Assignment 2
Model Evaluation, Testing (Unsupervised Learning Model)

Problem Statement -

Based on the available dataset, perform the ML analysis. Based on the method described in the class, model the data class. Use Unsupervised Learning algorithms (like K-means clustering OR Hierarchical Clustering.) Check the performance based on the algorithm used.

Submission –

The report (not exceeding 10 pages) shall be drafted. Posters combined into a single pdf document with each analysis included. It is up to the learners how they present well. But mandatorily it should include the presentation in terms of flow and the graphs based on the data. A3 posters should be presented well. Lerner can embed different fonts, pictures, good sketches, animated figures etc. The representation should be quite appealing.

Enrollment No- 09-13 (Hierarchical Clustering)

Submitted by: Vidushi Yadav
AU20B1013

\

Identification of Dataset

“Iris dataset”

It was first introduced by Sir Ronald Fisher, a British statistician and biologist, in his 1936 paper "The use of multiple measurements in taxonomic problems".

The dataset consists of 150 data points, each of which represents an iris flower and is characterized by four features:

Attribute Information:

- Class:

    -- Iris Setosa

    -- Iris Versicolour

    -- Iris Virginica

- Sepal length: the length of the sepal, which is the part of the flower that encloses the petals and forms a calyx
- Sepal width: the width of the sepal
- Petal length: the length of the petal, which is the part of the flower that is often brightly colored and attracts pollinators
- Petal width: the width of the petal

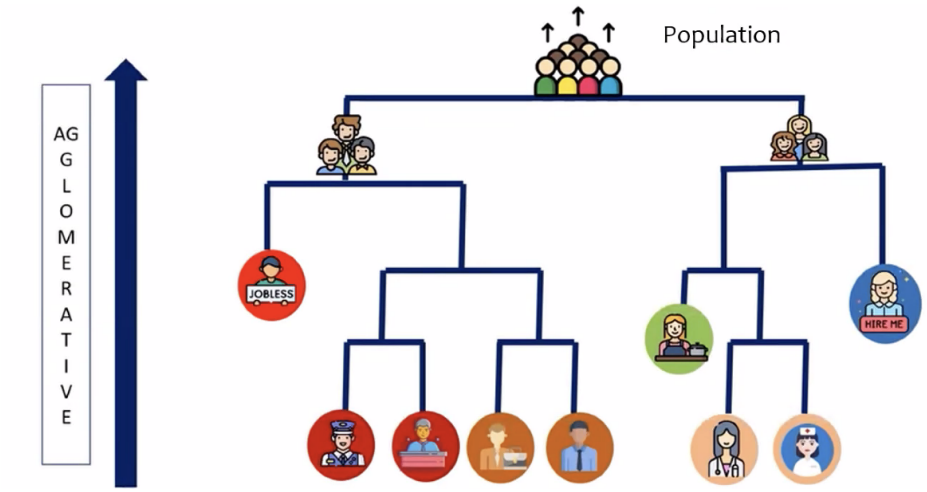
Model Observation

Hierarchical C: It has predetermined ordering from top to bottom

Ex: all files and folders are arranged in a hierarchical order.

1) **Agglomerative Clustering**

Working :We assign each observation to its own cluster, it's necessary which type of cluster we have.



2) **Divisive:** We assign all the observations to a single cluster, and then partition the cluster to 2 least similar clusters.

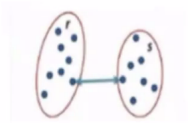
- Then we proceed recursively on each cluster until there is one cluster for each observation,
- Example : K means clustering.

Characteristics :

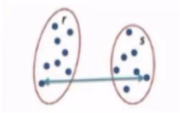
R & S are 2 clusters, same linkage, the length is based on min distance by 2 data points.

Proximity matrix

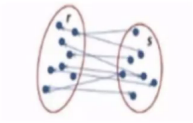
* Before any clustering performed, it is required to determine proximity matrix containing the distance between each point using a distance function.
* Then the matrix is updated to display the distance between each cluster.



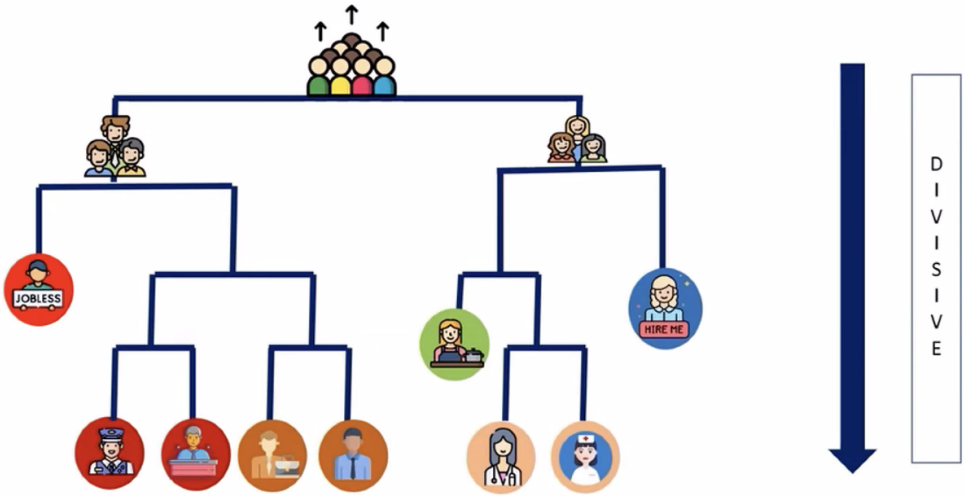
Single linkage
L(r,s) = min(D(Xn,Xy))



complete linkage
L(r,s) = max(D(Xn,Xy))

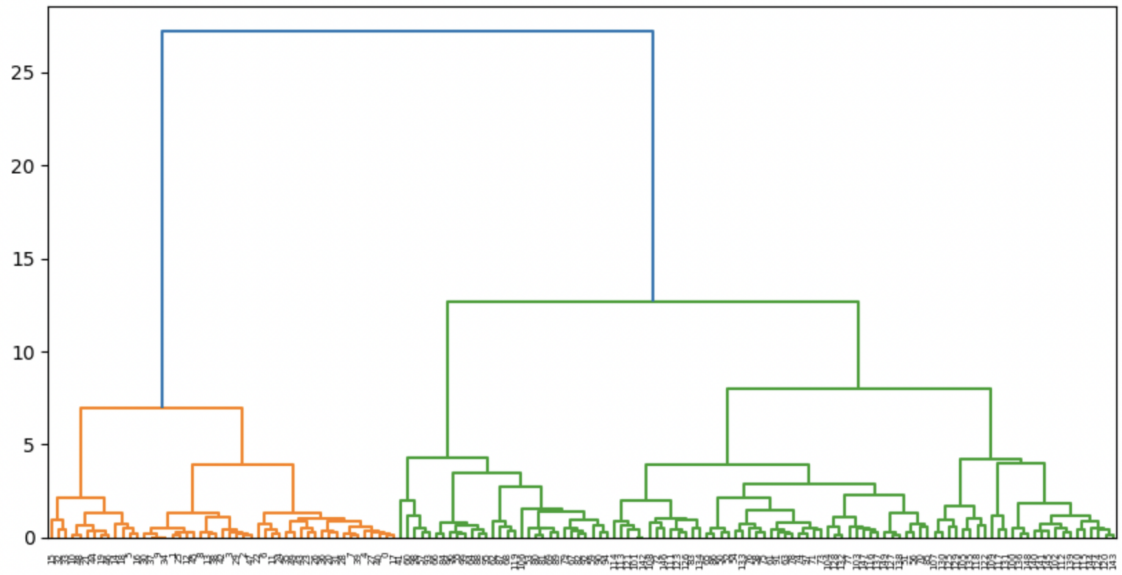


average linkage



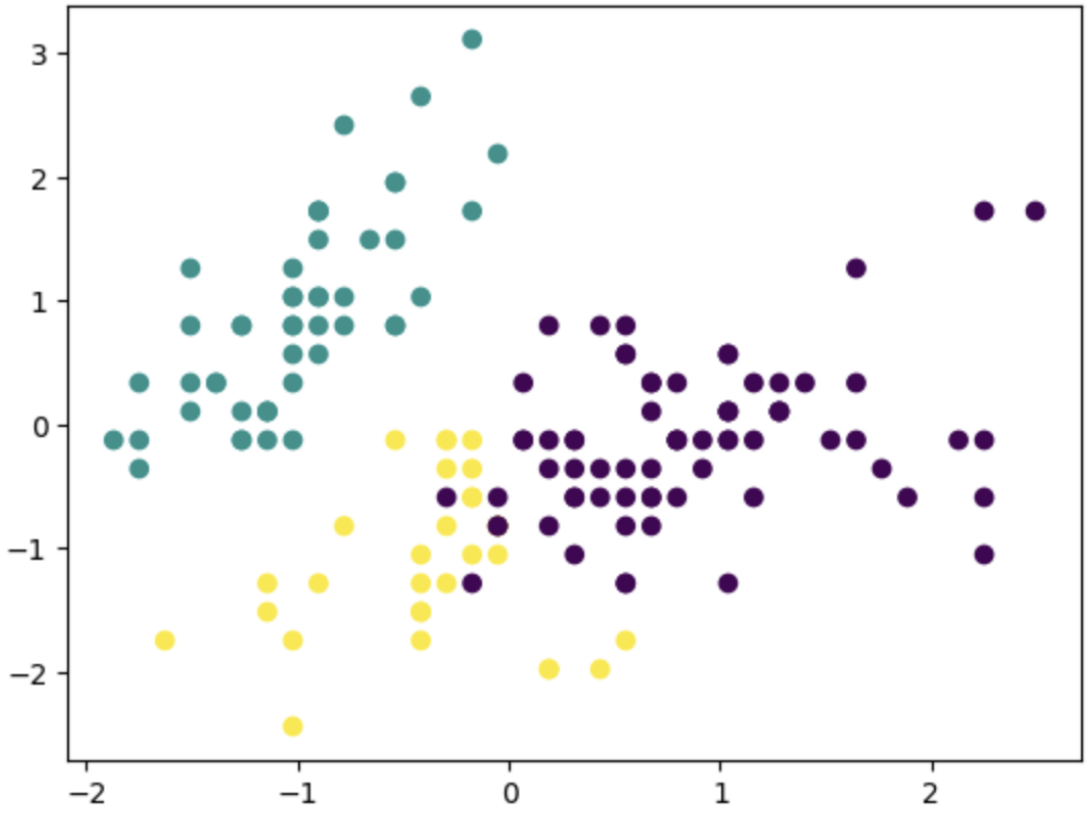
To plot a dendrogram using the hierarchical clustering model, you will need to first create a linkage matrix using the linkage function from scipy.cluster.hierarchy. This function takes the data matrix X as input and returns a linkage matrix that can be used to generate a dendrogram.

```
# Import necessary libraries
import pandas as pd
from sklearn.preprocessing import StandardScaler
from sklearn.cluster import AgglomerativeClustering
from scipy.cluster.hierarchy import linkage, dendrogram
import matplotlib.pyplot as plt

# Load the Iris dataset
df =
pd.read_csv('https://archive.ics.uci.edu/ml/machine-learning-databases/iri
s/iris.data',
                names=['sepal_length', 'sepal_width', 'petal_length',
'petal_width', 'target'])

# Preprocess the data
X = df.drop('target', axis=1)
X = StandardScaler().fit_transform(X)

# Model the data using hierarchical clustering
model = AgglomerativeClustering(n_clusters=3)
predictions = model.fit_predict(X)

# Create a linkage matrix
Z = linkage(X, method='ward')

# Plot a dendrogram of the hierarchical clustering results
plt.figure(figsize=(10, 5))
dendrogram(Z)
plt.show()
```



U16

```
Silhouette score: 0.4455395639920042
Calinski-Harabasz index: 220.2604374375408
```



Model Accuracy:

Hierarchical clustering is an unsupervised learning algorithm, which means that it does not use labeled data.

This code will perform the following steps:

1. Import the necessary libraries
2. Load the Iris dataset from the UCI Machine Learning Repository
3. Preprocess the data by scaling the features and separating the features and target into separate variables
4. Split the data into training and test sets
5. Train the model using logistic regression
6. Make predictions on the test set
7. Evaluate the model using classification accuracy

```
# Import necessary libraries
import pandas as pd
from sklearn.preprocessing import StandardScaler
from sklearn.linear_model import LogisticRegression
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score

# Load the Iris dataset
df =
pd.read_csv('https://archive.ics.uci.edu/ml/machine-learning-databases/iri
s/iris.data',
                names=['sepal_length', 'sepal_width', 'petal_length',
'petal_width', 'target'])
```

```
# Preprocess the data
X = df.drop('target', axis=1)
y = df['target']
X = StandardScaler().fit_transform(X)

# Split the data into training and test sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2,
random_state=42)

# Train the model using logistic regression
model = LogisticRegression()
model.fit(X_train, y_train)

# Make predictions on the test set
predictions = model.predict(X_test)

# Evaluate the model using classification accuracy
accuracy = accuracy_score(y_test, predictions)
print("Classification accuracy:", accuracy)
```

Output:

```
Classification accuracy: 1.0
```

Conclusion:

After completion of this assignment I got hands- on learning and unsupervised learning algorithm, Tree based structure and is based on priority, it has predefined order.

- To conclude the analysis of the data using the hierarchical clustering algorithm, I can now evaluate the performance of the model.
- The purpose of the analysis: What was the goal of the analysis? Did the model achieve this goal? The strengths and limitations of the model: What are the strengths and limitations of the hierarchical clustering algorithm? How did these affect the results of the analysis?

Overall, the hierarchical clustering algorithm is a useful tool for grouping data into clusters based on their similarity. It is particularly useful when you do not have prior knowledge of the structure of the data or when you want to explore the data to discover patterns and relationships. However, it is important to carefully evaluate the results of the analysis using appropriate evaluation metrics and consider the limitations of the model when interpreting the results.