Clustering Methods using scikit-learn

Srishti Lamba-202518003

M.Sc. Data Science, DAU Gandhinagar

Introduction

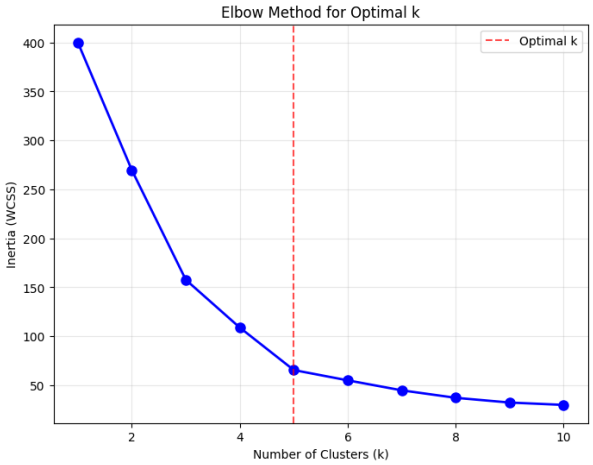
Customer segmentation is a key task in data-driven marketing. Clustering algorithms offer an unsupervised approach to grouping customers based on behavioral patterns. This report applies K-Means, Hierarchical Clustering, and DBSCAN to a mall customer dataset using scikit-learn, comparing their performance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

Data and Methods

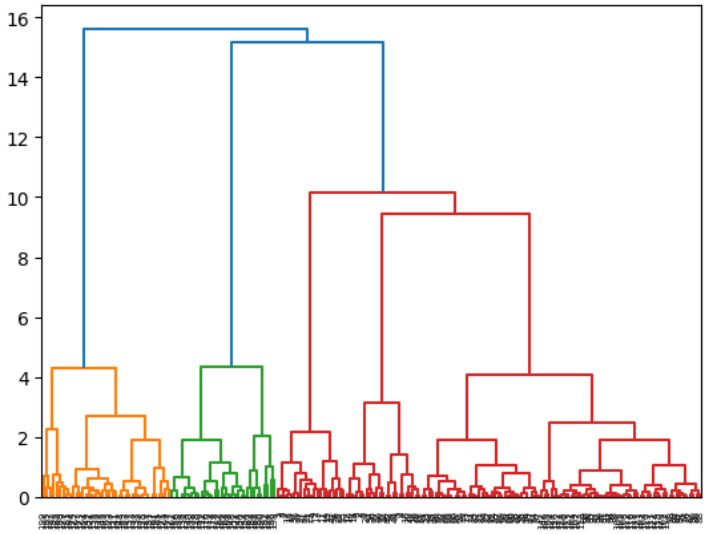
The dataset contains 200 mall customers with two features: Annual Income (in $k) and Spending Score (1–100). Features were standardized using StandardScaler.

- **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.**

Optimal Clusters: K-Means and Hierarchical Clustering

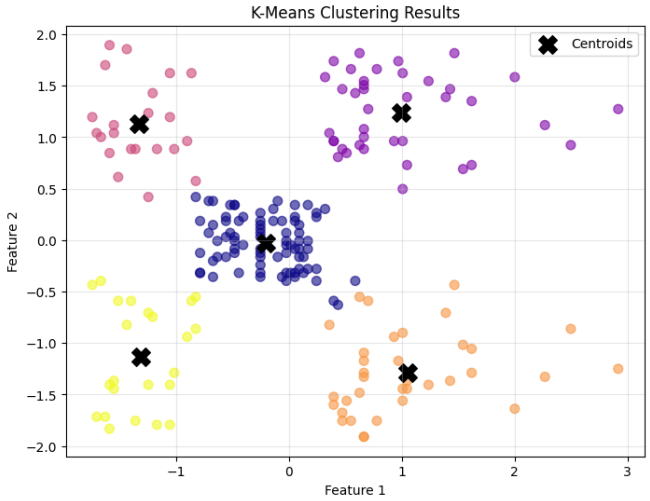


K-Means: The Elbow Method graph shows a sharp bend at k = 4, which means adding more clusters beyond this point doesn't improve the grouping much. So, 4 clusters is the best choice.

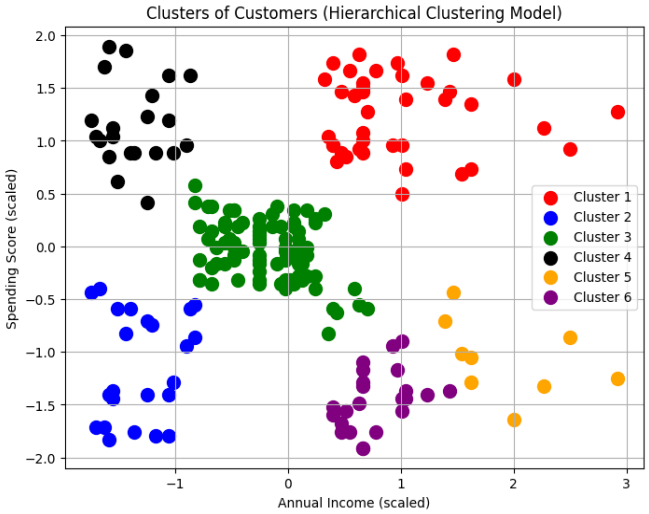


Hierarchical Clustering: The dendrogram suggests 6 clusters. This is based on where the longest vertical line cuts across the horizontal branches, showing six clear groups.

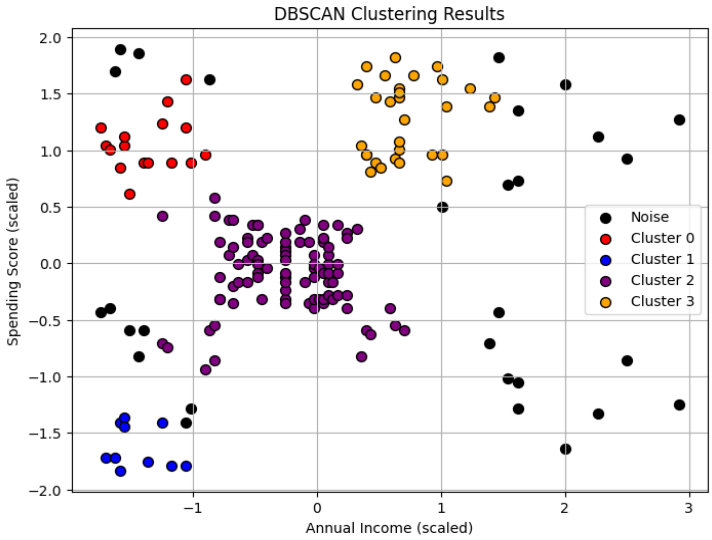
Cluster Comparison: K-Means vs Hierarchical vs DBSCAN



K-Means formed 4 compact, spherical clusters with clear centroids. The clusters were well-separated and easy to interpret.



Hierarchical Clustering resulted in 6 clusters, offering finer segmentation. However, some clusters overlapped slightly, and the boundaries were less distinct.



DBSCAN identified 4 clusters and also labeled several points as noise (black dots). These noise points were not assigned to any cluster, which makes DBSCAN unique compared to the other two methods.

Notable Differences:

- **K-Means and Hierarchical assign every point to a cluster.**
- **DBSCAN is density-based and can exclude outliers, making it more flexible for irregular shapes and sparse regions.**

DBSCAN Performance

DBSCAN performed well in identifying clusters of varying shapes and densities. It successfully detected noise points, which are data points that don't belong to any dense region. These were labeled as -1 in the output.

Compared to K-Means and Hierarchical Clustering, DBSCAN was better at handling outliers and non-spherical clusters. However, its performance was highly dependent on the choice of parameters: eps and min_samples.

With proper tuning, DBSCAN revealed meaningful clusters and isolated scattered data that the other methods misclassified.

Most Suitable Algorithm

Among the three algorithms, K-Means was the most suitable for this dataset. It produced clean, interpretable clusters and aligned well with the Elbow Method. The data distribution was fairly compact and spherical, which fits K-Means assumptions.

Hierarchical Clustering was useful for understanding the nested structure of the data and provided more detailed segmentation. However, it was less robust to noise and harder to scale.

DBSCAN was excellent for detecting outliers and handling non-uniform cluster shapes, but its sensitivity to parameter tuning made it less consistent for this dataset.

Real-World Application

In a real-world business scenario, the mall's marketing team could use clustering insights to target high-income customers with low spending scores more effectively. This segment likely has the financial capacity but lacks motivation or interest in current offerings. To engage them, the mall could introduce premium loyalty programs, personalized shopping experiences, or exclusive access to luxury brands and events. These strategies tap into their lifestyle preferences and create a sense of exclusivity. Additionally, targeted communication—such as curated recommendations or concierge services—can enhance their shopping experience and encourage higher spending. By aligning marketing efforts with the unique traits of this segment, the mall can convert potential into profit.

Data and code

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