Customer Segmentation in Marketing using K-Means Clustering

Divya Rani[1], Kumar Harsh[2]

[1,2]Department of Artificial Intelligence and Data Science, IIMT College of Engineering, Greater Noida

Abstract

In today's data-driven marketing world, understanding and segmenting customers based on their behavior and demographics is a crucial task. This paper presents a machine learning approach to customer segmentation using the K-Means clustering algorithm. The project uses the popular Mall_Customers.csv dataset and leverages Python with libraries such as pandas, matplotlib, seaborn, and scikit-learn. The Elbow Method is employed to determine the optimal number of clusters, followed by 2D and 3D visualization of the customer groups. The study demonstrates how segmentation can lead to targeted marketing and better business outcomes. This method empowers organizations to shift from generalized to personalized marketing tactics, enabling better allocation of resources, increased customer satisfaction, and improved sales conversions. Additionally, this approach proves to be scalable, allowing it to be adopted by businesses of various sizes and sectors.

Keywords: Customer Segmentation, K-Means, Machine Learning, Clustering, Marketing Analytics, Python, Data Science

Introduction

Customer segmentation refers to the process of dividing a customer base into groups that exhibit similar characteristics. These characteristics may include age, income level, gender, buying patterns, or spending behavior. Traditionally, segmentation was done manually using intuition or static rules. However, the growth of machine learning has made it possible to automate and enhance this process using data-driven approaches.

Unsupervised learning techniques such as K-Means clustering allow businesses to explore hidden patterns in their data and generate meaningful groups of customers. These insights are particularly valuable in designing personalized marketing campaigns, improving customer satisfaction, and maximizing revenue. This research applies K-Means clustering to a customer dataset with the aim of identifying patterns and defining customer segments that businesses can act upon.

Literature Review

Customer segmentation is a cornerstone of effective marketing. According to Han et al. in Data Mining: Concepts and Techniques [1], clustering algorithms like K-Means are widely used for market segmentation due to their simplicity and scalability.

A study published in Springer [2] highlights how machine learning improves the flexibility and adaptability of segmentation strategies. It compares traditional demographic segmentation with behavior-based clustering models. Another study from ResearchGate [3] demonstrates how large customer datasets can be segmented more accurately and automatically using unsupervised learning.

Datasets such as those found on Kaggle [4] provide a reliable benchmark for academic and industry experiments in segmentation, including customer data from retail, finance, and telecom domains.

MethodologyTo effectively understand customer behavior and enhance targeted marketing, a systematic approach to customer segmentation was implemented. To uncover valuable customer insights, a structured approach to segmentation was applied. This methodology ensures clarity, accuracy, and actionable results.

The following steps were followed to implement customer segmentation:

Step 1: Data Collection

The Mall_Customers.csv dataset was used. It includes customerID, gender, age, annual income, and spending score.

Step 2: Data Cleaning

The CustomerID column was dropped. Missing values were checked and handled (none were found).

Step 3: Exploratory Data Analysis

Data was visualized using histograms and violin plots to understand distributions across income and spending.

Step 4: Feature Selection

We selected Age, Annual Income, and Spending Score as the clustering attributes.

Step 5: Elbow Method

The optimal number of clusters was determined by plotting WCSS values and identifying the 'elbow' point, which was found at k = 5.

Step 6: K-Means Clustering

K-Means clustering was applied using scikit-learn. Clusters were formed based on the three selected features.

Step 7: Visualization Results were displayed using 2D scatter plots and 3D plots using matplotlib and seaborn.

This methodological framework not only ensures technical rigor but also supports scalability and reproducibility in real-world marketing environments. These insights empower businesses to tailor marketing strategies for distinct customer groups, enhancing personalization. Moreover, this segmentation approach can be further refined using advanced clustering techniques like hierarchical clustering or DBSCAN.

FLOWCHART

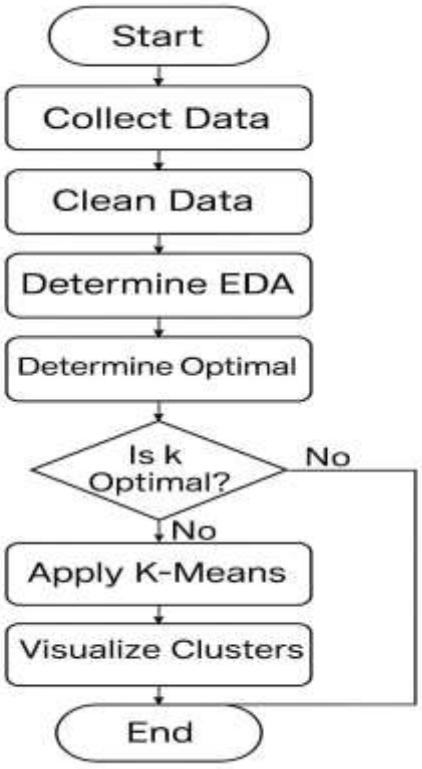


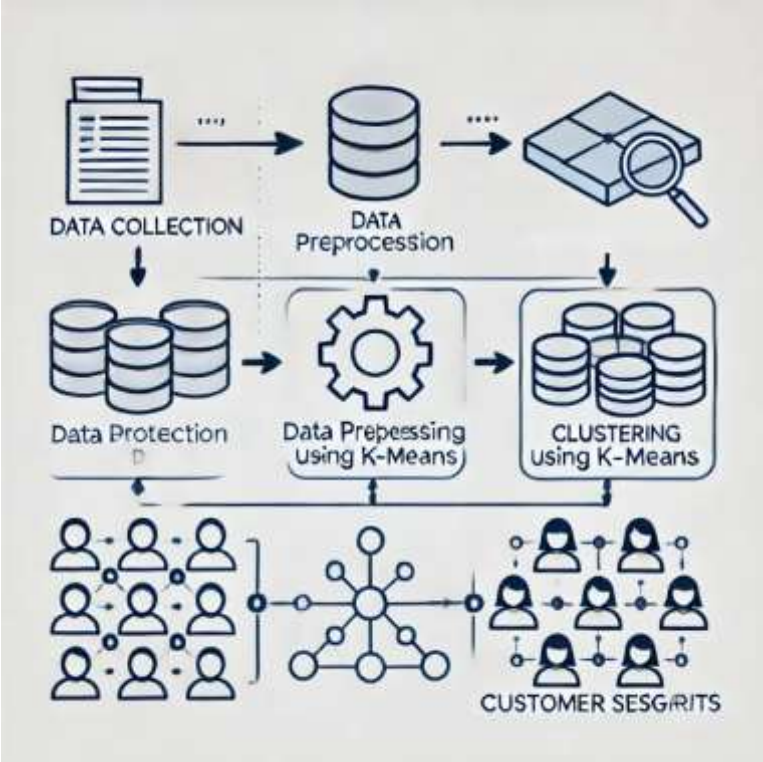
FIGURE 1: Flowchart representing the customer segmentation process using K-Means clustering.

System Design and Architecture

The system consists of the following modules:

- Data Input: Loads customer data from CSV files.
- Preprocessing: Removes nulls and irrelevant columns.
- Analysis: Visualizes distributions for gender, age, income, and spending score.
- Clustering Engine: Implements K-Means based on the selected k-value.
- Visualization: Plots clusters in 2D and 3D to enable interpretation by marketing teams.

This architecture supports integration into larger dashboards or customer relationship systems.



Implementation

The implementation of the proposed customer segmentation model was carried out using Python within the Jupyter Notebook environment, which offers an interactive interface ideal for data exploration, visualization, and model development.

Key libraries used include:

- pandas and numpy: For efficient data manipulation and numerical operations.
- seaborn and matplotlib: For creating insightful visualizations such as histograms, violin plots, and cluster scatter plots.
- scikit-learn: For executing the K-Means algorithm and evaluating the model using tools like the Elbow Method and WCSS.

The workflow began with data cleaning and exploratory analysis. After selecting the most relevant features (Age, Annual Income, and Spending Score), the Elbow Method was used to determine the optimal number of clusters. The K-Means algorithm was then applied, and the resulting clusters were visualized in both 2D and 3D formats to enable meaningful interpretation.

Results and Discussion

The following five customer segments were identified:

1. Young High Spenders
2. Mid-Income, Balanced Spenders
3. High Income, Low Spenders
4. Low Income, Low Spenders
5. Senior High Spenders

These segments were validated visually and statistically using cluster centroids. Businesses can apply these insights to:

- Target high spenders with premium offers
- Design loyalty programs for frequent low spenders
- Improve communication strategies based on income groups

Conclusion and Future Scope

Conclusion

This study confirms that K-Means clustering offers a practical and scalable approach for customer segmentation. Using just three features—Age, Annual Income, and Spending Score—the model successfully grouped customers into meaningful clusters, providing valuable insights for targeted marketing and improved customer engagement.

The simplicity and interpretability of K-Means make it an ideal choice for organizations seeking quick, actionable segmentation without complex infrastructure. The results demonstrate that even with limited features, businesses can derive substantial value and optimize their marketing strategies effectively.

Future Scope

- To enhance the model's performance and applicability, future work could focus on:
- Including behavioral variables such as purchase history and location for deeper segmentation.
- Applying clustering to real-time data for dynamic customer updates.
- Comparing with other algorithms like DBSCAN or hierarchical clustering for accuracy and flexibility.
- Deploying the model in an interactive web-based dashboard for real-time insights.

References

[1] J. Han, M. Kamber, and J. Pei, Data Mining: Concepts and Techniques, 3rd ed. Amsterdam: Elsevier, 2011.

[2] R. K. Yadav, M. Garg, and A. Goel, "Customer segmentation using K-Means clustering and neural networks," The European Journal of Health Economics, vol. 24, Sep. 2023. [Online]. Available: https://link.springer.com/article/10.1007/s10257-023-00640-4

[3] S. Iqbal and A. Yousuf, "Customer segmentation using machine learning," ResearchGate, Nov. 2021. [Online]. Available: https://www.researchgate.net/publication/356756320_Customer_Segmentation_Using_Machine_Learning

[4] Y. Hassan, "Customer Segmentation Dataset," Kaggle, Accessed May 2025. [Online]. Available: https://www.kaggle.com/datasets/yasserh/customer-segmentation-dataset