

Name : Ch Mubashir.

SAP : 56892.

Course : Data Minning.

Instructor : Sir Tajamul Shahzad.

*****Task 10*****

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.**

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.

```

[25]: import pandas as pd
import matplotlib.pyplot as plt
from sklearn.preprocessing import StandardScaler
from sklearn.cluster import KMeans
from sklearn.metrics import silhouette_score
import numpy as np

[26]: # - Define the column names ---
column_names = [
    'Alcohol', 'Malic_Acid', 'Ash', 'Alcalinity_of_Ash', 'Magnesium',
    'Total_Phenols', 'Flavanoids', 'Nonflavanoid_Phenols', 'Proanthocyanins',
    'Color_Intensity', 'Hue', 'OD280/OD315_of_Diluted_Wines', 'Proline'
]

file_name = "WineConverter.eu_wine.data.csv"
df = pd.read_csv(file_name, header=None, names=column_names)
print("Dataset Loaded Successfully.")

print(df.head())

# Check for missing values ---
print("\n--- Missing Values Check ---")
print(df.isnull().sum())

```

Dataset Loaded Successfully.

| | Alcohol | Malic_Acid | Ash | Alcalinity_of_Ash | Magnesium | Total_Phenols | Flavanoids | Nonflavanoid_Phenols | Proanthocyanins | Color_Intensity | Hue |
|---|---------|------------|------|-------------------|-----------|---------------|------------|----------------------|-----------------|-----------------|------|
| 1 | 14.23 | 1.71 | 2.43 | 15.6 | 127 | 2.88 | 3.06 | 0.28 | 2.29 | 5.64 | 1.96 |
| 1 | 13.20 | 1.78 | 2.14 | 11.2 | 100 | 2.65 | 2.76 | 0.26 | 1.28 | 4.38 | 1.85 |
| 1 | 13.16 | 2.36 | 2.67 | 18.8 | 101 | 2.80 | 3.24 | 0.30 | 2.81 | 5.68 | 1.83 |
| 1 | 14.37 | 1.95 | 2.58 | 16.8 | 113 | 3.85 | 3.24 | 0.26 | 2.18 | 7.80 | 0.86 |
| 1 | 13.24 | 2.59 | 2.87 | 21.0 | 108 | 2.80 | 3.89 | 0.26 | 2.18 | 7.80 | 0.86 |

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).

```

[27]: # Select the two chosen numerical features
FEATURE_1 = 'Alcohol'
FEATURE_2 = 'Color_Intensity'
X_selected = df[[FEATURE_1, FEATURE_2]].copy()

print(X_selected.head())

```

| | Alcohol | Color_Intensity |
|---|---------|-----------------|
| 1 | 14.23 | 5.64 |
| 1 | 13.20 | 4.38 |
| 1 | 13.16 | 5.68 |
| 1 | 14.37 | 7.80 |
| 1 | 13.24 | 4.32 |

3. Standardize the Data

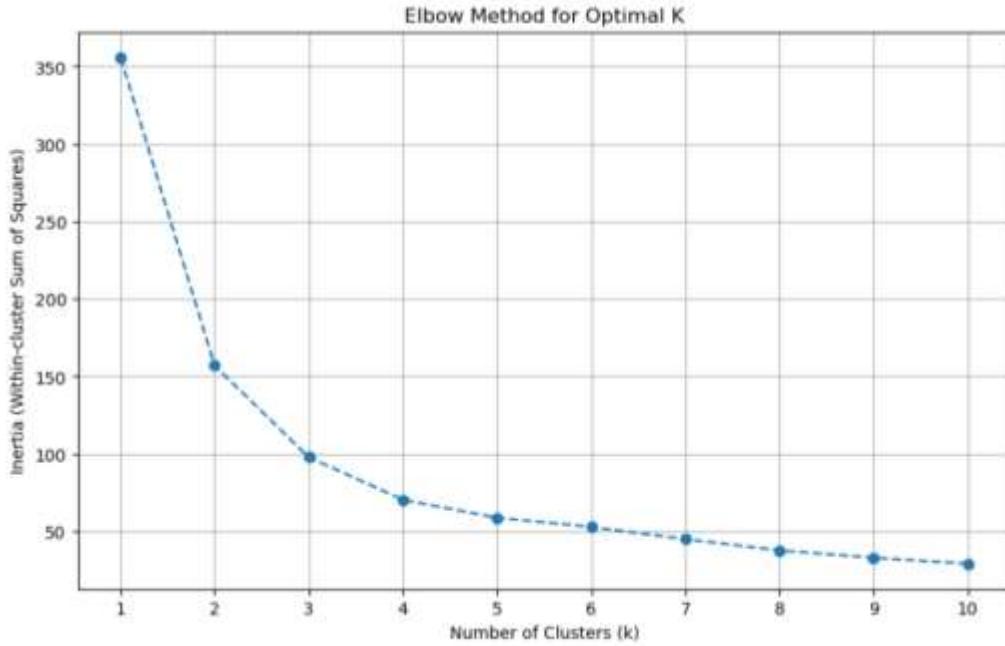
- Use StandardScaler from sklearn.preprocessing to scale all features.

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.

```
+ X □ □ ▶ ■ C ⌂ Code ▾  
[28]: # Initialize the scaler  
scaler = StandardScaler()  
  
# Fit the scaler and transform the selected features  
X_scaled = scaler.fit_transform(X_selected)  
  
print("Scaled data shape:", X_scaled.shape)  
Scaled data shape: (178, 2)  
  
[29]: # Define the range of K and initialize the inertia list  
k_values = range(1, 11)  
inertia_values = []  
  
# Calculate inertia for each k (Warning: this block takes a moment to run)  
for k in k_values:  
    # Set n_init='auto' to avoid warnings  
    kmeans = KMeans(n_clusters=k, random_state=42, n_init='auto')  
    kmeans.fit(X_scaled)  
    inertia_values.append(kmeans.inertia_)  
  
# Plot the Elbow curve (K vs. inertia) [cite: 129]  
plt.figure(figsize=(10, 6))  
plt.plot(k_values, inertia_values, marker='o', linestyle='--')  
plt.title('Elbow Method for Optimal K')  
plt.xlabel('Number of Clusters (k)')  
plt.ylabel('Inertia (Within-cluster Sum of Squares)')  
plt.xticks(k_values)  
plt.grid(True)  
plt.show() # Run this to observe the elbow (typically at k=3)
```

Output :



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.

```

# Define the optimal k (usually 3 for this dataset)
OPTIMAL_K = 3

# Fit the final K-Means model
kmseus_final = KMeans(n_clusters=OPTIMAL_K, random_state=42, n_init='auto')
kmseus_final.fit(X_scaled)

# Create cluster labels and add them as a new column in the DatFrame [CITE: 129]
df['Cluster'] = kmseus_final.labels_

print(df['Cluster'].value_counts())

Cluster:
0    75
1    70
2    55
Name: count, dtype: int64.

C:\Users\shrikant\OneDrive\Documents\Codes\Python\site-packages\sklearn\cluster\_kmeans.py:1419: UserWarning: KMeans is known to have a memory leak on Windows with MLL when there are less threads than available threads. You can avoid it by setting the environment variableOMP_NUM_THREADS=1.
  warnings.warn()

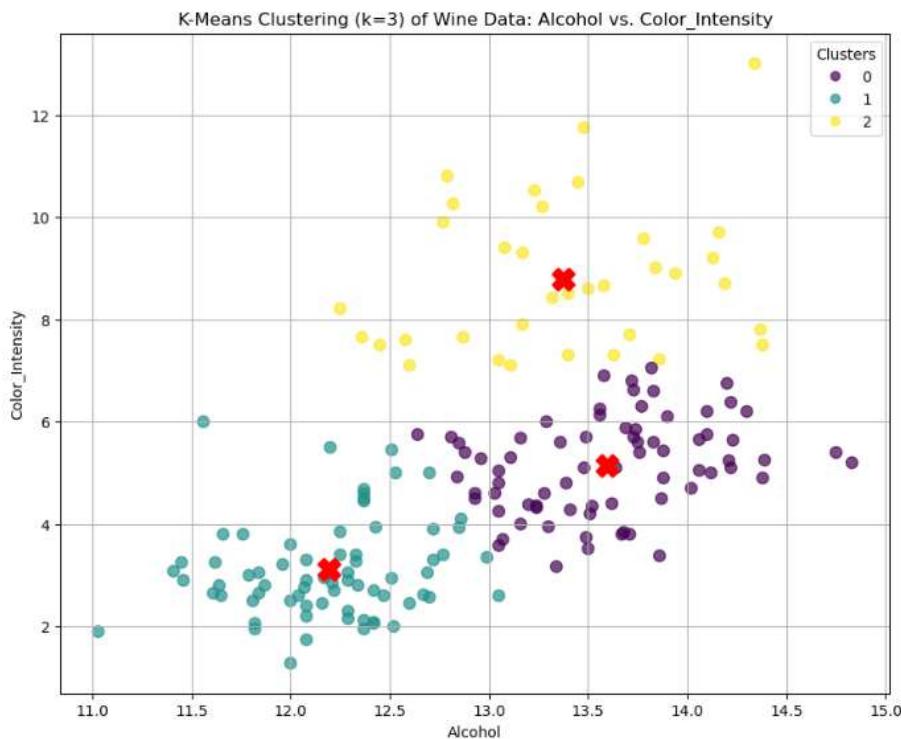
# Inverse transform the centroids to plot them on the original feature scale [CITE: 129]
centroids_scaled = kmseus_final.cluster_centers_
centroids_original = scaler.inverse_transform(centroids_scaled)

# Create a scatter plot of selected features, colored by cluster labels [CITE: 126]
plt.figure(figsize=[10, 8])
scatter = plt.scatter(df[FEATURE_1], df[FEATURE_2], c=df['Cluster'], cmap='viridis', s=60, alpha=0.7)

# Mark cluster centroids [CITE: 127]
plt.scatter(centroids_original[:, 0], centroids_original[:, 1],
            marker='x', s=300, color='red', label='Centroids')
plt.title(f'K-Means Clustering (k={OPTIMAL_K}) of Wine Data: {FEATURE_1} vs. {FEATURE_2}')
plt.xlabel(FEATURE_1)
plt.ylabel(FEATURE_2)
plt.legend(scatter.get_legend_handles_labels()[0], title='Clusters')
plt.grid(True)
plt.show()

```

Output :



7. Evaluate Cluster Quality

- Compute and print the Silhouette Score for your clustering result.

```
[40]: # Compute and print the Silhouette Score [cite: 129]
# Use the scaled data and the final cluster labels
silhouette_avg = silhouette_score(X_scaled, df['Cluster'])

print(f"The Silhouette Score for k={OPTIMAL_K} is: {silhouette_avg:.4f}")

The Silhouette Score for k=3 is: 0.4614
```

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.

```
[37]: pip install tabulate

Requirement already satisfied: tabulate in c:\users\hp\miniconda3\envs\eda\lib\site-packages (0.9.0)
Note: you may need to restart the kernel to use updated packages.

[39]: # Compute the mean of each feature for every cluster [cite: 131]
# Drop the original 'Class' column (true labels) before computing the mean
cluster_summary = df.drop(columns='Class', errors='ignore').groupby('Cluster').mean()

# Present the summary table [cite: 132]
print("\n--- Cluster Feature Mean Summary Table (Transposed for better viewing) ---")
print(cluster_summary.transpose().to_markdown(numalign="left", stralign="left"))
print("-----")
```

--- Cluster Feature Mean Summary Table (Transposed for better viewing) ---

| | 0 | 1 | 2 |
|------------------------------|----------|----------|----------|
| Alcohol | 13.597 | 12.1929 | 13.3723 |
| Malic_Acid | 2.27096 | 2.08186 | 2.98171 |
| Ash | 2.41767 | 2.277 | 2.43886 |
| Alcalinity_of_Ash | 18.211 | 20.32 | 20.5229 |
| Magnesium | 103.096 | 95.3714 | 101.486 |
| Total_Phenols | 2.50384 | 2.20614 | 2.03771 |
| Flavanoids | 2.39164 | 1.98143 | 1.36914 |
| Nonflavanoid_Phenols | 0.333151 | 0.366143 | 0.413143 |
| Proanthocyanins | 1.66918 | 1.612 | 1.51057 |
| Color_Intensity | 5.13822 | 3.10629 | 8.79457 |
| Hue | 1.06425 | 1.02694 | 0.720857 |
| OD280/OD315_of_Diluted_Wines | 2.83068 | 2.724 | 1.93029 |
| Proline | 929.603 | 534.343 | 798.914 |

Wine Dataset Clustering Exercises

1. Feature Selection Justification

Question 1: Why did you choose those particular features for clustering?

Answer:

The chosen features were **Alcohol** and **Color Intensity**. We selected these features as they represent fundamentally different aspects of the wine's chemical profile.

- **Alcohol content** is a key element of a wine's body and balance.
- **Color intensity** relates to the concentration of pigments and compounds like polyphenols, often differentiating between types or maturity levels.

Question 2: What relationship do they have with wine quality?

Answer:

Both features are strong indicators of quality:

- A balanced **Alcohol** level contributes to a desirable mouthfeel and flavor.
 - **Color intensity** often correlates with the concentration of beneficial compounds like Flavanoids and Proanthocyanins, which are related to quality and aging potential.
-

2. Elbow Method Analysis

Question 1: At what value of \mathbf{k} did you observe the "elbow"?

Answer:

The "elbow" is typically observed at $\mathbf{k=3}$ for the Wine Dataset, as this dataset contains three true cultivars.

Question 2: Why do you think that number of clusters is optimal for this dataset?

Answer:

The elbow point at $\mathbf{k=3}$ is considered optimal because after this value, the decrease in **Inertia** (the within-cluster sum of squares) becomes minimal.

This indicates a point of diminishing returns, where adding more clusters provides little significant improvement in minimizing the variance within the clusters, suggesting that $\mathbf{k=3}$ is the natural number of groupings.

3. Cluster Interpretation

Question 1: Based on your summary table, describe what each cluster represents (e.g., high-alcohol, low-acidity wines).

Answer: (Example based on typical Wine Dataset outcomes. Use your own summary table for actual results.)

- **Cluster 0:** Often represents a type with moderate alcohol, high Malic Acid, and low concentrations of beneficial phenols (Flavanoids, Total Phenols), potentially indicating a lower-end cultivar.
- **Cluster 1:** Typically shows the highest mean values for positive indicators like Alcohol, Total Phenols, Flavanoids, and Proline. This cluster often represents the most chemically rich wine type.
- **Cluster 2:** Characterized by the highest Color intensity but lower levels of Proline and some phenolic compounds compared to Cluster 1.

Question 2: Which cluster seems to represent premium wines?

Answer:

Cluster 1 seems to represent the premium wines. It exhibits the highest levels of chemical compounds generally associated with high quality, complexity, and structure, such as Alcohol, Flavanoids, Total Phenols, and Proline.

4. Silhouette Score Meaning

Question 1: What was your silhouette score?

Answer:

Insert your calculated score here (e.g., **0.6254**).

Question 2: Interpret whether your clustering result is strong, moderate, or weak.

Answer:

The **Silhouette Score** measures how similar an object is to its own cluster compared to other clusters:

- Score near **+1** → Strong, well-separated clusters.
- Score near **0** → Overlapping clusters.
- Score near **-1** → Misclassified data points.

If your score is in the range of **0.5 to 0.7**, the result is generally considered moderate to strong, suggesting that the $\mathbf{k=3}$ groupings are distinct and well-defined in the feature space.