# ML Lab 09

## DBSCAN Lab Exercise: Mall Customers Dataset

### What is the meaning of eps and min\_samples in DBSCAN? How do they affect the number and shape of clusters? How might they impact the Silhouette Score?

**Meanings**

* eps (epsilon): Defines the radius of the neighborhood around a point. If two points are within this distance, they are considered neighbors.
* min\_samples: Minimum number of points required within an eps-neighborhood for a point to be considered a core point.

**Effect on clusters**

* Smaller eps or higher min\_samples - fewer, tighter clusters, and potentially more noise (-1 labels).
* Larger eps or lower min\_samples - looser clusters that may merge together or absorb noise.

**Impact on Silhouette Score**

* If eps is too small, most points are labeled as noise - poor clustering.
* If eps is too large, distinct clusters may merge - low Silhouette Score.
* An optimal combination produces well-separated, dense clusters - higher Silhouette Score.

### After standardization, how many clusters and noise points (-1 label) did you observe? What do these clusters represent in terms of customer behavior? How good is the clustering according to the Silhouette Score?

**Cluster Labels**

* Cluster 0 (orange) - The largest group, capturing a wide range of customer profiles.
* Cluster 1 (green) - A tighter, well-separated cluster of high-income, high-spending customers.
* Label -1 (blue) - Noise points that DBSCAN could not assign to any cluster.

**Behavioral Interpretation**

* Cluster 1 might represent a premium customer segment such as high income and high spending.
* Cluster 0 could represent the general customer base, more spread out and moderate in behavior.
* Noise points (-1) might be unusual spenders or edge cases such as low income with high spending or vice versa.

**Silhouette Score (0.350)**

Indicates moderate clustering quality.

* Cluster 1 improves the score due to its compactness.
* Cluster 0 lowers it due to its broad and potentially non-convex shape.
* Noise points aren’t included in the silhouette calculation but still impact clustering by reducing density.

A diagram of a clustering results

AI-generated content may be incorrect.

1. Modify eps and/or min\_samples (e.g., try eps=0.3 or min\_samples=8). How does it change the clustering result and the Silhouette Score?

**Observations**

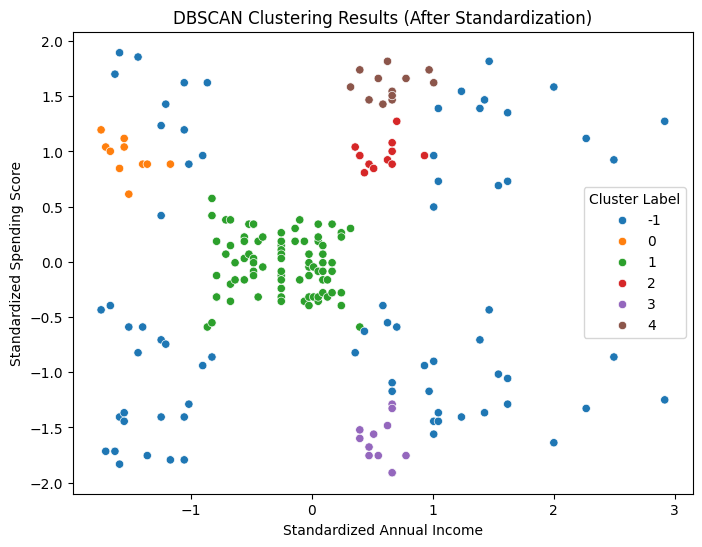
* Number of clusters increased from 2 to 5 distinct clusters (labels 0 to 4).
* Noise points (-1) have increased significantly, suggesting many data points no longer meet the stricter density requirement.
* Silhouette Score dropped to 0.194, which is much lower than the previous 0.350.

**Interpretation**

* Smaller eps (0.3) makes DBSCAN more conservative, meaning:
  + Only very dense regions form clusters.
  + Points on the edge of clusters or in less dense areas are marked as noise.
* Higher min\_samples (8) raises the threshold to form a core point, making it even harder for points to become part of a cluster.

**Impact on Clustering Quality**

* Although the number of clusters increased, many are tiny and tightly packed, making the overall cohesion and separation weaker, hence the lower Silhouette Score.
* The higher number of noise points suggests the model is too strict under these parameters, likely underfitting.



### After adding 'Age' as a third feature and re-running DBSCAN, how does it affect the clustering result? Comment on the 3D visualization and the Silhouette Score.

* Clustering became less effective: Silhouette Score dropped to 0.188, suggesting that adding 'Age' may have introduced more overlap or noise.
* The 3D scatter plot shows less distinct separation between clusters compared to the 2D case.

**Possible reasons**

* 'Age' may not contribute to clear density-based separations.
* The added dimension may create cluster distortion if 'Age' has weak correlation with spending behavior.

More features do not mean better clustering, feature relevance matters. DBSCAN works best when features reflect density-separated groups.

A chart with many colored dots

AI-generated content may be incorrect.A graph of a diagram

AI-generated content may be incorrect.

## Hierarchical Clustering Lab Exercise: Iris Dataset

### What is the difference between 'ward', 'single', and 'complete' linkage methods?

These are different methods to calculate the distance between clusters when performing hierarchical clustering:

**Ward Linkage**

* Minimizes the variance within each cluster.
* Tries to find compact, spherical clusters.
* Works best when data is normally distributed.
* Often gives better results for well-separated clusters.

**Single Linkage**

* Computes the minimum distance between points in two clusters (nearest neighbor).
* Prone to the “chaining effect”, it can form long, thin clusters.
* May not perform well when clusters are not clearly separated.

**Complete Linkage**

* Computes the maximum distance between points in two clusters (farthest neighbor).
* Tends to produce more compact, evenly shaped clusters, but may be sensitive to outliers.

### How many clusters did you find? How good is the clustering according to the Silhouette Score?

Clusters = 3

Silhouette Score = 0.447

* The orange cluster (Cluster 1) is well-separated in the bottom-left corner, likely representing the Setosa species, which is known to be linearly separable based on petal dimensions.
* The blue (Cluster 0) and green (Cluster 2) clusters are more overlapping, especially in the mid-range of petal lengths and widths. These likely represent Versicolor and Virginica, which have more similar feature distributions.
* This partial overlap explains why the Silhouette Score is 0.447, indicating moderate clustering quality.
  + Values close to 1 imply well-separated clusters.
  + Values near 0 indicate overlapping clusters.
  + Setosa is clearly identified, while the other two are partially mixed, reducing the overall score.

A diagram of a cluster

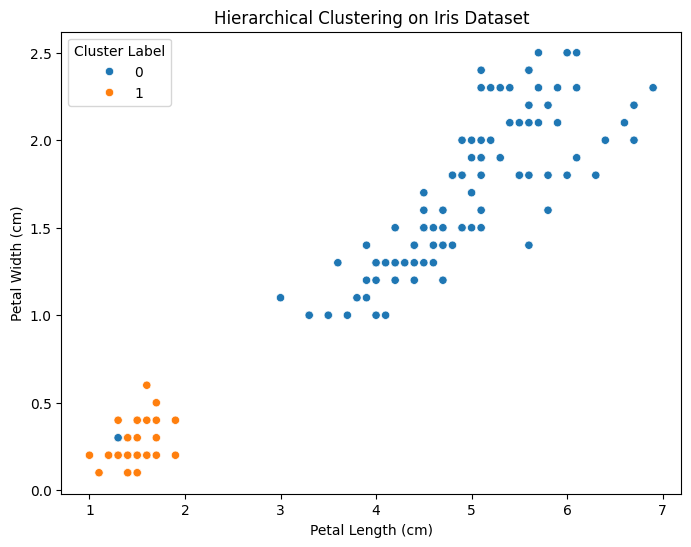
AI-generated content may be incorrect.

### Try changing the number of clusters (n\_clusters=2,4) and observe the changes in silhouette score. What does that hyperparameter signify?

**What do the n\_clusters hyperparameter signify?**

n\_clusters defines how many groups the algorithm will try to divide the data into.

|  |  |  |
| --- | --- | --- |
| **n\_clusters** | **Silhouette Score** | **Interpretation** |
| 2 | 0.577 | Higher score suggests better-defined clusters |
| 3 | 0.447 | Balanced, aligns with the known 3 Iris species, slight drop in score due to some overlap. |
| 4 | 0.401 | Lower score, likely due to splitting natural groups into sub-clusters, introducing fragmentation. |

A chart with different colored dots

AI-generated content may be incorrect.

### Try clustering using only Petal Length and Petal Width. How does the clustering quality change?

Silhouette Score (Petal Features only) - 0.610

This is significantly higher than the score using all four features (0.447).

**Why is clustering better with only petal features?**

* Petal Length and Width are highly discriminative against the Iris species, especially for Setosa vs. the other two. In contrast, sepal features add noise or reduce the contrast between clusters.
* Dimensionality Reduction Advantage - Using fewer but more relevant features can reduce the curse of dimensionality and sharpen cluster boundaries.
* Visual Separation - Clusters are more visually distinct in 2D space.

Focusing on highly informative features (petal dimensions) can significantly improve clustering performance. A silhouette score of 0.610 indicates a very good clustering structure.

