Dokuz Eylül Üniversitesi Mühendislik Fakültesi Fen ve Mühendislik Dergisi
Dokuz Eylul University Faculty of Engineering Journal of Science and Engineering
Elektronik/Online ISSN: 2547-958X

RESEARCH ARTICLE / ARAŞTIRMA MAKALESI

A Deep Dive Into Customer Segmentation Through Advanced Data Mining Techniques

Gelişmiş Veri Madenciliği Teknikleri Yoluyla Müşteri Segmentasyonuna Derin Bir Bakış

Vahid Sinap

Ufuk University, Faculty of Economics and Administrative Sciences, Department of Management Information Systems, Ankara, TÜRKİYE
Corresponding Author / Sorumlu Yazar : vahidsinap@gmail.com

Abstract

This study examines how data mining techniques are used to segment customers to reveal complex customer profiles in a grocery store's database. Customer segmentation is crucial to effectively tailor marketing strategies. This procedure makes it easier to create customized customer profiles, making it possible to create more targeted and effective marketing campaigns. The dataset used in the study was obtained from the database of a well-known grocery company and contains 2.240 data points with 29 different features. These features are grouped into four categories: customer demographics, product information, purchase channels and promotional response data. The study attempts to identify meaningful patterns and groupings among customers using advanced clustering techniques such as K-Means Clustering and Agglomerative Clustering. Another goal of the research is to demonstrate how data mining and machine learning techniques can be effectively applied to customer segmentation, a critical component of adapting to the ever-changing complexity of the market and changes in customer behavior. Within the scope of the research, four customer clusters emerged. Clusters represent meaningful subsets and trends among customers, encompassing a range of features such as demographics, purchasing patterns, and responses to marketing campaigns. The findings provide a useful framework for understanding the complexity of customer profiles and adapting marketing strategies accordingly.

Keywords: *Customer Segmentation, Data Mining, Consumer Behavior, Marketing Strategies, Clustering Methods*

Öz

Bu araştırma, bir market firmasının müşteri veritabanındaki müşteri profillerini veri madenciliği tekniklerini kullanarak detaylı bir şekilde ortaya çıkarmak için gerçekleştirdiği müşteri segmentasyon sürecini incelemektedir. Müşteri segmentasyonu, pazarlama stratejilerinin etkin bir şekilde uyarlanmasında kritik öneme sahiptir. Bu süreç, kişiselleştirilmiş müşteri profillerinin oluşturulmasına olanak tanıyarak daha verimli ve hedefe yönelik pazarlama çalışmalarına olanak sağlamaktadır. Çalışmada kullanılan veri seti, tanınmış bir market firmasının veri tabanından elde edilmiştir ve 29 farklı özelliğe sahip 2.240 veri noktası içermektedir. Bu özellikler müşteri demografisi, ürün bilgileri, satın alma kanalları ve promosyon yanıt verileri olmak üzere dört kategoride toplanmıştır. Çalışma, K-Means Kümeleme ve Aglomeratif Kümeleme gibi ileri kümeleme tekniklerini kullanarak müşteriler arasındaki anlamlı kalıpları ve gruplamaları keşfetmeyi amaçlamaktadır. Ayrıca, araştırmanın bir diğer amacı, dinamik pazar karmaşıklıklarına uyum sağlamanın ve tüketici davranışlarını değiştirmenin kritik bir yönü olan müşteri segmentasyonu için veri madenciliği ve makine öğrenmesi yöntemlerinin nasıl etkili bir şekilde kullanılabileceğini ortaya koymaktır. Araştırma kapsamında dört müşteri kümesi ortaya çıkmıştır. Bu kümeler, demografik bilgiler, satın alma davranışı ve promosyon faaliyetlerine verilen yanıtlar gibi çeşitli özelliklere dayalı olarak müşteriler arasındaki anlamlı gruplaşmaları ve kalıpları temsil etmektedir. Bulgular, müşteri profillerinin karmaşıklığını anlamak ve pazarlama stratejilerini buna göre ayarlamak için değerli bir çerçeve sunmaktadır.

Anahtar Kelimeler: *Müşteri Segmentasyonu, Veri Madenciliği, Tüketici Davranışı, Pazarlama Stratejileri, Kümeleme Yöntemleri*

1. Introduction

In today's competitive business world, companies are responsible for maintaining their market dominance and increasing customer satisfaction. This requires more than just selling products and services. Businesses must foster loyal customer bases and develop appropriate solutions to meet their individual needs. As businesses evolve, customer expectations and demands show an increasing trend towards diversity and specialization. This complexity, along with an ever-expanding customer base, increases the difficulty of efficiently meeting each customer's unique demands.

Businesses use analytical techniques such as data mining to explore the intricacies of this complexity and develop marketing strategies with maximum effectiveness [1]. Data mining techniques comprise a wide range of methods and algorithms that are carefully crafted to uncover important information concealed within large datasets [2]. These methods provide significant benefits to businesses in simplifying the examination of customer data, understanding customer behavior, and applying sophisticated procedures such as customer segmentation.

DOI: https://doi.org/10.21205/deufmd.2025278014

Geliş Tarihi / Received: 19.02.2024
Kabul Tarihi / Accepted: 06.10.2024

Through the use of data mining, individuals with similar traits can be grouped together into distinct clusters for the purpose of managing a large clientele [3]. This classification approach offers numerous possibilities for direct or indirect influence on marketing tactics. Such opportunities include assessing product compatibility in a particular segment, customizing marketing strategies for each group, providing special discounts, and uncovering undisclosed connections between customers and products [4]. By incorporating customer segmentation, grants companies the authority to acquire a profound understanding of customers' authentic purchasing behaviors. This understanding enables them to provide exceptional customer service, amplify customer gratification, identify their prime consumer demographic, and develop marketing strategies that optimize revenue generation [5].

Performing customer segmentation through data mining is a meticulous process that requires a precisely orchestrated series of actions. The initial step involves gathering and intricately organizing a wealth of demographic information, purchase records, behavioral tendencies, and individual preferences [6]. Afterwards, data mining techniques are utilized to extract relevant data from multiple sources and refine its structure for scrupulous analysis [7]. At this stage, techniques such as cluster analysis are used to categorize customers who exhibit comparable patterns. This facilitates the recognition of common characteristics and provides the basis for identifying customer segments as a component of the data analysis procedure [8]. Once the segments have been identified, they lay the groundwork for adjusting marketing strategies. Tailored approaches are devised for each segment, with products and services being revamped accordingly. This step is important to increase customer satisfaction, increase sales and concentrate revenue [9]. Lastly, elevated vocabulary and active voice can assist in the continuous monitoring and optimization of segmentation. Over time, customer behaviors may transform, potentially leading to the emergence of novel segments. Using data mining techniques can effectively detect such changes and help adapt marketing strategies accordingly.

In the field of customer segmentation, several challenges have been identified in the literature, resulting in the need for innovative solutions and methodologies. One of the main challenges lies in the complex nature of customer behavior, which is characterized by its dynamic and multifaceted qualities. The nuances of changing consumer preferences and behaviors often exceed the capabilities of traditional segmentation methods and require advanced approaches. A further challenge pertains to the selection of pertinent segmentation variables. The process of identifying the factors that truly define meaningful customer segments requires a nuanced understanding of the specific market dynamics and business context. In certain instances, the chosen variables may prove insufficient in providing a comprehensive view of customer behavior. In addition, keeping segments constantly up to date is an ongoing challenge. Rapid changes in customer preferences, market dynamics and external influences can cause existing segmentation models to become outdated. As a result, segmentation strategies need to be constantly monitored and adapted to ensure their sustainable effectiveness.

To overcome these challenges, advanced data mining and machine learning techniques are emphasized, placing them at the forefront of the segmentation process. These methodologies, characterized by their adaptability and complexity, can capture intricate patterns in customer behavior. Implementing dynamic and adaptive segmentation models addresses the challenge of maintaining relevance over time by ensuring that segmentation strategies evolve with changing trends and preferences.

A comprehensive approach that incorporates a variety of variables covering both traditional demographic factors and emerging behavioral indicators increases the robustness and accuracy of the segmentation process. In this context, the primary objective of this study is to utilize unsupervised clustering methods to analyze customer records obtained from a grocery firm's database. By employing techniques such as K-Means Clustering, and Agglomerative Clustering, the study aims to segment customers into distinct groups based on shared features, thereby unveiling patterns and similarities within the data. The goal is to optimize the strategic value of each customer for the business by identifying and categorizing customer profiles, which enables the firm to develop targeted marketing strategies and customized product offerings. By understanding the unique needs and behaviors of different customer segments, the firm can enhance customer satisfaction and loyalty, ultimately driving sales and business growth. This research seeks to provide insights into customer segmentation methods that can serve as a foundation for more effective, data-driven decision-making in the retail sector.

2. Research Background

Understanding consumer behavior is important for effective marketing as it enables businesses to competently appeal to their customer base and ensure profitability. Companies prioritize the creation of products or services that precisely address the unique needs and preferences of their customers and use their extensive knowledge as a tool to improve customer loyalty. Marketers must expertly acquire and incorporate data in order to deliver individually tailored services at each encounter, effectively mitigating any potential negative reactions from consumers [10].

Consumer behavior has changed greatly over time. Individuals have become more versatile and prone to changing their behavior and preferences. This dynamism poses a challenge for merchants or producers aiming to identify the varied needs and wants of consumers in large markets. As a result, segmentation – dividing the market into separate groups or subgroups – emerges as a solution to this obstacle. Utilizing market segmentation allows corporations to anticipate and mitigate consumer responses by considering the diverse preferences that individuals possess based on their profiles more accurately [11]. The approach to choosing segmentation methods depends on input variables, including geographic, demographic, behavioral and psychological profiles predicted by statistical or non-statistical means.

According to [12], segmentation is closely intertwined with product differentiation and homogeneity. In a market that is diverse, where customers have numerous options, manufacturers may encounter challenges in choosing or maintaining a customer base. To entice and retain customers, marketers frequently employ discerning tactics such as advertising or sales promotion. Although it is often overlooked, comprehending customer motivations plays a secondary role in broad mass-market methods. For marketers, customer segmentation becomes a deliberate decision to offer personalized merchandise or services with the goal of grouping customers according to their desires and inclinations. In 1956, Smith first introduced the notion of customer segmentation as an unconventional approach to differentiating products, defining segments as groups sharing similar demographic, psychological, and behavioral characteristics [13].

The challenges inherent in traditional customer segmentation methods are widely recognized in the academic environment. These traditional techniques, which rely on specific demographic features, geographic indicators, or specific behaviors, exhibit limitations in addressing the dynamic and complex nature of customer behavior. One primary limitation involves their struggle to proficiently acclimate to changing circumstances. As consumer preferences and behaviors evolve constantly, segmentation techniques that operate with a fixed viewpoint risk becoming obsolete [14]. Additionally, relying solely on salient features for segmentation may fall short of capturing the inherent differences between customer segments. The intricate nature of customer behaviors highlights the insufficiency of relying solely on demographic data in comprehending their needs and motivations [15]. In addition, traditional segmentation techniques, which are typically aimed at larger target groups, can hinder the ability to effectively address individual customer needs [16]. Consequently, there is a growing need for more customized, inventive, and flexible approaches to customer segmentation as emphasized in scholarly works. These challenges underline the limitations of traditional customer segmentation techniques and point to the need for research into the development of more effective and contemporary approaches.

Significant progress has been made in the field of segmentation techniques in the era of information and communication technology. There is a strong focus on data mining and database management systems (DBMS). With the advent of large datasets, traditional market forecasting methods fall short and pose a formidable obstacle even to established statistical approaches such as multivariate analysis and time series. Understanding the transformative impact of customer segmentation and Customer Relationship Management (CRM) in the current business landscape is imperative, necessitating the incorporation of recent advancements in data mining, machine learning, and artificial intelligence (AI). Recent academic studies by [17] and [18] have highlighted the impact of machine learning algorithms on predictive modeling of customer behavior. These works highlight the substantial advancements in machine learning techniques, pointing towards a paradigm shift in businesses' approach to predictive modeling. Moreover, these developments have transcended traditional boundaries and enabled a comprehensive understanding of complex patterns in customer behavior. This evolution has, in turn, granted businesses the capability to tailor their segmentation strategies with unprecedented precision and agility. The advanced sensitivity achieved through these sophisticated algorithms enables detailed responses to ever-changing fluctuations in consumer preferences [19].

AI applications, as evidenced in the research by [20], greatly aids in automating customer interactions. These applications, which contain advanced AI algorithms, facilitate internal processes and enable businesses to interact effectively with their customers. Through advanced algorithms based on machine learning and natural language processing, AI systems can analyze live customer data and precisely identify patterns and individual preferences. This enables the automation of responses and actions, such as those executed by AI-driven chatbots and virtual assistants. Through predictive modeling within AI applications, companies are able to forecast customer requirements, enabling them to proactively and individually engage with customers [21]. Additionally, the study provides comprehensive empirical validation, highlighting the significant impact of AI on personalization methodologies in CRM. The research illuminates how AI-driven systems, utilizing advanced algorithms and machine learning models, have reshaped personalization by actively contributing to advanced customer profiling. This involves complex analysis of customer behavior and provides a deeper understanding of individual preferences. This analysis highlights the important role of AI in developing personalization strategies in modern CRM applications by providing more precise content recommendations.

[22] comprehensively examine sentiment analysis, an area where AI has proven its effectiveness in CRM. Sentiment analysis empowers AI to discover and interpret customer perceptions and emotions, providing insights that go beyond traditional analytics. By leveraging sentiment analysis, businesses can gain a comprehensive understanding of their target audience's reactions to products, services, and marketing efforts [23]. This integration effectively augments the overall understanding of customer feedback and sentiments within CRM strategies, empowering businesses to make more informed decisions.

It is imperative to underline the synchronization between data mining strategies and overall business goals. Dynamic customer segmentation not only elevates ongoing marketing efforts and CRM but also serves as a vital component to an organization's lasting success and competitiveness [24]. Through aligning segmentation methods with broader goals, organizations can maximize resource allocation, amplify customer contentment, and establish a sturdy framework for continuous expansion. [25] suggest that actively incorporating customer insights from data into overarching business goals develops a holistic marketing approach and strengthens competitive advantage in ever-changing markets.

In today's cutthroat market, sellers strive to understand their customers' needs and trends and establish solid connections at every level of business transactions. As evidenced by recent studies [15,17,20-22,24,26], a paradigm shift in customer segmentation and CRM is occurring. CRM has become an important component of marketing tactics, aided by the proliferation of online platforms and various computational methods. Data mining, a computational mining technique, is emerging as a critical tool for uncovering and examining hidden insights embedded within vast amounts of customer information [27]. As customer preferences continue to change, ever-evolving customer segments require a constant flow of data. In this scenario, data mining plays a crucial role in organizing information and accelerating permanent customer segmentation [28]. Furthermore, the utilization of data mining techniques allows enterprises to predict future opportunities and behaviors, empowering them to adopt a more practical and insightful approach [29]. The application of these technologies in customer segmentation leads businesses to a state where segmentation is both predictive and adaptable, improving the overall customer experience and contributing to long-term business prosperity [30].

In the current literature, significant gaps exist in the application of dynamic and adaptable customer segmentation techniques that adequately address the evolving complexity of consumer behavior in large markets. Traditional segmentation approaches often rely on static demographic, geographic, or behavioral characteristics, which fail to capture the nuanced and constantly changing preferences of modern consumers. These limitations create a need for more sophisticated methods that can respond in real time to shifts in consumer behavior, preferences, and market conditions. This study aims to fill these gaps by leveraging advanced data mining and machine learning techniques, such as K-Means and Agglomerative Clustering, to develop a more responsive and accurate segmentation framework. By incorporating a diverse set of customer data points, including demographics, purchasing patterns, and promotional responses, this research demonstrates how innovative segmentation models can enhance personalization strategies, improve marketing effectiveness, and contribute to sustained customer loyalty. The findings provide a deeper understanding of customer profiles and offer actionable insights that help businesses adapt to dynamic market environments, addressing the limitations of traditional segmentation methods and advancing the understanding of customer behavior.

3. Role of Data Mining in CRM

Data mining in the CRM environment emerges as a powerful tool, especially through techniques such as clustering and association rule mining. The basic premise behind using data mining in CRM lies in the belief that historical data holds valuable insights that can be instrumental in shaping future strategies. This belief is based on the understanding that customer behavior in corporate data reflects different needs, preferences, tendencies, and behaviors rather than being random.

The overarching goal of data mining in CRM is to unearth patterns within historical data. However, this task is inherently challenging. Models are not always robust, and signals from customers are often mixed and noisy [31]. The complex role of data mining in this context is to effectively distinguish signals from noise. By making it easier to identify different customer segments, data mining can help develop specific products and offers that suit specific customer priorities. By offering insights into customer behavior, data mining becomes integral in crafting an effective CRM strategy, enabling personalized interactions and subsequently fostering satisfaction and profitable customer relationships through comprehensive data analysis [32].

Data mining operates on a variety of tasks in CRM, as shown in Table 1, and includes modeling tasks such as association, clustering, and classification. Versatile applications of these models range from market basket analysis to web usage analysis, segmentation and forecasting.

Table 1. Data mining tasks, modeling techniques, and example applications.

Data Mining Task	Modeling Techniques	Example Applications
Association	Apriori, FP-Growth	Market Basket Analysis, Cross-Selling Strategies
Sequence Mining	Apriori, FP-Growth	Web Usage Analysis, Clickstream Pattern Recognition
Clustering	K-Means, Hierarchical Models	Customer Segmentation, Behavioral Pattern Identification
Classification	Decision Trees, Neural Networks, Regression, Support Vector Machines	Customer Churn Prediction, Product Recommendation Systems

Table 1 provides a comprehensive overview of key data mining tasks, relevant modeling techniques, and their practical applications in CRM. One pivotal task is labeled association, focusing on the revelation of relationships within data. In this context, vital modeling techniques such as Apriori and FP-Growth come into play. Market Basket Analysis is a notable application that falls under the Association task. This requires rigorous analysis of customer purchase data to uncover relationships between products that are often purchased together. Identifying a consistent pattern where customers tend to purchase pasta and tomato sauce concurrently can be harnessed to devise targeted promotions and optimize product placements [33]. Cross-Selling Strategies represent a strategic extension of the insights gained from Market Basket Analysis. This approach aims to actively encourage customers to explore additional products closely related to their initial purchase [34]. If a customer acquires a camera, a well-crafted cross-selling strategy might recommend complementary items like camera accessories or memory cards.

Sequence mining is emerging as another important data mining task that focuses on identifying patterns within user clickstream histories on the web. To accomplish this task, modeling techniques such as Apriori and FP-Growth are actively employed. Web Usage Analysis involves carefully examining users' interactions with a website, investigating details such as pages visited, time spent on each page, and discernible click patterns [35]. The overarching goal is to gain profound insights into user behavior and preferences, laying the groundwork for informed decision-making in website optimization and content customization. Clickstream Pattern Recognition refers to the identification of recurring patterns in users' click stream data. The discernment that users frequenting specific pages exhibit a higher likelihood of making a purchase can be instrumental in shaping website optimization strategies and tailoring personalized content recommendations [36]. This holistic approach to sequence mining extracts meaningful patterns and transforms them into actionable insights to improve user experience and maximize engagement [37].

Another key data mining task is clustering, a process that facilitates targeted marketing initiatives and the delivery of personalized services by bringing together customers with similar behaviors. This task relies on modeling techniques like K-Means and hierarchical models to effectively group customers based on shared features. Customer segmentation is the systematic division of customers into different groups based on common attributes such as demographics, purchasing behavior, or preferences. This segmentation serves as a foundational strategy, allowing businesses to tailor their marketing approaches for each identified segment, thereby maximizing relevance and impact [38]. Behavioral pattern identification, a complement to customer segmentation, is a crucial initiative that involves examining and understanding patterns in customer behavior. This analysis spans a spectrum, encompassing considerations like frequent purchases, preferred products, and responses to promotions [39]. Armed with this knowledge, businesses can create personalized marketing approaches that suit individual customer preferences, ultimately elevating the overall customer experience.

The final data mining task, classification, is effective in customer relationship management with applications in areas such as customer churn prediction and product recommendation systems. Employing advanced modeling techniques such as Decision Trees, Neural Networks, Regression, and Support Vector Machines, this task significantly contributes to enhancing customer relations and optimizing business strategies. Customer churn prediction uses historical customer data to identify patterns indicative of customers at risk of abandoning a service, allowing businesses to implement proactive retention strategies [40]. Similarly, product recommendation systems analyze customer preferences and behavior, offering individualized product suggestions based on unique tastes. Platforms like Amazon leverage these systems to propose products aligned with users' past purchases and browsing history, enhancing the overall shopping experience and fostering customer loyalty. Strategic use of classification gives businesses the power to predict customer behavior, optimize service offerings and achieve sustainable success in a competitive market [41].

Marketers use direct marketing campaigns to deliver their messages through a variety of channels, including mail, internet, email, and telephone. The stages of these campaigns include data collection and cleaning, customer analysis and segmentation, targeted marketing campaign development, execution, and evaluation. Data mining plays a crucial role in identifying the right customers throughout these stages.

4. Materials and Methods

The main purpose of this study is to use unsupervised clustering methods to analyze customer records obtained from the database of a grocery firm. This involves using customer segmentation, which requires categorizing customers based on common features to identify patterns and similarities that exist within each cluster. The ultimate goal is to effectively stratify customers into distinct segments, thereby optimizing the strategic value of each individual for the business. The segmentation method enables the business to competently address the various concerns of its customers by attempting to customize products and services based on the different needs and actions demonstrated by various customer demographics [42].

The approach implemented in this study involves a comprehensive data mining process, incorporating unsupervised clustering methods to analyze customer records. The study employs K-Means Clustering and Agglomerative Clustering to segment customers based on their characteristics. To determine the optimal number of clusters, the Elbow Method is utilized.

4.1. Data mining

The practice of data mining is the process of extracting valuable patterns and information from extensive datasets [43]. As we navigate the vast digital landscape, vast amounts of information are constantly being generated. Data mining uses the most advanced algorithms and computational methods to meticulously examine this seemingly endless mass of data.

This technique focuses on uncovering hidden treasures within the data rather than merely obtaining it [44]. It attempts to uncover important correlations, trends, and patterns that would typically evade detection. Drawing on the fields of statistics and machine learning, data mining goes beyond traditional data examination by seeking to uncover underlying frameworks that serve as a compass for decision-making and predictive modeling.

4.2. Clustering methods

The field of data mining is very broad, and within this field, clustering methods are proving to be important tools in the field of unsupervised algorithms. Once the meticulous mining process uncovers a multitude of complex patterns, the subsequent task becomes to decipher the inherent complexity embedded in this abundance of information. Regarding unsupervised algorithms, clustering serves as a crucial factor to organize, and group discovered patterns according to their inherent similarity.

In the field of unsupervised algorithms, clusters serve as a systematic framework for understanding multifaceted patterns revealed through data extraction. Unlike their supervised counterparts, which rely on predetermined labels to guide learning, unsupervised algorithms, including clustering, operate without such restrictions [45]. Rather than isolating each piece of information, clustering is used to reveal internal connections and relationships that may not be noticed when examining the data in isolation.

4.2.1. K-means clustering

The selection of K-Means clustering as a data mining technique is underpinned by its effectiveness and widespread application in handling complex datasets. This method is renowned for its ability to segment large volumes of data into distinct, meaningful groups based on similarity measures. The iterative nature of K-Means, where it continuously refines cluster centers to minimize the total squared distances within each cluster, allows for precise identification of inherent patterns and structures in the data. Its widespread usage in areas such as market segmentation and image analysis highlight its robustness and practicality. Additionally, K-Means is favored for its simplicity and computational efficiency, making it an ideal choice for large datasets where rapid and effective segmentation is crucial. The methodological foundation of K-Means clustering thus provides a solid basis for uncovering actionable insights and patterns within complex datasets.

4.2.2. Agglomerative clustering

The choice of agglomerative clustering in this study is driven by its ability to offer a detailed hierarchical view of data structures, essential for uncovering complex patterns in unsupervised data analysis. Unlike flat clustering methods, agglomerative clustering provides a multi-level organization of data points by systematically merging the closest clusters. This process starts with each data point as an individual cluster and progressively combines them to form more comprehensive clusters, ultimately creating a dendrogram that illustrates the hierarchical relationships within the dataset [47]. This hierarchical approach facilitates a nuanced understanding of data patterns and connections, making it particularly valuable for exploring intricate relationships within complex datasets [48]. By offering a granular view of data relationships at various levels of detail, agglomerative clustering enhances the overall analysis and interpretation of unsupervised data.

4.2.3. Elbow method

The Elbow Method is a practical technique used in unsupervised clustering to determine the optimal number of clusters in a given dataset [49]. When implemented in algorithms such as agglomerative clustering, it requires executing the algorithm for various numbers of clusters and examining the relevant metrics. The critical concept lies in detecting the point at which additional clusters no longer contribute significantly to improving the clustering performance metrics and cause a noticeable “elbow” in the metric graph. This approach offers an informed and data-driven method to discern the optimal number of clusters, facilitating a compromise between capturing the complexity of the dataset and avoiding unnecessary detail.

4.3. Dataset

This study conducting a thorough analysis utilizing a dataset procured from the database of a renowned grocery firm. This dataset includes 2240 datapoints, each flaunting 29 features. These features are organized into four distinct categories: customer demographics, product information, place, and promotion data. Table 2 contains features and descriptions of data.

Table 2. Features and descriptions.

Category	Feature	Description
Customer Information	ID	Unique identifier for each customer
	Year_Birth	Year of the customer's birth
	Education	Level of education attained by the customer
	Marital_Status	Marital status of the customer
	Income	Annual household income of the customer
	Kidhome	Count of children in the customer's household
	Teenhome	Count of teenagers in the customer's household
	Dt_Customer	Date when the customer enrolled with the firm
	Recency	Number of days since the customer's last purchase
	Complain	1 if the customer has complained in the last 2 years, 0 otherwise
Products	MntWines	Amount spent on wine in the last 2 years
	MntFruits	Amount spent on fruits in the last 2 years
	MntMeatProducts	Amount spent on meat in the last 2 years
	MntFishProducts	Amount spent on fish in the last 2 years
	MntSweetProducts	Amount spent on sweets the in last 2 years
	MntGoldProds	Amount spent on gold in the last 2 years
Promotion	NumDealsPurchases	Number of purchases made with a discount
	AcceptedCmp1-AcceptedCmp5	1 if the customer accepted the offer in the (n)th campaign, 0 otherwise
	Response	1 if the customer accepted the offer in the last campaign, 0 otherwise
Place	NumWebPurchases	Number of purchases made through the firm's website
	NumCatalogPurchases	Number of purchases made using a catalogue
	NumStorePurchases	Number of purchases made directly in stores
	NumWebVisitsMonth	Number of purchases made directly in stores

Figure 1 shows the correlation matrix of the dataset, providing insights into the relationships between consumer behaviors, socio-demographics, and promotional engagements through a heatmap. The strength and direction of correlations are depicted through color intensities, with green indicating strong positive correlations and dark blue representing strong negative ones. The diagonal, marked by a perfect correlation of 1, naturally stands out, as each variable correlates fully with itself. Some key observations include strong positive correlations between variables such as the number of purchases (NumDealsPurchases, NumWebPurchases, NumCatalogPurchases, NumStorePurchases) and total spending (Total_Expenses). Additionally, the variable "AcceptedCmp1" through "AcceptedCmp5," representing the acceptance of marketing campaigns, shows significant correlations with promotional activity and consumer engagement. On the other hand, variables such as "Income" and "Education" display relatively lower correlations with these campaign variables, indicating that consumer acceptance of promotions may be more influenced by their purchasing habits rather than their income or educational background.

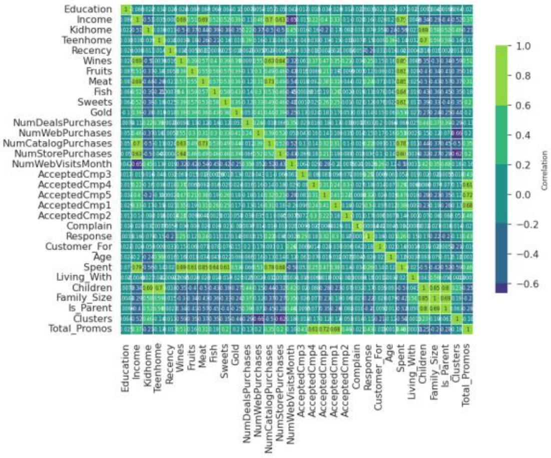

Figure 1. Correlation matrix of the dataset

4.4. Data preparation

The data preparation phase is crucial for ensuring the quality and effectiveness of the clustering analysis. In this phase, several essential steps were undertaken to refine the dataset and prepare it for clustering. Initially, missing values were addressed by removing rows with null income data, thereby maintaining dataset integrity. Following this, new features were engineered to enhance the dataset's richness and relevance. Redundant features were eliminated to streamline the dataset and improve its clarity. Additionally, outliers in the income and age features were identified and removed to ensure the accuracy and reliability of the analysis. These preparatory steps were designed to enhance the overall data quality, which is critical for achieving robust and meaningful results in the clustering analysis. Detailed methodologies for these steps are discussed in the subsequent sections.

4.4.1. Handling missing values

To ensure dataset integrity and the reliability of subsequent analyses, rows with missing income data were systematically removed. The process began with the identification of missing values through a comprehensive examination of the dataset. This involved using descriptive statistics to pinpoint entries with null or incomplete income information. Following identification, the extent of missing data was assessed to understand its potential impact on the overall dataset. Given that income is a critical variable for customer segmentation, rows with missing values were excluded to prevent bias and ensure the accuracy of the clustering results.

4.4.2. Feature engineering

In the feature engineering phase, several new attributes were introduced to enhance the dataset's descriptive power and relevance for clustering analysis. A novel feature, "Customer_For," was created to represent the time elapsed since each customer's enrollment in the firm's database, which provides insights into customer tenure and engagement. The "Age" feature was derived from the "Year_Birth" attribute to capture the current age of customers, a critical factor in understanding customer demographics. To consolidate expenditure information, a “Spent” feature was developed, aggregating total spending across multiple categories over a two-year period, reflecting overall customer expenditure behavior.

Additional features included “Living_With,” which was inferred from “Marital_Status” to indicate cohabitation status, providing context on customer living arrangements. The “Children” feature was calculated to account for the total number of children, including both young children and teenagers, offering a more detailed view of household composition. To gain insights into family dynamics, the “Family_Size” feature was established to represent the total number of individuals in the household. The “Is_Parent” feature was introduced as a binary indicator of whether customers are parents, enhancing the understanding of customer life stages. Lastly, the “Education” feature was simplified into three categories to improve interpretability and analysis efficiency.

4.4.3. Redundant feature removal

To streamline the dataset and enhance the efficiency of the clustering analysis, redundant features were systematically removed. This process involved evaluating the relevance and contribution of each feature to the overall analytical objectives. Features that were deemed irrelevant or provided minimal additional value were excluded to reduce dimensionality and mitigate potential noise in the data. By removing these redundant attributes, the dataset was optimized, allowing for a more focused and effective clustering process. This refinement improved computational efficiency and ensured that the remaining features provided meaningful and distinct information necessary for accurate customer segmentation.

4.4.4. Outliers treatment

Outliers in the income and age features were identified and removed to improve the accuracy and robustness of the analysis. The identification process involved utilizing the Interquartile Range (IQR) method, a robust technique for detecting outliers in continuous data. The IQR was calculated by subtracting the first quartile (Q1) from the third quartile (Q3) for both income and age distributions. Data points falling below Q1−1.5×IQR or above Q3+1.5×IQR were considered outliers. These extreme values were then excluded from the dataset.

4.4.5. Standardization and label encoding

Ensuring that data is uniformly processed and effectively aligned with clustering algorithms is crucial for obtaining reliable results. In this study, categorical features were first converted into numerical labels using one-hot encoding. This technique involves creating binary columns for each category, ensuring that each category is represented as a separate feature with values of either 0 or 1, which facilitates the uniform processing of categorical data. Following this, the entire feature set was scaled to a standard range using min-max normalization. This method transforms feature values to a range between 0 and 1 by subtracting the minimum value of the feature and then dividing by the range (i.e., the difference between the maximum and minimum values). This scaling process ensures that all features contribute equally to the clustering process, mitigating the impact of features with larger ranges or varying units.

4.5. Model setups

In this study, various clustering techniques were applied with specific configurations to optimize their performance. The K-Means Clustering algorithm was employed with a range of cluster numbers (k) determined through the Elbow Method, which identified four as the optimal value. To enhance the accuracy of centroid initialization, the k-means++ method was utilized. For Agglomerative Clustering, the ward linkage method was chosen to minimize within-cluster variance, while Euclidean distance was selected as the distance metric to ensure effective cluster formation. Principal Component Analysis (PCA) was used to reduce dimensionality, with the number of components (n_components) set to three, which effectively captured the most significant variance in the data while simplifying the dataset. The Elbow Method also guided the determination of the optimal number of clusters by analyzing the distortion score (WCSS), confirming four clusters as the most appropriate. Model parameters were finely tuned, with K-Means set to a maximum of 300 iterations (max_iter) and Agglomerative Clustering using a Euclidean distance metric.

5. Experimental Study and Findings

The experimental study aimed to explore and validate the effectiveness of various clustering techniques on the prepared dataset. Dimensionality reduction techniques were applied to address challenges posed by high feature counts, with PCA reducing the dataset to three principal components for simplified analysis and visualization. Clustering techniques, including K-Means and Agglomerative Clustering, were employed to segment the data into distinct groups. The optimal number of clusters was determined using the Elbow Method, leading to the identification of four clusters. Visualizations were employed to interpret the clustering results, revealing patterns and characteristics of the identified clusters.

The large number of features creates difficulties, especially with redundant and interconnected features. To overcome this problem, dimensionality reduction methods are applied to reduce data loss while increasing understandability. PCA plays a crucial role in this setting and serves as an important technique [50]. PCA through ordinal measurements is applied to reduce dimensionality and condense the feature space into three principal components. This simplifies calculations and enables the identification of key variables that govern the structure of the dataset. Subsequent visualization of the reduced data frame through plotting helps in understanding inherent patterns in the data.

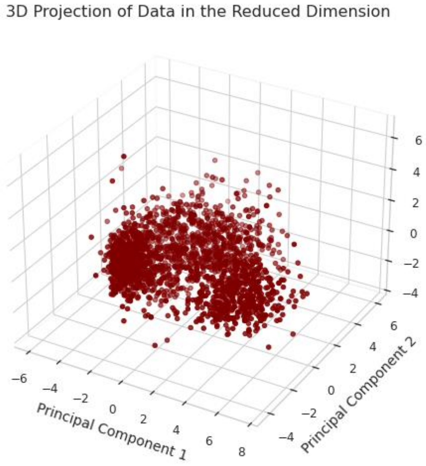

Figure 2. 3D Projection of data using PCA.

Figure 2 shows a salient feature of three-dimensional projection with reduced dimensionality. Compact alignment of points indicates the presence of distinctive subcategories in the data. Similar features or behaviors between observations lead to their proximity, creating distinguishable patterns. Specifically, in consumer segmentation analyses, combining individuals with comparable customer profiles enables more meticulous development of marketing tactics. The proximity between elements signifies the compactness of a particular characteristic or behavioral pattern, as well as their inclination to congregate. By reducing data dimensionality, this nearness offers an enhanced comprehension of the underlying dataset structure and aids in recognizing notable patterns.

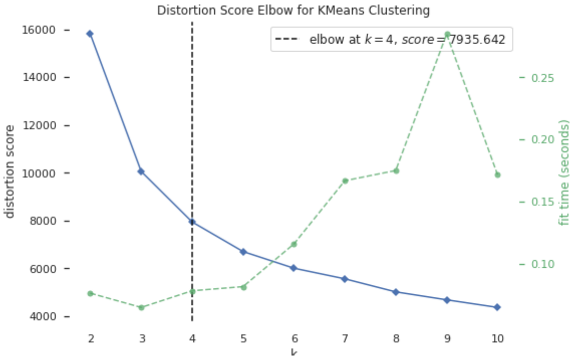

Figure 3. Distortion score elbow for K-Means.

The determination of the optimal number of clusters is aimed at enhancing the accuracy and interpretability of the clustering model. Accordingly, Figure 3 showcases the utilization of the Elbow Method, proposing that four clusters would be most suitable for segmenting the dataset. This decision is founded on identifying where there is a point of diminishing returns in WCSS, indicating a harmonious balance between minimizing inter-cluster distance and avoiding excessive fragmentation. After determining the optimal number of clusters through K-Means Clustering, we utilize the Agglomerative Clustering Model to solidify the final structure for clustering.

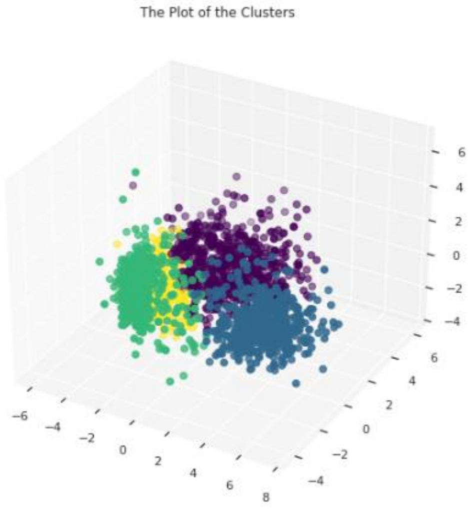

Figure 4. 3D distribution of the clusters.

The Agglomerative Clustering Model is implemented to yield the final clusters, and the visual representation of these clusters is provided in Figure 4. The decision to employ this model is driven by its hierarchical methodology that amalgamates data points according to their resemblances, imparting a holistic understanding of the underlying structures within the dataset. The illustration in Figure 4 proves to be a valuable visual tool, showcasing both the spatial arrangement and cohesion of data points within each identified cluster. Clusters and their constituent points are in close proximity, implying a constricted aggregation of analogous observations. This juxtaposition attests to the success of the Agglomerative Clustering Model in identifying and consolidating data points with shared features or traits. The spatial cohesion within each cluster denotes a heightened level of similarity among grouped observations, thus bolstering the soundness of the segmentation methodology.

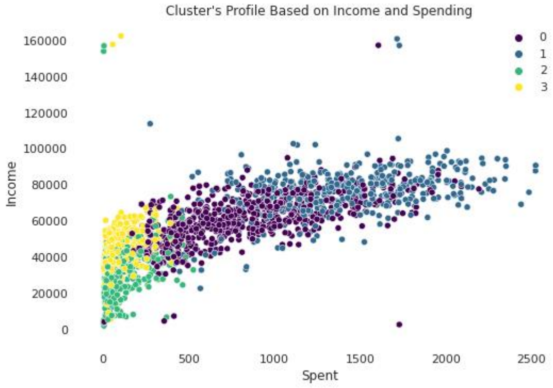

Figure 5. Cluster's profile based on income and spending.

Figure 5 depicts the profiles of clusters based on both income and expenditure. Each colored dot is linked to a particular cluster, their placement symbolizing the correlation between income tiers and spending behaviors. This graph displays four discernible clusters. The first group comprises individuals with high spending habits and moderate-income levels, while the second group is composed of individuals with high spending and high income. The third group denotes those who exhibit low spending patterns and low incomes, whereas the fourth group consists of individuals with high spending but low income.

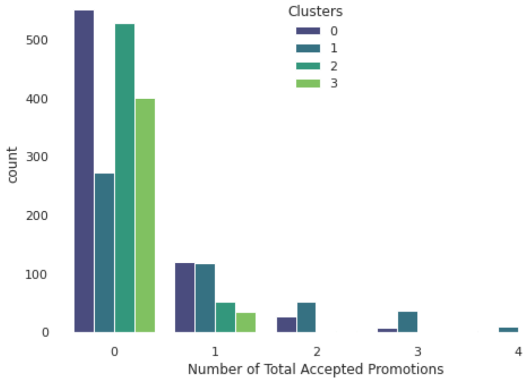

Figure 6. Count of promotion accepted.

The count of promotions accepted is graphically represented in Figure 6, with the intention of illuminating the participation patterns observed in diverse promotional campaigns across distinct clusters. Each bar in the plot signifies the number of participants based on the total accepted promotions. The graph's observations indicate a relatively low response to the campaigns, with an overall scarcity of participants. Moreover, none of the participants engaged in all five campaigns. These findings imply that the current promotional strategies may not be effectively resonating with the customer base. The lack of participants who embrace all campaigns suggests the necessity for more focused and meticulously strategized promotional endeavors to augment the overall efficacy of the campaign and, consequently, elevate sales.

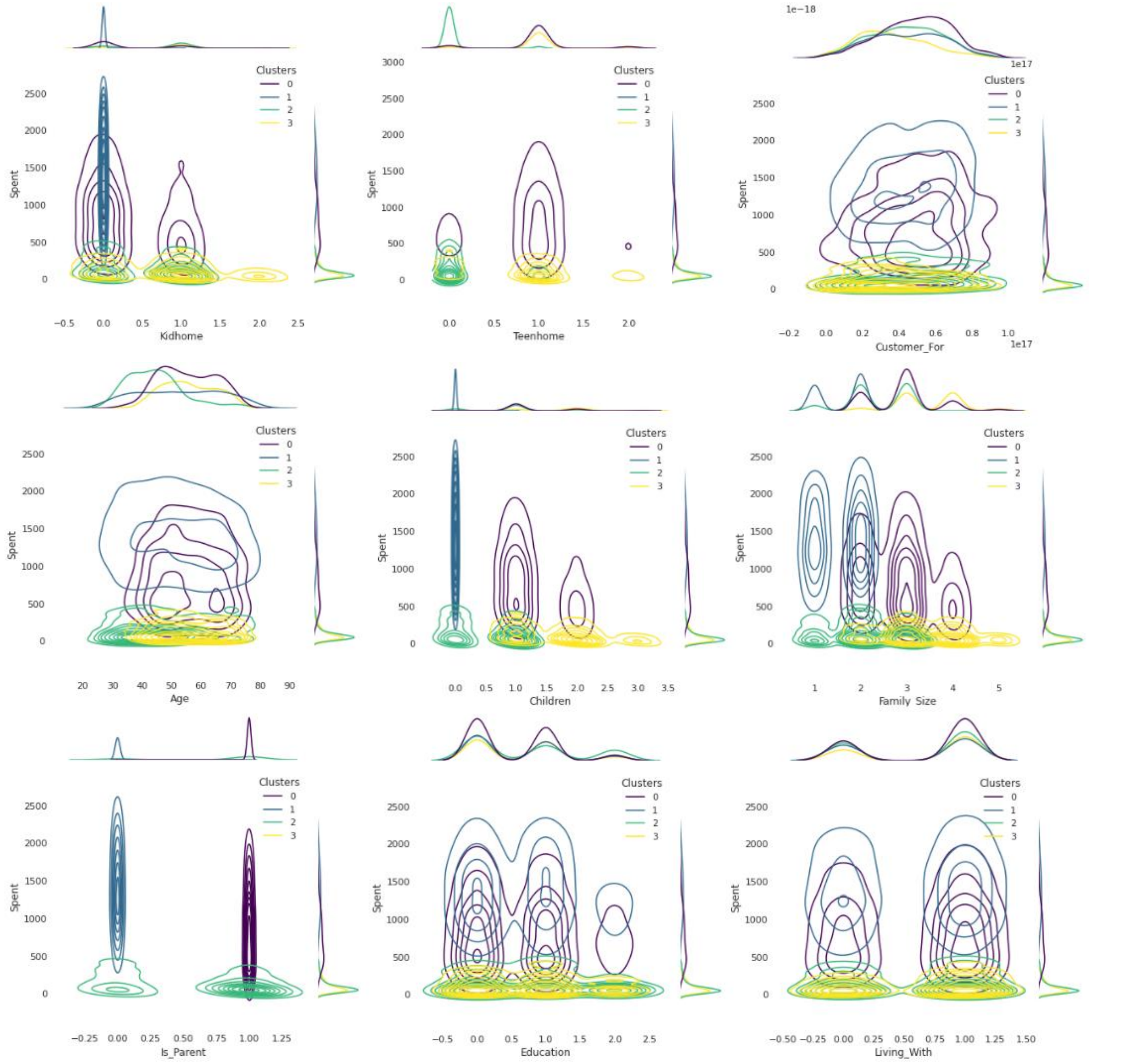

Figure 7. Personal traits impacting spending patterns.

Figure 7 demonstrates specific features that indicate customers' personal characteristics linked to the clusters formed. These visual representations are designed to understand the profiling of individuals within the clusters. Each graph exhibits the expenditure distribution among customers in a particular cluster based on a specific personal trait such as age, education level, number of children, and more.

When examining the research findings and presented graphs, the characteristics of the four customer clusters identified are as follows:

- Cluster 0: This group predominantly consists of individuals who are unequivocally parents. The size of their families varies from at least two members to a maximum of four, with single parents making up a significant subset within this cluster. A large portion of the households in this grouping have teenagers living at home, and the age range skews towards an older demographic.
- Cluster 1: Distinctly characterized as non-parents, this cluster typically has a smaller family size, with a maximum of 2 members. In this group, couples surpass single individuals in number, covering a wide range of ages. Notably, this cluster is associated with a higher income demographic.
- Cluster 2: The majority of individuals in this cluster are parents, with family sizes ranging up to a maximum of 3 members. Families in this cluster typically have one child, generally not in the teenage demographic. The distribution of ages in this category tends to lean towards a relatively younger demographic.
- Cluster 3: Similar to Cluster 0, this group is predominantly comprised of individuals who are

undoubtedly parents. Family sizes range from a minimum of 2 members to a maximum of 5, with a significant portion having teenagers at home. The age distribution, like Cluster 0, skews relatively older, and this cluster is associated with a lower-income demographic.

6. Discussion

Investigating data mining applications in customer segmentation in the retail industry provides valuable information, and discussing the findings in the context of existing literature enriches our understanding of the subject. This discussion aims to contextualize the implications of the research, aligning the results of the study with relevant academic contributions.

The data preparation process, which includes converting categorical features into labels, scaling and dimensionality reduction, is emphasized as a very important aspect, in line with established practices in the literature [51]. Optimization of data before cluster analysis is considered very important as emphasized by various researchers. The dimensionality reduction strategy using PCA complies with accepted methodologies in this field [52]. Similar to other studies, PCA is considered a crucial method to simplify computational demands and extract key variables that govern the underlying structure of the dataset. By making it compatible with established techniques, robustness and reproducibility are added to cluster analyzes in the broader context of data mining. Furthermore, determining the optimal number of clusters through the application of the Elbow Method is in line with the methodologies advocated by [53] and [54]. The rationale behind this approach coincides with the need to strike a harmonious balance between minimizing inter-cluster distance and avoiding excessive fragmentation, which is a fundamental consideration in the development of a robust clustering model. Reliability is increased by leveraging these established methodologies.

The Agglomerative Clustering Model used in the study offers significant advantages in deciphering complex data structures. The hierarchical structure of this approach facilitates a detailed understanding of how individual data points combine into clusters, contributing to a richer understanding of underlying patterns. This is consistent with [55]'s observations, which highlight the model's effectiveness in handling diverse datasets, especially those characterized by complex relationships between data points. Additionally, the ability of the hierarchical clustering method to reveal nested structures in the data aligns with [56] propositions and demonstrates its suitability for exploring multilevel relationships between observations.

The analysis of customer profiles in terms of income, spending behavior and responses to promotional activities reflects the views proposed by various researchers in consumer behavior analysis. Correlations between income tiers and spending habits among identified customer clusters are consistent with the findings of [57] and [58], both of whom highlight the importance of income as a determining factor in consumer behavior. Additionally, similar to the observation of [59], the relatively low response to promotional campaigns underscores the importance of improving promotional strategies to better resonate with different customer segments.

Additionally, while this study identifies different customer clusters based on demographic characteristics, the results are in line with studies by [60] and [61] that highlight the value of personalized marketing strategies tailored to specific consumer segments. This aligns with the idea that understanding the unique characteristics of customer clusters allows businesses to customize marketing tactics and optimize promotional efforts for greater engagement and therefore business success.

Despite the valuable contributions of this study, there are some limitations related to the challenges inherent in [62]'s data mining applications. Focusing on a single grocery firm requires caution in generalizing the findings. In line with [63]'s recommendation for longitudinal studies, future research can examine the temporal dynamics of customer segments and their responses to changing marketing strategies.

7. Conclusions

The study explores the complex procedure of customer segmentation, which is a crucial component in formulating a successful marketing strategy. The main goal is to provide a detailed understanding of the unique consumer profiles in a grocery firm's database. The research aims to uncover important patterns and clusters among customers using advanced clustering techniques. Within the scope of the research, it was concluded that four different customer clusters emerged through the discovery of the effective application of data mining and machine learning techniques in customer segmentation.

The identified customer clusters provided information of great importance for strategic decision-making in the areas of marketing and business operations. By understanding the unique characteristics of each cluster, businesses can customize their marketing tactics to suit specific consumer segments and thus achieve the best results from their promotional efforts. For instance, the delineation of Cluster 0 as predominantly composed of unequivocal parents with a focus on households with teenagers and relatively older age groups suggests a target audience with distinct needs and preferences. This understanding allows marketers to create custom promotions and ads that resonate more effectively with this specific customer segment.

Similarly, the identification of Cluster 1 as a group characterized by non-parents with smaller family sizes and a higher income demographic provides strategic insights. This segment may respond more positively to marketing efforts that appeal to couples and individuals with higher disposable income. Additionally, data on these clusters provides the opportunity to focus marketing efforts on a specific demographic group through individualized messages. Results from customer demographics provide businesses with a way to improve marketing strategies, increase customer engagement, and ultimately increase overall business success.

The results of this research improve our understanding of customer segmentation by highlighting the critical importance of data mining methods in developing marketing tactics for the retail industry. The findings highlight the value of customer segmentation to optimize advertising campaigns, increase customer engagement and ultimately advance the prosperity of retail businesses. The presented approach addresses the immediate challenges facing retailers, provides a foundation for future research in this field, and underlines the enduring importance of data mining in shaping successful marketing approaches.

8. Limitations and Suggestions

This study offers valuable insights into customer segmentation through the use of advanced data mining and machine learning techniques; however, several limitations should be noted. First, the analysis is based on a dataset from a single grocery firm, which may not capture the full spectrum of consumer behavior across different industries or geographic regions. Consequently, the findings may have limited generalizability to other market contexts or sectors. Additionally, the study primarily employs K-Means and Agglomerative Clustering techniques. While these methods are effective, they may not account for all the complexities in consumer behavior. The choice of clustering algorithms could influence the results, and exploring alternative techniques such as DBSCAN or hierarchical clustering with different distance metrics might provide different insights.

Another limitation is the static nature of the analysis. The study represents a snapshot of customer segments at a specific point in time and does not account for temporal changes in consumer behavior or market conditions. This omission could impact the relevance and stability of the identified clusters over time. Moreover, external factors such as economic shifts, seasonal variations, and emerging consumer trends are not considered in the analysis. These factors could affect customer behavior, and the effectiveness of targeted marketing strategies derived from the study.

To address these limitations and advance the field of customer segmentation, future research should explore several avenues. Expanding the dataset to include multiple firms across various industries and geographic locations would enhance the generalizability of the findings. Incorporating diverse data sources can offer a more comprehensive understanding of customer segmentation. Additionally, investigating alternative clustering techniques and advanced methods, such as DBSCAN, Gaussian Mixture Models, or deep learning-based approaches, could provide a broader perspective on customer segmentation and uncover more intricate patterns. Longitudinal studies that track changes in customer behavior over time would offer insights into the stability of customer segments and the adaptability of marketing strategies to evolving consumer preferences. Incorporating external variables, such as economic indicators, seasonal effects, and emerging trends, into the segmentation analysis would provide a more dynamic and context-sensitive understanding of customer behavior. Furthermore, leveraging real-time data analytics and machine learning algorithms could enhance the responsiveness and precision of segmentation strategies. Future research could investigate how real-time data integration impacts customer segmentation and marketing effectiveness.

Ethics committee approval and conflict of interest statement

This article does not require ethics committee approval. This article has no conflicts of interest with any individual or institution.

References

[1] Hung, P. D., Lien, N. T. T., Ngoc, N. D. 2019. Customer Segmentation Using Hierarchical Agglomerative Clustering, 2nd International Conference on Information Science and Systems, 16 – 19 March-2019, Tokyo, Japan, pp.33-37.

[2] Huang, S. 2014. Method for Customer Segmentation Based on Three-Way Decisions Theory-Journal of Computer Applications, Vol. 34, No. 1, p.244.

[3] Tabianan, K., Velu, S., Ravi, V. 2022. K-means Clustering Approach for Intelligent Customer Segmentation Using Customer Purchase Behavior Data-Sustainability, Vol. 14, No. 12, p.7243.

[4] Goyat, S. 2011. The Basis of Market Segmentation: A Critical Review of Literature-European Journal of Business and Management, Vol. 3, No. 9, p.45-54.

[5] Thakur, R., Workman, L. 2016. Customer Portfolio Management (CPM) for Improved Customer Relationship Management (CRM): Are Your Customers Platinum, Gold, Silver, or Bronze?-Journal of Business Research, Vol. 69, No. 10, pp.4095-4102.

[6] Khandpur, N., Zatz, L. Y., Bleich, S. N., Taillie, L. S., Orr, J. A., Rimm, E. B., Moran, A. J. 2020. Supermarkets in Cyberspace: A Conceptual Framework to Capture the Influence of Online Food Retail Environments on Consumer Behavior-International Journal of Environmental Research and Public Health, Vol. 17, No. 22, p.8639.

[7] Diba, K., Batoulis, K., Weidlich, M., Weske, M. 2020. Extraction, Correlation, and Abstraction of Event Data for Process Mining-Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, Vol. 10, No. 3, p.e1346.

[8] Arora, P., Varshney, S. 2016. Analysis of K-means and K-medoids Algorithm for Big Data-Procedia Computer Science, Vol. 78, p.507-512.

[9] Cosenz, F., Bivona, E. 2021. Fostering Growth Patterns of SMEs Through Business Model Innovation. A Tailored Dynamic Business Modelling Approach-Journal of Business Research, Vol. 130, pp. 658-669.

[10] Lefait, G., Kechadi, T. 2010. Customer Segmentation Architecture Based on Clustering Techniques. 2010 Fourth International Conference on Digital Society, 10-16 February-2010, St. Maarten, Netherlands Antilles, 243-248.

[11] Steenkamp, J. B. E., Ter Hofstede, F. 2002. International Market Segmentation: Issues and Perspectives-International Journal of Research in Marketing, Vol. 19, No. 3, pp.185-213.

[12] Smith, W. R. 1956. Product Differentiation and Market Segmentation as Alternative Marketing Strategies-Journal of Marketing, Vol. 21, No. 1, pp.3-8.

[13] Tynan, A. C., Drayton, J. 1987. Market Segmentation-Journal of Marketing Management, Vol. 2, No. 3, pp.301-335.

[14] Zhang, J. Z., Chang, C. W. 2021. Consumer Dynamics: Theories, Methods, and Emerging Directions-Journal of the Academy of Marketing Science, Vol. 49, p.166-196.

[15] Shahid, S., Paul, J. 2021. Intrinsic Motivation of Luxury Consumers in An Emerging Market-Journal of Retailing and Consumer Services, Vol. 61, p.102531.

[16] Beauvisage, T., Beuscart, J. S., Coavoux, S., Mellet, K. 2023. How Online Advertising Targets Consumers: The Uses of Categories and Algorithmic Tools by Audience Planners-New Media & Society, Vol. 46, p.14614448221146174.

[17] Surendro, K. 2019. Predictive Analytics for Predicting Customer Behavior, 2019 International Conference of Artificial Intelligence and Information Technology (ICAIIT), 13-15 March-2019, Yogyakarta, Indonesia, pp.230-233.

[18] Kotras, B. 2020. Mass Personalization: Predictive Marketing Algorithms and the Reshaping of Consumer Knowledge-Big Data & Society, Vol. 7, No. 2, p.2053951720951581.

[19] Soni, V. 2021. Deep Learning and Computer Vision-Based Retail Analytics for Customer Interaction and Response Monitoring-Eigenpub Review of Science and Technology, Vol. 5, No. 1, pp.1-15.

[20] Verma, R. K., Kumari, N. 2023. Generative AI as a Tool for Enhancing Customer Relationship Management Automation and Personalization Techniques-International Journal of Responsible Artificial Intelligence, Vol. 13, No. 9, pp.1-8.

[21] Bharadiya, J. P. 2022. Driving Business Growth with Artificial Intelligence and Business Intelligence-International Journal of Computer Science and Technology, Vol. 6, No. 4, pp.28-44.

[22] Capuano, N., Greco, L., Ritrovato, P., Vento, M. 2021. Sentiment Analysis for Customer Relationship Management: An Incremental Learning Approach-Applied Intelligence, Vol. 51, pp.3339-3352.

[23] Verma, S. 2022. Sentiment Analysis of Public Services for Smart Society: Literature Review and Future Research Directions-Government Information Quarterly, Vol. 39, No. 3, pp.101708.

[24] Yang, J., Xiu, P., Sun, L., Ying, L., Muthu, B. 2022. Social Media Data Analytics for Business Decision Making System to Competitive Analysis. Information Processing & Management, Vol. 59, No. 1, p.102751.

[25] Zhang, C., Wang, X., Cui, A. P., Han, S. 2020. Linking Big Data Analytical Intelligence to Customer Relationship Management Performance. Industrial Marketing Management, Vol. 91, pp.483-494.

[26] Amarasinghe, H. 2023. Transformative Power of AI in Customer Relationship Management (CRM): Potential Benefits, Pitfalls, and Best Practices for Modern Enterprises. International Journal of Social Analytics, Vol. 8, No. 8, pp.1-10.

[27] Dasu, T., Johnson, T. 2003. Exploratory Data Mining and Data Cleaning. John Wiley & Sons, Hoboken.

[28] Carnein, M., Trautmann, H. 2019. Customer Segmentation Based on Transactional Data Using Stream Clustering. Advances in Knowledge Discovery and Data Mining: 23rd Pacific-Asia Conference, 14-17 April-2019, Macau, China, pp.280-292.

[29] Punhani, R., Arora, V. P. S., Sabitha, A. S., Shukla, V. K. 2021. Segmenting E-Commerce Customer Through Data Mining Techniques-Journal of Physics: Conference Series, Vol. 1714, No. 1, p.012026.

[30] Hermanto, H., Sulistyan, R. B., Touati, H. 2022. Service Satisfaction Based on Performance Index and Importance Performance Analysis (IPA)-Innovation Business Management and Accounting Journal, Vol. 1, No. 2, pp.41-52.

[31] Müller, N. M., Markert, K. 2019. Identifying Mislabeled Instances in Classification Datasets. 2019 International Joint Conference on Neural Networks (IJCNN), 14-19 July-2019, Budapest, Hungary, pp.1-8.

[32] Mach-Król, M., Hadasik, B. 2021. On a Certain Research Gap in Big Data Mining for Customer Insights-Applied Sciences, Vol. 11, No. 15, p.6993.

[33] Hossain, M., Sattar, A. S., Paul, M. K. 2019. Market Basket Analysis Using Apriori and FP Growth Algorithm. 2019 22nd international conference on computer and information technology (ICCIT), 18-20 December-2019, Dhaka, Bangladesh, pp.1-6.

[34] Laxmi, K. R., Srivastava, S., Madhuravani, K., Pallavi, S., Dewangan, O. 2022. Modified Cross-Sell Model for Telecom Service Providers Using Data Mining Techniques-Data Mining and Machine Learning Applications, pp.195-207.

[35] Kumar, S., Kar, A. K., Ilavarasan, P. V. 2021. Applications of Text Mining in Services Management: A Systematic Literature Review-International Journal of Information Management Data Insights, Vol. 1, No. 1, p.100008.

[36] Koehn, D., Lessmann, S., Schaal, M. 2020. Predicting Online Shopping Behaviour from Clickstream Data Using Deep Learning-Expert Systems with Applications, Vol. 150, p.113342.

[37] Olmezogullari, E., Aktas, M. S. 2022. Pattern2vec: Representation of Clickstream Data Sequences for Learning User Navigational Behavior-Concurrency and Computation: Practice and Experience, Vol. 34, No. 9, p.e6546.

[38] Anitha, P., Patil, M. M. 2022. RFM Model for Customer Purchase Behavior Using K-Means Algorithm-Journal of King Saud University-Computer and Information Sciences, Vol. 34, No. 5, pp.1785-1792.

[39] Safa, N. S., Ghani, N. A., Ismail, M. A. 2014. An Artificial Neural Network Classification Approach for Improving Accuracy of Customer Identification in E-Commerce-Malaysian Journal of Computer Science, Vol. 27, No. 3, pp.171-185.

[40] Ahmad, A. K., Jafar, A., Aljoumaa, K. 2019. Customer Churn Prediction in Telecom Using Machine Learning in Big Data Platform-Journal of Big Data, Vol. 6, No. 1, pp.1-24.

[41] Dash, A., Chakraborty, A., Ghosh, S., Mukherjee, A., Gummadi, K. P. 2021. When The Umpire is Also a Player: Bias in Private Label Product Recommendations on E-Commerce Marketplaces. 2021 ACM Conference on Fairness, Accountability, and Transparency, 3-10 March-2021, Online, pp.873-884.

[42] Ernawati, E., Baharin, S. S. K., Kasmin, F. 2021. A Review of Data Mining Methods in RFM-Based Customer Segmentation-Journal of Physics: Conference Series, Vol. 1869, No. 1, p.012085.

[43] Kamath, C. 2001. On Mining Scientific Datasets-Data Mining for Scientific and Engineering Applications, p.1-21.

[44] Raj, P., Kumar, S. A. 2017. Big Data Analytics Processes and Platforms Facilitating Smart Cities-Smart cities: Foundations, Principles, and Applications, pp.23-52.

[45] Grira, N., Crucianu, M., Boujemaa, N. 2004. Unsupervised and Semi-Supervised Clustering: A Brief Survey-A Review of Machine Learning Techniques for Processing Multimedia Content, Vol. 1, No. 2004, p.9-16.

[46] Gasch, A. P., Eisen, M. B. 2002. Exploring the Conditional Coregulation of Yeast Gene Expression Through Fuzzy K-Means Clustering-Genome Biology, Vol. 3, No. 11, pp.1-22.

[47] Mutihac, L., Mutihac, R. 2008. Mining in Chemometrics-Analytica Chimica Acta, Vol. 612, No. 1, pp.1-18.

[48] Bouguettaya, A., Yu, Q., Liu, X., Zhou, X., Song, A. 2015. Efficient Agglomerative Hierarchical Clustering-Expert Systems with Applications, Vol. 42, No. 5, p.2785-2797.

[49] Yuan, C., Yang, H. 2019. Research on K-Value Selection Method of K-Means Clustering Algorithm-J — Multidisciplinary Scientific Journal, Vol. 2, No. 2, pp.226-235.

[50] Alkhayrat, M., Aljnidi, M., Aljoumaa, K. 2020. A Comparative Dimensionality Reduction Study in Telecom Customer Segmentation Using Deep Learning And PCA-Journal of Big Data, Vol. 7, pp.1-23.

[51] Zheng, A., Casari, A. 2018. Feature Engineering for Machine Learning: Principles and Techniques for Data Scientists. O'Reilly Media, Inc, California, USA.

[52] Reddy, G. T., Reddy, M. P. K., Lakshmanna, K., Kaluri, R., Rajput, D. S., Srivastava, G., Baker, T. 2020. Analysis of Dimensionality Reduction Techniques on Big Data-IEEE Access, Vol. 8, pp.54776-54788.

[53] Shi, C., Wei, B., Wei, S., Wang, W., Liu, H., Liu, J. 2021. A Quantitative Discriminant Method of Elbow Point for the Optimal Number of Clusters in Clustering Algorithm-Eurasip Journal on Wireless Communications and Networking, Vol. 2021, No. 1, pp.1-16.

[54] Kodinariya, T. M., Makwana, P. R. 2013. Review on Determining Number of Cluster in K-Means Clustering-International Journal, Vol. 1, No. 6, pp.90-95.

[55] Kamvar, S. D., Klein, D., Manning, C. D. 2002. Interpreting and Extending Classical Agglomerative Clustering Algorithms Using a Model-Based Approach-ICML, Vol. 2, pp.283-290.

[56] Murtagh, F., Contreras, P. 2017. Algorithms for Hierarchical Clustering: An Overview-II. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, Vol. 7, No. 6, p.e1219.

[57] Mihić, M., Čulina, G. 2006. Buying Behavior and Consumption: Social Class Versus Income-Management: Journal of Contemporary Management Issues, Vol. 11, No. 2, pp.77-92.

[58] Roy, G., Debnath, R., Mitra, P. S., Shrivastava, A. K. 2021. Analytical Study of Low-Income Consumers' Purchase Behaviour for Developing Marketing Strategy-International Journal of System Assurance Engineering and Management, Vol. 12, No. 5, pp.895-909.

[59] Kallier, S. M. 2017. The Influence of Real-Time Marketing Campaigns of Retailers on Consumer Purchase Behavior-International Review of Management and Marketing, Vol. 7, No. 3, pp.126-133.

[60] Behera, R. K., Gunasekaran, A., Gupta, S., Kamboj, S., Bala, P. K. 2020. Personalized Digital Marketing Recommender Engine-Journal of Retailing and Consumer Services, Vol. 53, p.101799.

[61] Olson, E. M., Olson, K. M., Czaplewski, A. J., Key, T. M. 2021. Business strategy and the Management of Digital Marketing-Business Horizons, Vol. 64, No. 2, pp.285-293.

[62] Ezugwu, A. E., Ikotun, A. M., Oyelade, O. O., Abualigah, L., Agushaka, J. O., Eke, C. I., Akinyelu, A. A. 2022. A comprehensive Survey of Clustering Algorithms: State-Of-The-Art Machine Learning Applications, Taxonomy, Challenges, and Future Research Prospects-Engineering Applications of Artificial Intelligence, Vol. 110, pp.104743.

[63] Goić, M., Levenier, C., Montoya, R. 2021. Drivers of Customer Satisfaction in The Grocery Retail Industry: A Longitudinal Analysis Across Store Formats-Journal of Retailing and Consumer Services, Vol. 60, p.102505.