electronic transactions and demographic information (case study: Refah Bank). Journal of Management, Advertising, and Sales. DOR: JR_BUMARA-2-3_004. [In Persian]

Alamsyah, P. E. P., Prasetyo, S., Sunyoto, S., Bintari, S. H., Saputro, D. D., Rohman, S., & Pratama, R. N. (2022). " Customer Segmentation Using the Integration of the Recency Frequency Monetary Model and the K-Means Cluster Algorithm", Scientific Journal of Informatics, Vol. DOI: 10.15294/sji.v9i1.29127

Alnuaimi, A. F. A. H., & Albaldawi, T. H. K. (2024). "An overview of machine learning classification techniques," BIO Web of Conferences, 97, 00133. DOI: 10.1051/bioconf/20249700133.

Ayodele, E., & Sodeinde, V. (2024). "Customer segmentation using the K-means clustering algorithm," Ilaro Journal of Science and Technology (IJST), 4, 1-6.

Budiyono, M., Tho'in, M., Muliasari, D., & Putri, S. A. R. (2021). "An Analysis of Customer Satisfaction Levels in Islamic Banks Based on Marketing Mix as a Measurement Tool", Annals of R.S.C.B., Vol. 25, Issue 1, pp. 2004-2012. ISSN: 1583-6258.

Guerola-Navarro, V., Gil-Gomez, H., Oltra-Badenes, R., & Sendra-García, J. (2021). "Customer relationship management and its impact on innovation: A literature review," Journal of Business Research, 129, 83-87. DOI: 10.1016/j.jbusres.2021.02.052

Guerola-Navarro, V., Gil-Gomez, H., Oltra-Badenes, R., & Soto-Acosta, P. (2024). "Customer relationship management and its impact on entrepreneurial marketing: A literature review," International Entrepreneurship and Management Journal, 20, 507-547. DOI: https://doi.org/10.1007/s11365-022-00800-x

Hariyanto, H. T., & Trisunarno, L. (2020). " Putra, D. P., Suprihartini, L., & Kurniawan, R. (2021). " Celebrity Endorser, Online Customer Review, Online Customer Rating on Purchasing Decision with Trust as an Intervening Variable on Tokopedia Marketplace", Jurnal Bahtera Inovasi, Vol. 5, No. 1, pp.ISSN 2747-0067.", JURNAL TEKNIK ITS, Vol. 9, No. 2, pp. DOI:

https://doi.org/10.31629/bi.v5i1.3800

Kamisa, N., Devita, A., & Novita, D. (2022). "The influence of online customer reviews and online customer ratings on customer trust IN FORE COFFEE PRODUCTS AT SUN", Journal of Economic and Business Research, Vol. 2, No. 1, pp. 21-29. DOI:10.54443/ijebas.v3i4.964

Li, Y., Chu, X., Tian, D., Feng, J., & Mu, W. (2021). "Customer segmentation using K-means clustering and the adaptive particle swarm optimization algorithm," Applied Soft Computing, 113(B), 107924.

Li, Y., Meng, C., Tian, J., Fang, Z., & Cao, H. (2024). "Data-Driven Customer Online Shopping Behavior Analysis and Personalized Marketing Strategy", Journal of Organizational and End User Computing, Vol. 36, Issue 1, pp. DOI: 10.4018/JOEUC.346230

Lone, H., & Warale, P. (2022). " Cluster Analysis: Application of K-Means and Agglomerative Clustering for Customer Segmentation", Journal of Positive School Psychology, Vol. 6, No. 5, pp. 7798–7804.

Rachman, F. P., Santoso, H., & Djajadi, A. (2024). "Machine learning mini-batch k-means and business intelligence utilization for credit card customer segmentation," Proceedings of the International Conference on Data Science and Business Intelligence, 1-6. DOI:10.14569/IJACSA.2021.0121024

Wu, S., Yau, W.-C., Ong, T.-S., & Chong, S.-C. (2021). "Integrated churn prediction and customer segmentation framework for telco business," IEEE Access, 9, 113456-113467. DOI: 10.1109/ACCESS.2021.3073776

Xiao, Z., Zhao, J., Li, Y., Shindou, R., & Song, Z.-D. (2024). "Spin space groups: Full classification and applications," Journal of Quantum Materials, 1(1), 1-6. DOI: 10.1103/PhysRevX.14.031037

Xiahou, X., & Harada, Y. (2022). "B2C E-Commerce Customer Churn Prediction Based on K-Means and SVM", J. Theor. Appl. Electron. Commer. Res., 17, pp. 458–475. DOI: https://doi.org/10.3390/jtaer17020024

Tabianan, K., Velu, S., & Ravi, V. (2024). "K-means clustering approach for intelligent customer segmentation using customer purchase behavior data," Sustainability, 16(1), 1-10. DOI: https://doi.org/10.3390/su14127243

Putra, D. P., Suprihartini, L., & Kurniawan, R. (2021). "Celebrity Endorser, Online Customer Review, Online Customer Rating on Purchasing Decision with Trust as an Intervening Variable on Tokopedia Marketplace", Jurnal Bahtera Inovasi, Vol. 5, No. 1, pp.ISSN 2747-0067.