Lab Activity: K-Means Clustering on Wine Dataset

Objectives

1. Implement the **K-Means clustering** algorithm on a new dataset.
2. Apply the **Elbow Method** to find the optimal number of clusters.
3. Interpret the clusters in terms of **wine chemical composition**.
4. Visualize and explain the cluster results.

Tools

- **Language:** Python
- **Libraries:** Scikit-learn, Pandas, Matplotlib
- **Dataset:** Wine Dataset (UCI Machine Learning Repository)
- **Dataset Link:** Wine Dataset (UCI)

Dataset Description

This dataset contains the results of a chemical analysis of wines grown in the same region in Italy but derived from three different cultivars.

Each row corresponds to a different wine sample and includes:

- **Alcohol**
- **Malic acid**
- **Ash**
- **Alcalinity of ash**
- **Magnesium**
- **Total phenols**
- **Flavanoids**
- **Nonflavanoid phenols**
- **Proanthocyanins**
- **Color intensity**
- **Hue**
- **OD280/OD315 of diluted wines**
- **Proline**

Lab Tasks

1. **Load the Dataset**
   - Download the dataset from the link above.
   - Load it into a Pandas DataFrame.
   - Assign proper column names based on the description given on the UCI website.
2. **Select Features for Clustering**
   - Choose two or three numerical features (e.g., "Alcohol" and "Color intensity").
   - Explain why you selected these features (students can justify using correlation or domain reasoning).
3. **Standardize the Data**
   - Use `StandardScaler` from `sklearn.preprocessing` to scale all features.
4. **Apply the Elbow Method**
   - Try values of `k` from 1 to 10.
   - Plot the **Elbow curve** (k vs. inertia).
   - Identify the optimal number of clusters from the curve.
5. **Run K-Means Algorithm**
   - Fit a KMeans model using your chosen `k`.
   - Obtain cluster labels and add them as a new column in the dataset.
6. **Visualize the Clusters**
   - Create a **scatter plot** of your selected features, colored by cluster labels.
   - Mark cluster centroids using red "X" markers.
7. **Evaluate Cluster Quality**
   - Compute and print the **Silhouette Score** for your clustering result.
8. **Cluster Summary**
   - Compute the mean of each feature for every cluster (to describe each wine type).
   - Present the results in a small summary table.

Exercises (Students Must Answer)

1. **Feature Selection Justification:**
   Why did you choose those particular features for clustering?
   What relationship do they have with wine quality?
2. **Elbow Method Analysis:**
   At what value of `k` did you observe the "elbow"?
   Why do you think that number of clusters is optimal for this dataset?
3. **Cluster Interpretation:**
   Based on your summary table, describe what each cluster represents (e.g., high-alcohol, low-acidity wines).
   Which cluster seems to represent premium wines?

4. **Silhouette Score Meaning:**
   What was your silhouette score?
   Interpret whether your clustering result is strong, moderate, or weak.
5. **Comparison with Known Classes (Optional):**
   The Wine dataset actually has **3 true wine types (classes)**.
   Compare your predicted clusters with these classes.
   Did K-Means correctly discover the natural groupings?