Clustering Analysis on Customer Dataset

Nikita Sharma
M.Sc. Data Science, DAIICT, Gandhinagar, India

September 26, 2025

Abstract

Abstract. This paper presents an unsupervised learning approach using K-Means, Agglomerative Hierarchical Clustering, and DBSCAN on the Mall Customers dataset. The aim is to analyze customer behavior and preferences using clustering algorithms to identify meaningful customer segments and provide actionable insights for marketing applications.

1 Introduction

Customer segmentation is a vital task in retail analytics. It helps businesses identify and categorize customers based on their behavior and preferences. This paper applies clustering algorithms: K-Means, DBSCAN, and Agglomerative Hierarchical Clustering, to the Mall Customers dataset, focusing on annual income and spending score. The goal is to uncover distinct customer groups that can enhance marketing strategies and engagement.

2 Data and Methods

The dataset contains 200 entries with features: CustomerID, Gender, Age, Annual Income (k$), and Spending Score (1–100). For clustering, only Annual Income and Spending Score were used. These features were standardized using `StandardScaler` to ensure equal contribution to distance-based algorithms.

2.1 K-Means Clustering

The Elbow Method was used to determine the optimal number of clusters. K-Means was applied with $k = 5$, and cluster centroids were visualized. The Within-Cluster Sum of Squares (WCSS) was plotted to identify the elbow point.

2.2 Agglomerative Hierarchical Clustering

A dendrogram was generated using Ward linkage to determine the optimal number of clusters. Agglomerative Clustering was applied with $n = 5$ clusters, and results were visualized using scatter plots.

2.3 DBSCAN Clustering

DBSCAN was applied with parameters `eps=0.5` and `min_samples=5`. The algorithm identified core points, border points, and noise. Cluster labels were visualized, with noise points marked distinctly.

3 Results

3.1 Optimal Clusters

The Elbow Method and dendrogram both suggested 5 clusters as optimal. K-Means and Agglomerative clustering produced similar groupings.

3.2 Cluster Comparison

K-Means and Hierarchical clustering showed compact, well-separated clusters. DBSCAN revealed noise points and was sensitive to parameter tuning.

3.3 DBSCAN Performance

DBSCAN identified several outliers and non-spherical clusters. Unlike K-Means, it did not force every point into a cluster, offering more flexibility.

3.4 Tables and Figures

Algorithm	**Clusters**	**Silhouette Score**
K-Means	5	0.55
Hierarchical	5	0.554
DBSCAN	Variable	0.388

Table 1: **Clustering performance comparison.** Silhouette scores for each algorithm. DBSCAN's cluster count varies due to its density-based nature.

4 Conclusion

K-Means was the most interpretable and stable for this dataset. DBSCAN provided insights into outliers and non-spherical clusters. These clusters can help mall managers tailor promotions—for example, targeting high-income, low-spending customers with premium loyalty programs.

Availability of data and software code

The dataset is available on Kaggle.