Lab 11 Student Assignment: Customer Segmentation Using Unsupervised Learning

Dataset Included • Unsupervised Clustering • Real-World Scenario

Dataset

We will use a real customer dataset from Kaggle:

Mall Customers Dataset (CSV)

Download link (free):
https://www.kaggle.com/datasets/shwetabh123/mall-customers

File needed: **Mall_Customers.csv**

Objectives

1. Load a real-world dataset and perform preprocessing.
2. Apply **k-Means**, **DBSCAN**, and **Hierarchical Clustering**.
3. Compare clustering performance on real customer segmentation data.
4. Visualize clusters and interpret meaningful patterns.

Tools

- Pandas
- Scikit-learn
- Matplotlib
- SciPy

Supporting Content Overview

k-Means: Partitions data into *k* clusters using distance to centroids. Works best with spherical clusters.

DBSCAN: Density-based method useful for detecting oddly shaped clusters and outliers.

Hierarchical Clustering: Builds a tree (dendrogram) showing merges of clusters at different distances.

LAB TASKS

Task 1 — Load Dataset

1. Download dataset from Kaggle: https://www.kaggle.com/datasets/shwetabh123/mall-customers
2. Load `Mall_Customers.csv` using pandas.
3. Select only these features for clustering:
    - **Annual Income (k$)**
    - **Spending Score (1–100)**
4. Standardize the features using `StandardScaler`.

Task 2 — Apply k-Means

1. Pick `k = 5` clusters.
2. Fit the model and predict cluster labels.
3. Plot the clusters with different colors.
4. Explain whether the clusters look meaningful or not.

Task 3 — Apply DBSCAN

1. Use the same standardized data.
2. Try different values of:
    - `eps = 0.2 → 0.5`
    - `min_samples = 4 → 10`
3. Find a combination that results in **at least 3 meaningful clusters**.
4. Plot the clustering result.
5. Count how many points were labeled as **-1 (noise/outliers)**.

Task 4 — Apply Hierarchical Clustering

1. Use SciPy's `linkage` with `method='ward'`.
2. Plot a **clear dendrogram** using:
    - `truncate_mode='lastp'`
    - `p = 10`
3. Based on the dendrogram, decide:
    - How many clusters is the dataset naturally forming?

Task 5 — Compare All Three Models

Write a brief comparison of:

- k-Means
- DBSCAN
- Hierarchical

Which algorithm is better for this dataset and why?

EXERCISES

Exercise 1

Why is it important to standardize the features before clustering?

Exercise 2

In k-Means, what is the effect of choosing a larger value for **k**?
Does it always improve clustering? Why or why not?

Exercise 3

DBSCAN marks some points as **–1**.
What does this label represent, and why might a point be marked this way?

Exercise 4

In DBSCAN tuning, how does decreasing **eps** affect cluster formation?

Exercise 5

Look at your dendrogram from the hierarchical clustering step.
At what distance threshold would you cut the dendrogram to produce exactly **3 clusters**?

Exercise 6

Among the three methods you used (k-Means, DBSCAN, Hierarchical),
which one created clusters that make the most sense for customer segmentation?
Explain your reasoning.