Evaluating Customer Segmentation Techniques in the Retail Sector

Nur Diyabi[1], Duygu Çakır[2], Ömer Melih Gül[1, 3]*, Tevfik Aytekin[1], Seifedine Kadry[4]

[1] Department of Computer Engineering, Bahcesehir University, Istanbul (Türkiye)
[2] Department of Software Engineering, Bahcesehir University, Istanbul (Türkiye)
[3] Informatics Institute, Istanbul Technical University, Istanbul (Türkiye)
[4] Department of Computer Science and Mathematics, Lebanese American University, Beirut (Lebanon)

* Corresponding author: omgul@itu.edu.tr

Received 26 January 2025 | Accepted 24 May 2025 | Early Access 28 May 2025

LA UNIVERSIDAD
EN INTERNET

Abstract

In the current competitive corporate landscape, understanding client preferences and adapting marketing strategies accordingly has become crucial. This study evaluates the effectiveness of four machine learning algorithms (K-Means, Density-Based Spatial Clustering of Applications with Noise (DBSCAN), Gaussian Mixture Models (GMM), and Self-Organizing Maps (SOM)) for customer segmentation in the Turkish retail market. Two datasets were analyzed: a large-scale Turkish market sales dataset and a focused marketing campaign dataset. The research employed a comprehensive methodology encompassing data preparation, algorithm application, and performance evaluation using metrics such as the Calinski-Harabasz Index and Davies-Bouldin score. Results indicate that K-Means demonstrated superior performance in terms of interpretability and statistical validity. DBSCAN showed strengths in identifying non-spherical clusters, while GMM and SOM provided more granular segmentation. The findings offer actionable insights for Turkish retailers to optimize marketing strategies and enhance customer relationship management. This study contributes to the field of retail analytics by providing a methodological framework for evaluating customer segmentation techniques in specific market contexts.

Keywords

Clustering Algorithms, Customer Segmentation, Machine Learning, Retail Analysis, Unsupervised Learning.

DOI: 10.9781/ijimai.2025.05.001

I. Introduction

In the dynamic and competitive landscape of modern retail, understanding and effectively segmenting customers has transcended from being a mere advantage to becoming an absolute necessity. Customer segmentation, the meticulous process of grouping customers with similar characteristics and purchasing behaviors, has emerged as a critical strategy for businesses to navigate this complex terrain [1]. Traditional approaches to customer segmentation, while valuable, often fall short in capturing the patterns hidden within vast and complex datasets. The advent of machine learning techniques has opened new avenues for more sophisticated and accurate customer segmentation. These methods promise to uncover hidden patterns and insights that go beyond basic demographics, potentially revolutionizing how businesses understand and interact with their customers [2].

This paradigm shift is particularly evident in the Turkish retail market, where local supermarkets face the dual challenge of intense competition and rapidly evolving consumer preferences. The Turkish retail sector, characterized by its diversity and rapid growth, presents a unique context for studying customer segmentation. With major players like A101, BIM, CarrefourSA, and Migros dominating the field of supermarkets in Türkiye, the need for sophisticated customer insights has never been more pressing. These retailers are increasingly turning to data analytics to gain a competitive edge, with customer segmentation at the forefront of their strategies [3].

The primary goal of this study is to evaluate the effectiveness of four machine learning algorithms (K-Means clustering, Density-Based Spatial Clustering of Applications with Noise (DBSCAN), Gaussian Mixture Models (GMM), and Self-Organizing Maps (SOM)) for customer segmentation in the Turkish retail market. This evaluation involves comparing these algorithms across two datasets to identify their strengths and weaknesses in uncovering actionable customer segments. The study also assesses the algorithms using robust metrics and explores their practical implications for targeted marketing and customer relationship management, ultimately developing a framework for selecting the most suitable segmentation technique based on specific data and business needs.

Please cite this article as:
N. Diyabi, D. Çakır, Ö. M. Gül, T. Aytekin, S. Kadry. Evaluating Customer Segmentation Techniques in the Retail Sector, International Journal of Interactive Multimedia and Artificial Intelligence, vol. 9, no. 3, pp. 175-190, 2025, http://dx.doi.org/10.9781/ijimai.2025.05.001

This study contributes to the broader field of retail analytics by providing a methodological framework for evaluating customer segmentation techniques in specific market contexts. As businesses worldwide grapple with the challenges of data-driven decision-making, the findings offer insights that can inform strategy development and implementation across various retail environments. By identifying the most effective segmentation techniques for the Turkish retail market, local supermarkets can be equipped with the tools to make data-driven decisions about customer targeting strategies. This has far-reaching implications for enhancing customer satisfaction, optimizing marketing return of investment (ROI), tailoring product offerings, and ultimately fostering long-term customer loyalty in a highly competitive market.

In the following sections, the theoretical details of each segmentation technique will be analyzed, the methodology for comparison will be outlined, findings will be presented, and their implications for both practice and future research will be discussed. Through this comprehensive analysis, the understanding of customer segmentation in retail can be advanced, and actionable insights for businesses seeking to leverage data for competitive advantage in the dynamic world of modern retail can be provided.

This study is significant for advancing customer segmentation techniques by comparing advanced machine learning algorithms, with a focus on the Turkish retail market. It provides valuable insights for both academics and practitioners, offering practical implications for retailers to optimize their marketing strategies, enhance customer experiences, and improve decision-making through data-driven approaches. Additionally, the study contributes to the broader field of applied machine learning, with potential economic benefits for retail sector.

II. Literature Review

In customer segmentation, various methodologies have been developed across industries. This study focuses on four key approaches: Density-Based Spatial Clustering, Gaussian Mixture Models, Self-Organizing Maps, and K-means Clustering. These methods have been extensively studied and applied in customer segmentation, each offering distinct advantages and challenges.

The literature emphasizes the critical role of effective customer segmentation in enhancing marketing strategies, improving customer satisfaction, and driving business performance. It is stated that segmentation enables businesses to tailor offerings and communications to specific customer groups, leading to more efficient resource allocation and improved customer relationships [4].

Recent years have witnessed a shift towards machine learning-based approaches in customer segmentation. These methods have proven effective in identifying complex patterns within large datasets, a valuable capability in the current data-rich business environment. However, the effectiveness of these methods can vary depending on context and data characteristics. Mehrabi *et al.* [5] emphasize that factors such as data quality, algorithmic bias, and result interpretability must be carefully considered in real-world applications .

The following subsections will examine each approach in detail, focusing on their theoretical foundations, practical applications, and reported effectiveness in customer segmentation tasks. This review aims to provide a foundation for understanding the comparative analysis conducted in this study.

A. K-means Clustering Approach

K-means clustering has been widely recognized as a popular and effective unsupervised machine learning algorithm for customer segmentation. Its ability to cluster data points based on similarity without requiring labeled data has made it particularly useful in scenarios where customer labels are not readily available.

The application of K-means in customer segmentation has been extensively documented across various industries. In the retail sector, Kansal et al. [6] reported a case study where K-means was used to segment customers into four groups based on their shopping habits: high-value, medium-value, low-value, and at-risk customers. In the banking and financial services industry, Mohit [7] described the use of K-means to segment bank customers into three risk categories: low-risk, medium-risk, and high-risk, based on their risk profiles. Additionally, in telecommunications, Rungruang *et al.* [9] presented a study where K-means was employed to segment telecom customers into four groups based on usage patterns: heavy users, medium users, light users, and inactive users.

The process of applying K-means clustering for customer segmentation typically involves several key steps, as outlined in the literature. Determining the optimal number of clusters (k) is a crucial step that often involves trial and error, with various k values being assessed based on domain knowledge and business goals [7]. The initialization of centroids is also important, with advanced techniques like k-means++ being used to distribute centroids more evenly across the data [10]. The choice of distance metric, such as Euclidean or Manhattan distance, can significantly impact the clustering results [11]. Finally, the iterative process of updating centroids and reassigning data points continues until convergence is reached, indicating the successful identification of distinct customer segments [6].

While K-means has been widely applied and proven effective, several limitations have been identified in the literature. The algorithm's sensitivity to the initial placement of centroids can lead to suboptimal segmentation results [10]. Additionally, K-means assumes that clusters are spherical, which may not always align with real-world data distributions [11]. Furthermore, the requirement to pre-specify the number of clusters (k) beforehand can be challenging and may necessitate domain expertise or trial-and-error approaches [9].

Despite these limitations, K-means clustering is a popular choice for customer segmentation due to its simplicity, efficiency, and effectiveness in many practical cases.

B. Density-Based Spatial Clustering

Density-Based Spatial Clustering of Applications with Noise (DBSCAN) has emerged as a powerful tool for customer segmentation, especially in scenarios where clusters have irregular shapes and varying densities. The algorithm's ability to detect clusters of any shape and its robustness against noise have been widely recognized in the literature [12]. DBSCAN operates by grouping together points that are closely packed in space, marking points that lie alone in low-density regions as outliers. This approach is particularly valuable in customer segmentation, where traditional centroid-based methods may fail to capture complex relationships between customers.

The process of DBSCAN clustering typically involves several key steps: constructing a neighborhood graph, where each node represents a data point and edges connect points within a specified distance (epsilon); identifying core points, which have at least a minimum number of points (MinPts) within their neighborhood; expanding clusters from core points to density-reachable points; and labeling points not belonging to any cluster as noise [13].

The effectiveness of DBSCAN in customer segmentation has been demonstrated across various industries. In a case study of a retail company, DBSCAN clustering resulted in the identification of three distinct customer groups [13]. These groups were characterized by different purchasing behaviors and demographic profiles, providing meaningful analysis for targeted marketing strategies.

One of the key advantages of DBSCAN, is its ability to handle outliers effectively [15]. In the context of customer segmentation, this translates to the ability to identify niche customer groups or unusual purchasing patterns that might be overlooked by other methods.

However, challenges associated with DBSCAN have also been identified in the literature. The selection of appropriate values for the epsilon and MinPts parameters can be critical to the algorithm's performance, as highlighted by Schubert *et al.* [16]. This selection often requires domain knowledge and can impact the resulting segmentation.

Despite these challenges, DBSCAN has been widely adopted for customer segmentation tasks, particularly in scenarios where the shape of clusters is not known *a priori*. Its ability to identify clusters of varying densities and shapes makes it a valuable tool in the increasingly complex landscape of customer behavior analysis.

C. Gaussian Mixture Model Customer Segmentation

Gaussian Mixture Models (GMMs) have been increasingly applied in customer segmentation due to their ability to model complex, multi-modal data distributions. As described by Scientific [17], GMMs model the data as a mixture of Gaussian distributions, with each distribution potentially representing a distinct customer group.

The application of GMMs in customer segmentation has been documented across various industries. In the retail sector, Zakrzewska and Murlewski [8] reported the use of GMMs to categorize retail customers into four segments: high-value, medium-value, low-value, and at-risk, based on purchasing habits, demographics, and other characteristics, where they utilized a hybrid GMM-fuzzy logic model to segment bank customers into three risk categories: low-risk, medium-risk, and high-risk, based on account activity and other variables.

The effectiveness of GMMs in customer segmentation has been attributed to their ability to capture complex, multi-dimensional relationships in customer data. Naga's study [17] on a dataset including customer age, demographics, gender, income, and purchase history reported an accuracy of 70% in customer segmentation using GMM.

However, several challenges associated with GMM-based segmentation have been noted in the literature. Determining the optimal number of clusters can be difficult, as selecting too many Gaussian components may lead to overfitting [17]. The computational cost of training GMMs on large datasets can be high, potentially limiting their use in real-time customer segmentation scenarios [18]. Additionally, the complexity of GMMs can make it challenging for businesses to interpret the resulting customer segments and translate them into actionable marketing strategies [19]. GMMs are also sensitive to the initial values of model parameters, which can affect performance [20]. Lastly, GMMs operate under the assumption that the data is generated from a mixture of normal distributions, an assumption that may not always hold in real-world scenarios [21].

Related to probabilistic clustering approaches, fuzzy clustering methods have also shown promise in customer segmentation applications. A recent study by Saadi et al. [22] demonstrated how fuzzy clustering can improve retrieval performance in case-based reasoning systems, suggesting potential synergies between fuzzy methods and customer segmentation tasks. While our study focuses on GMM, future research could explore the comparative performance of fuzzy clustering approaches in the Turkish retail context.

D. Self-Organizing Maps Clustering

Self-Organizing Maps (SOMs) [27], an unsupervised machine learning algorithm, have been widely applied in customer segmentation due to their ability to handle complex, high-dimensional data. The effectiveness of SOMs in visualizing and analyzing such data has been documented in studies [23].

The utility of SOMs in customer segmentation has been demonstrated in several case studies. Üstebey et al. presented a case study on airline passengers, where SOMs were used to segment customers based on attributes such as ticket type, fare type, travel date, and total fare paid [26]. The study resulted in the identification of four distinct customer groups, providing observations for targeted marketing strategies.

Key advantages of SOMs in customer segmentation, as highlighted in the literature, include their ability to visualize high-dimensional data by projecting it onto a lower-dimensional space while preserving relationships between data points, which aids in understanding complex customer behaviors. Additionally, SOMs facilitate feature extraction from complex datasets, potentially uncovering hidden patterns in customer behavior. Moreover, SOMs are capable of handling non-linear relationships within the data, making them particularly suitable for complex customer datasets.

However, several challenges associated with the use of SOMs have been identified. SOMs can be sensitive to the initial state of the algorithm, potentially leading to different segmentation results based on initialization [28]. Additionally, the computational cost of training SOMs on large datasets can be significant, which may limit their applicability in certain scenarios. Furthermore, the complexity of SOMs can sometimes make it difficult for businesses to interpret the resulting customer segments and translate them into actionable marketing strategies.

To address these challenges, various techniques have been proposed in the literature. For instance, Valova *et al.* [28] suggested using multiple initialization techniques and selecting the model that yields the best results. Liu *et al.* [23] and Lundberg & Lee [20] proposed methods to enhance the interpretability of SOMs, including feature selection and the application of visualization techniques.

III. Problem Definition

In the Turkish retail market, the need for sophisticated customer segmentation techniques has become increasingly apparent. The problem addressed in this study is the evaluation and comparison of various machine learning algorithms for customer segmentation, with a focus on their applicability and effectiveness in the Turkish retail sector.

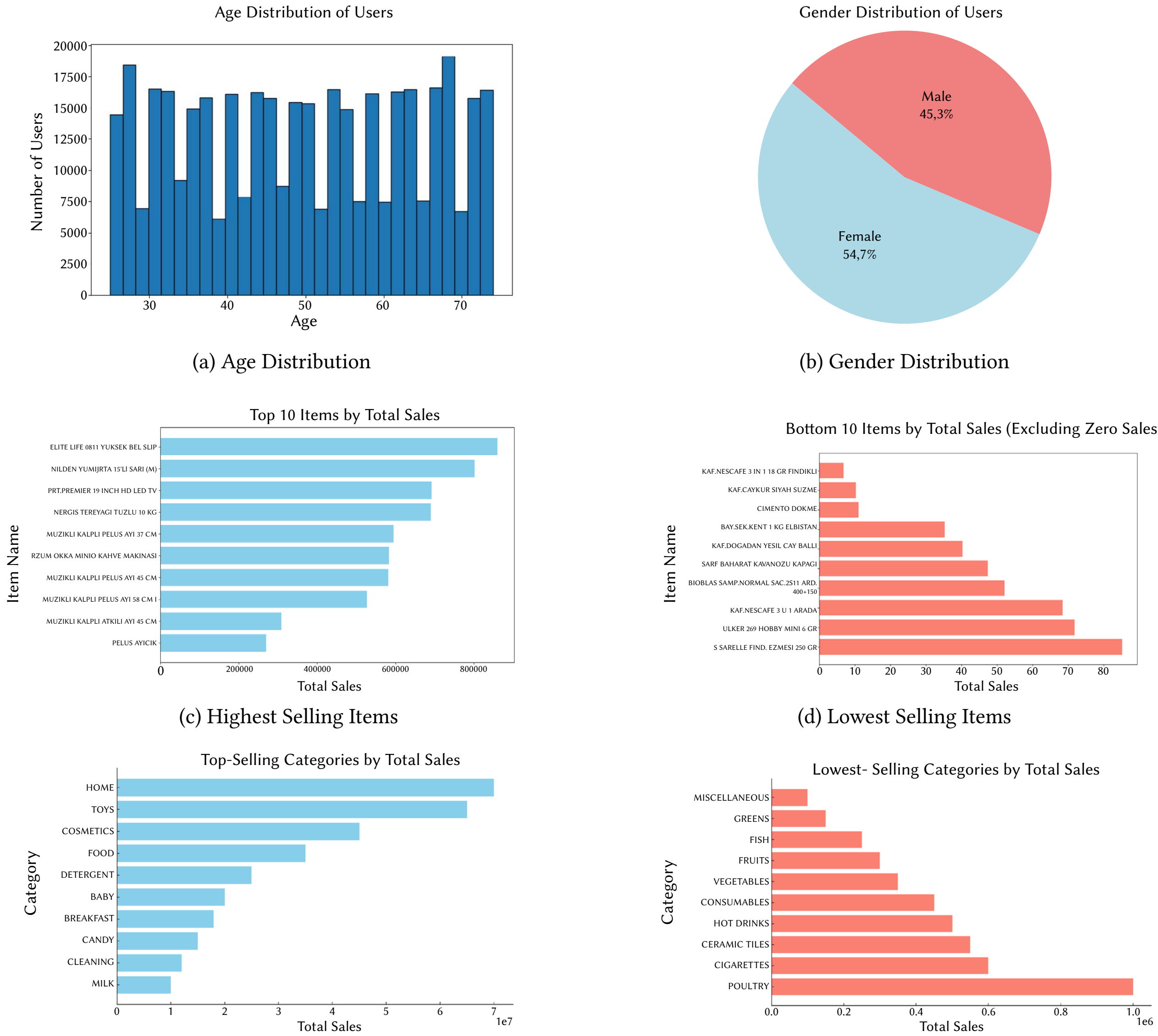
The primary challenge lies in determining which of the four selected algorithms is the most effective and actionable customer segmentation technique for the unique characteristics of the Turkish retail sector: K-Means, Density-Based Spatial Clustering of Applications with Noise (DBSCAN), Gaussian Mixture Models (GMM), and Self-Organizing Maps (SOM). This problem is compounded by the diverse nature of available data and the specific characteristics of the Turkish market.

Two distinct datasets are utilized in this study to provide a comprehensive evaluation:

1. Turkish Market Sales Dataset: A large-scale Turkish market sales dataset comprising 10 million rows.
2. Marketing Campaign Dataset: A more focused marketing campaign dataset.

The use of these two datasets allows for the assessment of the algorithms' performance across different data scales and characteristics, which is crucial for understanding their practical applicability in various retail scenarios.

The problem addresses several critical aspects, including the identification of the most effective algorithms for segmenting customers based on their purchasing behavior and other relevant attributes. It also involves assessing the algorithms' ability to manage both large-scale



(a) Age Distribution

(b) Gender Distribution

(c) Highest Selling Items

(d) Lowest Selling Items

(e) Top Sales by Category

(f) Lowest 10 Sales

Fig. 1. Demographic, economic, and consumer behavior analysis on the TMS (first) dataset.

data and more focused datasets, while evaluating the interpretability and actionability of the resulting customer segments. Furthermore, the problem includes determining the computational efficiency and scalability of each algorithm, as well as assessing the robustness of the segmentation results across different data characteristics.

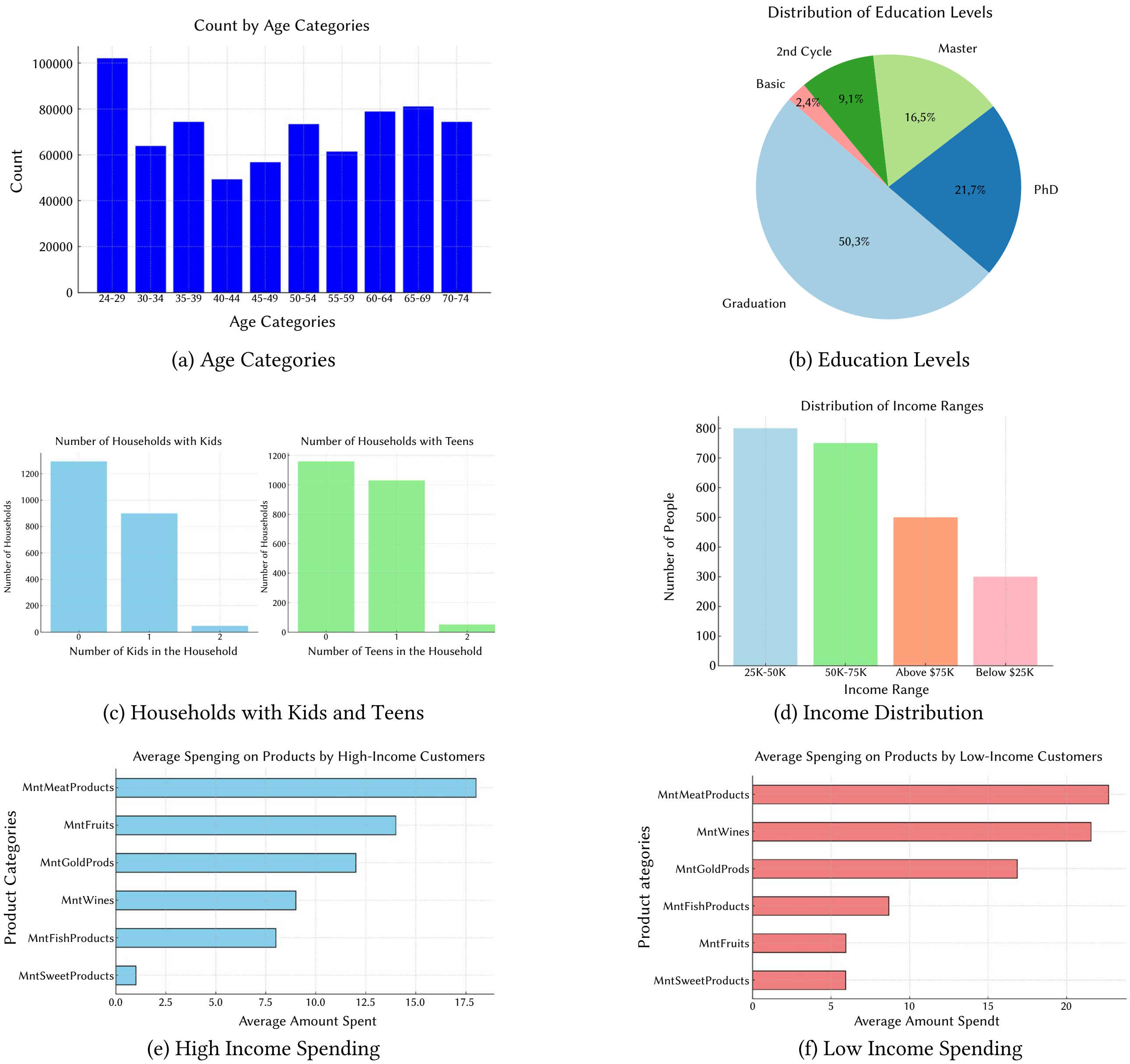
By addressing these aspects, this study aims to provide Turkish retailers to help find out the most suitable customer segmentation techniques for their specific needs and data characteristics. The ultimate goal is to enable more effective targeted marketing strategies, improved customer relationship management, and enhanced business decision-making in the Turkish retail sector.

A. Turkish Market Sales (TMS) Dataset

The first dataset utilized in this study is a comprehensive Turkish Market Sales Dataset, which provides a wealth of information about customer transactions at a local supermarket in Türkiye. This dataset is characterized by its large scale, comprising 10 million rows of transaction data [31].

This dataset is notable for its scale, consisting of 10 million rows of transaction data, which provides a substantial volume for evaluating algorithm performance on large-scale retail datasets. Each row represents an individual customer transaction, offering a detailed view of purchasing behavior at the transaction level. The dataset is also rich in features, encompassing various aspects of customer behavior and transaction characteristics. These features include customer demographics such as age and gender, product information like product category and brand, transaction details including purchase amount, date and time of purchase, payment method, and store location. The temporal aspect of the dataset, marked by the inclusion of transaction dates, enables the analysis of purchasing patterns over time, which is essential for understanding seasonal trends and the customer lifecycle. Additionally, the dataset is multi-dimensional, combining customer, product, and transaction data to provide a comprehensive view of customer behavior, facilitating complex segmentation analyses. Fig. 1 plots some of these insights from the dataset.

The use of this dataset presents several challenges and opportunities. The large scale of the dataset demands efficient data processing and analysis techniques, while the richness of available features requires careful consideration in selecting the most relevant ones for segmentation. The presence of categorical variables, such as product categories and payment methods, requires the application of appropriate pre-processing and encoding techniques.



(a) Age Categories

(b) Education Levels

(c) Households with Kids and Teens

(d) Income Distribution

(e) High Income Spending

(f) Low Income Spending

Fig. 2. Demographic, economic, and consumer behavior analysis on the MC (second) dataset.

The TMS dataset provides a realistic representation of the complexity and scale of data that large retailers in Türkiye might encounter, making it an excellent dataset for evaluating the scalability and effectiveness of different segmentation algorithms in the real-world retail scenario.

B. Marketing Campaign (MC) Dataset

The second dataset utilized in this study is the Marketing Campaign (MC) Dataset, which offers a focused examination of customer responses to various marketing initiatives [32]. Although this dataset is smaller in scale (2240 rows) compared to the Turkish Market Sales Dataset (10M rows), it provides a rich set of features that are particularly relevant for marketing campaign analysis and customer profiling. This dataset is more manageable in size, allowing for a detailed analysis of individual customer attributes and behaviors. Each entry represents an individual customer, giving a holistic view of customer characteristics, including demographics, behavioral data, purchase history, customer value metrics, and campaign response data. The dataset's diversity in features, such as birth year, education level, website visits, accepted deals, and specific purchase histories (e.g., wine and fruits), enriches the analysis. Additionally, it includes derived metrics like recency and customer tenure, which offer deeper insights into customer behavior and value.

The MC Dataset presents unique opportunities and challenges in the context of marketing analysis. Its multi-faceted customer profiles allow for the creation of detailed segments, facilitating more refined customer segmentation. The inclusion of campaign response data is particularly valuable for evaluating segmentation algorithms, enabling an assessment of how well these algorithms can identify customer groups with similar response patterns. The dataset's mix of numerical, categorical, and ordinal data requires careful pre-processing, making the handling of these varied data types a critical aspect of the analysis. Furthermore, the broad range of features necessitates a focused approach to determining feature importance, which is key to effective segmentation. The interaction between demographic factors and behavioral data within the dataset allows for a subtle exploration of how these elements define customer segments.

The MC Dataset complements the broader Turkish Market Sales Dataset by providing a detailed view of individual customers and their interactions with marketing campaigns. It enables the evaluation of segmentation algorithms in a context directly relevant to marketing strategy development and campaign optimization, thus playing a crucial role in this study's analysis. Demographic, economic, and consumer behavior analysis on the MC (second) dataset can be seen in Fig. 2.

While the two datasets provide complementary perspectives on Turkish retail customers, we acknowledge potential selection bias in our dataset selection. The Turkish Market Sales dataset may over-represent urban areas where such data collection is more feasible, while the Marketing Campaign dataset may have self-selection bias from customers who choose to participate in marketing programs. These limitations should be considered when generalizing our findings to the broader Turkish retail market.

IV. Methodology

In this study, a comprehensive methodological approach has been adopted to evaluate and compare four distinct machine learning algorithms for customer segmentation in the context of Turkish retail markets. The selected algorithms (K-Means, Density-Based Spatial Clustering of Applications with Noise (DBSCAN), Gaussian Mixture Models (GMM), and Self-Organizing Maps (SOM)) were applied to two different datasets: a large-scale Turkish market sales dataset and a focused marketing campaign dataset. The selection of these four algorithms was based on their representation of different clustering paradigms commonly used in customer segmentation research. K-means represents centroid-based approaches (the most widely used baseline), DBSCAN represents density-based methods (suitable for non-spherical clusters), GMM represents probabilistic models (capable of handling overlapping clusters), and SOM represents neural network-based approaches (excellent for high-dimensional data visualization).

The methodology encompasses several key stages, including data pre-processing, feature selection, algorithm implementation, and evaluation. Each stage is carefully designed to ensure a rigorous and fair comparison of the algorithms' performance in customer segmentation tasks.

A. Proposed Approach

The proposed approach for this study involves a systematic comparison of four machine learning algorithms for customer segmentation. The primary method for initiating the segmentation process is K-means clustering, which serves as a benchmark against which the performance of other algorithms is measured. The approach can be outlined as follows:

1. ***Implementation of K-means Clustering:***

   The optimal number of clusters is determined using the Elbow method, followed by the application of K-means clustering on the pre-processed data. The resulting clusters are then analyzed and interpreted.

2. ***Application of Alternative Algorithms:***

   DBSCAN, GMM, and SOM are implemented on the same pre-processed data, with appropriate parameter tuning techniques employed for each algorithm.

3. ***Comparative Analysis and Context-Specific Evaluation:***

   The results from all algorithms are compared based on the performance metrics mentioned in subsection IV.G, and the interpretability and actionability of the resulting segments are assessed. Additionally, the computational efficiency and scalability of each algorithm are evaluated. The performance of each algorithm is assessed in the context of the Turkish retail market, and the applicability of the resulting segments to real-world marketing strategies is considered.

The proposed approach is designed to not only identify the most effective algorithm for customer segmentation but also to provide insights into the specific conditions under which each algorithm performs best. This information can be valuable for retailers in selecting the most appropriate segmentation technique based on their specific data characteristics and business objectives.

B. Data Preparation and Preprocessing

Data pre-processing is a crucial step in ensuring the quality and reliability of the customer segmentation results. For both the large-scale Turkish market sales dataset and the focused marketing campaign dataset, the following pre-processing steps were undertaken:

1. ***Data Cleaning***: Missing values were identified and handled appropriately, using techniques such as row deletions for the variables income and age.
2. ***Feature Engineering***: New features were created, such as deriving a "total spending" feature from individual transaction amounts in the sales dataset.
3. ***Encoding of Categorical Variables***: Categorical variables were encoded using appropriate techniques, with one-hot encoding applied to nominal categorical variables and ordinal encoding used for ordinal variables. For high-cardinality categorical variables, techniques such as frequency encoding or target encoding were considered to reduce dimensionality.
4. ***Feature Scaling***: Numerical features were scaled to ensure that all variables contributed equally to the analysis, with standardization (z-score normalization) applied to bring all numerical features to a common scale.
5. ***Data Type Conversion and Dimensionality Reduction***: Data types were converted as necessary to ensure compatibility with the chosen algorithms. For example, categorical variables were converted to numerical types for algorithms that require numerical inputs. Principal Component Analysis (PCA) has been applied to both datasets with n_components=3, to reduce the number of features while retaining most of the information.

C. Implementation of Clustering Algorithm

The core of the segmentation procedure is the k-means clustering algorithm. Based on their similarity, it repeatedly divides data points into a set number of clusters (*k*). Cluster centroids are initialized by the algorithm, either at random or with predetermined values. Our approach initialized clusters randomly and used a predetermined number of clusters obtained with the Elbow method. After that, each data point is assigned to the closest cluster centroid based on the Euclidean distance metric. Specifically, the algorithm assigns each data point to the cluster whose centroid has the minimum Euclidean distance from that point. The centroids are updated by taking the mean of all the data points in that cluster after the data points have been assigned. The centroids are updated and data points are assigned again until convergence is reached, at which point there is no more noticeable movement in the centroids.

K-means clustering served as the primary method for initiating the customer segmentation process. The approach involved (1) determining the optimal number of clusters using the Elbow method and Silhouette analysis,(2) initializing cluster centroids, and (3) assigning data points to the closest cluster centroid. Centroids were then (4) updated by taking the mean of all data points within each cluster, with steps three and four repeated until convergence was reached.

Determining the ideal cluster count (*k*) is essential for significant segmentation. The elbow method and silhouette analysis are two popular techniques.

Elbow method: Plotting the within-cluster sum of squares (WCSS) against the number of clusters (*k*) is the elbow method's method of analysis. The elbow point, where the WCSS begins to drop quickly and then stabilizes, is found to be the ideal number of clusters; this suggests that adding more clusters does not appreciably enhance the clustering result. We calculated WCSS for different k values and visualized it to identify the elbow point manually. Additionally, it leverages the function `KElbowVisualizer` in the library Yellowbrick, to automate the elbow method visualization, aiding in the selection of the most suitable number of clusters.

Silhouette analysis: determines how similar an object is to its own cluster versus other clusters. The silhouette score ranges between -1 and 1, with a higher score indicating that the object is well matched to its own cluster but poorly matched to neighboring clusters. The silhouette coefficient is calculated for each sample by taking the mean intra-cluster distance (a) and the mean nearest-cluster distance (b). A silhouette score close to 1 indicates that the sample is far away from the neighboring clusters; a score of 0 indicates that the sample is on or very close to the decision boundary between two neighboring clusters; and a score of -1 indicates that the samples may have been assigned to the incorrect cluster. We calculated silhouette scores for various cluster counts and visualized them to manually determine the best number of clusters. Furthermore, we use the function `SilhouetteVisualizer` in the library Yellowbrick, to automate the silhouette method visualization, assisting in the selection of the optimal number of clusters. Silhouette Analysis indicates that the optimal value for *k* is 3.

D. Density Based Spatial Clustering of Applications With Noise (DBSCAN)

DBSCAN was implemented by considering two important features of every data point: a distance threshold (`epsilon`) and minimum neighbors (`MinPts`). The algorithm identifies clusters as high-density areas separated by low-density areas.

Consider a landscape with data points strewn all over it. High-density areas, such as busy city centers, are recognized by DBSCAN as clusters, which are made up of points that are close to one another and have lots of neighbors. Noisy areas are those with few data points, such as rural or suburban areas. DBSCAN looks at two important features of every data point. After fitting the model to the reduced-dimension customer data, cluster labels for each data point were extracted from the fitted model's labels attribute. These labels represent cluster membership or -1 for outliers. Then, the total number of clusters created, the number of outliers discovered, and the distribution of points within each cluster were examined.

E. Gaussian Mixture Model

Following the same steps as those employed in K-means clustering, including data preparation and cleaning, the GMM model was implemented. GMM provides granular segmentation by modeling the data as a mixture of Gaussian distributions, where each customer can have probabilistic membership across multiple clusters rather than hard assignments. This allows for more detailed understanding of customer segments, as some customers may exhibit characteristics of multiple segments. The code iterated through a range of cluster sizes, from 1 to 10, applying a GMM model to the data for each cluster size and computing the BIC score, a model selection metric. The number of clusters corresponding to the lowest BIC score was selected, indicating the most appropriate number of clusters for the data according to this metric. Subsequently, a GMM model was trained using the optimal number of clusters.

The algorithm estimates the parameters of these Gaussian distributions using the Expectation-Maximization (EM) algorithm [29], which iteratively refines the cluster assignments and distribution parameters until convergence. The trained model was then utilized to predict cluster labels for each customer data point, effectively assigning each customer to a specific segment based on their attributes. A new column labeled 'Cluster' was added to the data frame to include these cluster labels. The distribution of customers across segments was displayed by counting the occurrences of each cluster. To visualize the data, Principal Component Analysis (PCA) was employed, reducing the dimensionality of the data. Scatter plots were generated, with colors representing the various clusters, allowing for a visual inspection of how customers were classified based on their characteristics.

F. Self-Organizing Maps

The customer data is clustered using Self-Organizing Maps (SOM), an unsupervised learning technique for visualizing and analyzing high-dimensional data. Initially, missing values are handled, and numerical features are scaled to ensure consistency. The data is then converted into a format suitable for SOM analysis. The optimal SOM grid size is determined by comparing quantization errors across various grid sizes, with the appropriate grid size selected based on minimizing the quantization error as indicated by plotted results.

The SOM is trained on the data using the chosen grid size, and each data point is assigned to a cluster based on the winning neuron. The dataset is subsequently labeled into clusters for further analysis. The segmentation results are analyzed by computing the mean values of different attributes within each cluster, providing insights into distinct customer segments.

G. Evaluation Metrics

To assess the performance of the clustering algorithms, several evaluation metrics were employed:

- **Calinski-Harabasz Index**: Compares the ratio of between-cluster variance to the average within-cluster variance.
- **Davies-Bouldin Index:** Measures the ratio of within-cluster scatter to between-cluster separation.

While our study employs the state-of-the-art Calinski-Harabasz and Davies-Bouldin indices, recent advances in clustering validation have introduced new metrics such as the S-Divergence-Based Internal Clustering Validation Index [30], which provides an alternative approach to measuring cluster quality. Future work could benefit from incorporating these newer validation metrics.

V. Findings

The analysis of the Turkish market sales datasets using the four clustering algorithms resulted in several key findings, providing insights into customer segmentation within the Turkish retail market.

Two distinct datasets were employed in this study, each offering unique perspectives on customer behavior:

- A large-scale dataset comprising 1,000,000 rows and 28 columns, providing a comprehensive view of Turkish market sales. This dataset offered a broad spectrum of customer interactions and transactions, allowing for in-depth analysis of purchasing patterns across a wide customer base.
- A more focused marketing campaign dataset of 2240 rows and 29 columns, which, while smaller in scale, provided targeted information on customer responses to specific marketing initiatives. This dataset was particularly valuable for understanding the effectiveness of various marketing strategies and customer engagement levels.

A. Turkish Market Sales Dataset Results - Using K-Means Clustering

Initial analysis of the TMS dataset revealed key demographic distributions as shown in Fig. 1, and the correlation matrix of the dataset can be found in Fig. 3.

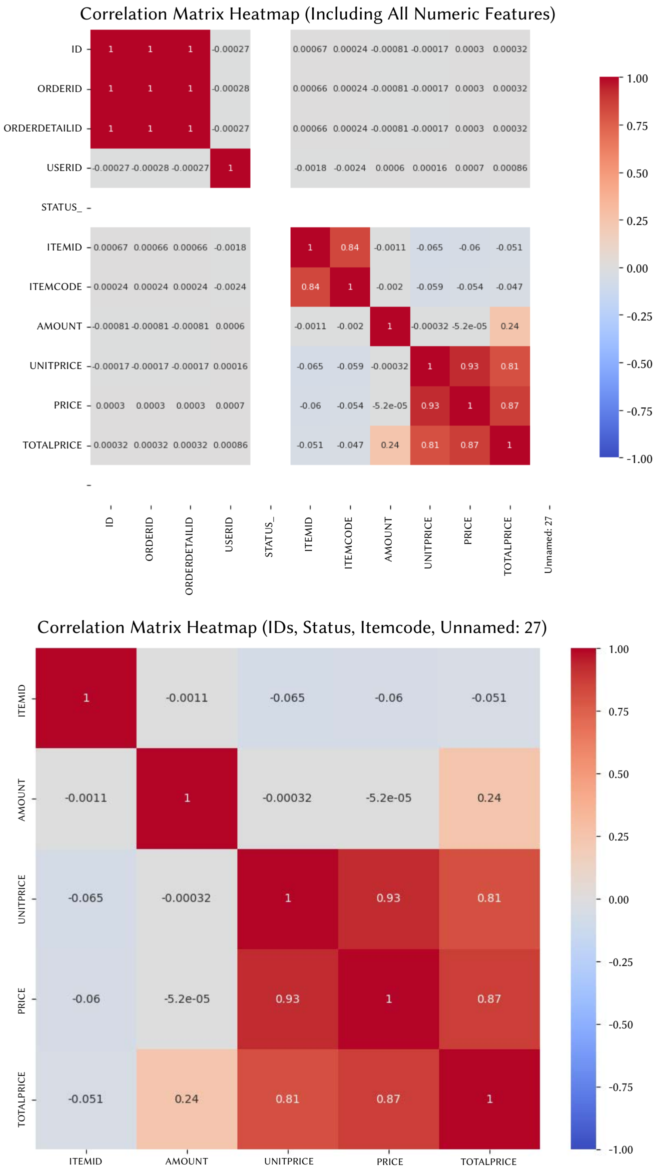


Fig. 3. Heatmap including (a) all numeric features, and (b) relevant features on the Turkish Market Sales Dataset.

The application of K-means clustering to the large-scale dataset revealed several important findings:

- Through the application of the Elbow method, it was determined that the optimal number of clusters for this dataset was 6 as illustrated in Fig. 4. This suggests that the Turkish retail market can be effectively segmented into six distinct customer groups, each with unique characteristics and behaviors (Fig. 5).
- To visualize these clusters, Principal Component Analysis (PCA) was employed. This technique allowed for the reduction of the high-dimensional data into a more manageable form, revealing clear and distinct customer segments. The visualization highlighted the separation between these segments, providing a clear picture of the market structure (Fig. 6).

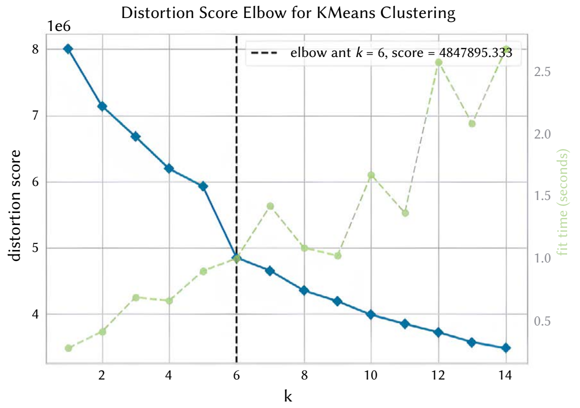


Fig. 4. Optimal cluster count detection using the Elbow method on the Turkish Market Sales Dataset.

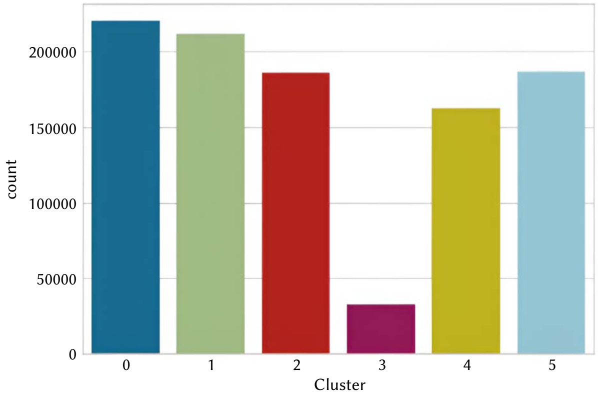


Fig. 5. Cluster analysis on the Turkish Market Sales dataset.

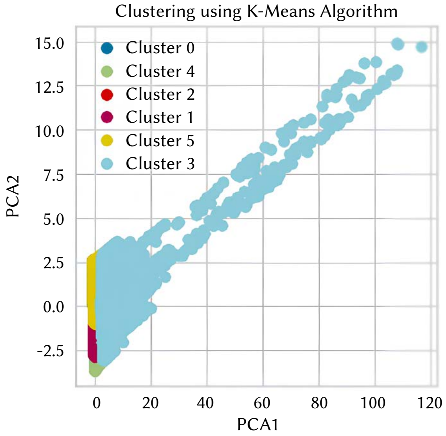


Fig. 6. PCA Dimension Extraction on the Turkish Market Sales dataset.

- The quality of the clustering was assessed using the Davies-Bouldin score, which was calculated to be less than 2 (1.6020). This low score is indicative of well-separated clusters, suggesting that the identified customer segments are distinctly different from one another. This clear separation is crucial for developing targeted marketing strategies for each segment.

Titles, labels, and a legend were added to the plot to ensure clarity and understanding. The visualization, created through the combined efforts of K-Means and PCA, offers valuable insights into the underlying structure of the data. The distribution of data points within each cluster can be observed, revealing potential groupings and unique characteristics. Subsequently, the original data was merged with the cluster labels assigned by K-means, resulting in a new DataFrame that incorporates these cluster labels. This enhancement allows for the analysis of features within each cluster, comparison of characteristics across groups, and a deeper understanding of the data's structure. It effectively tags each data point with a *group membership*, facilitating further exploration.

B. Marketing Campaign Dataset Results

Initial analysis of the Marketing Campaign (MC) dataset revealed key demographic distributions as shown in Fig. 2 and the correlation matrix of the dataset can be found in Fig. 7. The analysis of the focused marketing campaign dataset using various clustering algorithms provided detailed insights into customer segmentation.

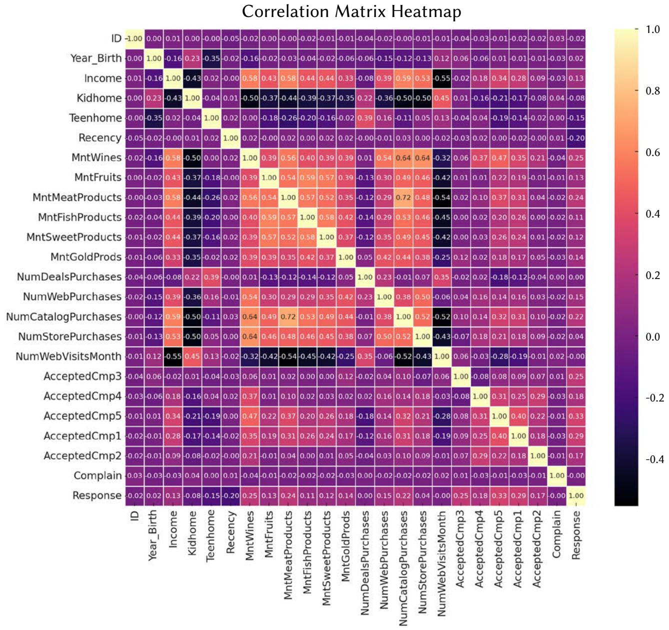
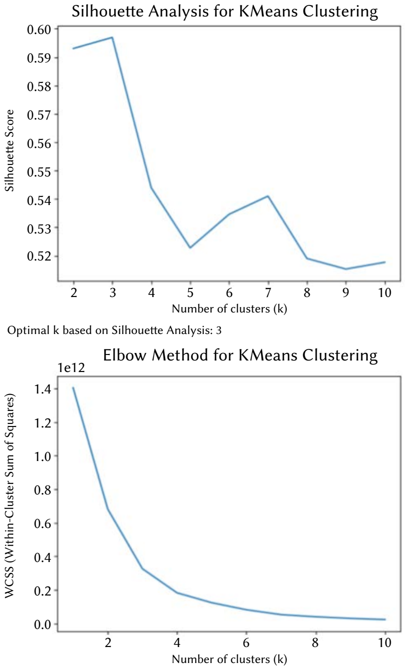


Fig. 7. Correlation matrix of the Marketing Campaign dataset.



Optimal k based on Silhouette Analysis: 3

Fig. 8. Silhouette Analysis and Optimal Cluster Count Detection using the Elbow Method on the Marketing Campaign Dataset.

A Silhouette analysis and the Elbow method were performed, which identified the optimal number of clusters as three. This result assisted in determining the appropriate value for k, as illustrated in Fig. 8.

1. K-Means

The K-means algorithm identified three distinct clusters within the dataset, indicating three primary customer segments in the context of marketing campaign responses, as shown in Fig. 9. Notably, Cluster 1 emerged with the highest customer count, suggesting it as a dominant segment that could be a key target for marketing efforts, as illustrated in Fig. 10. Significant differences were observed across the clusters in terms of income levels, frequency of website purchases, and responsiveness to deal purchases, offering valuable insights for tailoring marketing strategies to the specific preferences and behaviors of each segment.

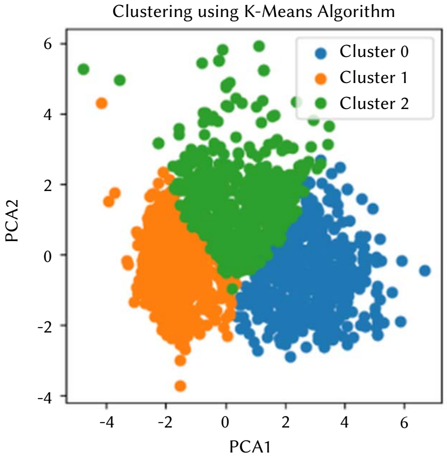


Fig. 9. PCA Dimension Extraction on the Marketing Campaign dataset.

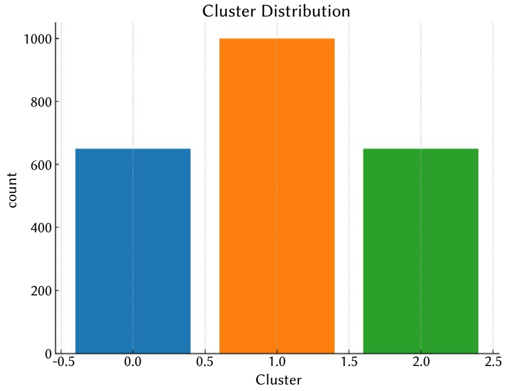
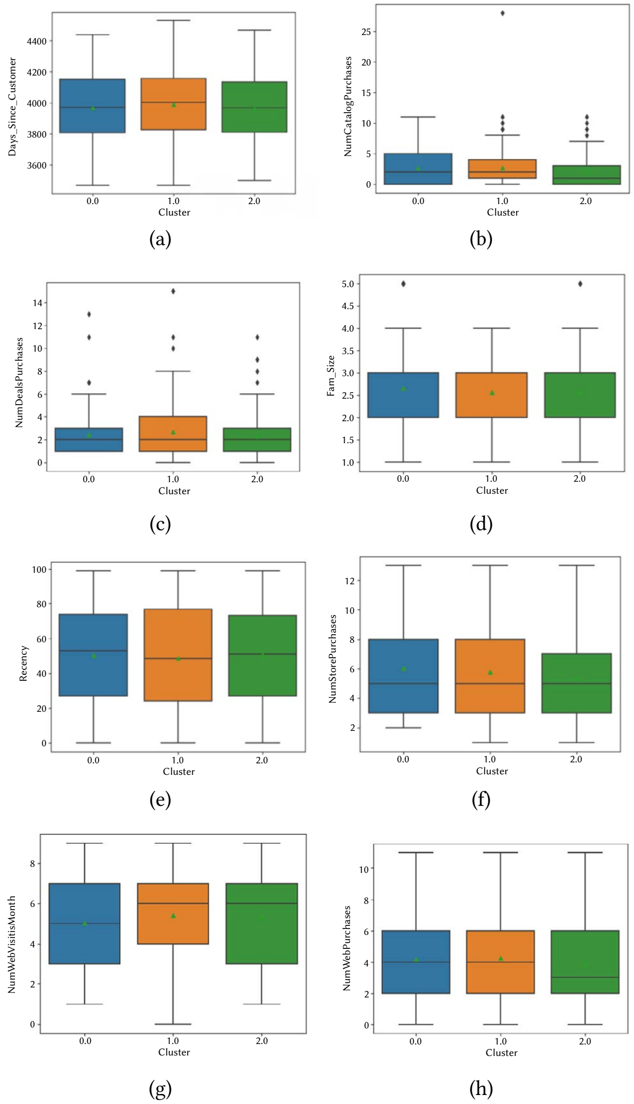


Fig. 10. Cluster analysis on the Marketing Campaign dataset.

Each cluster was visualised according to the results for each feature, as presented in Fig. 11, which provides valuable insights into the characteristics of the clusters. (a) presents the number of days since a member became a customer, identifying Clusters 0 and 2 as the oldest and most loyal customer groups. (b) highlights the number of purchases made through the catalogue, where Cluster 0 exhibited the strongest response, corresponding to the second cluster in the DBSCAN clustering approach Fig. 12. (c) shows the number of purchases made through supermarket deals, indicating that Clusters 1 and 2 are more responsive to deals, according to K-means results. (d) illustrates the family size of customers within each cluster, showing that most clusters consist of families with 2 or 3 members, indicating that family size was not a significant differentiating factor for clustering. (e) demonstrates the recency of purchases, with cluster 1 having the highest recency. (f) shows the number of in-store purchases, with Clusters 0 and 1 showing the highest purchase rates, emphasising the distinct purchasing behaviours of each cluster as identified by K-means. (g) displays the number of website visits, which corresponds to Cluster 9 in the GMM method shown in Fig. 13, providing insights into the patterns of website usage across clusters, which could inform future marketing strategies. (h) illustrates the number of purchases made through the website, revealing that Clusters 0 and 1 purchase from the website more frequently than Cluster 2, a pattern that is consistent with the K-means method but not clearly defined in GMM V.B.3 and SOM V.B.4 clustering algorithms.



(a)

(b)

(c)

(d)

(e)

(f)

(g)

(h)

Fig. 11. Box-plot analyses on the Marketing Campaign dataset. Details are in section 1, under K-means.

2. DBSCAN

The DBSCAN algorithm, known for its ability to identify clusters of arbitrary shape, revealed three main clusters and additionally identified outliers. This suggests the presence of niche customer groups that might be overlooked by other methods. Across these clusters, notable variations were observed in spending patterns, family size, and responsiveness to marketing campaigns. These insights offer a more detailed and subtle understanding of customer behavior, potentially uncovering unique market segments.

DBSCAN clustering results across the Marketing Campaign dataset can be found in Fig. 12. DBSCAN produced a relatively better Calinski-Harabasz score compared to other methods, suggesting a favorable balance between cluster density and separation.

The customer clusters were analyzed across key attributes, including spending patterns, family size, and purchasing behaviors. Total spending, shown in (a), reveals expenditure differences, with Cluster 0 as the highest spender. Family size (b), shows that most clusters have 2 to 3 members, with Cluster 0 having the largest families, indicating DBSCAN's consideration of this factor. Deal responsiveness in (c) highlights Cluster 2 as the most responsive, while Cluster 0 is the least, demonstrating the influence of deal purchases on clustering.

Store purchases, illustrated in (d), are highest in Cluster 2, consistent with GMM V.B.3 and SOM V.B.4 results. Website visits, shown in (e), indicate minimal differences, with Cluster 0 leading. Catalog purchases, in (f), also peak in Cluster 0, aligning with DBSCAN's focus on purchasing behavior.

Website purchases, shown in (g), place Cluster 2 as active online spenders. Age distribution in (h) shows Cluster 2 as middle-aged, while Cluster 1 is older. Cluster 2 also leads in accepted campaigns (i), reflecting their engagement. Income distribution, shown in (j), identifies Cluster 0 with the lowest income. Recency of purchases, in (k), shows Cluster 1 as the most recent buyers, and tenure (l), suggests Cluster 2 has the longest customer relationship. DBSCAN's clustering primarily focused on purchasing behavior while only slightly considering attributes like tenure.

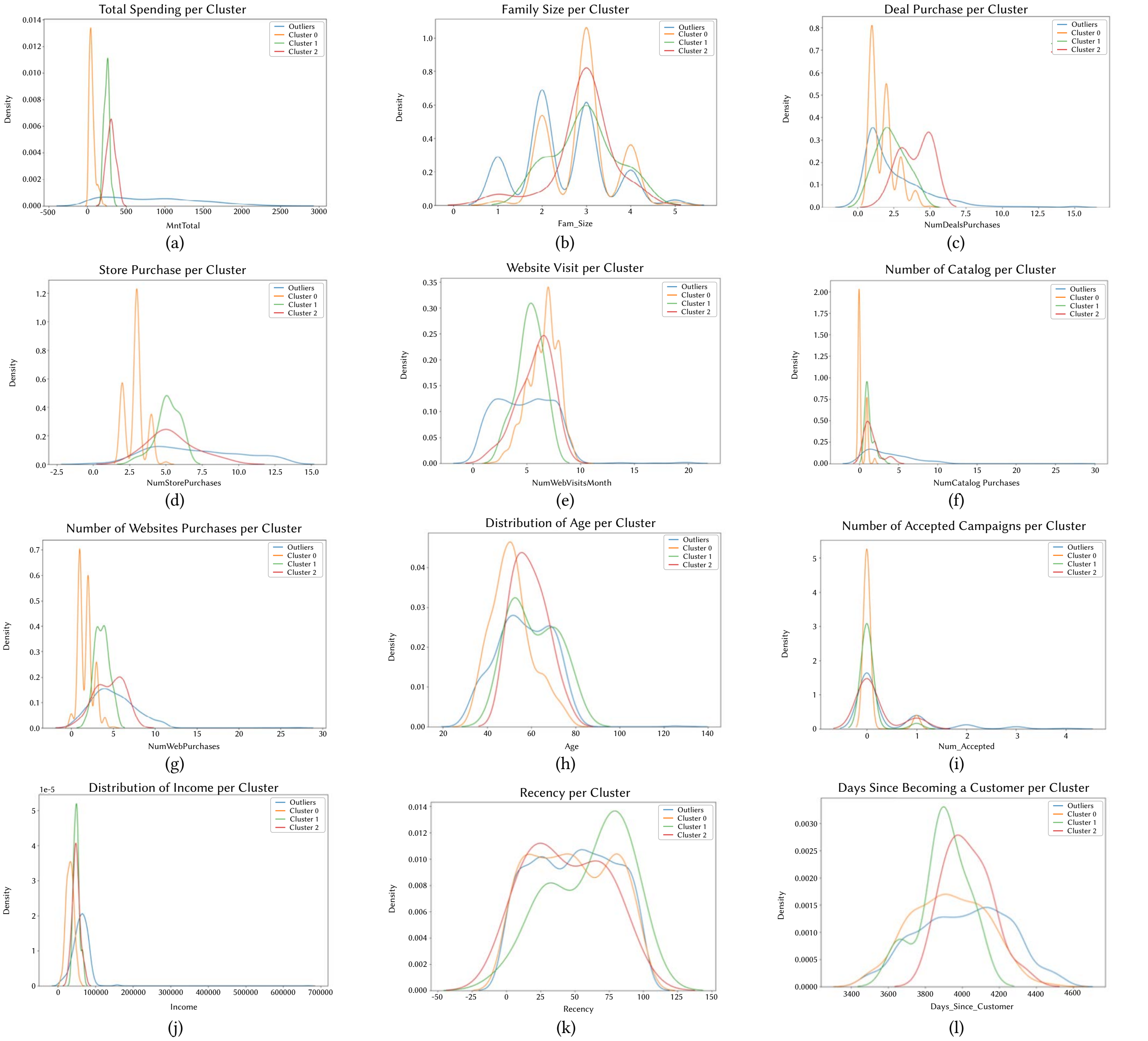
3. GMM

The Gaussian Mixture Model approach identified 10 distinct clusters, providing a more granular segmentation of the customer base. Each cluster exhibited varying characteristics in terms of spending habits, family size, and purchasing behavior, offering a highly detailed view of customer segments. Fig. 13 contains the results of each cluster analysis using the GMM clustering method according to the features.

The analysis and visualization of the ten customer clusters revealed distinct spending patterns, demographic characteristics, and purchasing behaviors. Cluster 6 showed the highest spending rate, identifying its members as the most active spenders, though spending varied within the cluster (a). Family sizes ranged mainly between 2 to 3 members, with Cluster 0 having larger families averaging 4 members, while Cluster 7 averaged 2 members (b). Cluster 0 also had the highest deal purchase response, marking these customers as the most receptive to promotions (c). Store purchases were highest in Cluster 1, while Cluster 7 had the least activity, including a group with zero store purchases (d). Cluster 0 led in website visits, aligning with its high deal purchase rate (e), and catalog purchases were strong in Clusters 3 and 6 (f). Cluster 1 dominated in website purchases, highlighting GMM's effectiveness in defining clusters based on purchase behaviors (g). Age distribution showed Cluster 2 as the oldest group, while Clusters 9 and 7 were the youngest (h). Campaign acceptance was highest in Cluster 6 (i), and Cluster 2 had the highest income, making it the wealthiest group (j). Cluster 4 had the most recent purchases (k), and Cluster 0 had the longest customer tenure (l), marking them as the oldest customer segment within the supermarket.

4. SOM

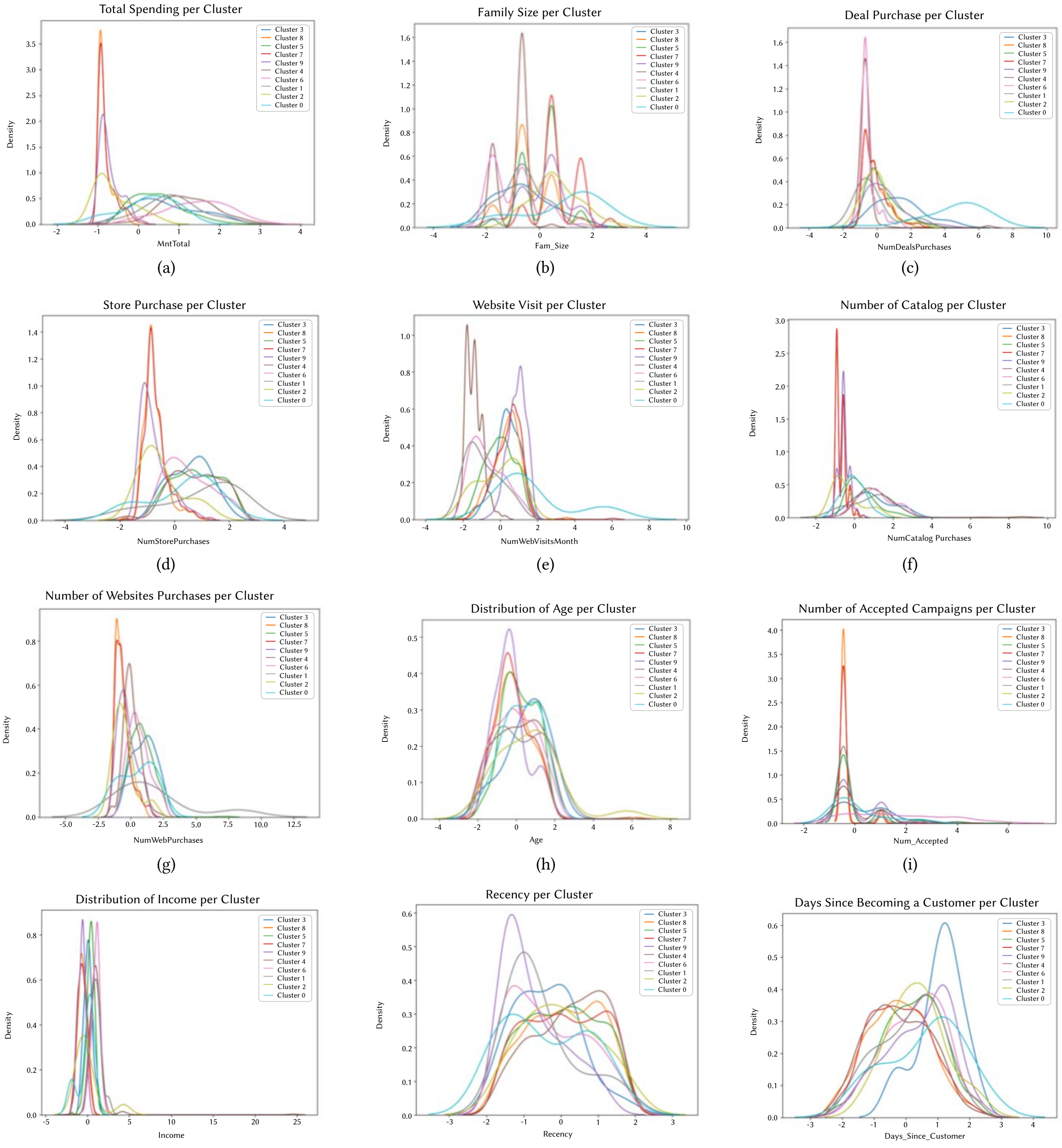
The Self-Organizing Maps technique resulted in the identification of 17 clusters, the highest number among all methods used. Each of these clusters presented unique characteristics based on factors such as age, income, and purchasing patterns. While this high number of clusters provides extremely detailed segmentation, it may present challenges in terms of practical application in marketing strategies. Cluster 1: Customers in this cluster tend to have an average family



(a) (b) (c) (d) (e) (f) (g) (h) (i) (j) (k) (l)

Fig. 12. DBSCAN clustering results across the Marketing Campaign dataset. Details are in section V.B.2, under DBSCAN.

size of approximately three members and are older, with an average age of around 56. They have a relatively low average income and a low total spending on products. These customers often make infrequent purchases, especially of meat and fish products, and are not very responsive to marketing campaigns. Cluster 2: These customers typically have medium incomes and moderate spending habits. They are mostly middle-aged, around 47 years old, with an average family size of about three members. Their purchase frequency is moderate, particularly for wines and sweets, and they show limited engagement with marketing efforts. They visit web stores quite often and are relatively consistent in their purchasing patterns. Cluster 3: This group has a slightly higher average age of 50 and consists of families with approximately three members. They have moderate incomes and spending levels, especially on wine, meat, and sweets. These customers show moderate responsiveness to marketing campaigns and have a balanced approach to both online and in-store purchases. Cluster 4: comprises smaller families, often single individuals or couples, with the lowest average income among the clusters. These customers are relatively young, around 41 years old, and have minimal spending, particularly on non-essential items like sweets and gold products. They show low engagement with marketing campaigns and visit online stores moderately. Cluster 5: With an average family size of nearly three, these customers have medium incomes and spending habits. They are generally middle-aged, around 50 years old, and exhibit moderate purchasing patterns, especially for wine and sweets. Their responsiveness to marketing campaigns is average, and they balance their shopping between online and physical stores. Cluster 6: This cluster consists of slightly larger families, around three members, with higher incomes and spending, particularly on wine and gold products. These customers are typically older, averaging 54 years in age, and show moderate engagement with marketing campaigns. They make frequent purchases both online and in-store, reflecting their active shopping behavior. Cluster 7: Customers in this cluster are older, averaging 58 years, with nearly three family members. They have high incomes and significant spending, particularly on wine, meat, and gold products. Their responsiveness to marketing campaigns is higher than



(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

(j) (k) (l)

Fig. 13. GMM clustering results across the Marketing Campaign dataset. Details are in section V.B. 3, under GMM.

average, and they frequently shop both online and in-store, making them highly valuable customers. Cluster 8: These customers, averaging around 52 years old, have a slightly larger family size of about three members. Their income and spending levels are moderate, with a focus on wine and gold products. They show limited responsiveness to marketing campaigns but are consistent in their purchasing patterns, both online and in physical stores. Cluster 9: This cluster consists of families with approximately two to three members, averaging 45 years old. They have moderate incomes and spending, particularly on wine and gold products. These customers show a low engagement with marketing campaigns and have a balanced approach to online and in-store shopping. Cluster 10: Customers in this cluster are older, around 58 years, with moderate family sizes. They have high incomes and spend significantly, particularly on meat and fish products. Their engagement with marketing campaigns is above average, and they frequently shop both online and in physical stores, reflecting their active consumer behavior. Cluster 11: This group has high-income customers, typically around 54 years old, with smaller families. They exhibit high spending, especially on wine and gold products, and show moderate responsiveness to marketing campaigns. These customers visit online stores frequently and have a consistent purchasing pattern. Cluster 12: Customers in this cluster are older, averaging 57 years,

TABLE I. Comparison of Clustering Methods

Method	Dataset	Clusters	Calinski-Harabasz	Davies-Bouldin
K-means	Turkish Market Sales	6	–	1.60
K-means	Marketing Campaign	3	617.33	1.85
DBSCAN	Marketing Campaign	3	302.34	1.26
GMM	Marketing Campaign	10	184.92	2.29
SOM	Marketing Campaign	17	611.83	0.63

TABLE II. Summary of Clustering Results Across Algorithms (SP: Spending Patterns, FS: Family Size, PB: Purchasing Behaviors, AD: Age Distribution, MR: Marketing Responsiveness)

Attribute	K-means	DBSCAN	GMM	SOM
SP	High: Cluster 0 Medium: Cluster 1 Low: Cluster 2	High: Cluster 0 Medium: Cluster 2 Low: Cluster 1	High: Cluster 6 Medium: Clusters 1, 3, 5 Low: Clusters 0, 4, 7	High: Clusters 6, 7, 10, 12, 14 Medium: Clusters 2, 3, 5, 8, 9 Low: Clusters 1, 4
FS	Large: Cluster 1 Medium: Cluster 2 Small: Cluster 0	Large: Cluster 0 Medium: Cluster 2 Small: Outliers	Large: Cluster 0 Medium: Most clusters Small: Cluster 7	Large: Clusters 13, 15, 17 Medium: Most clusters Small: Clusters 4, 11, 14
PB	Online: N/A In-store: Clusters 0, 1 Catalog: Cluster 0	Online: Cluster 1 In-store: Cluster 2 Catalog: Cluster 1	Online: Cluster 1 In-store: Cluster 1 Catalog: Clusters 3, 6	Online: Clusters 2, 7, 12, 14 In-store: Clusters 3, 6, 10 Catalog: Varied
AD	Oldest: Cluster 2 (39) Middle: Cluster 1 (38) Youngest: Cluster 0 (32)	Oldest: Cluster 2 Middle: Cluster 1 Youngest: N/A	Oldest: Cluster 2 Middle: Most clusters Youngest: Clusters 7, 9	Oldest: Cluster 15 (60) Middle: Most clusters Youngest: Cluster 4 (41)
MR	High: Cluster 0 Medium: Cluster 1 Low: Cluster 2	High: Cluster 1 Medium: Outliers Low: Clusters 0, 2	High: Cluster 6 Medium: Clusters 3, 5, 8 Low: Clusters 0, 4, 7	High: Clusters 7, 12, 14 Medium: Clusters 3, 5, 8, 13 Low: Clusters 1, 4, 16

with smaller family sizes. They have high incomes and substantial spending, particularly on wine and gold products. They are highly responsive to marketing campaigns and exhibit frequent shopping behavior both online and in physical stores, making them highly valuable. Cluster 13: These customers are older, averaging around 58 years, with larger family sizes. They have moderate incomes and spending levels, particularly on wine and sweets. Their engagement with marketing campaigns is average, and they balance their shopping between online and physical stores. Cluster 14: This cluster consists of relatively younger customers, around 56 years old, with smaller family sizes. They have high incomes and significant spending, especially on wine, meat, and gold products. Their responsiveness to marketing campaigns is very high, and they frequently shop both online and in-store, making them among the most valuable customers. Cluster 15: Customers in this cluster are older, averaging 60 years, with larger family sizes. They have moderate incomes and spending levels, particularly on wine and sweets. Their engagement with marketing campaigns is average, and they balance their shopping between online and physical stores. Cluster 16: This group consists of middle-aged customers, around 49 years old, with medium family sizes. They have moderate incomes and spending levels, particularly on wine and sweets. Their responsiveness to marketing campaigns is low, and they show consistent purchasing patterns, both online and in physical stores. Cluster 17: These customers are older, around 53 years, with larger family sizes. They have moderate incomes and spending levels, particularly on wine and gold products. Their engagement with marketing campaigns is above average, and they balance their shopping between online and physical stores.

C. Comparative Analysis of Clusters

The application of four distinct clustering algorithms to the Marketing Campaign dataset revealed unique insights into customer segmentation:

- **K-means**: Provided a clear, income-based segmentation with three distinct clusters.
  - Uniquely identified a high-income, young (average 32 years) customer segment with high marketing responsiveness.
  - Revealed an inverse relationship between income and recency of purchases.
  - Emphasized the importance of purchasing behavior in defining customer segments.
  - Highlighted the correlation between high spending, frequent website visits, and marketing campaign acceptance.
- **DBSCAN**: Excelled in identifying outliers and non-spherical clusters.
  - Uncovered a distinct group of moderate-income, highly engaged customers (Cluster 2).
  - Uniquely categorized customers with variable spending patterns as outliers, potentially identifying niche market segments.
- **Gaussian Mixture Model (GMM)**: Provided the most granular segmentation with 10 distinct clusters.
  - Revealed subtle variations in customer behavior, particularly in the high-income segments.
  - Identified a unique cluster (Cluster 0) combining large family size, high deal responsiveness, and frequent website visits.
- **Self-Organizing Maps (SOM)**: Offered the most detailed age-based segmentation with 17 clusters.
  - Provided nuanced insights in age and consumer behavior (Cluster 6) as described in Section V.4.
  - Uniquely identified several high-value, older customer segments with distinct purchasing preferences.

The trade-off between interpretability and complexity is evident across our results. K-means provides straightforward, easily interpretable segments ideal for immediate business application, as evidenced by its clear three-cluster structure that retail managers can readily understand and act upon. DBSCAN maintains reasonable interpretability while adding the capability to identify outliers, offering a balance between simplicity and advanced clustering capabilities. In contrast, GMM (10 clusters) and SOM (17 clusters) offer significantly more granular segmentation but require additional analytical expertise to translate into actionable strategies. This increasing complexity allows for more nuanced understanding of customer behavior but may challenge practical implementation in retail environments where quick decision-making is essential.

Cross-algorithm comparisons revealed several key insights:

1. Income and age consistently emerged as primary factors in customer segmentation across all algorithms.
2. The inverse relationship between income and purchase recency was a common finding, particularly evident in K-means and DBSCAN results.
3. While K-means provided a broad overview with three clusters, GMM and SOM offered more granular insights, potentially useful for highly targeted marketing strategies.
4. DBSCAN's ability to identify outliers provided unique insights into niche customer groups that other algorithms might have overlooked.

Table I displays the results of various clustering methods applied to 2 different datasets, evaluating their performance based on the number of clusters identified, Calinski-Harabasz Index, and Davies-Bouldin score. The performance evaluation, using metrics such as the Calinski-Harabasz Index and the Davies-Bouldin Index, indicated that K-Means achieved the highest scores. Although GMM and SOM also yielded respectable scores, the highest CH score was achieved by K-Means, affirming its effectiveness for the datasets and objectives of this study.

Table II summarizes the clustering result across algorithms in a standardized notion. The composite attribute definitions used in the table are as follows:

- **SP: Spending Patterns**
  - **High**: Customers who spend the most money (high-value customers)
  - **Medium**: Customers with moderate spending levels
  - **Low**: Customers who spend the least (low-value customers)
- **FS: Family Size**
  - **Large**: Customers with big families
  - **Medium**: Customers with average-sized families (typically 2-3 members)
  - **Small**: Customers with small families or single-person households
- **PB: Purchasing Behaviors**
  - **Online**: Customers who prefer to shop through websites/ online platforms
  - **In-store**: Customers who prefer to shop at physical store locations
  - **Catalog**: Customers who prefer to shop through catalogs (mail-order)
- **AD: Age Distribution**
  - **Oldest**: The older customer segments
  - **Middle**: Middle-aged customer segments
  - **Youngest**: The younger customer segments
- **MR: Marketing Responsiveness**
  - **High**: Customers who frequently respond to marketing campaigns, deals, and promotions
  - **Medium**: Customers with moderate response to marketing efforts
  - **Low**: Customers who rarely respond to marketing campaigns or promotions

VI. Conclusion and Future Work

This study conducted a comprehensive comparative analysis of four machine learning algorithms (K-means, DBSCAN, GMM, SOM) for customer segmentation in the Turkish retail market. Using two distinct datasets, a large-scale Turkish market sales dataset and a focused marketing campaign dataset, this research aimed to identify the most effective and actionable customer segmentation techniques for the unique characteristics of the Turkish retail sector.

K-means demonstrated the most robust performance, offering a balance between interpretability and statistical validity. DBSCAN showed strengths in identifying non-spherical clusters and handling outliers, while GMM and SOM provided more granular segmentation at the cost of increased complexity.

These findings have shown significant implications for Turkish retailers, enabling more targeted marketing strategies and improved customer relationship management. However, the study's limitations, including its focus on specific datasets, suggest caution in generalizing results.

An important consideration for retailers is the trade-off between model interpretability and complexity. Our findings demonstrate that while simpler algorithms like K-means offer highly interpretable results that can be readily implemented by marketing teams, more complex methods such as GMM and SOM provide deeper insights that may require specialized expertise to leverage effectively. Organizations must balance their need for sophisticated customer understanding against their capacity to interpret and act upon complex segmentation results.

Future work should explore the potential of deep learning techniques and hybrid models that combine traditional clustering approaches with neural networks, which could provide more sophisticated pattern recognition and potentially uncover complex, non-linear relationships in customer behavior data. These advanced approaches might include autoencoders for dimensionality reduction, deep clustering methods, or ensemble approaches that leverage the strengths of multiple algorithms.

Additionally, future research should explore the application of these algorithms across diverse retail sectors in Turkiye, investigate the long-term effectiveness of resulting marketing strategies, and examine how Turkish cultural norms, regional differences, and consumer behavior patterns influence segmentation strategies by analyzing how factors such as traditional shopping habits, family structures, and regional economic differences affect the interpretation and application of clustering results.

References

[1] P. Sharma, The ultimate guide to K-means clustering: Definition, methods and applications. Retrieved from https://www.analyticsvidhya.com/blog/2019/08/comprehensive-guide-k-means-clustering/. Accessed on March, 2024.

[2] S. P. Nguyen, "Deep customer segmentation with applications to a Vietnamese supermarkets' data," *Soft Computtng*, vol. 25, no. 12, pp. 7785-7793, 2021.

[3] İ.Kabasakal, "Customer segmentation based on recency frequency monetary model: A case study in E-retailing," *Bilişim Teknolojileri Dergisi*, vol. 13, no. 1, pp. 47-56, 2020.

[4] G. Armstrong and P. Kotler, *Marketing: an introduction*, Pearson Educación, 2003.

[5] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, A. Galstyan, "A survey on bias and fairness in machine learning," *ACM computing surveys (CSUR)*, vol. 54, no. 6, pp. 1-35, 2021.

[6] T. Kansal, S. Bahuguna, V. Singh, and T. Choudhury, "Customer segmentation using K-means clustering," *2018 international conference on computational techniques, electronics and mechanical systems (CTEMS)*, pp. 135-139, IEEE, 2018.

[7] G. Mohit, *Customer Segmentation using Machine Learning applied to Banking Industry* (Doctoral dissertation, Hochschule Neu Ulm), 2023.

[8] D. Zakrzewska and J. Murlewski, "Clustering algorithms for bank customer segmentation," In *5th International Conference on Intelligent Systems Design and Applications (ISDA'05)*, pp. 197-202, IEEE, 2005.

[9] C. Rungruang, P. Riyapan, A. Intarasit, K. Chuarkham, and J.Muangprathub, "RFM model customer segmentation based on hierarchical approach using FCA," *Expert Systems with Applications*, vol. 237, 121449, 2024.

[10] 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, 7243, 2022.

[11] A. Ashabi, S. B. Sahibuddin, and M. Salkhordeh Haghighi, "The systematic review of K-means clustering algorithm," In *Proceedings of the 2020 9th International Conference on Networks, Communication and Computing*, pp. 13-18, 2020.

[12] M. Ester, H. P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters," in large spatial databases with noise. In *kdd*, Vol. 96, No. 34, pp. 226-231, 1996.

[13] V. Kachroo, "Customer segmentation and profiling for e-commerce using DBSCAN and fuzzy C-means," *Proceedings on Engineering*, vol. 5, no. 3, pp. 539-544, 2023.

[14] E. A. Laksana, and M. M. Fahrezi, "Customer segmentation and analysis based on Gaussian mixture model algorithm," In *Widyatama International Conference on Engineering 2024 (WICOENG 2024)*, pp. 67-75, Atlantis Press, 2024.

[15] D. Birant and A. Kut, "ST-DBSCAN: An algorithm for clustering spatial-temporal data," *Data & knowledge engineering*, vol 60, no. 1, pp. 208-221, 2007.

[16] E. Schubert, J. Sander, M. Ester, H. P. Kriegel, X. Xu, "DBSCAN revisited, revisited: why and how you should (still) use DBSCAN," *ACM Transactions on Database Systems (TODS)*, vol. 42, no. 3, pp. 1-21, 2017.

[17] L. L. Scientific, "Data segmentation using mixture regression models with generalized Gaussian distribution and K-means," *Journal of Theoretical and Applied Information Technology*, vol. 103, no. 8, 2025.

[18] E. A. Laksana and M. M. Fahrezi, "Customer segmentation and analysis based on Gaussian mixture model algorithm," In *Widyatama International Conference on Engineering 2024 (WICOENG 2024)*, pp. 67-75, Atlantis Press, 2024.

[19] M. A. Camilleri and, M. A. Camilleri, *Market segmentation, targeting and positioning*, pp. 69-83, Springer International Publishing, 2018.

[20] S. M. Lundberg and S. I. Lee, "A unified approach to interpreting model predictions," *Advances in neural information processing systems*, 30, 2017.

[21] G. J. McLachlan, S. X. Lee, and S. I. Rathnayake, Finite mixture models. *Annual review of statistics and its application*, vol. 6, no. 1, pp. 355-378, 2019.

[22] F. Saadi, B. Atmani, and F. Henni, "Improving retrieval performance of case based reasoning systems by fuzzy clustering," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 9, no. 1, pp. 84-91, 2024.

[23] Y. C. Liu, M. Liu, and X. L. Wang, "Application of Self- Organizing Maps in text clustering," *Applications of Self- Organizing Maps*, 205, 2012.

[24] D. Barman and N. Chowdhury, "A novel approach for the customer segmentation using clustering through self-organizing map," *International Journal of Business Analytics (IJBAN)*, vol. 6, no. 2, pp. 23-45, 2019.

[25] R. Vohra, J. Pahareeya, A. Hussain, F. Ghali, and A. Lui, "Using self organizing maps and K means clustering based on RFM model for customer segmentation in the online retail business," In *Intelligent Computing Methodologies: 16th International Conference, ICIC, Bari, Italy, Proceedings*, Part III 16, pp. 484-497, Springer International Publishing, 2020.

[26] S. Üstebey, İ. Yelmen, and M. Zontul, "Customer segmentation based on self-organizing maps: a case study on airline passengers," *Havacılık Ve Uzay Teknolojileri Dergisi*, 2020.

[27] T. Kohonen, "The self-organizing map," *Proceedings of the IEEE*, vol. 78, no. 9, pp. 1464-1480, 1990.

[28] I. Valova, G. Georgiev, N. Gueorguieva, and J. Olson, "Initialization issues in self-organizing maps," *Procedia Computer Science*, vol. 20, pp. 52-57, 2013.

[29] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the royal statistical society: series B (methodological)*, vol. 39, no. 1, pp. 1-22, 1977.

[30] K. Kumar Sharma, A. Seal, A. Yazidi, and O. Krejcar, "S- divergence-based internal clustering validation index," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 8, no. 4, pp. 127-139, 2023.

[31] O. Colakoglu, 10 Million Rows Turkish Market Sales Dataset (MSSQL). Retrieved from https://www.kaggle.com/datasets/omercolakoglu/10million-rows-turkish-market-sales-dataset. Accessed: March 12, 2024.

[32] R. Saldanha, Marketing Campaign. Retrieved from https://www.kaggle.com/datasets/rodsaldanha/ arketing-campaign. Accessed on March, 2024.

Nur Diyabi

She received a B.Sc. degree in computer engineering with a minor in strategic public relations management from Bahcesehir University in Türkiye in 2022 and M.Sc. degree in big data analytics from Bahcesehir University in Türkiye, in 2024. She is currently working as a Security Engineer at Paramount Computer Systems in UAE.

Duygu Çakır

She was born in Istanbul, Turkiye. She received her BSc, MSc, and PhD degrees in computer engineering department in Bahcesehir University (BAU), Turkiye. She also finished her under-grad level double major in mathematics and computer sciences. From 2007 to 2012, she was a Research Assistant in computer and software engineering departments respectively. Since 2012, she has been working in the department of Software Engineering. She worked in government projects, managed many software and artificial intelligence related projects, and has experience in introductory and advanced programming, data structures, computer graphics, and data science courses in Istanbul as a full time Assistant Professor, in Berlin-Germany and Jelgava-Latvia as a visiting professor. Her research interests include facial action unit and facial expression analysis as well as automated machine learning in computer vision and she holds a national patent on eye tracking in mobile devices, another national patent (pending) on generating virtual agents by converting the input voice directly to a non-existing synthetic face.

Ömer Melih Gül

He received BSc., MSc., and PhD. degrees from the Department of Electrical and Electronics Engineering at Middle East Technical University (METU), Ankara, Türkiye, in 2012, 2014, and 2020, respectively by also working as a research assistant at the same department. His research interests include AI/machine learning applications, wireless security, networking, scheduling, IoT, UAV, robotics, and blockchain. He has co-authored over 50 papers and 4 book chapters. He was awarded third place in the 2019 Lance Stafford Larson Outstanding Student Paper Award by the IEEE Computer Society. He was also awarded third place in the poster competition at 2021 IEEE Rising Stars Global Conference. In 2022, he worked as a postdoctoral fellow at School of Electrical Engineering and Computer Science at University of Ottawa, Canada. He is a recipient of the best paper award at 48th Wireless World Research Forum (WWRF) in 2022. In 2023, he worked as an Assistant Professor in the Department of Computer Engineering at Bahcesehir University, Istanbul, Türkiye, where he supervised 4 MSc theses and co- supervised 1 thesis. Since March 2024, he has been working as an associate professor in the Informatics Institute at Istanbul Technical University (ITU), Istanbul, Türkiye, where he is supervising 1 PhD and 4 MSc students. He serves as an Editor in IEEE Open Journal of Computer Society, (Elsevier) Sustainable Computing: Informatics and Systems, (Springer) Telecommunication Systems, Wireless Networks, Cluster Computing and also the International Journal of Interactive Multimedia and Artificial Intelligence. As cochair, he organized CIEAI workshop at IEEE Fog and Mobile Edge Computing (FMEC) 2023 in Estonia. Moreover, he became the Publicity Chair in the IEEE iThings 2024. He organized two editions of EAI International Conference on Robotic Sensor Networks (ROSENET 2023, ROSENET 2024) as general chair.

Tevfik Aytekin

He is an Associate Professor in the Department Computer Engineering at Bahçeşehir University, in İstanbul, Turkey. He received his Ph.D. in Cognitive Science from Middle East Technical University, M.Sc. in Computer Science from Hacettepe University, and B.Sc. in Computer Science from Bilkent University. His current research interests include data mining, machine learning, and recommender systems. In addition to his academic work, he actively consults on AI/ML projects across various industries.

Seifedine Kadry

He has a bachelor's degree in 1999 from Lebanese University, MS degree in 2002 from Reims University (France) and EPFL (Lausanne), PhD in 2007 from Blaise Pascal University (France), HDR degree in 2017 from Rouen University. He is a Full Professor of Data Science at Lebanese American University, Lebanon. At present, his research focuses on Data Science, education using technology, system prognostics, stochastic systems, and applied mathematics. He is an ABET program evaluator for Computing and an ABET program evaluator for Engineering Tech. He is a Fellow of IET, Fellow of IETE, and Fellow of IACSIT. He was a distinguished speaker of the IEEE Computer Society.