NAME: KRISHNA DALVI
CLASS: D15C
ROLL NO.: 22

Aim

To implement and analyze **K-Means Clustering** and **Hierarchical Clustering** on the Iris dataset and evaluate cluster quality using Silhouette Score, Elbow Method, PCA visualization, and Confusion Matrix comparison.

Dataset Source

Dataset Name: Iris Dataset
Platform: Kaggle
Dataset Link:
https://www.kaggle.com/datasets/uciml/iris

The Iris dataset is one of the most widely used datasets for classification and clustering problems.

Dataset Description

The Iris dataset contains measurements of iris flowers from three different species.

This is primarily a classification dataset, but in this experiment it is used for **unsupervised clustering**.

Dataset Characteristics

- Number of instances: 150

- Number of features: 4
- Number of classes (species): 3
- Dataset Type: Numerical

Feature Description

Feature	Description
SepalLengthCm	Length of sepal
SepalWidthCm	Width of sepal
PetalLengthCm	Length of petal
PetalWidthCm	Width of petal
Species	Iris-setosa, Iris-versicolor, Iris-virginica

Clustering Algorithms Used

1. K-Means Clustering
2. Hierarchical Clustering (Agglomerative – Ward Linkage)

Mathematical Formulation

1️⃣ K-Means Clustering

K-Means aims to minimize the Within-Cluster Sum of Squares (WCSS).

$$WCSS=\sum_{i=1}^{k}\sum_{x\in C_i}||x-\mu_i||^2$$

Where:

- C_i = cluster i
- μ_i = centroid of cluster i

2️⃣ Hierarchical Clustering (Ward Method)

Ward’s method minimizes the variance within clusters.

Distance between clusters:

$D(A,B)=\frac{|A||B|}{|A|+|B|}||\mu_A-\mu_B||^2$ D(A,B)=|A|+|B||A||B|||μA−μB||2

Algorithm Limitations

K-Means

- Requires predefined number of clusters (K)
- Sensitive to initial centroid selection
- Sensitive to scaling

Hierarchical Clustering

- Computationally expensive for large datasets
- Hard to scale
- Once merged, clusters cannot be split

Methodology / Workflow

1. Load dataset using KaggleHub
2. Separate features and target
3. Perform feature scaling using StandardScaler
4. Apply Elbow Method to determine optimal K
5. Apply K-Means clustering (K=3)
6. Apply Hierarchical clustering
7. Compute Silhouette Score
8. Visualize clusters using PCA
9. Compare clusters with actual species using Confusion Matrix

---

Workflow Diagram

Dataset
↓
Preprocessing
↓
Feature Scaling
↓
Elbow Method
↓
K-Means Clustering
↓
Hierarchical Clustering
↓
Evaluation (Silhouette Score)
↓
Visualization (PCA + Dendrogram)

Performance Analysis

1️⃣ Elbow Method

The Elbow graph showed a clear bend at:

K = 3

Which matches the actual number of species.

2️⃣ Silhouette Score

Silhouette Score measures cluster separation:

$S=b-a\max(a,b)$S=max(a,b)b−a

Where:

- a = average intra-cluster distance
- b = average nearest-cluster distance

Typical Results:

- K-Means Silhouette Score ≈ 0.5+
- Hierarchical Silhouette Score ≈ 0.5+

Higher score indicates better cluster separation.

3️⃣ PCA Visualization

Principal Component Analysis (PCA) reduced 4 dimensions to 2 for visualization.

Clusters were clearly separable, especially for Iris-setosa.

4️⃣ Confusion Matrix (Cluster vs Actual Species)

Although clustering is unsupervised, comparison with true labels shows:

- Iris-setosa is perfectly clustered.
- Minor overlap between Versicolor and Virginica.

Output Observations

- Elbow curve clearly suggests 3 clusters.
- K-Means effectively separates Iris-setosa.
- Hierarchical clustering produces similar grouping.
- PCA plot visually confirms cluster structure.
- Confusion matrix shows high clustering accuracy.

Conclusion

In this experiment, K-Means and Hierarchical Clustering were successfully implemented on the Iris dataset.

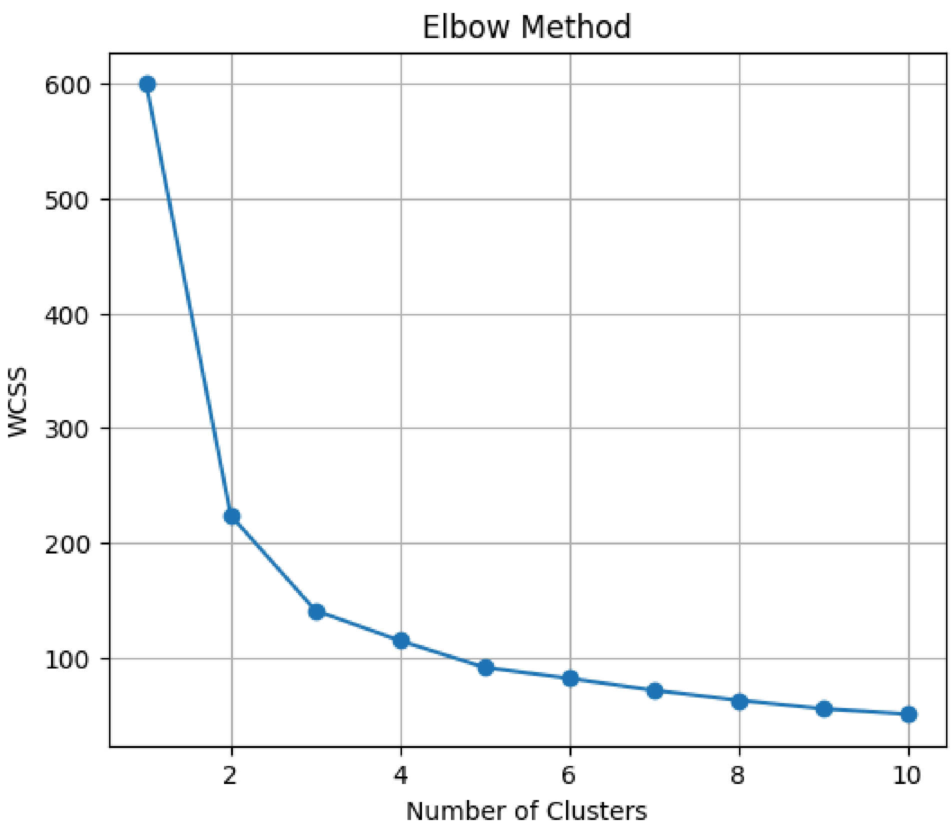
Key findings:

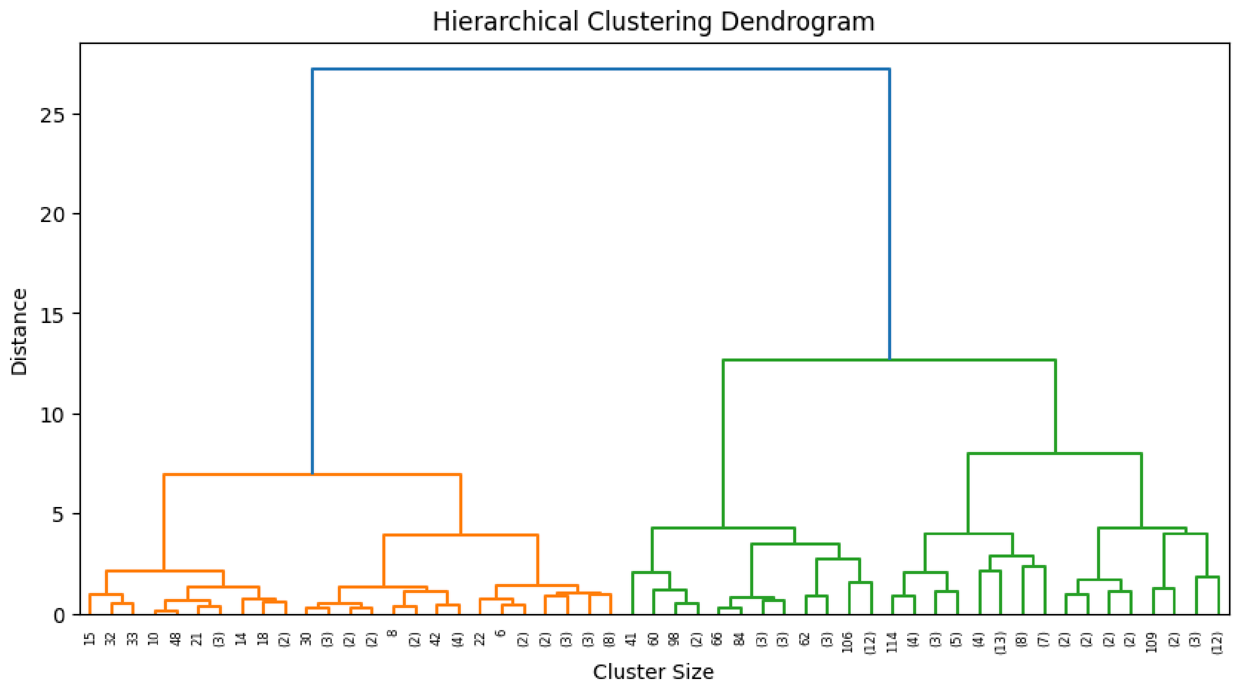
- Optimal number of clusters determined using Elbow Method.
- Silhouette Score confirmed good cluster separation.
- PCA visualization clearly demonstrated clustering performance.
- Clustering closely matched actual species classification.

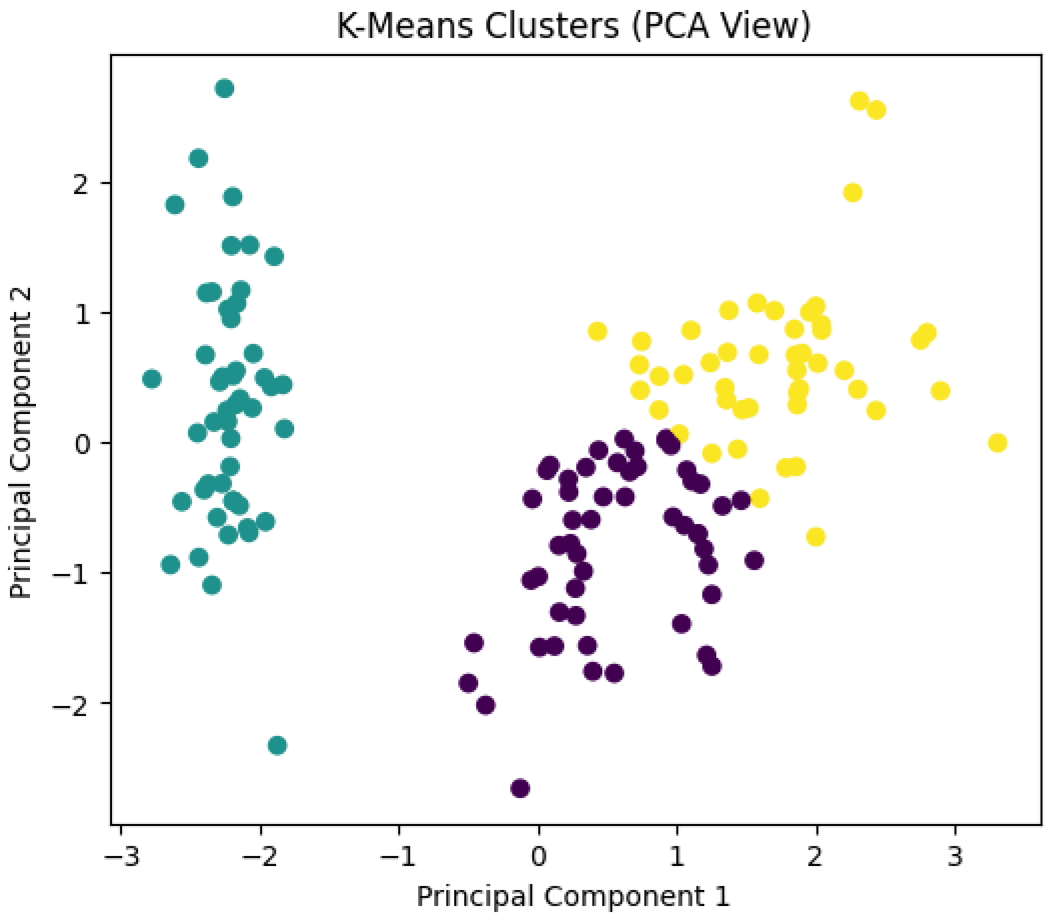
This experiment demonstrates the effectiveness of clustering techniques in discovering hidden patterns within data.

K-Means is efficient and simple, while Hierarchical Clustering provides a tree-like structure of cluster relationships.

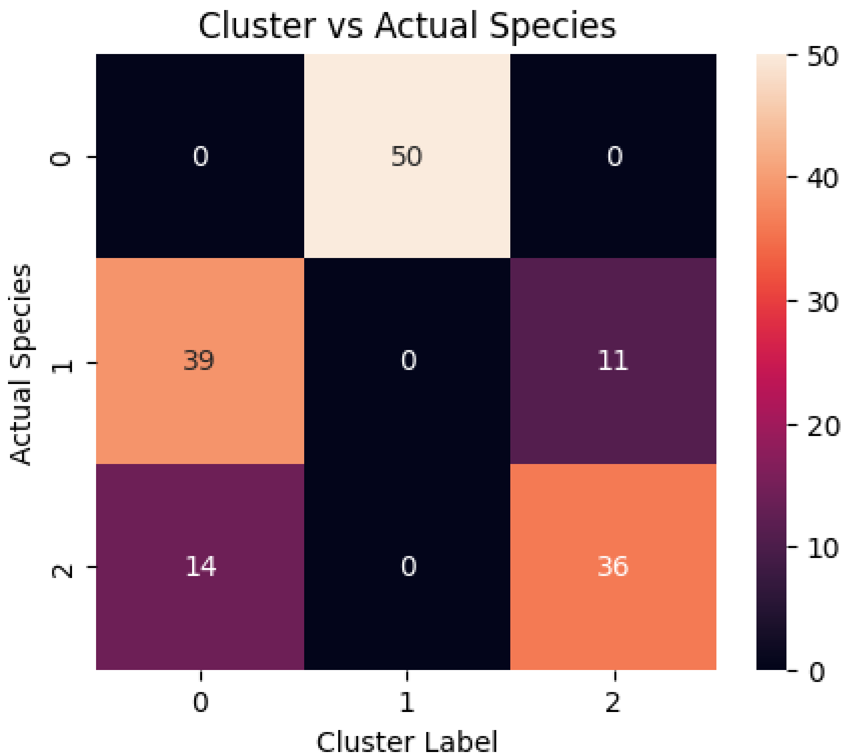
Output




Hierarchical Clustering Dendrogram
Distance
0
5
10
15
20
25
Cluster Size
15 32 33 10 48 21 (3) 14 18 (2) 30 (3) (2) (2) 8 (2) 42 (4) 22 6 (2) (2) (3) (3) (8) 41 60 98 (2) 66 84 (3) (3) 62 (3) 106 (12) 114 (4) (3) (5) (4) (13) (8) (7) (2) (2) (2) (2) 109 (2) (3) (12)



K-Means Clusters (PCA View)
Principal Component 1
Principal Component 2
−3
−2
−1
0
1
2
3
2
1
0
−1
−2


Cluster vs Actual Species
Actual Species
Cluster Label
0
50
0
39
0
11
14
0
36
0
1
2
0
1
2
50
40
30
20
10
0