Assignment#05: K-means and Hierarchical Clustering

Dataset Summary

The Iris dataset is a dataset in machine learning, used for classification and clustering tasks. It contains measurements of different physical characteristics of three species of Iris flowers: Iris-setosa, Iris-versicolor, and Iris-virginica. Each sample in the dataset consists of four features that describe the flowers' physical characteristics.

- **Number of samples**: 150
- **Number of features**: 4
- **Number of classes**: 3 (one for each species of Iris flower)

Description of Variables

Input Variables (Features)

These features are numeric measurements taken from each Iris flower:

1. **Sepal Length**: Length of the sepal in centimeters.
2. **Sepal Width**: Width of the sepal in centimeters.
3. **Petal Length**: Length of the petal in centimeters.
4. **Petal Width** : Width of the petal in centimeters.

These input features are used to describe the physical properties of each Iris flower.

Output Variable (Target)

1. **Species** : The class label of the Iris flower. There are three possible species in the dataset: Iris-setosa, Iris-versicolor and Iris-virginica.

We have omitted the species label because clustering is an unsupervised technique, meaning it doesn’t use labeled output data to guide the clustering. However, we can compare the clusters to these known species labels after the clustering analysis.

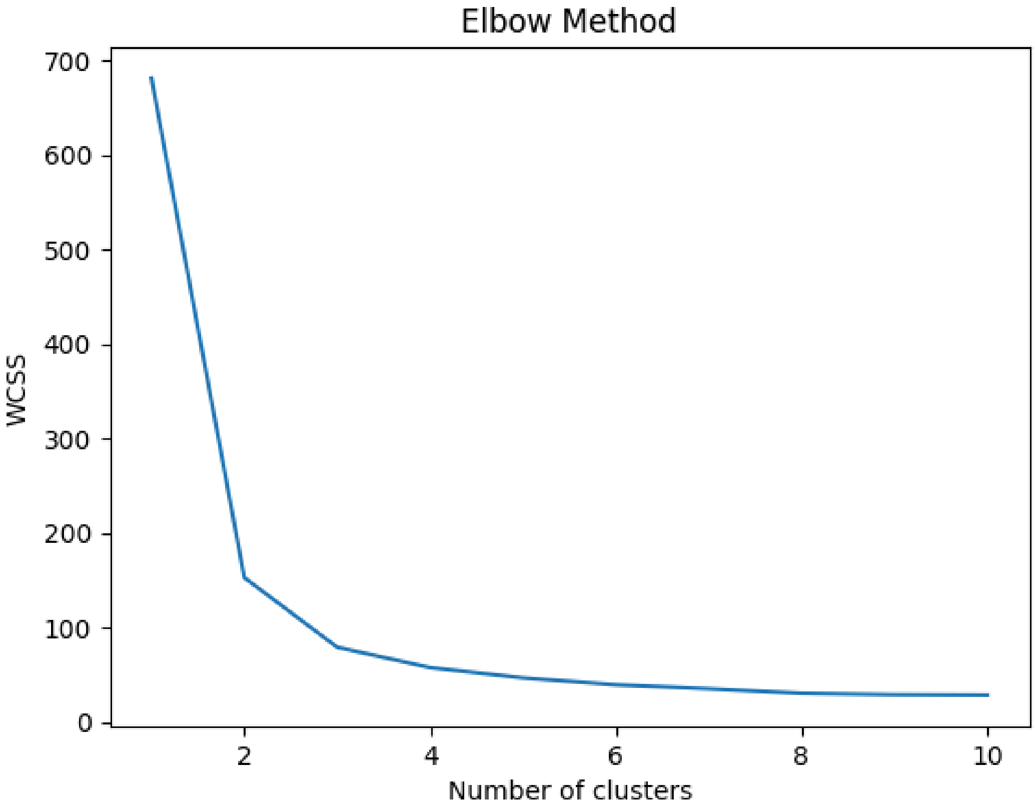
Analysis

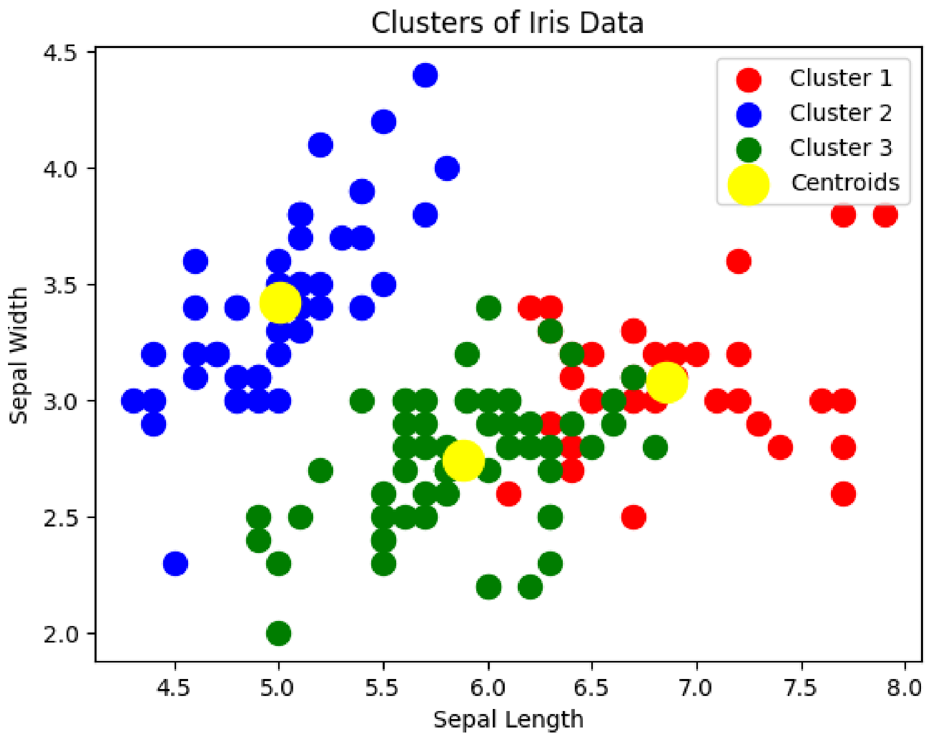
K-means Clustering

Using the **Elbow Method**, we determined that the optimal number of clusters is **3**, which aligns with the number of Iris species in the dataset. After applying K-means clustering with k=3, the data was grouped into three clusters based on the similarities in Sepal and Petal measurements.

The K-means clustering grouped data points by minimizing the within-cluster variance. When visualized on a 2D plot of Sepal Length vs. Sepal Width, we observed that K-means was able to capture the general distribution of the data points across three clusters, although there was some overlap.

After clustering, we compared the clusters with the actual species labels. The results showed a high degree of alignment between the clusters and the true species, indicating that the K-means algorithm effectively identified natural groupings that approximate the actual species divisions.



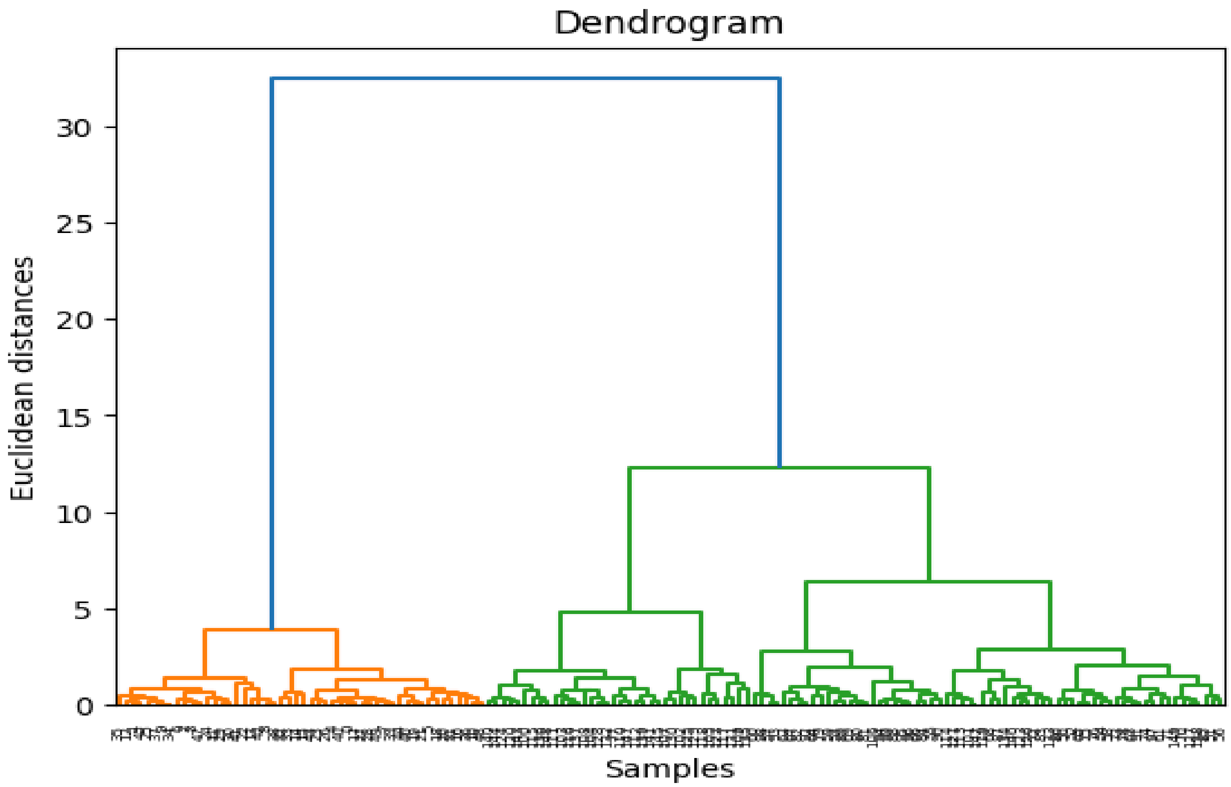


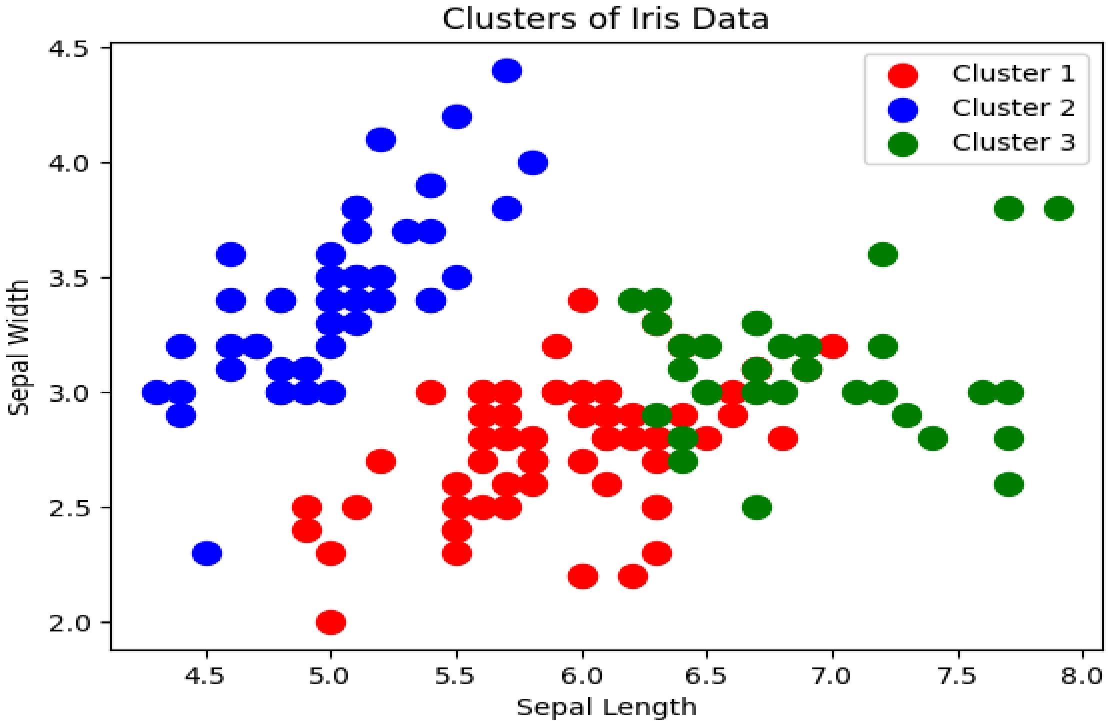
Hierarchical Clustering

We also used **Hierarchical Clustering** with n_clusters = 3, determined from analyzing the **dendrogram**. Hierarchical clustering organizes data into a tree-like structure, which allows for visual interpretation of relationships between clusters.

The dendrogram showed clear separations at a threshold that suggested three primary clusters, consistent with our findings from K-means. This tree structure allowed us to view how individual samples are related within and across clusters.

Plotting the clusters on Sepal Length vs. Sepal Width showed a similar distribution of clusters to K-means. Hierarchical Clustering also effectively separated the data into three main groups that correlate with the three species, though some points overlapped.


Dendrogram


Clusters of Iris Data

Conclusion

K-means and Hierarchical Clustering algorithms effectively identified three distinct groups in the Iris dataset, corresponding closely to the actual species labels.