Clustering Methods using scikit-learn

Srishti Lamba
M.Sc. Data Science, DAU Gandhinagar

September 26, 2025

Abstract

This project presents a comparative analysis of three clustering algorithms—K-Means, Hierarchical Clustering, and DBSCAN—applied to mall customer data using scikit-learn. The dataset includes annual income and spending score features, enabling segmentation based on consumer behavior. Optimal clusters were determined to be five for both K-Means and Hierarchical models, guided by the Elbow Method and dendrogram analysis. DBSCAN was tuned to identify density-based clusters and noise points, offering a flexible alternative to centroid-based methods. Visualizations and performance comparisons highlight the strengths and limitations of each approach. These insights support real-world marketing strategies by enabling targeted engagement with distinct customer segments.

1 Introduction

Customer segmentation is a key task in data-driven marketing. Clustering algorithms offer an unsupervised approach to grouping customers based on behavioral patterns. This report applies K-Means, Hierarchical Clustering, and DBSCAN to a mall customer dataset using scikit-learn, comparing their performance and suitability for identifying meaningful customer groups.

2 Data and Methods

The dataset contains 200 mall customers with two features: Annual Income (in $k) and Spending Score (1–100). Features were standardized using `StandardScaler`.

- **K-Means**: Applied with `n_clusters=5`, guided by the Elbow Method.
- **Hierarchical Clustering**: Used Ward linkage and Euclidean distance; cluster count determined from dendrogram.
- **DBSCAN**: Tuned with `eps=0.5`, `min_samples=5` to detect density-based clusters and noise.

Visualizations were generated using Matplotlib.

3 Results

1. Optimal Clusters

The Elbow Method showed a sharp drop in WCSS up to $k = 5$, indicating five optimal clusters for K-Means. The dendrogram also suggested five clusters based on the longest vertical linkage before merging.

2. Cluster Comparison

K-Means and Hierarchical Clustering produced similar, compact clusters. DBSCAN identified five clusters with irregular shapes and labeled several points as noise, offering a more flexible view of the data.

3. DBSCAN Performance

DBSCAN successfully detected outliers and non-spherical clusters. Unlike the other methods, it did not force every point into a cluster, making it suitable for identifying anomalies.

4. Algorithm Suitability

K-Means was most suitable due to its simplicity and alignment with the data's compact structure. Hierarchical Clustering was comparable but less scalable. DBSCAN was useful for outlier detection but sensitive to parameter tuning.

5. Real-World Application

The identified segments can guide mall marketing strategies:

- **High Income, Low Spending Score**: Offer exclusive services and loyalty programs.
- **Low Income, High Spending Score**: Promote budget bundles and seasonal discounts.
- **High Income, High Spending Score**: Introduce premium memberships and early access to luxury brands.
- **Low Income, Low Spending Score**: Use sampling campaigns and flash sales.

4 Conclusion

All three clustering algorithms revealed five meaningful customer segments. K-Means was the most effective for this dataset. DBSCAN added value by identifying noise points. These insights can help businesses personalize marketing strategies and improve customer engagement.

Data and code

- Dataset: `https://www.kaggle.com/datasets/vjchoudhary7/customer-segmentation-tutorial`
- Code Repository: `https://github.com/srishti7103/clustering-using-scikit-learn`