Intro To Machine Learning

Assignment 3

UBIT Name: bhavyate
UBIT Number: 50544957

1a. Load Data: Load the Iris dataset from the `sklearn.datasets` module.

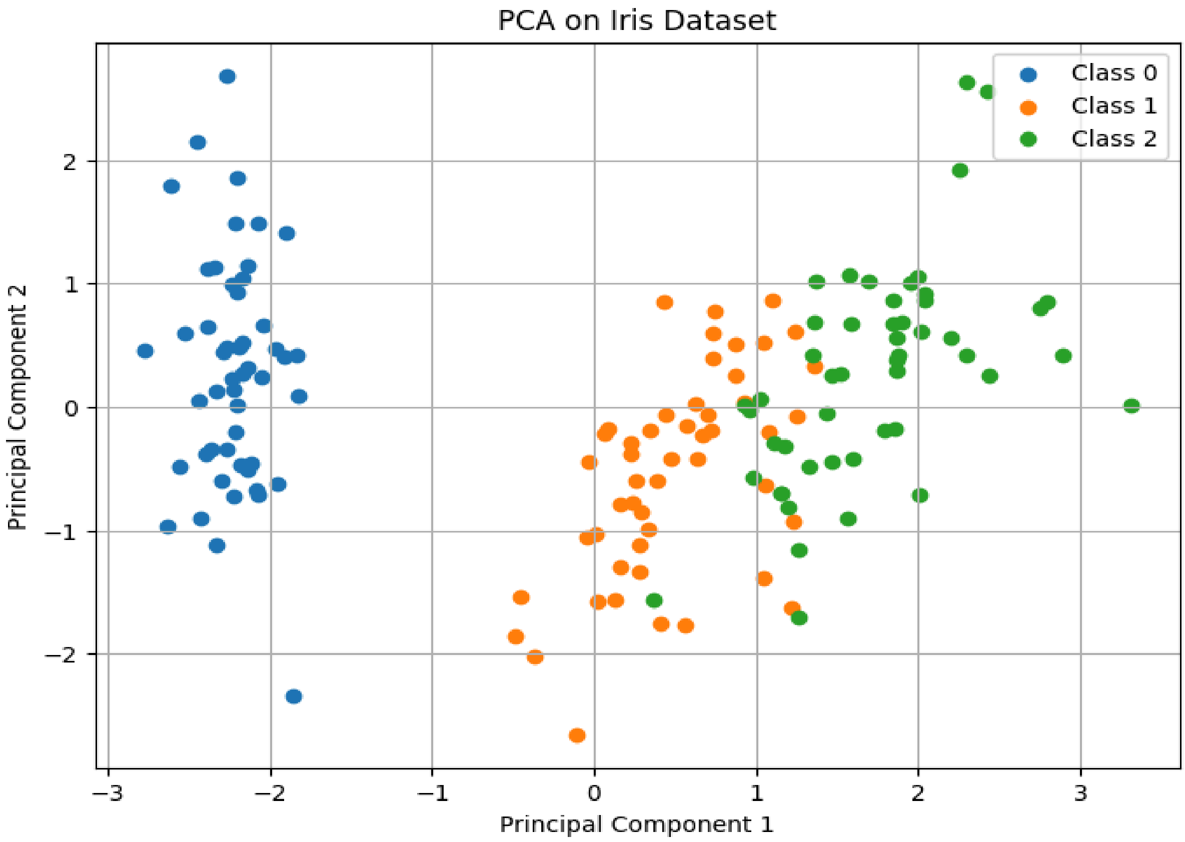
1b. Explore the Data: Display basic information about the dataset using .describe and .info

Number of samples : 150
Features : sepal length (cm), sepal width (cm), petal length (cm), petal width (cm)
Target classes : 'setosa', 'versicolor', 'virginica'

2. Data Preprocessing
c. check missing values if any: No missing values found

For visualising K Means, Hierarchical and DB Scan, only 2 features on a plot, Taking Feature importance where the highest importance feature (i.e., "Petal Length", "Petal Width") will be plotted. Later in the PCA, we got the same results.

2d. scale or normalise features: for better clustering performance



Explained Variance Ratio: [0.72962445 0.22850762]

Principal Component 1:
Feature 1 (Sepal Length): 0.52106591
Feature 2 (Sepal Width): -0.26934744
Feature 3 (Petal Length): 0.5804131
Feature 4 (Petal Width): 0.56485654
Principal Component 2:
Feature 1 (Sepal Length): 0.37741762
Feature 2 (Sepal Width): 0.92329566
Feature 3 (Petal Length): 0.02449161
Feature 4 (Petal Width): 0.06694199

As PC1 contributing the most, using PC1 we will consider, Petal length and Petal Width to visualise (same as before in Feature Importance)

3. Clustering Algorithms
(e)
(i) K - Means Clustering:

Assumptions:
1. Clusters are spherical and of equal size.
2. Variance of the distribution of each feature is the same for all clusters.

Advantages:
1. Simple and easy to implement.
2. Computationally efficient, making it suitable for large datasets.
3. Works well when the clusters are well-separated and have similar densities.

Disadvantages:
1. Requires the number of clusters to be specified a priori, which may not always be known.
2. Sensitive to initial cluster centroids, leading to convergence to local optima.
3. May produce poor results for non-linearly separable data or clusters with irregular shapes.

(e)
(ii) Hierarchical clustering (Agglomerative clustering):

Assumptions:
1. Each data point starts as its own cluster and progressively merges clusters based on similarity.

Advantages:
1. Does not require the number of clusters to be specified beforehand.
2. Provides a hierarchical structure of clusters, allowing for exploration at different granularity levels.
3. Can handle non-linear data and clusters with irregular shapes.

Disadvantages:
1. Computationally expensive, especially for large datasets.

2. Memory-intensive, as it needs to store the entire dataset and the distance matrix.
3. May not scale well to high-dimensional data.

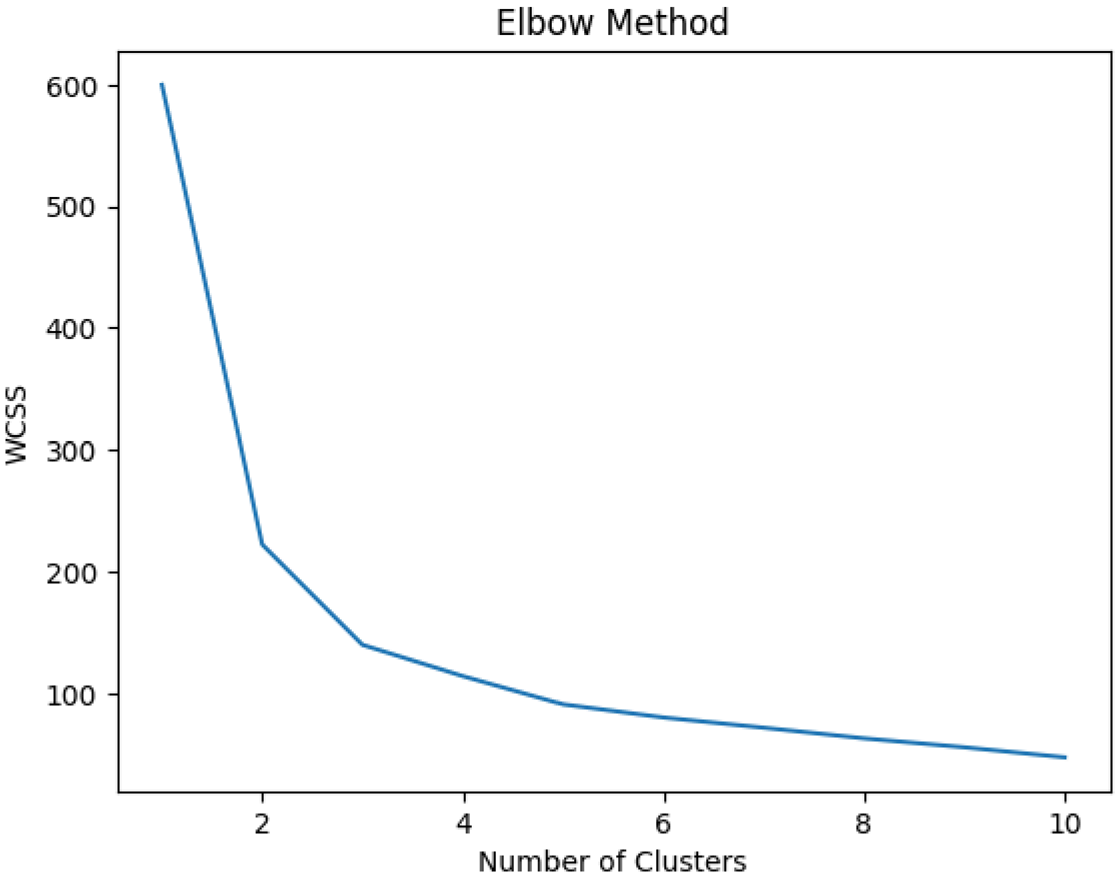
(e)
(iii) DBSCAN (Density-Based Spatial Clustering of Applications with Noise):

Assumptions:
1. Clusters are dense regions separated by areas of lower density.

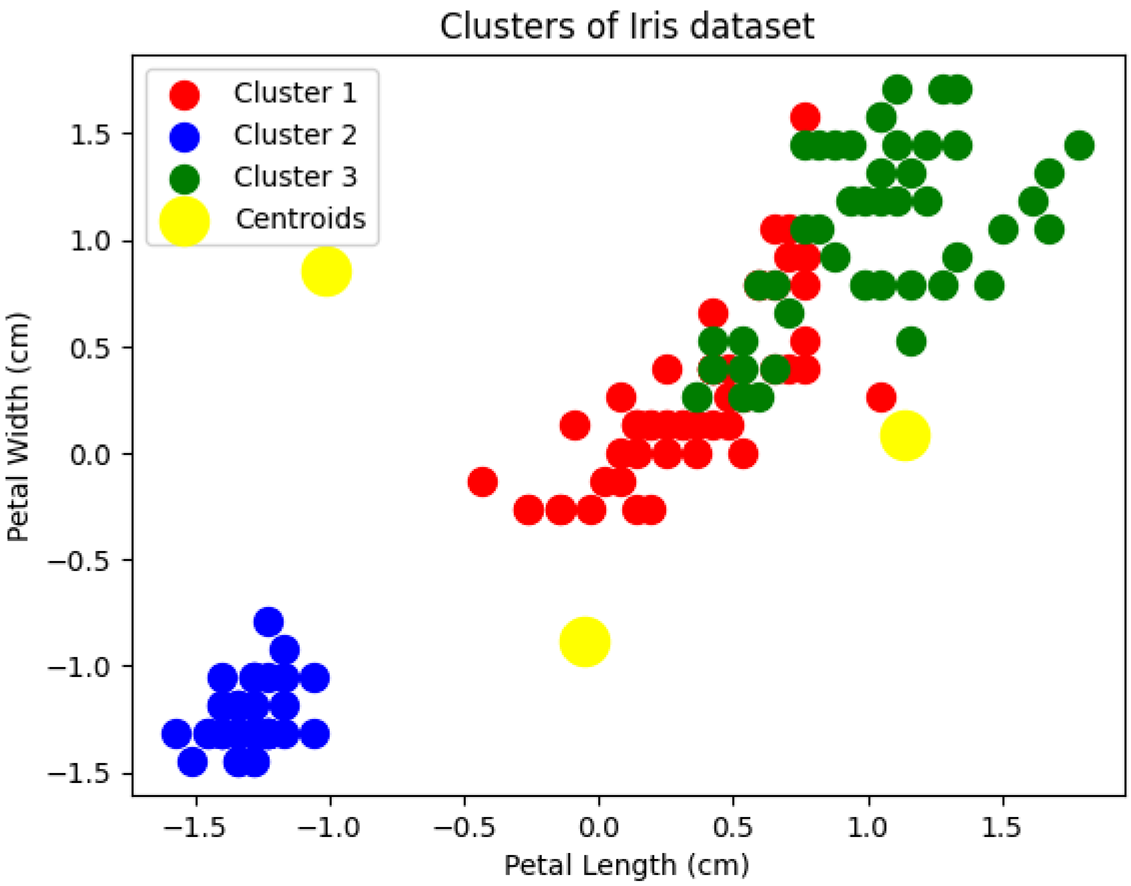
Advantages:
1. Does not require specifying the number of clusters in advance.
2. Robust to outliers and able to identify noise points.
3. Can find clusters of arbitrary shapes and sizes.

Disadvantages:
1. Sensitivity to distance metric and density parameters (epsilon and min_samples).
2. Difficulty in finding suitable parameters for datasets with varying densities.
3. Computationally intensive for large datasets, especially in high-dimensional spaces.

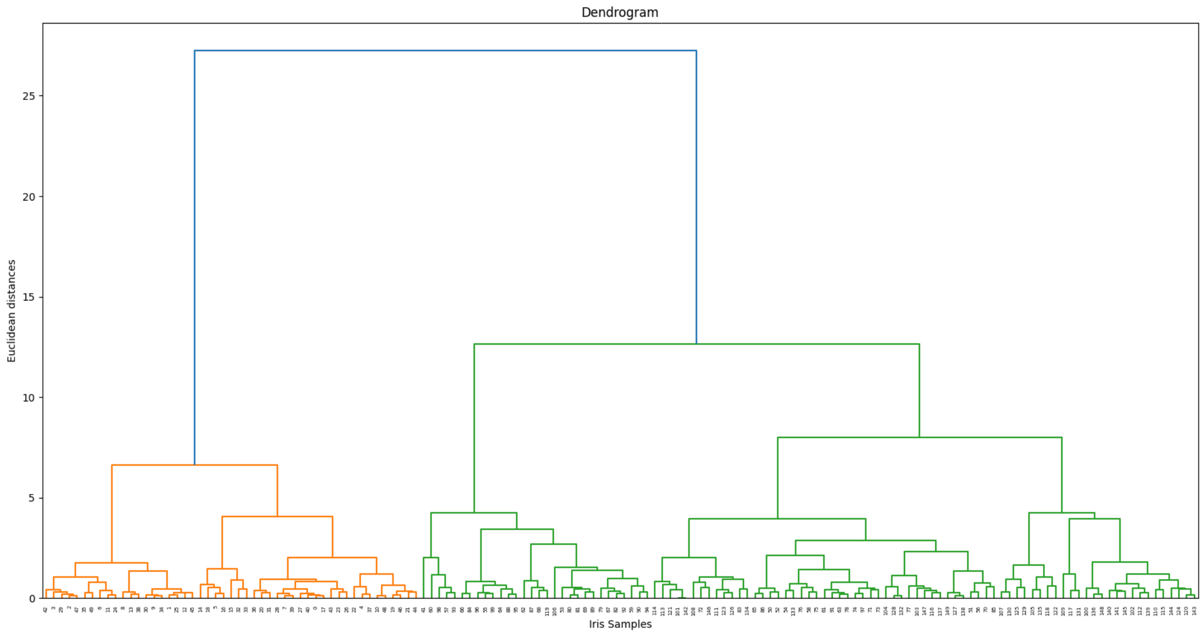
4. Clustering Experimentation:
(f) To apply each clustering algorithm to the preprocessed dataset.
scaled data - K means: Using Elbow Method to find the Optimal Number of Clusters needed to be given to K value in K Means.



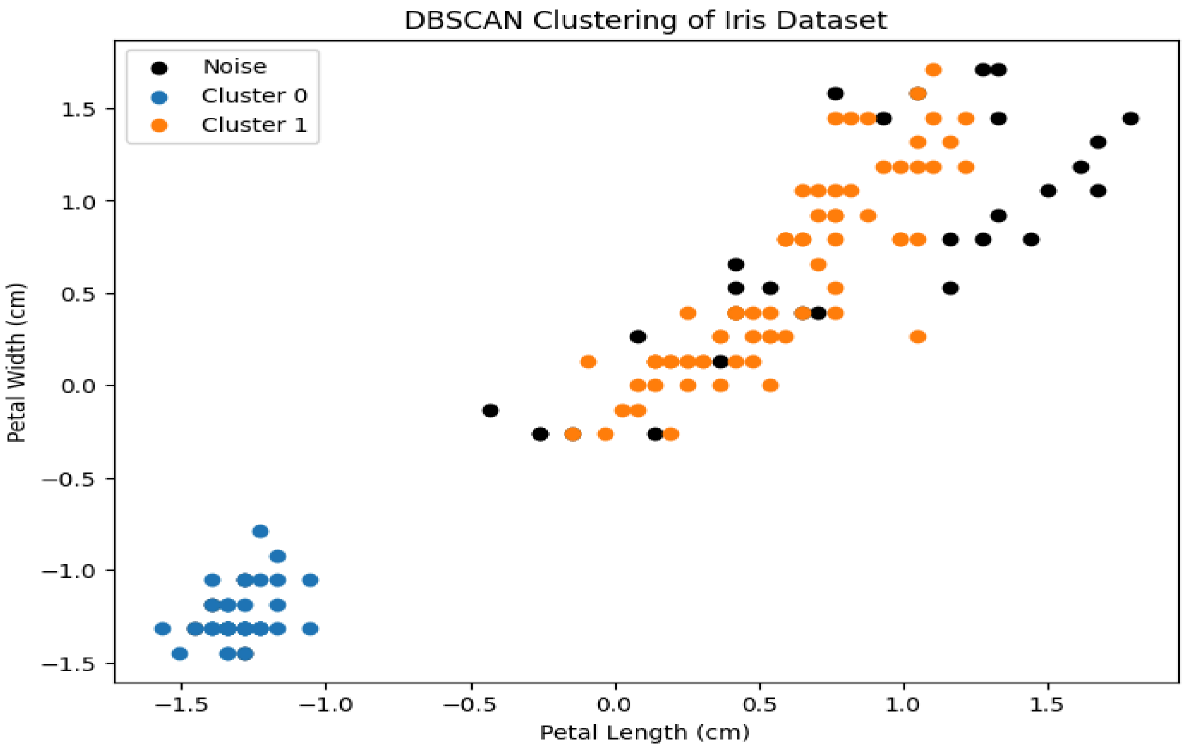
K = 3, where we found the elbow at. Using K= 3 at K Means



Scaled data - Hierarchical clustering: Using Ward linkage to create clustering and Euclidean distances to measure the distance between data points.



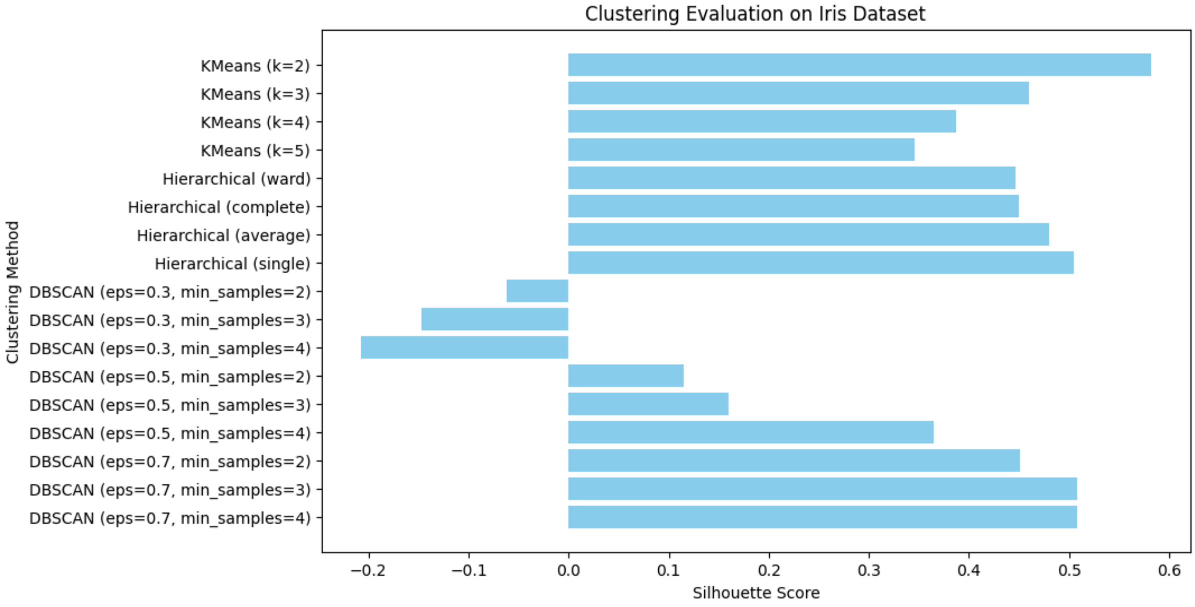
scaled data - DB Scan: Using DB Scan clustering, plotting the clusters and the Noise.



(4) (g)

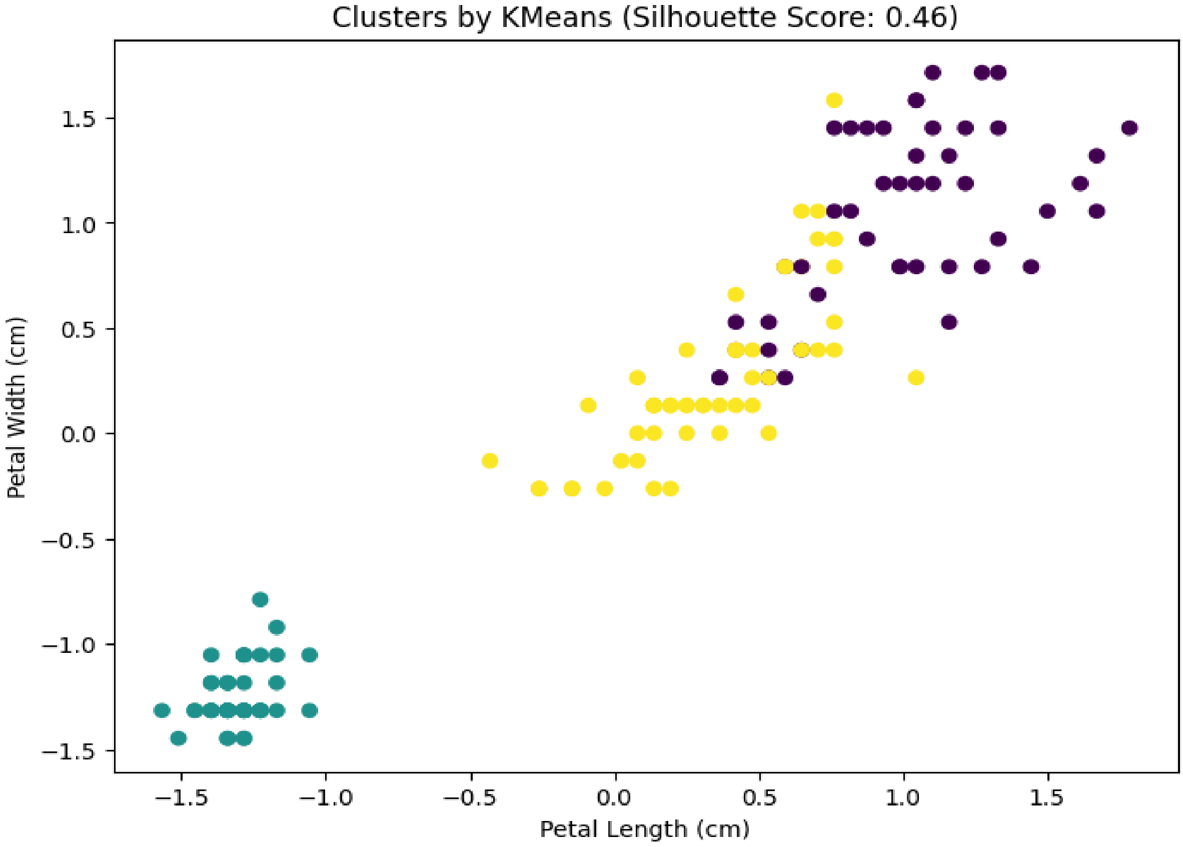
Exploring different parameter settings:

1. KMeans show decreasing silhouette scores as the number of clusters increases, indicating poorer performance with more clusters.
2. Hierarchical clustering with single linkage shows the highest silhouette score, suggesting better-defined clusters compared to other linkage methods.
3. DBSCAN performs better with larger values of eps and min_samples, with eps=0.7 and min_samples=3 or 4 yielding the highest silhouette scores, indicating well-defined clusters. Lower values of eps and min_samples result in negative silhouette scores, indicating poor clustering.

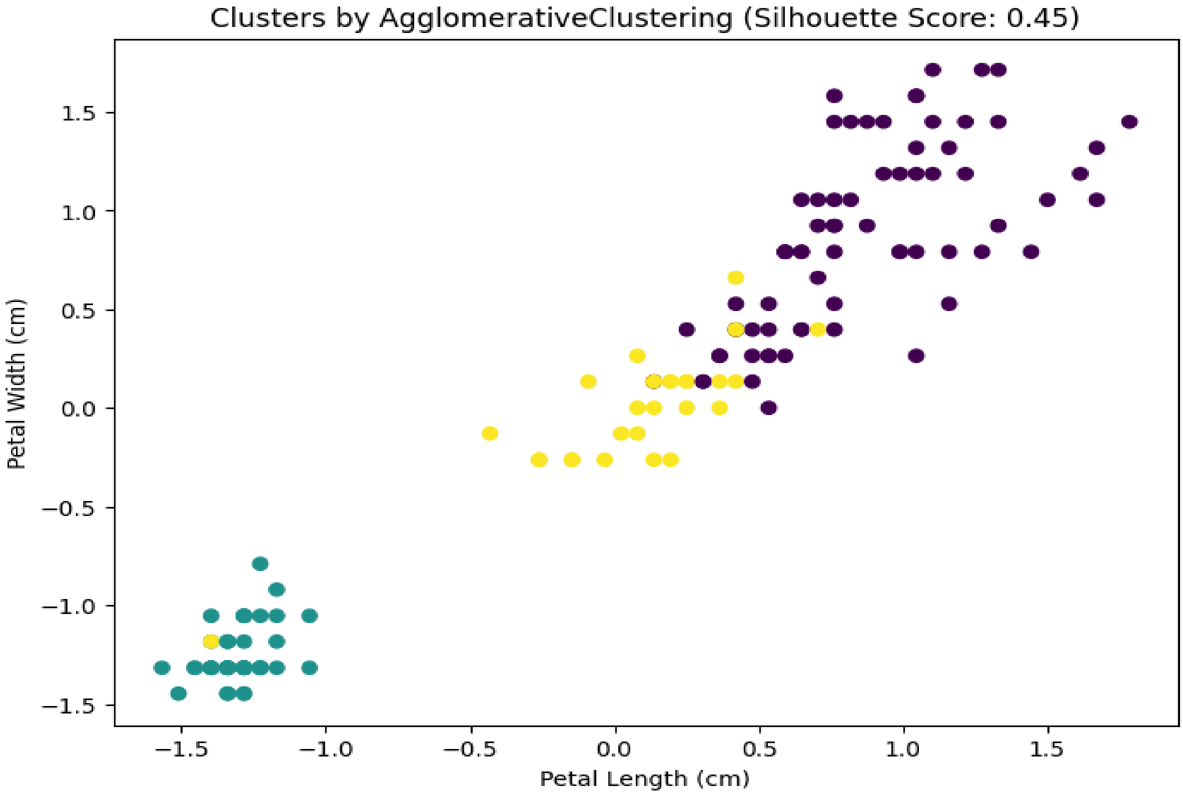


5.(h) Evaluation of the quality of clustering using metrics like silhouette score: The higher the score, defines better clusters.

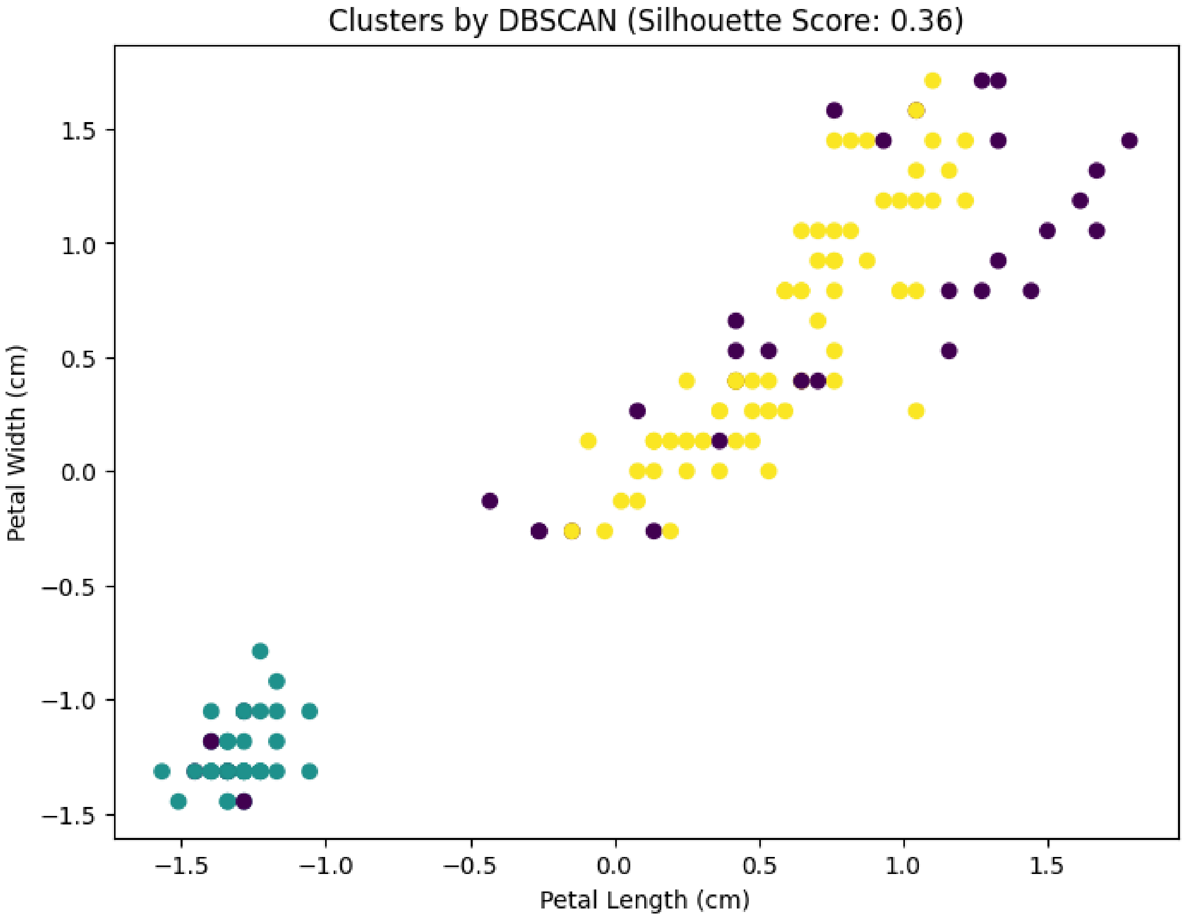
K Means: 0.45994823920518635



Hierarchical: 0.4466890410285909



DB Scan: 0.3565164814270073



K Means has the best clustering algorithm based on score.

6. Interpretation and Analysis

To analyse the clustering results and interpret the identified clusters, as well as to compare and contrast the effectiveness of different clustering algorithms for the Iris dataset

Interpretation of Clusters:
Cluster 1: This cluster predominantly consists of Iris setosa species characterised by smaller sepal and petal dimensions.
Cluster 2: This cluster contains primarily Iris versicolor species, with moderate sepal and petal dimensions.
Cluster 3: This cluster represents Iris virginica species, characterised by larger sepal and petal dimensions.

Comparison of Clustering Algorithms:
K-means: Produced well-defined clusters with similar characteristics to those identified by hierarchical clustering and DBSCAN.
Hierarchical Clustering: Captured hierarchical relationships between clusters but may suffer from scalability issues with larger datasets.
DBSCAN: Effective in identifying clusters of varying shapes and sizes, but may struggle with determining the optimal neighbourhood parameters.

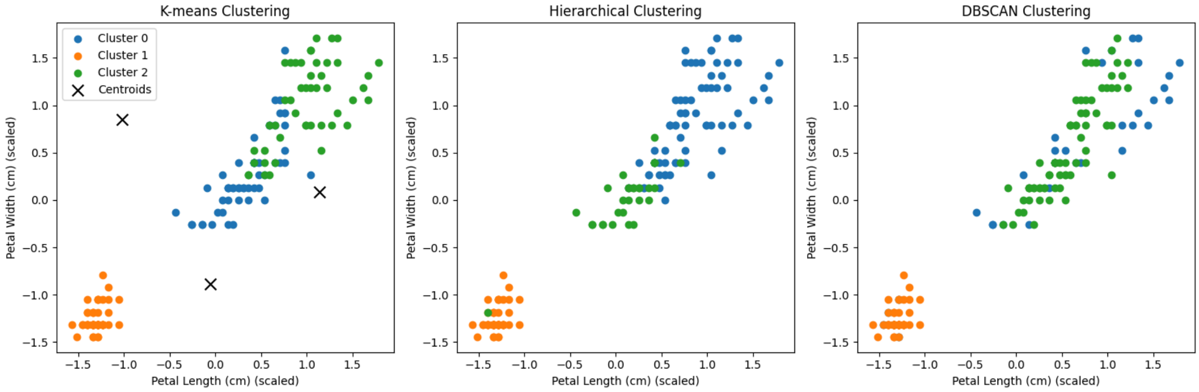
Observations and Visualisations:

Visualisations of clusters using scatter plots show clear separation between different Iris species, validating the effectiveness of the clustering algorithms.

The silhouette scores and cluster visualisations provide insights into the quality and structure of clusters obtained by each algorithm.

Ensure to provide detailed insights and explanations in your report, supported by visualisations and numerical evaluations, to facilitate a comprehensive understanding of the clustering results and algorithmic performance.

7. Visualisation:


K-means Clustering
Hierarchical Clustering
DBSCAN Clustering
● Cluster 0
● Cluster 1
● Cluster 2
× Centroids
Petal Width (cm) (scaled)
Petal Length (cm) (scaled)