Lab 3

Clustering

Clustering is an essential task in data mining, which aims to group data into clusters so that all similar data points are grouped together, while dissimilar ones belong to the different groups. This lab focuses on implementing and analyzing the following clustering algorithms:

1. **Partitioning Approach: K-Means**
2. **Hierarchical Method: Agglomerative Clustering**
3. **Density-Based: DBSCAN**

By completing this lab, students will:

1. Understand the working principles of clustering algorithms.
2. Implement clustering techniques using Python libraries.
3. Analyze the results to derive meaningful insights from clustering outcomes.

1 Dataset

The dataset used for this lab is the **Wine recognition dataset**, which contains the results of chemical analysis of 178 wine samples, each derived from one of three different cultivars.

This dataset contains 13 numerical attributes describing the physicochemical properties and detailed composition of wines: Alcohol, Malic acid, Ash, Alcalinity of ash, Magnesium, Total phenols, Flavanoids, Nonflavanoid phenols, Proanthocyanins, Color intensity, Hue, OD280/OD315 of diluted wines, and Proline.

The dataset can be loaded directly using `scikit-learn` as follows:

```
from sklearn.datasets import load_wine
# Load the dataset
wine = load_wine()
# Access features and target
X = wine.data
y = wine.target
```

Students are required to preprocess the dataset using standard normalization to ensure that the dataset is ready for use with clustering algorithms.

2 Requirements

Implement the following clustering algorithms using Python libraries (e.g., `scikit-learn`, `scipy`).

2.1 Partitioning Approach: K-Means

- Perform K-Means clustering for different values of k ($1 \leq k \leq 10$).
- For each value of k, compute and display the **inertia**. Find the optimal number of clusters by applying the **Elbow Method** and explain your choice.

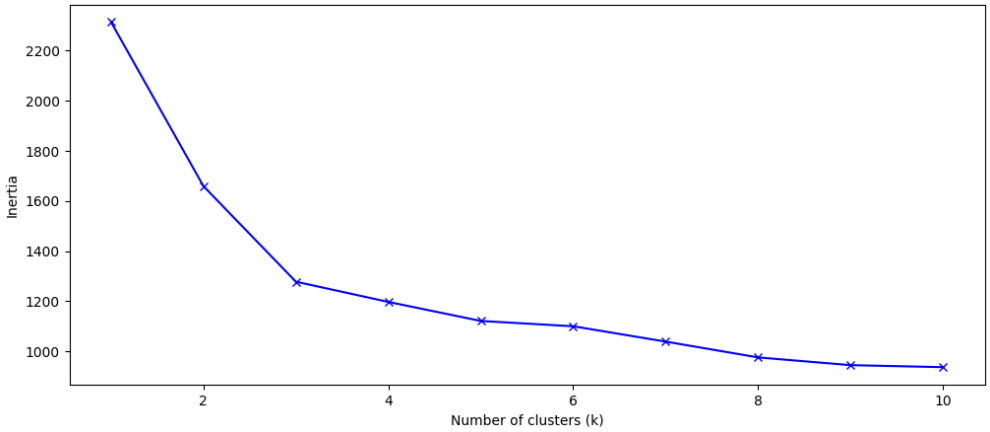

Figure 1: Inertia with different values of cluster numbers.

- Using the optimal k, visualize the resulting clusters on three dimensions using **PCA**.

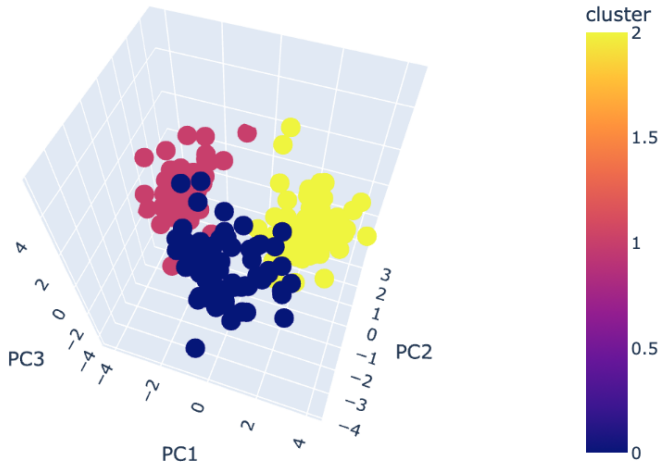

Figure 2: Wine Clusters using K-Means (visualized by plotly).

2.2 Hierarchical Method: Agglomerative Clustering

- Perform clustering using **Agglomerative Clustering** and experiment with different linkage (`ward`, `average`, `complete`, `single`). Visualize the resulting clusters for each linkage method and provide your insights based on the observed outcomes.

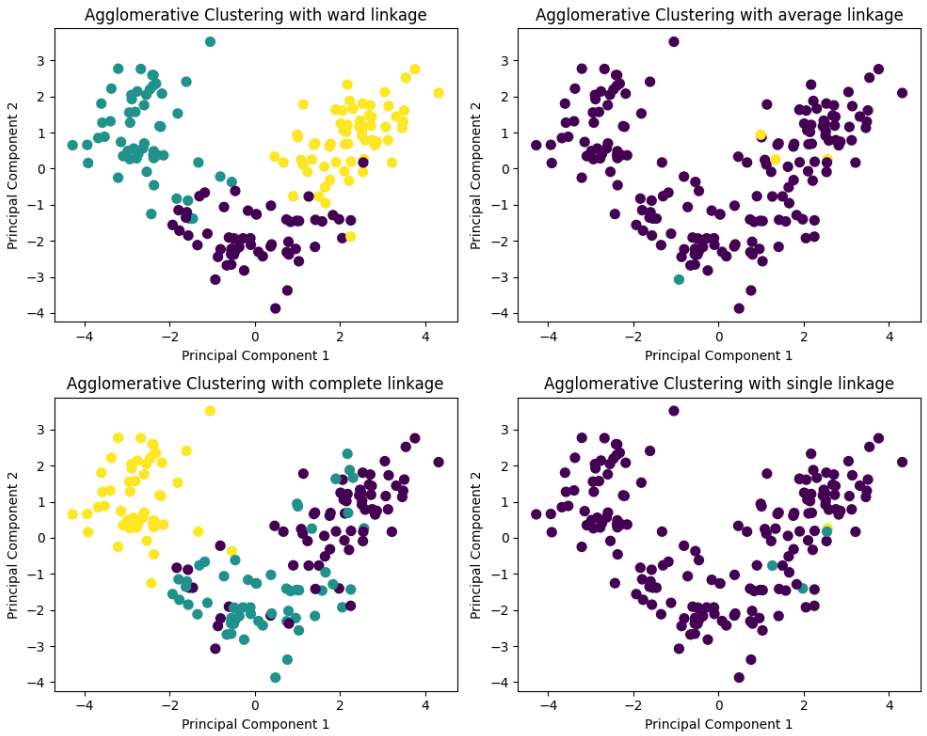

Figure 3: Wine Clusters using Agglomerative Clustering on 2D PCA.

- Create a **dendrogram** to visualize the **clustering hierarchy** using any linkage method. Discuss the insights and information that can be derived from the dendrogram.

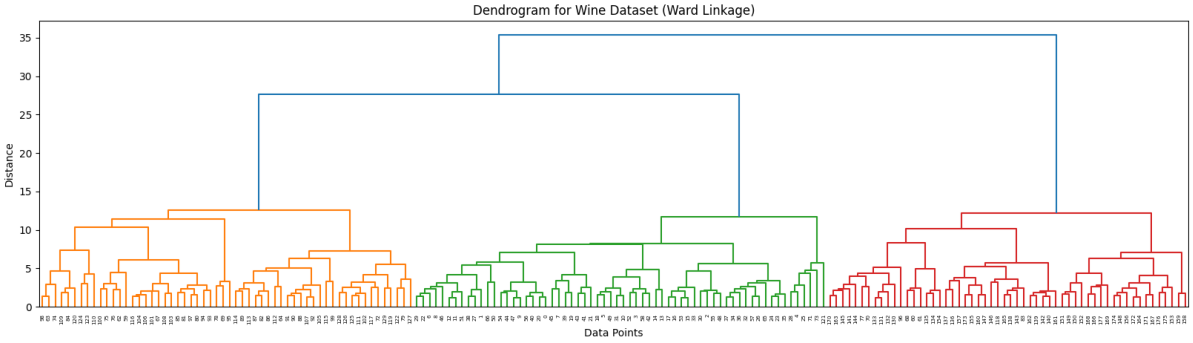

Figure 4: Dendogram with Ward linkage.

2.3 Density-Based: DBSCAN

- Use the **DBSCAN** to detect clusters and noise in the data. Experiment with different values for the parameters `eps` (the radius of the neighborhood) and `min_samples` (the minimum number of points required to form a cluster). Visualize the resulting clusters and noise points.

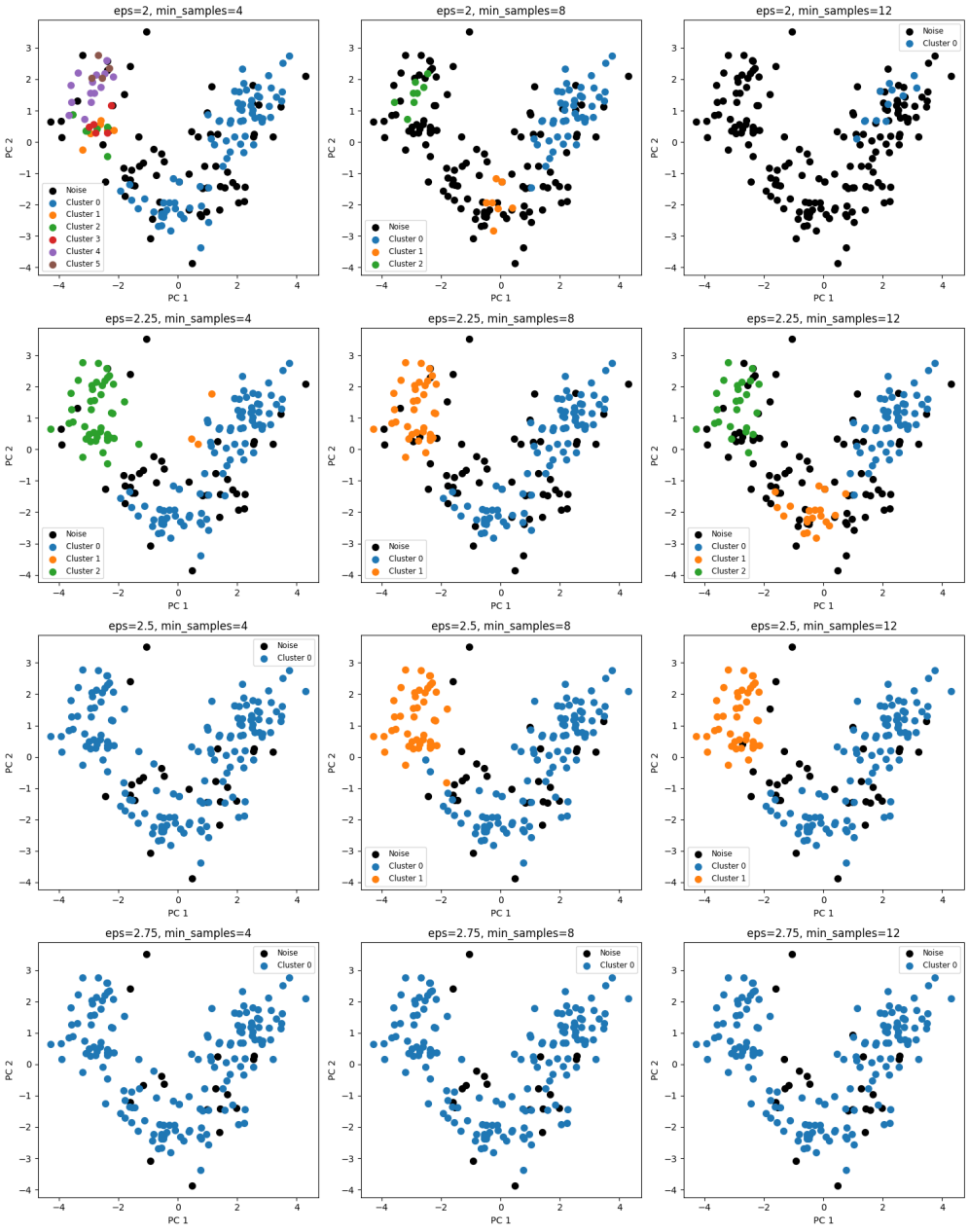

Figure 5: DBSCAN Clustering Results on 2D PCA.

- How does the change in `eps` and `min_samples` affect the density necessary to form a cluster?

3 Report (Jupyter Notebook)

The source code, result will be reported in a Jupyter Notebook with the following requirements:

- Student information (Student ID, full name, etc.).
- Self-evaluation of the assignment requirements.
- Detailed explanation of each step. Illustrative images, diagrams and equations are required.
- Each processing step must be fully commented, and results should be printed for observation.
- The report needs to be well-formatted.
- Before submitting, re-run the notebook (`Kernel` → `Restart & Run All`).
- References (if any).

4 Assessment

No.	Details	Score
1	Partitioning Approach: K-Means	35%
2	Hierarchical Method: Agglomerative Clustering	35%
3	Density-Based: DBSCAN	30%

5 Notices

Please pay attention to the following notices:

- This is an **INDIVIDUAL** assignment.
- Duration: about 2 weeks.
- Any plagiarism, any tricks, or any lie will have a 0 point for the course grade.

The end.