Using K-Means Clustering to Customer Segmentation

Le Van Sy and Do Cong Chi

University of Information Technology, Linh Trung, Thu Duc, Ho Chi Minh

Abstract: In today's age of competition and innovation, need to adapt their corporate strategy to current conditions. With countless potential customers uncertain about their purchasing decisions, novel ideas have become the lifeblood of successful businesses. However, businesses often find it difficult to assess their target market independently. This is where machine learning comes into play, using a variety of algorithms to uncover hidden patterns in data and facilitate better decision-making. One such machine learning technique is clustering, which involves comparing data points from different groups. Its applications span a wide variety of fields, including market research, medical data analysis, search optimization, pattern recognition, and image processing. In market research, an important aspect is customer segmentation, which entails classifying consumers into groups based on common characteristics. In today's business landscape, it is important for companies to segment their customers based on factors such as age, gender, geography, and other relevant attributes. This allows businesses to focus their efforts on specific customer segments most likely to buy their products, giving them a competitive edge over competitors. The main goal of this project is to use the K-means algorithm to group customers into groups based on their attributes. Using the mean as a primary indicator, we can determine which grouping new customers will likely be in, based on data from different clusters. Through the successful implementation of machine learning techniques, businesses can gain valuable insights to enhance their operations and gain a competitive edge in the marketplace.

Keywords: Customer segmentation, Machine learning, Python, K-means algorithm

1 INTRODUCTION

To remain competitive in the face of increasing competition from new businesses, established companies must adopt effective marketing tactics. In today's dynamic society, the rule of thumb for marketing is "adapt or fade away." As the customer base expands, businesses are facing greater challenges in meeting the diverse demands of their consumers. To address this issue, data mining emerges as a valuable tool for uncovering hidden patterns within a company's database. One such approach is client segmentation, which involves dividing the customer base into distinct groups based on various characteristics such as gender, age, interests, and purchasing habits [1]. Essentially, customers are grouped together based on shared qualities. This segmentation can directly or indirectly influence marketing strategies by revealing new avenues of exploration. For example, it can identify the ideal segment for a particular product, facilitate the customization of marketing plans for each segment, offer targeted discounts, and unveil previously unknown relationships between customers and products [2]. Implementing a customer segmentation strategy allows companies to focus on specific consumer groups, leading to more efficient allocation of marketing resources and increased potential for cross-selling and upselling. When customized communications are tailored to meet the unique needs of a particular group as part of a tailored marketing mix, companies find it easier to create enticing offers that encourage customers to spend more. Additionally, customer segmentation can enhance customer loyalty and retention by improving the overall customer service experience. Personalized marketing materials that employ customer segmentation are more valued and appreciated by consumers compared to impersonal brand communications that overlook purchase history or fail to acknowledge existing customer relationships [3]. Customer segmentation greatly benefits from clustering techniques, which fall under the realm of unsupervised learning and allow for the discovery of clusters within unlabeled datasets. Popular clustering methods include K-means, hierarchical clustering, DBSCAN clustering, and others. The primary objective of this approach is to employ data mining strategies, specifically the K-means clustering algorithm, to identify consumer groups by partitioning the data. The silhouette method proves to be particularly effective in determining the optimal number of clusters.

1.1. Customer Segmentation

Due to intense competition in the business field, companies have faced the need to enhance their profitability and expand their customer base over time. Meeting customer expectations and attracting new clients requires a significant effort, as it is challenging and time-consuming to identify and address the unique needs of everyone. This difficulty arises from the fact that customers have diverse goals, interests, and preferences. To address this complexity, customer segmentation has emerged as a valuable alternative to the traditional "one-size-fits-all" approach. By grouping customers based on similar characteristics or behaviors, customer segmentation allows for a more tailored marketing strategy. It involves dividing the market into distinct, homogeneous groups, enabling companies to target their efforts more effectively. Customer segmentation relies on various factors, including regional circumstances, economic patterns, demographic trends, and behavioral patterns. These factors provide the basis for categorizing customers into specific groups. By employing a customer segmentation technique, companies can optimize their marketing resources and improve their overall effectiveness [4].

1.2. Machine Learning

Machine learning plays a significant role in various industries, including prominent companies like Facebook and YouTube. Facebook utilizes machine learning to help us recognize ourselves and our friends, while YouTube employs it to assist us in discovering new content. Additionally, machine learning is categorized into two types: supervised learning and unsupervised learning.[5] Supervised learning is employed to tackle problems such as classification and regression. In this approach, the data is labeled or targeted, enabling us to make future predictions. For instance, it can be used to evaluate a student's performance or estimate monthly expenses.

On the other hand, unsupervised learning does not necessarily require predefined labels or specific objectives. Instead, it is based on mathematical models, like clustering, which aims to identify patterns or groupings within the data. For example, we can cluster students based on their learning interests or analyze customer purchase behavior. In the marketing industry, especially in malls, there is fierce competition to attract more customers and generate substantial profits.[6] To achieve this goal, many retailers and marketplaces are already leveraging machine learning. Malls and shopping centers gather customer data to develop machine-learning models that precisely target the right individuals. This not only enhances revenue and visitor numbers but also improves overall business efficiency.

2 MATERIALS AND METHODS

2.1 Clustering

Clustering is an approach used to identify comparable groups within a vast dataset. Each group consists of members who share more similarities with each other compared to members of other groups. Cluster-based segmentation has gained popularity in data analysis, especially in the field of marketing, since the 1970s. However, it's important to note that clustering is not a systematic data analysis approach. Its effectiveness heavily relies on the specific data or sample being utilized, as previously noted. One statistical strategy employed in cluster analysis investigations is the "tandem technique," which combines factor analysis and cluster analysis as two sequential processes. However, this approach has faced criticism due to a fundamental flaw: early factor analysis can potentially erase existing cluster formations. In cases involving binary variables, hierarchical cluster analysis may be used as an alternative to the tandem cluster analysis method.[7]

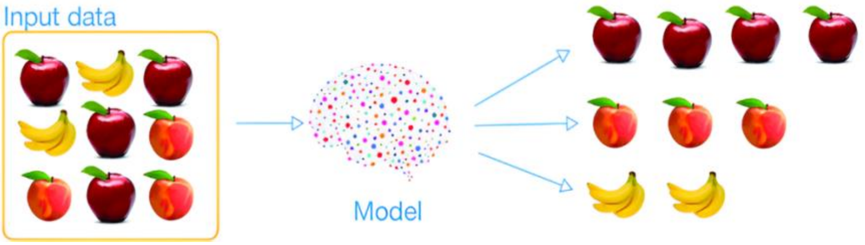


Figure 1: Clustering

2.2 Customer Segmentation and Machine Learning

One approach to client segmentation involves the utilization of machine learning algorithms. By employing machine learning for consumer segmentation, marketers can unlock valuable insights and uncover groups that may otherwise be challenging to identify. Establishing a feedback loop between the segmentation model and campaign outcomes allows marketers to witness continuous improvement in their customer groups. This iterative process enables the machine learning model to refine segment definitions and assess which segments outperform others, ultimately maximizing marketing effectiveness. [8]

2.3 Customer Segmentation using K-Means Clustering

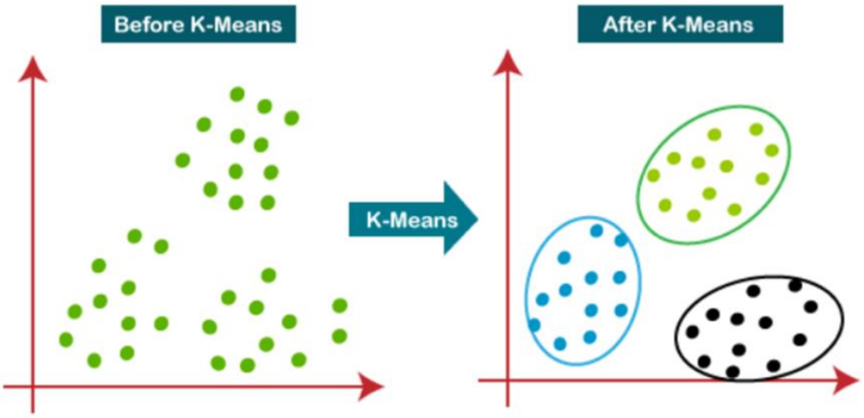


Figure 2 Using K-Means Clustering to customer segmentation.

K-Means Clustering stands out as the most popular unsupervised partitioning method for clustering. This technique, often referred to as the centroid-based approach, categorizes data into distinct non-hierarchical groups [8]. In this type of partitioning, the dataset is divided into a set of k groups, where k represents the predetermined number of clusters. The cluster center is constructed in such a way that the distance between data points within a cluster is minimized when compared to other cluster centroids.

2.4 Algorithm

Step-1: Select the number K to decide the number of clusters.

Step-2: Select random K points or centroids. (It can be other than the input dataset).

Step-3: Assign each data point to the closest centroid, forming K clusters**.**

Step-4: Calculate the new centroids by taking the mean of the data points in each cluster.

Step-5: Repeat steps 3 and 4 until convergence or a maximum number of iterations is reached.

Step-6: If convergence has not been reached, go back to step 5..

Step-7: The model is ready.

The flowchart below shows how k-means clustering works.

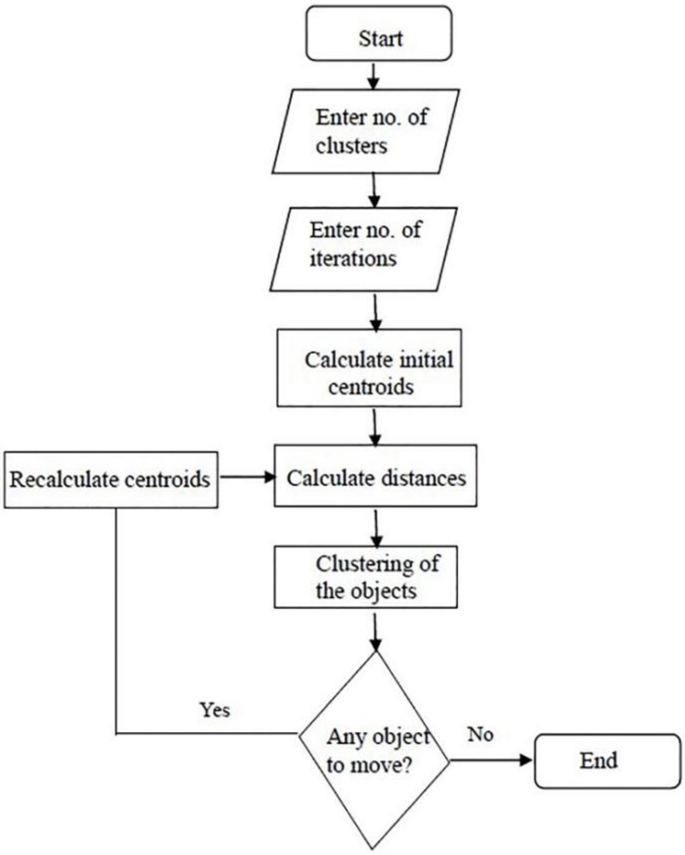


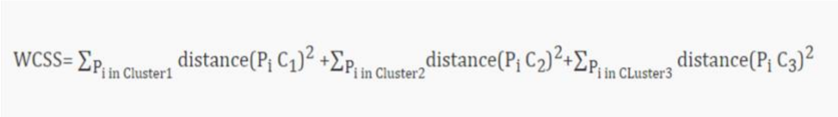
Figure 3: Flow of K-Means algorithm

To begin we need to find K. The number of clusters needs to be specified before running the algorithm. Choosing an appropriate value for K is often a subjective decision based on domain knowledge or through techniques like the elbow method, silhouette analysis, and gap statistic method.

Here we will use the elbow method:

Elbow Method: The Elbow Method is a technique used in data clustering analysis to determine the optimal number of clusters in a dataset. It involves plotting the variance explained by the clusters against the number of clusters. The "elbow" in the plot represents a point of diminishing returns, where the addition of more clusters does not significantly improve the clustering performance. This point is considered the optimal number of clusters for the given dataset.

The Elbow method is one of the most popular ways to find the optimal number of clusters. This method uses the concept of WCSS value. **WCSS** stands for **Within Cluster Sum of Squares**, which defines the total variations within a cluster. The formula to calculate the value of WCSS (for 3 clusters) is given below:



$$\text{WCSS}= \sum_{P_i \text{ in Cluster1}} \text{distance}(P_i\ C_1)^2 + \sum_{P_i \text{ in Cluster2}} \text{distance}(P_i\ C_2)^2 + \sum_{P_i \text{ in CLuster3}} \text{distance}(P_i\ C_3)^2$$

In the above formula of WCSS,

$\sum_{P_i \text{ in Cluster1}}$ distance $(P_i\ C_1)^2$: It is the sum of the square of the distances between each data point and its centroid within a cluster1 and the same for the other two terms.

To measure the distance between data points and the centroid, we can use any method such as Euclidean distance or Manhattan distance.

To find the optimal value of clusters, the elbow method follows the below steps:

Step 1. Choose a range of values for the number of clusters (K) that you want to evaluate.

Step 2. Perform the K-means clustering algorithm for each value of K and calculate the within-cluster sum of squares (WCSS) or variance.

Step 3. Calculate the WCSS value for each value of K.

Step 4. Plot a curve of the WCSS values against the number of clusters (K).

Step 5. Identify the elbow point on the plot, which represents a trade-off between a low number of clusters (underfitting) and a high number of clusters (overfitting), and consider it as the optimal number of clusters.

3 RELATED WORK

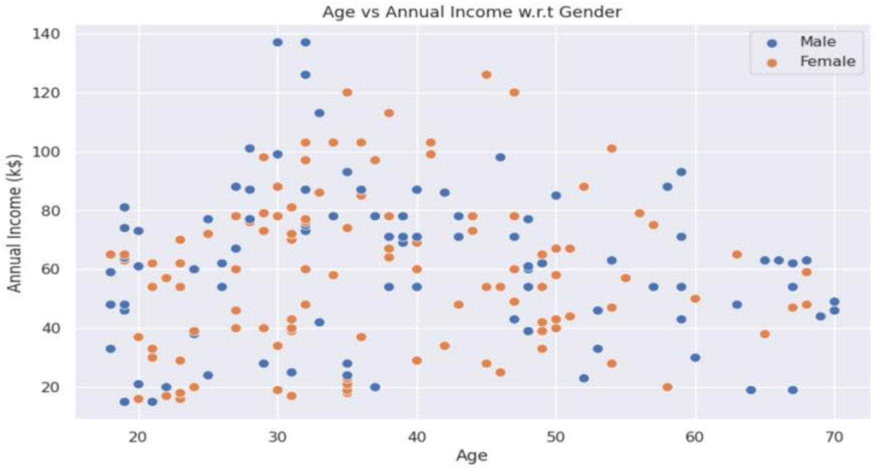
This research utilized data from an online retail store to demonstrate the application of the k-Means clustering technique for consumer segmentation. In this model, customers were classified into distinct and non-overlapping groups, specifically three clusters in this case.

The authors opted for internal clustering validation instead of external clustering verification due to the imbalanced nature of the dataset. External clustering verification relies on external data, such as labels. On the other hand, internal cluster validation enables the selection of the most suitable clustering algorithm for the dataset and ensures appropriate data clustering within the clusters. [9]Machine learning approaches excel in evaluating customer data and revealing meaningful patterns and insights. Artificially intelligent models serve as powerful decision-making tools, enabling precise definitions of client categories, a task that is considerably more challenging to accomplish manually or with traditional analytical methods. Various flavors of machine learning algorithms exist, each tailored to specific contexts. In this study, the model is constructed using the widely adopted k-means clustering technique, a popular algorithm for addressing client segmentation challenges.[10]

4 MODEL

Outline of Existing Model

The original base paper employs the "Elbow method" to determine the optimal number of clusters for K-means clustering. However, the Elbow method may not always yield effective results, as illustrated by the following scatter plot.



Although humans can identify five distinct clusters in the data, perceiving such patterns becomes challenging in high-dimensional data. In the case presented on the left, the Elbow Method would likely suggest k = 3 and k = 5, erroneously merging two clusters due to their proximity.

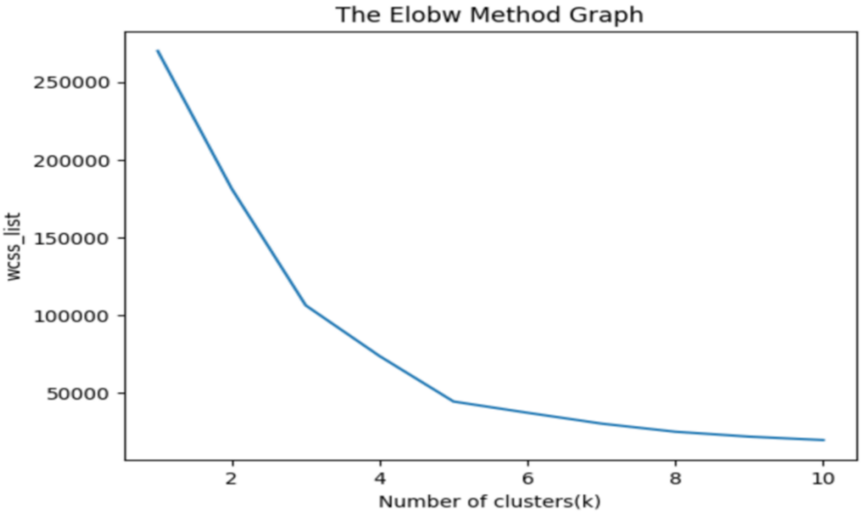
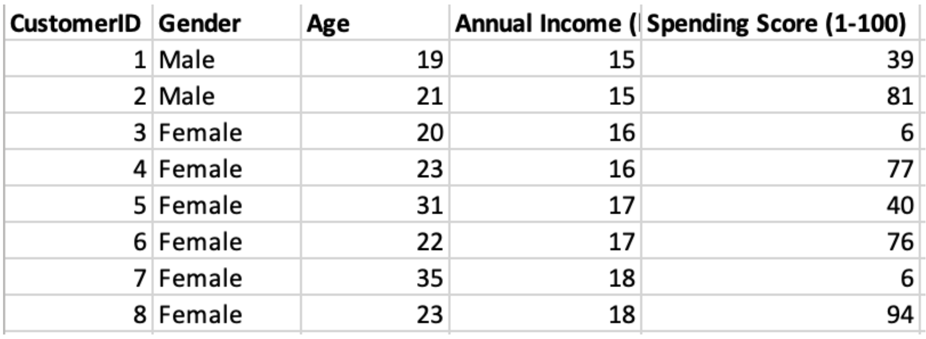


Figure 4: Elbow plot

This occurs because placing a centroid between these clusters diminishes the relative distance between data points. Consequently, a more accurate, rigorous, and dependable approach is required to determine the appropriate number of clusters for our clustering task. Currently, the silhouette score is utilized for this purpose.

5 METHODOLOGY

A dataset provided by a store in a shopping center was used for clustering using the K-means algorithm. The dataset consists of 200 tuples representing information about 200 consumers, with five attributes. These attributes include CustomerId, gender, age, yearly income (k$), and spending score on a scale of 1-100.



CustomerID	Gender	Age	Annual Income (	Spending Score (1-100)
1	Male	19	15	39
2	Male	21	15	81
3	Female	20	16	6
4	Female	23	16	77
5	Female	31	17	40
6	Female	22	17	76
7	Female	35	18	6
8	Female	23	18	94

To get started, our first step is to determine the type of data we will be working with (refer to Table 1 for the dataset). We utilize a comprehensive dataset that includes customer ID, gender, age, yearly income, and purchase score. The expenditure score, ranging from 1 to 100, represents the value of a customer's shopping or spending at the mall (where higher scores indicate greater amounts spent). The dataset has been properly formatted, and no null values are present.

In case a dataset contains null values, duplicates, or other noisy data, it is necessary to perform data cleaning. Data cleansing ensures that the information is reliable, usable, and available for analysis. Once we have the clean data, we can visualize it by comparing the annual income and spending score, considering gender. Based on the study, there are five distinct types of plots that demonstrate groups of customers engaged in various activities. Additionally, these plots reveal customer behaviors associated with yearly income and expenditure scores

.

- Cluster 1: Shows the customers with average salary and average spending so we can categorize these customers as
- Cluster 2: Shows the customer has a high income but low spending, so we can categorize them as careful.
- Cluster 3: Shows the low income and also low spending so they can be categorized as sensible.

- Cluster 4: Shows the customers with low income with very high spending so they can be categorized as careless.
- Cluster 5: Shows the customers with high income and high spending so they can be categorized as targets, and these customers can be the most profitable customers for the mall.

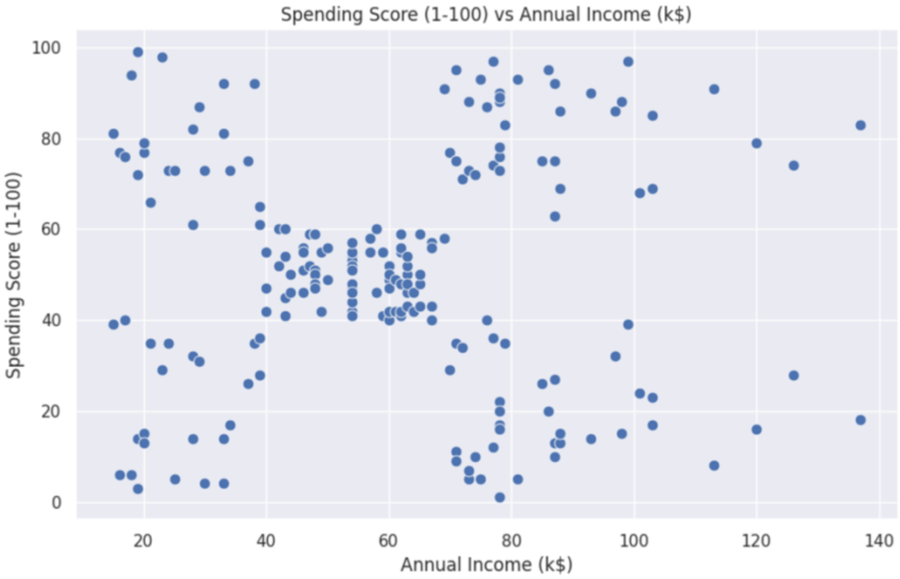


Figure.5. Annual Income vs Spending Score

We are now able to construct a K-means model considering the presence of numerous groups, although not with extensive specifics. The clustering task utilizing k-means employs the silhouette coefficient approach, which involves experimenting with various numbers of clusters (from 1 to 10, for instance). For each value, we estimate the sum of squared distances between each data point and its corresponding cluster center. By evaluating the silhouette scores, we can determine the optimal number of clusters that yields the highest score.

We can divide the plot into various groups, determine clusters can be prioritized, and then assign a label to each using the method stated above. The K-means approach can be used to decide which of the five clusters should be targeted, namely clients with Moderate Income- Moderate Spending Score, High Income- High Spending Score, and Low Income- High Spending Score. The required consumers have been located, as shown in Figure 6.

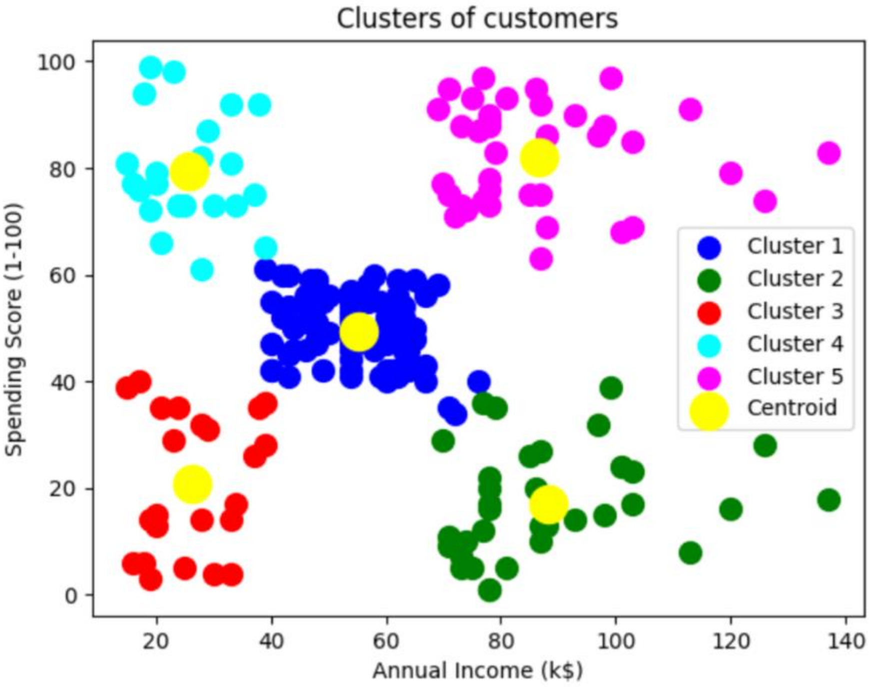


Figure.6. Final cluster of customers

6 EXPERIMENT RESULTS

There is a group known as the blue group, consisting of individuals who possess substantial wealth but spend very little. This presents an interesting scenario with multiple factors contributing to the formation of such a group. Presumably, these individuals enjoy shopping but find the current offerings and facilities at the mall unsatisfactory. They are potential targets for engagement, but it is crucial to understand the reasons behind their limited spending. The mall manager or authorities could consider designing or constructing a facility that would attract this group and fulfill their needs.

Contrary to the blue group, the orange group represents individuals with average earnings and expenditures. These individuals may not always make purchases but have a strong inclination to spend, despite financial constraints. As a manager, it is important to minimize marketing strategies targeting this group as they do not significantly contribute to the mall's revenue. However, employing various data analysis techniques may help increase their spending habits.

Next, we have the violet-colored group, comprising people with low incomes but high spending scores. Despite their limited financial resources, individuals in this group are interested in spending money. This may be attributed to their satisfaction with the mall's services, prompting them to spend due to a positive shopping experience.

The fifth group, represented by green, consists of individuals with low annual incomes and poor spending habits. It is reasonable to assume that these individuals have a restricted budget and seek cost-saving measures wherever possible, even if it means making sensible decisions based on their circumstances. Mall managers should prioritize this group the least.

7 CONCLUSION

This research showcases the feasibility of client segmentation within shopping malls, even though this machine-learning approach holds significant value in the market. By identifying distinct clusters of customers, mall managers can devote their undivided attention to each group and fulfill their specific needs. It is crucial for mall managers to comprehend customer requirements and, more importantly, learn how to address them effectively. They should analyze customers' purchasing patterns, establish regular interactions, and create an environment that ensures customer comfort, ultimately meeting their demands.

References

[1] “Customer segmentation based on survival character,” IEEE, Jul. 2003.

[2] “Customer Segmentation Using K Means Clustering,” Towards Data

[3] Tim Ehrens, “Customer segmentation,” TechTarget.

[4] V. Vijilesh, “CUSTOMER SEGMENTATION USING MACHINE LEARNING,” International Research Journal of Engineering and Technology (IRJET), vol. 08, no. 05, May 2021.

[5] Expert Systems with Applications, vol. 100, Feb. 2018, "Retail Business Analytics: Customer Visit Segmentation Using Market Basket Data."

[6] “Cluster analysis.”, Wikipedia.

[7]"CUSTOMER SEGMENTATION USING MACHINE LEARNING," IJCRT, AMAN BANDUNI and ILAVENDHAN A, vol. 05, 2018.

[8] Author Dhiraj Kumar, “Implementing Customer Segmentation Using Machine Learning [Beginners Guide],” Neptune blog, Dec. 13, 2021.

[9] AMAN BANDUNI and ILA VENDHAN A, “CUSTOMER SEGMENTATION USING MACHINE LEARNING,” IJCRT, vol. 05, 2018.

[10] Dhiraj Kumar, “Implementing Customer Segmentation Using Machine Learning”, July 10th, 2021.