Customer Segmentation Analysis Using K-Means Clustering Algorithm

Saleti Sumalatha
Assistant Professor
Computer Science and Engineering
SRM University AP
sumalatha.s@srmap.edu.in

Neelofar Shaik
Computer Science and Engineering
SRM University AP
neelofar_shaik@srmap.edu.in

***Abstract*—In today's era of advanced technology, the landscape of retail is evolving rapidly, driven by sophisticated machine learning algorithms and data analytics tools. This project delves into the transformative potential of customer segmentation in the context of modern technology, with a particular focus on leveraging the power of KMeans Clustering. Harnessing the wealth of customer data available to businesses, we explore how customer segmentation can revolutionize the way companies engage with their clientele. Through the lens of machine learning, we uncover hidden patterns within our data, enabling us to segment customers into distinct groups based on their preferences, behaviors, and purchasing habits. With the aid of cutting-edge algorithms and analytical techniques, customer segmentation becomes not only feasible but also remarkably intuitive and efficient. Armed with actionable insights derived from our segmentation analysis, businesses can craft targeted marketing strategies that resonate with specific customer**

segments, leading to enhanced customer satisfaction, increased engagement, and ultimately, greater business success. In conclusion, leveraging the KMeans clustering algorithm has provided us with invaluable insights into customer behavior and preferences. By identifying distinct customer groups, businesses can tailor their marketing efforts, gain a deeper understanding of their customer base, and drive strategic decision-making to new heights.

Keywords:- Customer Segmentation; Mall Customers; Machine learning; K-Means Clustering; Elbow method; Silhouette method, Python.

I. Introduction

your offerings with their needs and desires is the key to unlocking meaningful success." This perspective underscores the importance of customer-centric approaches in modern business practices. In the era of technology-driven retail, where businesses harness advanced tools and methodologies to gain profound insights into their customer base. Central to this transformation is the concept of customer segmentation, a strategic approach that entails categorizing customers into distinct groups based on shared characteristics or behaviors. It plays a pivotal role in deepening the understanding of the intricate relationship between customers and products within a market. By segmenting their customer base, businesses can tailor their products, services, and marketing strategies to meet the specific needs and preferences of each segment. This targeted approach not only enhances customer satisfaction but also maximizes efficiency and effectiveness in resource allocation. At the heart of customer segmentation lies the KMeans clustering algorithm, a powerful tool that enables businesses to analyze vast amounts of customer data and identify meaningful patterns and clusters. By leveraging KMeans clustering, businesses can uncover hidden insights within their data, enabling them to make informed decisions and drive strategic initiatives. Moreover, customer segmentation facilitates cross-selling and upselling initiatives by enabling businesses to identify complementary products or services that align with the preferences of each segment. By analyzing purchasing behavior across segments, businesses can strategically bundle products or recommend related items, thereby maximizing the value of each customer interaction and driving revenue growth. Furthermore, the ability to tailor products and services according to the preferences of each segment allows firms to meet the specific needs of their customers more effectively. This customization enhances customer satisfaction, fosters brand loyalty, and cultivates long-term relationships, ultimately benefiting the bottom line. In our analysis, we utilized the KMeans clustering algorithm alongside the Elbow and Silhouette methods for customer segmentation. KMeans partition data into K clusters based on similarity, while the Elbow Method identifies the optimal cluster number, and the Silhouette Method assesses clustering quality. These techniques ensure precise segmentation, enabling informed decisions, targeted marketing, and personalized customer experiences.

II. Literature Review

Customer segmentation is a widely studied topic in contemporary business literature, reflecting its pivotal role in shaping marketing strategies and driving business success. Numerous studies underscore the importance of understanding customer

behavior and preferences to effectively target marketing efforts and improve overall customer satisfaction. Studies by Smith and Wind (2017) and Kumar and Reinartz (2018) highlight its importance in optimizing marketing campaigns and resource allocation. The adoption of machine learning algorithms, particularly KMeans clustering, has emerged as a notable trend in customer segmentation research. KMeans clustering in identifying distinct customer segments based on various demographic and behavioral attributes. Furthermore, Zhao et al. (2019) discuss the importance of employing techniques such as the Elbow Method and Silhouette Method to determine the optimal number of clusters for segmentation purposes. Overall, customer segmentation, particularly with KMeans clustering, enables businesses to gain insights, refine strategies, and enhance competitiveness in the market.

III. Dataset Description

The dataset ”Mall_Customers.csv” is a comma-seperated file containing information that is collected from customers at a retail mall. It consists of the following attributes

	CustomerID	Gender	Age	Annual Income (k$)	Spending Score (1-100)
0	1	Male	19	15	39
1	2	Male	21	15	81
2	3	Female	20	16	6
3	4	Female	23	16	77
4	5	Female	31	17	40

Fig. 1. Dataset.

1. **CustomerID:** - Unique identifier assigned to each customer.
2. **Gender:** - Gender of the customer (Male/Female).
3. **Age:** - Age of the customer in years.
4. **Annual Income:** - Annual income of the customer in thousands of dollars
5. **Spending Score (1-100)** - A score assigned to the customer based on their spending behavior and purchasing data, ranging from 1 to 100

This dataset provides insights into the demographic characteristics, income levels, and spending behaviors of the mall customers. It can be utilized for various analyses, including customer segmentation, market research, and targeted marketing strategies.

Below is a brief description of the key features present in the dataset:-

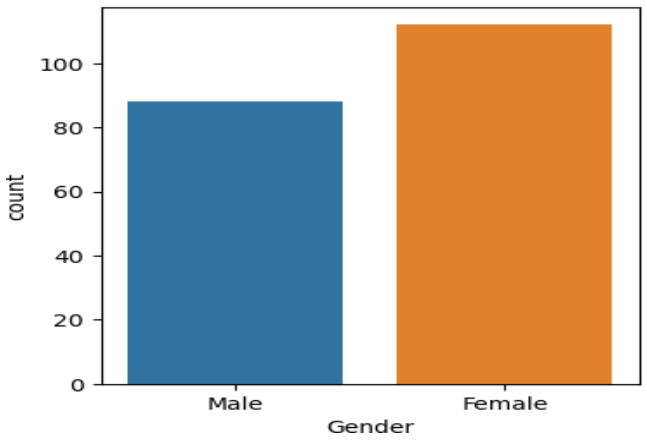

Fig. 2. Gender

Gender: Gender of the customer (Male/Female).

Fig. 3. Male and Female customers

Information on the “distribution of male and female customers” at the retail mall.

Presents the distribution of numerical features, including “age, annual income, and spending score”, among customers at the retail mall.

Showcases the distribution of numerical features, “including annual income, spending score, and age”, among customers at the retail mall.

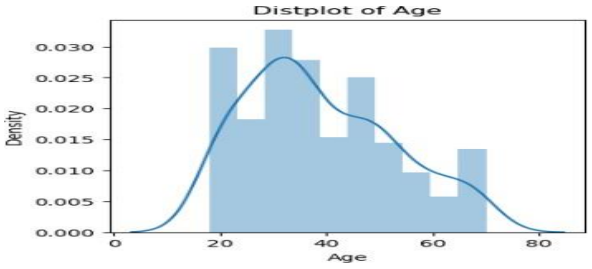

Fig. 4. Age vs Density.

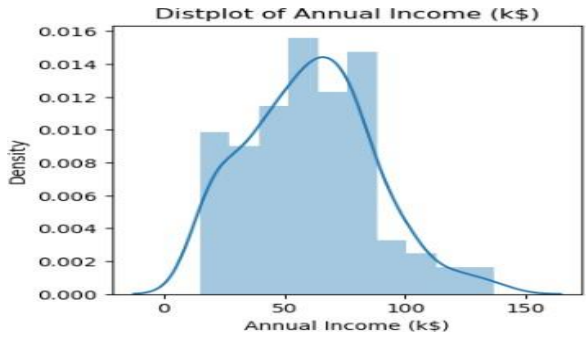

Fig. 5. Annual Income vs Density.

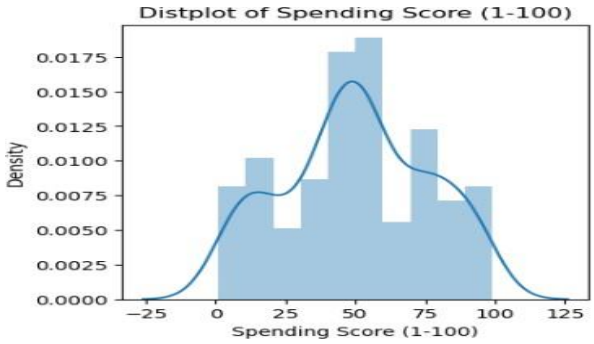

Fig. 6. spending sccore vs Density.

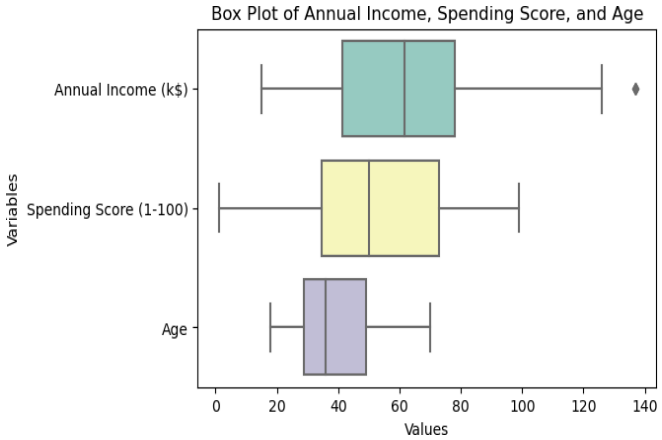

Fig. 7. Box Plot of Annual Income, Spending Score, and Age.

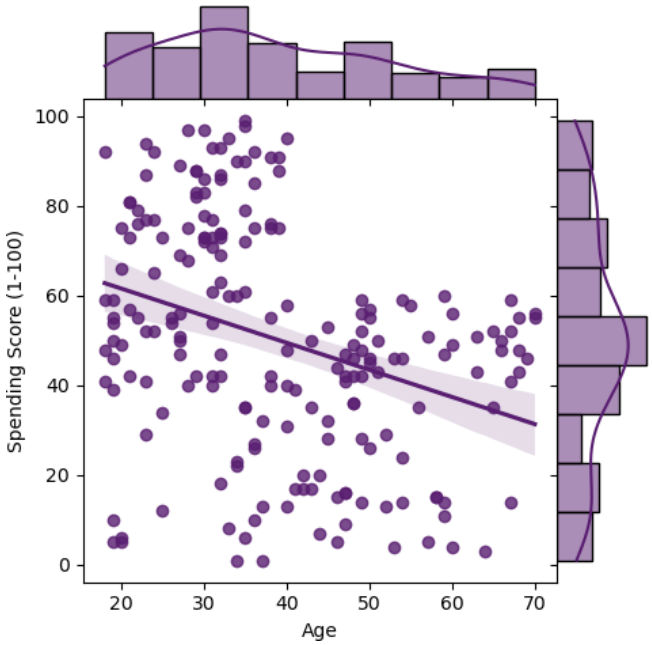

Fig. 8. Age vs Spending score(1-100).

Represents the relationship between two numerical features, “age and spending score”, among customers at the retail mall.

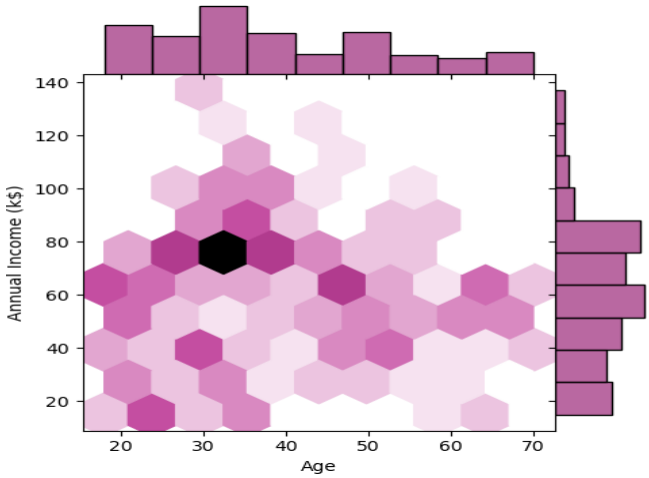

Fig. 9. Age vs Annual Income

The joint plot visualizes the relationship between two numerical features, “age and annual income”, among customers at the retail mall.

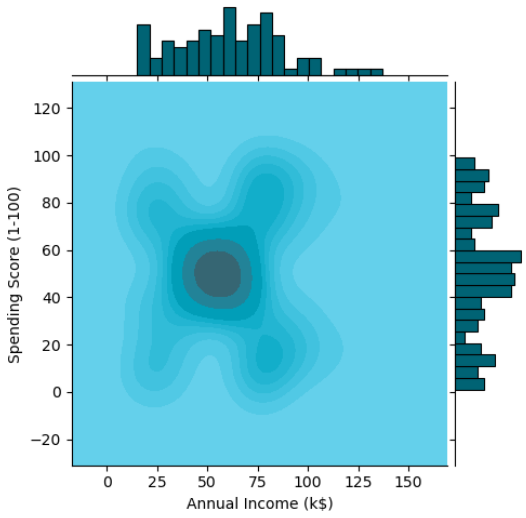

Fig. 10. Spending vs Annual Income

The JointGrid visualization presents the relationship between "annual income and spending score" among customers at the retail mall.

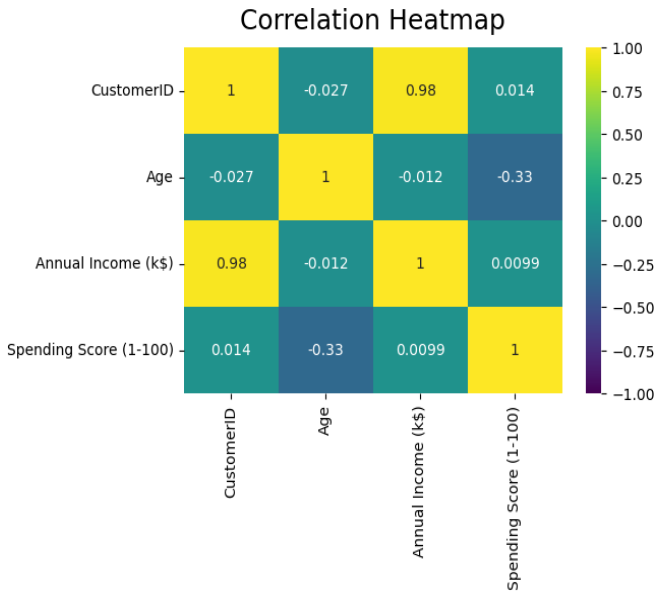

Fig. 11. Correlation

The heatmap visualizes the correlation matrix of the dataset, displaying the pairwise correlation coefficients between numerical features.

IV. Clustering

Clustering is a fundamental task in data analysis, involving the grouping of unlabeled data or data points into distinct clusters based on their similarities. The objective of clustering is to organize data points into clusters where members within each cluster share common characteristics or traits, while being distinct from data points in other clusters. Essentially, the goal of clustering is to identify inherent patterns and structures within the dataset, allowing for meaningful insights and interpretations.

V. Customer Segmentation using K-means K-

means is an iterative algorithm used for clustering data points into K distinct groups. It begins by randomly selecting K initial cluster centroids from the dataset. Each data point is then assigned to the nearest centroid, forming clusters. The centroids are updated iteratively by recalculating their positions based on the mean of the data points assigned to each cluster. This process continues until convergence, where the centroids stabilize, indicating minimal change in cluster assignments. K-means aims to minimize the within-cluster variance, resulting in well-separated and homogeneous clusters.

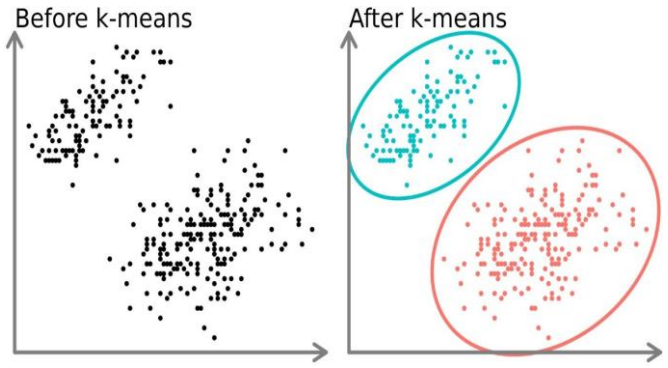

Fig. 12.

VI. Algorithm

Step-1: Start by randomly selecting the initial centroids, which serve as the starting points for each cluster.

Step-2: Perform iterative calculations to optimize the positions of the centroids. This involves assigning data points to the nearest centroid and updating the centroids' positions based on the mean of the data points in each cluster.

Step-3: Continue iterating until one of the stopping criteria is met:

Step-3.1 Centroids have stabilized: There is no significant change in their values, indicating successful clustering.

Step-3.2 The defined number of iterations has been achieved, indicating the algorithm has run for a sufficient number of steps.

VII. CLUSTER EVALUATION: ELBOW METHOD AND SILHOUETTE SCORE ANALYSIS

In this project we have used the Elbow method to find the optimal 'K' value. The elbow method helps find the optimal number of clusters (K) for K-means clustering. By plotting WCSS against K, it identifies the "elbow point" where the rate of WCSS decrease slows significantly. This point represents the optimal K value, balancing clustering accuracy and simplicity.

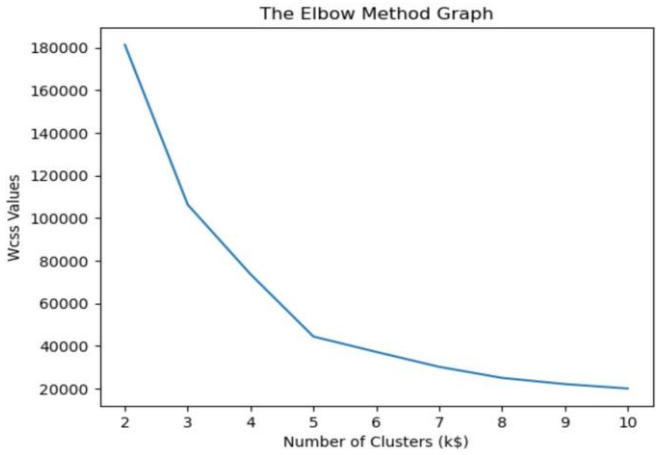

Fig. 13. Elbow Graph

Observation:

- Plot reduces drastically from culter number 1 to 3.
- Slows down till 5.
- Flattens from 6 to 10.
- We are getting an elbow where k=5.
- So, the optimal number of clusters will be 5.

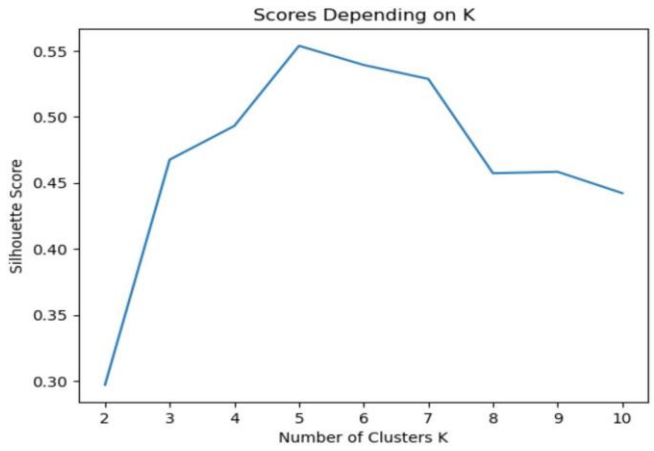

Fig. 14. Scores

The Silhouette Score Analysis complements the elbow method by providing a quantitative measure of cluster quality. Utilizing the silhouette score for each K value and appending these scores to the wcss list. This process enables an evaluation of how well-defined and distinct the clusters are for different values of K

```
Silhouette Score for K=2: 0.2968969162503008
Silhouette Score for K=3: 0.46761358158775435
Silhouette Score for K=4: 0.4931963109249047
Silhouette Score for K=5: 0.553931997444648
Silhouette Score for K=6: 0.53976103063432
Silhouette Score for K=7: 0.5281944387251989
Silhouette Score for K=8: 0.4575689106804838
Silhouette Score for K=9: 0.4605043439759829
Silhouette Score for K=10: 0.4416208208785718
```

Observation:

- Max Silhouette Score for K=5 is 0.553931997444648

Fig. 15. Silhouette Scores

Higher silhouette scores indicate better-defined clusters, with data points closely aligned with their respective clusters and distinct from neighboring clusters.

VIII. TRAINING MODEL USING K-MEANS

The visualization of K-means clustering results in a pair of features. After fitting the K-means model with five clusters to the dataset, the scatterplot showcases the distribution of customers within each cluster based on their annual income and spending score.

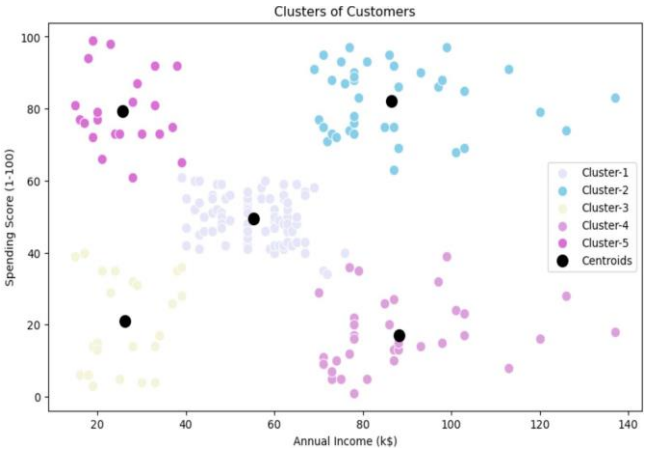

Fig. 16. Clusters of Customers

Each cluster is distinguished by a unique color, facilitating visual differentiation. Additionally, centroids representing the mean values of each cluster are depicted as black points,

serving as central reference points. This visualization provides valuable insights into the segmentation of customers into distinct groups.

```
Cluster Number : 0
       Annual Income (k$)  Spending Score (1-100)
count           39.000000              39.000000
mean            86.538462              82.128205
min             69.000000              63.000000
max            137.000000              97.000000


Cluster Number : 1
       Annual Income (k$)  Spending Score (1-100)
count           22.000000              22.000000
mean            25.727273              79.363636
min             15.000000              61.000000
max             39.000000              99.000000


Cluster Number : 2
       Annual Income (k$)  Spending Score (1-100)
count                35.0              35.000000
mean                 88.2              17.114286
min                  70.0               1.000000
max                 137.0              39.000000


Cluster Number : 3
       Annual Income (k$)  Spending Score (1-100)
count           23.000000              23.000000
mean            26.304348              20.913043
min             15.000000               3.000000
max             39.000000              40.000000


Cluster Number : 4
       Annual Income (k$)  Spending Score (1-100)
count           81.000000              81.000000
mean            55.296296              49.518519
min             39.000000              34.000000
max             76.000000              61.000000
```

Fig. 17. Cluster profiles

The output provides a succinct summary of the cluster profiles generated by the K-means algorithm. Each cluster is identified by a unique number, and for each cluster, the average annual income and spending score of its members are presented.

IX. Results

Cluster 0 - (lavender): Customers in this cluster exhibit a moderate annual income and a relatively high spending score. They are likely to be frequent shoppers who are willing to spend more on purchases. The firm should prioritize this segment for targeted marketing efforts and personalized promotions to capitalize on their propensity to spend.

Cluster 1 -(skyblue): This group comprises customers with a relatively high annual income but a low spending score. They might be more cautious or frugal with their spending habits despite having the financial means to spend more. While this segment offers potential for higher sales, the firm may need to implement strategies to incentivize spending and increase engagement.

Cluster 2 - (beige): Customers in this category demonstrate a low annual income and a moderate spending score. They might be budget-conscious shoppers who prioritize affordability over extravagant spending. While this segment may not contribute significantly to immediate sales, the firm can focus on building brand loyalty and providing value-added services to cultivate long-term relationships.

Cluster 3 - (plum): This cluster consists of customers with a low annual income but a high spending score. Despite their limited financial resources, they are willing to splurge on purchases, possibly indicating a preference for certain products or experiences. The firm should concentrate on this segment by offering flexible payment options and tailored promotions to maximize their spending potential.

Cluster 4 - (orchid): Customers in this segment display a moderate annual income and a moderate spending score. They strike a balance between their income and spending habits, neither being overly thrifty nor extravagant in their purchases. While this segment provides a stable revenue stream, the firm should focus on maintaining customer satisfaction and fostering brand loyalty to drive repeat business.

By analyzing the data, we can gain insights into customer behavior based on their Annual Income and Spending Score. This cluster analysis can inform various consumer marketing strategies. Targeting customers with high income and high spending scores is crucial as they contribute significantly to profit margins. These customers are attracted to the Mall Supermarket due to its wide range of offerings. Customers with lower income and spending scores can be enticed with promotions and discounts, encouraging them to spend more frequently. Cluster analysis helps identify customer preferences, enabling more targeted marketing efforts. In this scenario, customers in clusters 3 and 4 emerge as potential targets for strategic marketing initiatives.

X. CONCLUSION

Customer segmentation using K-means clustering provides crucial insights into consumer behavior and preferences. Analyzing demographic and behavioral data enables businesses to identify distinct customer segments and customize their marketing strategies accordingly. Techniques like the elbow method and silhouette score analysis help determine optimal cluster configurations, enhancing segmentation accuracy. Understanding different segment profiles allows businesses to prioritize marketing efforts on high-potential segments. Ultimately, leveraging customer segmentation facilitates targeted marketing, personalized experiences, and boosts business performance and competitiveness in the market.

REFERENCES

[1] Prof. Nikhil Patankar, Soham Dixit, Akshay Bhamare, Ashutosh Darpel and Ritik Raina. (2021, December). Customer Segmentation Using Machine Learning. [Online].

[2] Dr. C K Gomathy, Ms. D. Abirami, Mr. J. Balamurugan, Ms. Dondapati Tejaswi. (2021, September). The Customer Data Analysis Using Segmentation With Special Reference Mall. [Online].

[3] Medium. Robert Baker. (2023, May 30). Customer Segmentation using K-Means Clustering. [Online]. Available at: https://medium.com/@robertb909/k-means-clustering-a64f859a1074

[4] Medium. Camilo Gonçalves. (2024, April 24). Customer Segmentation using K-Means clustering. [Online]. Available at: https://medium.com/@camilolgon/customer-segmentation-using-k-means-clustering-9e5e11a3165a

[5] Analytics Vidhya. Pulkit Sharma. (2024, May 02). The Ultimate Guide to K-Means Clustering: Definition, Methods and Applications. [Online]. Available at: https://www.analyticsvidhya.com/blog/2019/08/comprehensive-guide-k-means-clustering/

[6] Geeksforgeeks. (2024, March 11). K means Clustering – Introduction. [Online]. Available at: https://www.geeksforgeeks.org/k-means-clustering-introduction/?ref=lbp

[7] Analytics Vidhya. Basil Saji. (2024, January 07). Elbow Method for Finding the Optimal Number of Clusters in K-Means. [Online]. Available at: https://www.analyticsvidhya.com/blog/2021/01/in-depth-intuition-of-k-means-clustering-algorithm-in-machine-learning/

[8] Medium. Adria Binte Habib. (2021, February 20). Elbow Method vs Silhouette Coefficient in Determining the Number of Clusters. [Online]. Available at: https://adria708.medium.com/elbow-method-vs-silhouette-co-efficient-in-determining-the-number-of-clusters-33baff2fbeee

[9] 365 Data Science. Natassha Selvaraj. (2023, June 5). How to Build Customer Segmentation Models in Python. [Online]. Available at: https://365datascience.com/tutorials/python-tutorials/build-customer-segmentation-models/

[10] Medium. Evgeniy Ryzhkov. (2020, July 23). 5 Stages of Data Preprocessing for K-means clustering. [Online]. Available at: https://medium.com/@evgen.ryzhkov/5-stages-of-data-preprocessing-for-k-means-clustering-b755426f9932: :text=K