**Problem statement:**

Apply DIANA (single linkage, complete linkage, average linkage) on Iris dataset for clustering.

**Dataset Description:**

The Iris dataset is one of the most famous datasets in machine learning and statistics. It is widely used for classification and clustering tasks. The dataset consists of measurements of 150 iris flowers, where each sample corresponds to a single flower from one of three species.

Details of the Iris Dataset:

Number of samples: 150 (50 samples from each of 3 species)

Number of features: 4 features (measurements of the flowers)

Number of classes (species): 3 species

Species: Setosa, Versicolor, Virginica

Features (Columns):

Sepal Length (cm) – The length of the sepal.

Sepal Width (cm) – The width of the sepal.

Petal Length (cm) – The length of the petal.

Petal Width (cm) – The width of the petal.

**Procedure:**

Step 1: Start with all data points as one cluster.

Step 2: Compute the dissimilarity or distance matrix for all data points.

Step 3: Identify the two most dissimilar (or least similar) observations within the cluster (using the dissimilarity matrix).

Step 4: Split the cluster into two smaller clusters. This can be done by selecting the most dissimilar point and assigning it to one of the new clusters, while the remaining points are assigned to the other cluster.

Step 5: Once the cluster is split, update the dissimilarity matrix for the remaining clusters.

Step 6: Repeat steps 3-5 iteratively for the newly created clusters. At each step, find the most dissimilar cluster, and then split it into two smaller clusters.

Step 7: The process continues until all individual data points are in their own clusters.

Step 8: The output of DIANA is a hierarchical clustering tree (dendrogram), which shows how clusters were split at each step.

**SOURCE CODE:**

import numpy as np

import pandas as pd

from scipy.cluster.hierarchy import dendrogram, linkage

import matplotlib.pyplot as plt

from sklearn.preprocessing import StandardScaler

# Load the Iris dataset

df = pd.read\_csv('Iris.csv')

# Drop the 'Id' and 'Species' columns as they are not part of the feature set

X = df.drop(columns=['Id', 'Species'])

# Handle missing values by filling with mean values

X.fillna(X.mean(), inplace=True)

# Remove duplicate rows

X.drop\_duplicates(inplace=True)

# Standardize the features

scaler = StandardScaler()

X\_scaled = scaler.fit\_transform(X)

# Linkage methods to evaluate

linkage\_methods = ['single', 'complete', 'average']

# Plot dendrograms for each linkage method

plt.figure(figsize=(15, 5))

for i, method in enumerate(linkage\_methods, 1):

# Perform hierarchical clustering

Z = linkage(X\_scaled, method)

# Plot dendrogram

plt.subplot(1, 3, i)

dendrogram(Z)

plt.title(f'Hierarchical Clustering Dendrogram ({method.capitalize()} Linkage)')

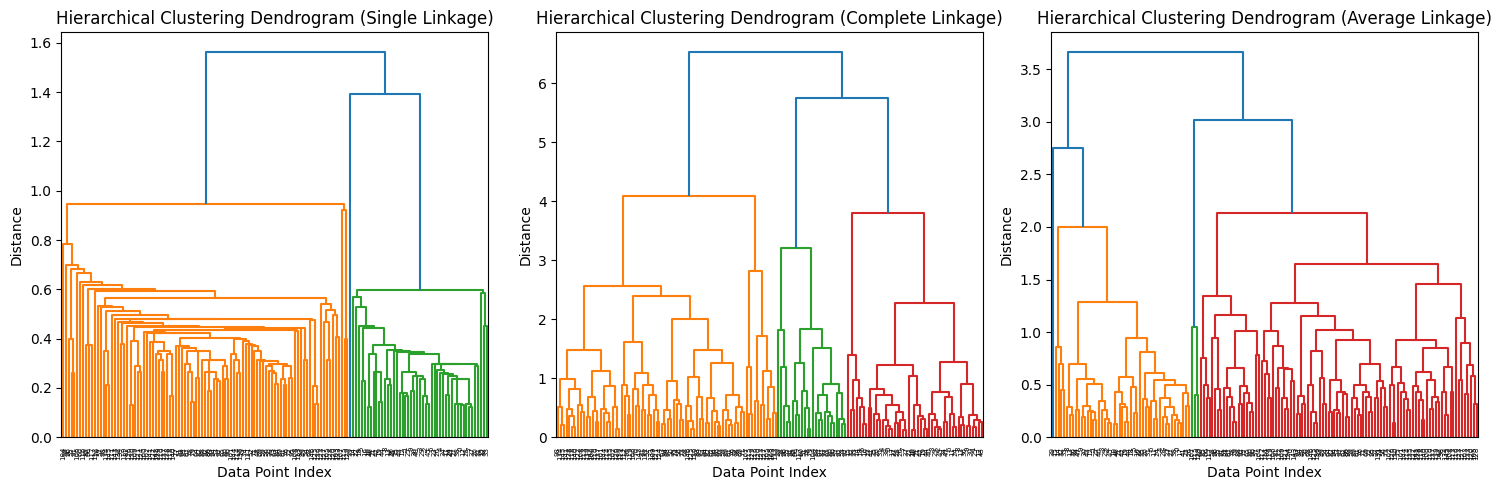
plt.xlabel('Data Point Index')

plt.ylabel('Distance')

plt.tight\_layout()

plt.show()

**Output:**



**Discussion:**

Divisive Analysis clustering (DIANA) is a hierarchical clustering algorithm. Here we have used PCA (Principal component analysis) component to reduce the dimension of the problem and make visualization easier. You can achieve same with Euclidian distance but you cannot visualize the dataset.