Ilaro Journal of Science and Technology (IJST) 2024
Published Online October 2024 (https://sciencetechjournal.federalpolyilaro.edu.ng)
Volume 4, 2024 Edition
ISSN: 2437-1505

Customer Segmentation Using the K-Means Clustering Algorithm

Emmanuel Ayodele
The Federal Polytechnic Ilaro, Department of Computer Science
ayodele.emmanuel@federalpolyilaro.edu.ng

Victor Sodeinde
The Federal Polytechnic Ilaro, Department of Computer Science
victor.sodeinde@federalpolyilaro.edu.ng

Abstract
Customer segmentation is an important tool for modern businesses because it allows them to personalize their experiences to different client categories, optimize their resource allocation, and improve their marketing efforts. The K-means clustering algorithm is an excellent strategy for identifying different consumer segments based on shared criteria. By lowering the sum of diagonal distances across points of data and cluster centroids, the K-means algorithm repeatedly splits datasets into clusters. This approach facilitates the identification of homogeneous client groups that share traits, tastes, and behaviors that are essential for efficient segmentation. Through unsupervised learning, K-means reveals hidden patterns in customer data, enabling organizations to create personalized communications, product recommendations, and marketing plans. However, a comprehensive evaluation of such obstacles is necessary for the K-means algorithm to be deployed in consumer segmentation in an efficient manner. Utilizing strategies like K-means++ initialization may be essential to reduce the likelihood of less-than-ideal results because of its sensitivity to initial centroid locations. In order to predict future consumer trends, Businesses must categorize their client in the modern business world according to factors like age, gender, and other attributes. This enables companies to concentrate on certain customers that are most likely to buy their products. If they can successfully apply machine learning to improve their operations, it will provide them a competitive advantage over their rivals. Using the K-means clustering approach in the context of customer segmentation produces informative results that enable businesses to have a thorough grasp of their customer base. Analyzing the results of the aforementioned tests, most machine learning methods perform well; nevertheless, the k-means strategy had the highest likely cluster accuracy rate, at 94.5%.

Keywords: Machine learning, Customer segmentation, competitive, recommendations, clustering.

Introduction

In today's fiercely competitive company world, understanding consumer behavior and preferences is critical for developing effective marketing strategies and sustaining customer happiness. Customer segmentation is a popular strategy for learning about customer behavior. Client segmentation is the process of breaking a large client base into smaller, more manageable groups with comparable characteristics. This allows businesses to target these groups more effectively with personalized products, services, and marketing campaigns. (Kotler, 2017). The K-means clustering algorithm stands out among these techniques as a reliable and adaptable tool for client segmentation, enabling businesses to draw insightful conclusions from their data. Businesses need to leverage customer resources to meet customer requests in highly competitive product landscapes and customize tactics for various consumer segments (Jagani & Chauhan, 2020). In order to facilitate effective enterprise development, it is imperative to conduct a preliminary analysis of the target market's requirements, which is followed by the identification and examination of discrete consumer groups inside the system via customer segmentation.

Customer segmentation is a key component of contemporary marketing tactics. It refers to the process of dividing up diverse clientele into distinct, uniform groups based on characteristics they have in common. This enables businesses to tailor their messaging, exchanges, and offerings to different customer categories, thereby increasing customer satisfaction, brand loyalty, and income. By identifying groups of customers with similar behaviors, interests, and purchase habits, organizations may enhance their targeting, develop customized advertising initiatives, and allocate resources more effectively.

In unsupervised learning, clustering presents a major problem since it focuses on finding patterns in unlabeled data. One of the best analytical techniques in business for comprehending consumer behavior and classifying it is customer segmentation (Chandrashekhar *et al*, 2020). Customers who display comparable behavioral patterns are put together into coherent clusters through the use of clustering algorithms, which aids in strategic decision making (Monil, 2020). One popular data mining technique is cluster analysis, which looks into dataset distribution features to help reach strategic goals.

Clustering is mostly applied to enterprise data analysis, but it is also essential for structuring search results into discrete groups that address different facets of the desired information. In the field of unsupervised machine learning, the K-means clustering algorithm has received significant interest due to its ability to effectively divide data into coherent groups. A dataset is divided into K clusters, and each data point is assigned to the cluster with the closest centroid, according to the basic principle of K-means. Minimizing the sum of squared distances between data points and cluster centroids is the fundamental idea behind the technique.

In-depth analysis of the K-means clustering technique's application to consumer segmentation is provided in this work. It explores the core ideas of the algorithm, clarifies how iteratively it works, and shows how adaptable it is to different kinds of customer data. In addition, it clarifies the benefits of using K-means to segment customers, demonstrating how companies may use this strategy to improve decision-making and cultivate long-lasting client connections. However, as with other analytical instruments, efficient application of the K-means clustering algorithm necessitates a thorough comprehension of its nuances and possible drawbacks. Two important factors that require careful analysis are the algorithm's susceptibility to the initial centroid placements and the crucial job of figuring out the ideal number of clusters.

In the next sections, we explore the K-means clustering algorithm's theoretical foundations, real-world applications, and implementation nuances in relation to customer segmentation. Our goal is to provide businesses and scholars with the knowledge they need to fully utilize K-means for customer segmentation, which will in turn impact their marketing plans and client interaction techniques. We shed light on this method's potential as well as its drawbacks, promoting a thorough comprehension. To improve the effectiveness of the clustering technique, we also examine its pros and cons. Geographical, demographic, psychographic, and behavioral factors are all taken into account when segmenting clients; this allows for a more thorough classification (Monil, 2020).

Literature Review

Due to intense rivalry in the business sector, companies have had to increase their profitability over time in order to match customer expectations and draw in new customers based on their preferences. Customer data can be effectively divided into discrete groups using K-means clustering, where each cluster corresponds to a separate customer group. Customers are categorized by the algorithm according to shared characteristics, actions, or inclinations. These clusters give companies a thorough grasp of their clientele, empowering them to see patterns and distinguish between distinct demographic groupings. For example, a merchant might identify discount-focused, infrequent buyers, and regular shoppers' consumer categories using K-means clustering.

The customer segmentation approach uses data based on several characteristics, such as economic patterns, demographic trends, and behavioral patterns, to separate customers into different categories. A company's marketing resources can be more effectively employed with the aid of a client segmentation strategy (Vijilesh *et al.,* 2021). Through customer segmentation, companies can create discrete groups within their client base according to common traits. Businesses can develop focused marketing efforts by having a thorough understanding of these markets. The likelihood of getting the correct message in front of the right audience is increased by this tailored strategy by using customer segmentation, you may put clients into groups according to traits they have in common, such hobbies, spending patterns, and preferences. Your marketing and sales teams can successfully customize their efforts by being aware of the particular needs of each segment. Customers feel that you are concerned about their demands when you communicate with them according to their particular interests. This individualized strategy builds brand loyalty by promoting recurring business and enduring relationships with your company. Understanding the preferences, needs, and pain points of your customers allows you to create goods and services that precisely meet their needs. Finding trends and opportunities for innovation is facilitated by segmentation. Customizing communications for every group improves the client experience as a whole. Product recommendations, loyalty programs, or tailored mailings are just a few examples of how segmentation makes sure that customers are respected and understood. Rather than distributing resources randomly, companies can concentrate on high-potential markets. Better returns on investment and cost reductions are the results of this efficiency.

Customer segmentation aids in resource allocation optimization, Segments having a higher potential for conversion can be targeted with marketing resources more effectively. This wise distribution maximizes profits while minimizing waste. For instance, a commercial company might assign customer support agents to the group of clients who inquire about services frequently.

As a subset of artificial intelligence, machine learning has transformed several industries, including business, by offering strong tools for task automation, data analysis, prediction, and process optimization. instead of depending exclusively on gut feeling or prior knowledge, machine learning allows organizations to make well-informed judgments by analyzing data. Machine learning algorithms are able to find patterns, trends, and insights that humans would miss by examining vast amounts of data. frequently uses supervised learning to solve issues like regression and classification, suggesting that the data in this instance is targetable and that we wish to plan for in the future, like

determining the monthly cost or the value of a student (Griva *et al*., 2018). Companies can provide recommendations, offers, and experiences that are tailored to each individual consumer by using machine learning algorithms to evaluate customer data and identify preferences, habits, and buying patterns. Higher conversion rates, greater customer happiness, and loyalty are the results of this, k means clustering is one of the machine learning tools that can be use in solve so many problems facing our today businesses and help our businesses in effective business decision making by using historical data, machine learning algorithms are able to predict future patterns, demand, and results. Businesses can save costs and increase efficiency by using this skill to optimize pricing strategies, inventory management, resource allocation, and customer forecasting. Recommendation engines can make personalized suggestions for items, movies, music, or articles are powered by machine learning and are utilized by streaming services, e-commerce platforms, and content websites. This tailored strategy boosts user engagement, boosts revenue, and increases user retention. Machine learning is already being used by numerous shops and other markets to accomplish this. Shopping centers and malls use the data they gather from patrons to build machine learning models that target the right people. This improves business efficiency in addition to revenue and visitor counts.

The application of the K-means clustering approach for consumer segmentation has garnered significant attention from academics and industry professionals in the last several years. In the context of consumer segmentation, this section provides a thorough summary of relevant research that examines the application, expansions, difficulties, and improvements of the K-means clustering technique.

Previous studies set the foundation for using K-means in customer segmentation. K-means was discussed in a landmark study on data clustering methods (Jain et al., 1999), which covered its uses, benefits, and drawbacks in the context of consumer segmentation. Moreover, Ullah *et al.* (2020) provided evidence of the effectiveness of K-means-based segmentation in e-commerce by showing how the algorithm skillfully divides online shoppers into discrete groups according to their shopping habits.

A study by Serpil *et al.* (2020) investigated consumer segmentation based on self-organizing maps using a case study involving airline passengers. Customer satisfaction ratings for firms are directly impacted by the practice of consumer segmentation, which involves grouping customers according to shared characteristics. Gaining a better understanding of customers makes it easier to use the proper techniques to reach the right customers. Airlines need to reevaluate their consumer segmentation tactics in light of the current market dynamics. They should move away from a social-demographic approach and toward a more nuanced behavioral one that takes into account the complete travel experience and contacts with airlines at every touchpoint. This study used data from airline ticket sales to construct a customer segmentation approach that focused on two important factors: customer loyalty and return behavior.

Deep learning and Principal Component Analysis (PCA) were used in a comparative study on dimensionality reduction in telecom consumer segmentation by (El-Bana *et al.*, 2020). Telecom businesses often record user behaviors, creating enormous datasets full of valuable information about user behavior and preferences. But the tremendous sparsity and high dimensionality of these datasets typically present difficulties. By comparing customer clustering in reduced and latent spaces to the original feature space, this study examines dimensionality reduction strategies on a real telecom dataset in an effort to improve clustering accuracy. 220 characteristics, or 100,000 customers, are included in the original dataset.
Dynamic segmentation algorithms have gained attention from researchers due to the temporal dynamics of consumer behavior. (Sheng et al., 2020) proposed a K-means based methodology that takes into account both previous and present purchase behaviors in order to permit dynamic client segmentation across time.

Samber *et al.* (2020) investigated the use of data mining for customer segmentation and clustering in their study. This article explores the growing rivalry between companies trying to keep customers. Through the process of conducting numerous analyses on big, complicated datasets and condensing them into meaningful summaries, data mining emerges as a powerful tool that enables firms to obtain a competitive edge in e-commerce and various industries. Because databases are so large and intricate, managing data in online retailers may be very difficult. Keeping clients comes out as the main goal of this effort.
Industry research has shown that K-means clustering is useful for customer segmentation. (Li et al., 2016) used call detail records and K-means to classify mobile users in the telecom industry, making resource allocation and focused marketing techniques easier. Similar to this, as online commerce platforms have grown in popularity, so has the use of K-means clustering for e-commerce client segmentation. The application of K-means in e-commerce consumer clustering was investigated by Roshan and Chandrashekara (2019), who also looked at the implications for tailored recommendations and marketing tactics.

Onur *et al.* (2020) used process mining visualization approaches and intuitionistic fuzzy clustering to study the segmentation of indoor customer journeys. Numerous studies and approaches have been proposed in the literature to understand the needs and actions of customers at every stage of their journey. All the same, there are issues with path analysis's intricate structure because there are so many different routes that different customers can choose.

Ilaro Journal of Science and Technology (IJST) 2024
Published Online October 2024 (https://sciencetechjournal.federalpolyilaro.edu.ng)
Volume 4, 2024 Edition
ISSN: 2437-1505

To sum up, the collection of relevant studies highlighting the application of K-means clustering for consumer segmentation highlights the technique's ongoing significance and development. Researchers have improved its functions by combining temporal considerations, industry-specific applications, hybrid approaches, and dimensionality reduction techniques. The aforementioned experiments highlight the K-means algorithm's versatility and efficacy in identifying significant customer insights, directing marketing campaigns, and cultivating customized customer experiences.

Methodology

The Himra Shopping store Ilaro provided the dataset for clustering using the K-means technique. It is composed of 225 tuples, each with five properties that represent the data of 225 consumers. Customer ID, age, gender, yearly income (in thousands of dollars), and a spending score ranging from 1 to 100 are some of these attributes.

	Customer ID	Age	Gender	Income (USD)	Spending Score (0-100)
2	1	35	Male	50000	70
3	2	45	Female	60000	65
4	3	30	Male	70000	80
5	4	25	Female	40000	50
6	5	50	Male	80000	75
7	6	55	Female	90000	90
8	7	40	Male	55000	60
9	8	28	Female	48000	55
10	9	38	Male	65000	70
11	10	48	Female	75000	85

Table 1 dataset from Himra store

Finding out what kind of data we'll be dealing with is the first step (see table 1 for the dataset). We use a simple yet complete dataset that includes purchase score, yearly income, gender, age, and customer ID. An expenditure score, which runs from 1 to 100, represents the value of the customer's store purchases or spending. The amount spent increases with the number. The structure of the dataset has been presented accurately, all anomaly removed and no null values are present. Data cleansing is required if a dataset has null values, duplicates, or other noisy data. Data cleansing makes sure that the information is trustworthy, practical, and open for examination. When the data is available, we can compare the gender-specific annual income and spending score to illustrate the data. The study found that groups of consumers that participate in the following activities and customer behaviors associated with annual income and spending scores are represented by five distinct types of plots:

1. Low Spending Score /High Income
2. Low Income / A high expenditure index
3. Despite having low income, a high expenditure score
4. Average Income - Score for Average Spending

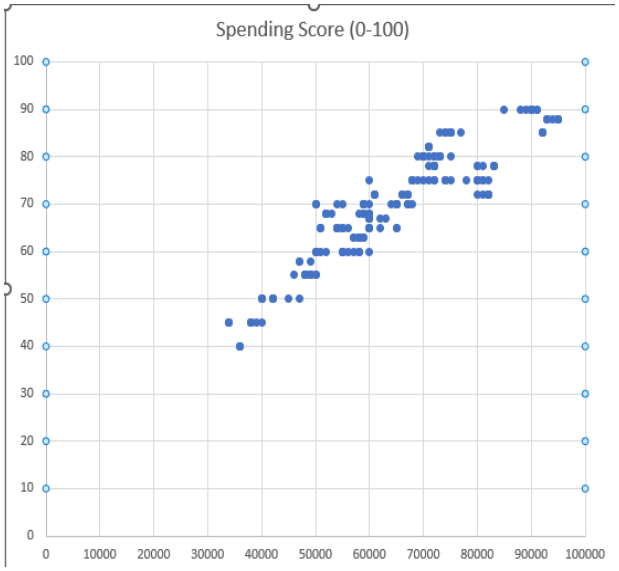
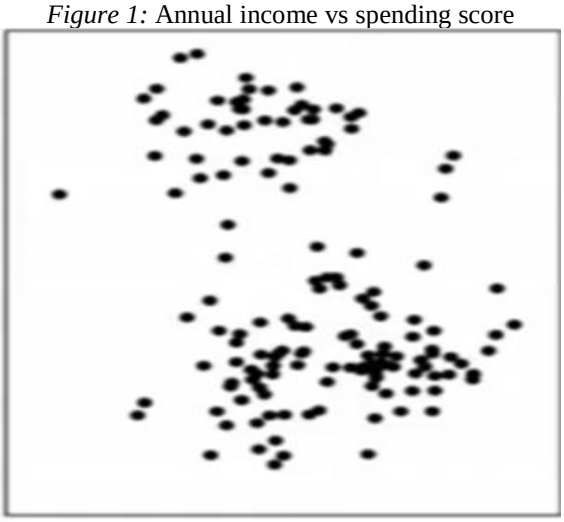


Figure 1: Annual income vs spending score



The fact that there are many groups allows us to now, though not in great detail, construct a K-means model. The silhouette coefficient approach is used to estimate the sum of square distances from each point to its designated center for each value and to perform k-means clustering for a range of k clusters (let's say 1 to 10). Choose how many clusters would provide you with the highest silhouette score. This explains the methodology used to compute the silhouette score. We observed that there is no more quick movement in WCSS (Within Cluster Sum of Squares) once K=4 is reached. Furthermore, K=4 will be the appropriate amount of clusters given the number of clusters we now have.

silhouette score.

2 : 0.389
3 : 0.347
4 : 0.257
5 : 0.299
6 : 0.287
7 : 0.282
8 : 0.325
9 : 0.289

Ilaro Journal of Science and Technology (IJST) 2024
Published Online October 2024 (https://sciencetechjournal.federalpolyilaro.edu.ng)
Volume 4, 2024 Edition
ISSN: 2437-1505

10 : 0.296
The maximum silhouette scores for 4 clusters: 0.299

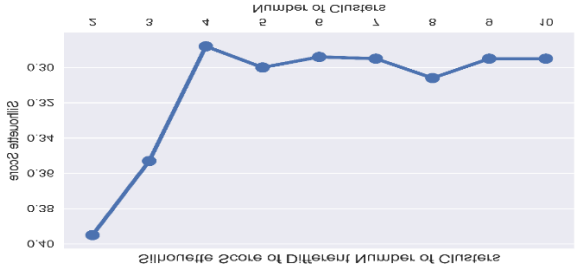


Figure 3: silhouette scores result.

The plot can be divided into different groups, each of which can then be given a label using the previously mentioned procedure. We can then decide which cluster should be prioritized. Which of the five clusters—clients with Moderate Income-Moderate Spending Score, High Income-High Spending Score, and Low Income-High Spending Score—should be targeted can be determined using the K-means approach. The necessary customers have been found.

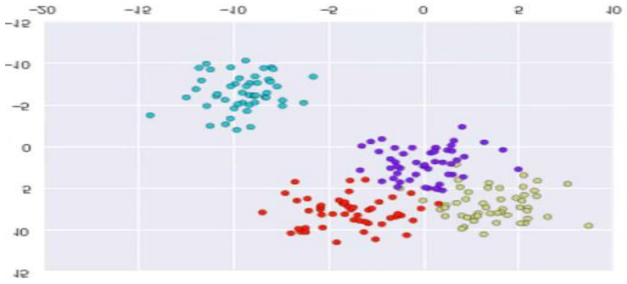


Figure 4: Final customer cluster

Results from the analysis

Four categories can be used to categorize mall patrons based on their annual income and purchasing patterns. First off, the grey group is made up of individuals with high incomes and spending scores; this is a fantastic example of a mall or other retail establishment that would be a desirable target. since they are the most lucrative clients. This individual might be a regular customer at a mall, where mall security could easily catch them.

Conversely, the light blue group is made up of people who are extremely wealthy but barely make any purchases. The fact that there are several reasons why a club of this kind developed makes this an interesting case. Assume that although they enjoy shopping, they are not happy with the mall's current amenities or offerings. These are also worthwhile goals, but we'll need to ascertain why they're not spending as much. The department manager or the authorities of the mall may create a facility that would draw these groups in and meet their demands.

The red group illustrates their average earnings and expenses based on the information we currently have. We can assume that these are people who don't always purchase things but who, in spite of their limited resources, have a great desire to spend. As a manager, my goal is to steer clear of marketing tactics that specifically target this demographic as much as possible, as they don't contribute significantly to the mall's bottom line. However, they might use a variety of data analysis methods to support their increased spending.

The violet-colored group consists of individuals with low incomes but high spending scores; in spite of their low incomes, members of this group find enjoyment or interest in making purchases. This could also occur if patrons of the mall are satisfied with the services provided and are hence inclined to make purchases.

Based on the Annual Income and Spending Score of the consumer, we may forecast their behavior by evaluating the data. Numerous customer marketing strategies can benefit from the application of this cluster analysis. Because they generate the largest profit margin, we would prefer to retain our target audience, which consists of people with high incomes and high spending scores. The large selection of goods at the Mall Supermarket will draw clients in due to their lifestyle expectations for a high income and low spending score. Less Income Less Spending Scores are more likely to receive promotions and discounts, which will encourage them to make purchases. To find out what kinds of products and services customers want to buy, a cluster analysis can be performed, which will help marketers focus their efforts. In this case, the individuals in clusters 3 and 4 are the possible clients. Using the K-means clustering approach in the context of customer segmentation produces informative results that enable businesses to have a thorough grasp of their customer base. Analyzing the results of the aforementioned tests, most machine learning methods perform well; nevertheless, the k-means strategy had the highest likely cluster accuracy rate, at 94.5%. Machine learning techniques become useful tools for data mining in large noisy databases when applied to marketing difficulties. These methods improve the accuracy of future and forecasting models used to attract crowds to the performing arts and provide researchers with additional avenues for studying consumer preferences.

Conclusion

This study demonstrates that customer segmentation in shopping malls is feasible, even though this kind of machine learning application is highly beneficial in the market. Each recognized cluster can receive all of a manager's attention and needs when they are fully attended to. Mall managers need to know what the demands of their patrons are and, more significantly, how to satisfy them. To meet their needs, examine their purchase patterns and create routine interactions with clients that they find comfortable. Utilizing the K-means clustering algorithm for client segmentation is a revolutionary approach in contemporary marketing strategies. A detailed analysis of the K-means algorithm's methodology, advantages, drawbacks, and associated literature makes it abundantly evident that this algorithm has the power to fundamentally

alter how organizations perceive, engage with, and cater to their wide range of customers.

K-means clustering for consumer segmentation Debates have demonstrated that the algorithm's ease of use and adaptability are essential to ensuring its broad acceptance. Its ability to manage data of all types, sizes, and shapes ensures its adaptability to a wide range of enterprises and scenarios. Because of its ease of use and efficiency in computing, it may be applied to both short- and long-term decision-making processes. This algorithm's continued relevance and evolution are demonstrated by the fact that it is still being researched and employed. The literature demonstrates how adaptable K-means is to complex segmentation scenarios. It encompasses both earlier studies that investigated dynamic and industry-specific applications and foundational academic efforts that created the framework.

References

Chandrashekhar, Y., Alexander, T., Mullasari, A., Kumbhani, D. J., Alam, S., Alexanderson, E. & Narula, J. (2020). *Resource and infrastructure-appropriate management of ST-segment elevation myocardial infarction in low-and middle-income countries. Circulation, 141(24), 2004-2025.*

El-Bana, S., Al-Kabbany, A., & Sharkas, M. (2020). A multi-task pipeline with specialized streams for classification and segmentation of infection manifestations in COVID-19 scans. *PeerJ Computer Science, 6, e303.*

Griva, A., Bardaki, C., Pramatari, K., & Papakiriakopoulos, D. (2018). Retail business analytics: *Customer visit segmentation using market basket data. Expert Systems with Applications, 100, 1-16.*

Jagani, K., Oza, F. V., & Chauhan, H. (2020). Customer Segmentation and Factors Affecting Willingness to Order Private Label Brands: *An E-Grocery Shopper's Perspective. In Improving Marketing Strategies for Private Label Products (pp. 227-253). IGI Global.*

John MacQueen (1967). A few techniques for categorizing and analyzing multivariate observations. *Pages. 281-297 in Volume 1, No. 14, "Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability."*

Kotler, P. (2017). Marketing Management (15th ed.). Pearson.

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

Monil, P., Darshan, P., Jecky, R., Vimarsh, C., & Bhatt, B. R. (2020). Customer segmentation using machine learning. *International Journal for Research in Applied Science and Engineering Technology (IJRASET), 8*(6), 2104-2108.

Onur Dogan, Basar Oztaysi and Carlos Fernandez-Llatas (2020). "Segmentation of Indoor Customer Paths Using Intuitionistic Fuzzy Clustering", *Process Mining Visualization; Journal of Intelligent & Fuzzy Systems 38 (1), 675-684.*

Samber, D. D., Ramachandran, S., Sahota, A., Naidu, S., Pruzan, A., Fayad, Z. A., & Mani, V. (2020). Segmentation of carotid arterial walls using neural networks. *World Journal of Radiology*, *12*(1), 1.

Serpil Ustebay, İlkay Yelmen and Metin Zontul (2020). "Customer Segmentation Based onSelf- Organizing Maps: A Case Study on Airline Passengers", *Journal of Aeronautics & Space Technologies/Havacilik ve Uzay Teknolojileri Dergisi 13 (2).*

Sheng, W., Wang, X., Wang, Z., Li, Q., Zheng, Y., & Chen, S. (2020). A differential evolution algorithm with adaptive niching and k-means operation for data clustering. *IEEE Transactions on Cybernetics*, *52*(7), 6181-6195.

Vijilesh, V., Harini, A., Dharshini, M. H., & Priyadharshini, R. (2021). Customer Segmentation using Machine Learning. *International Research Journal of Engineering and Technology (IRJET), 8*(05), 821-825.

Ullah, I., Boreli, R., & Kanhere, S. S. (2020). Privacy in targeted advertising: A survey. *arXiv preprintarXiv:2009.06861*.