Name :	Dipangshu Roy
Roll No :	001811001014
Department :	Information Technology
Class :	4th Year 1st Sem
Subject :	Machine Learning Lab

Assignment 4

1. Apply the below clustering algorithms using Python:
   a. Partition based: K-means, K-medoids/PAM
   b. Hierarchical: Dendrogram, AGNES, BIRCH
   c. Density based: DBSCAN, OPTICS

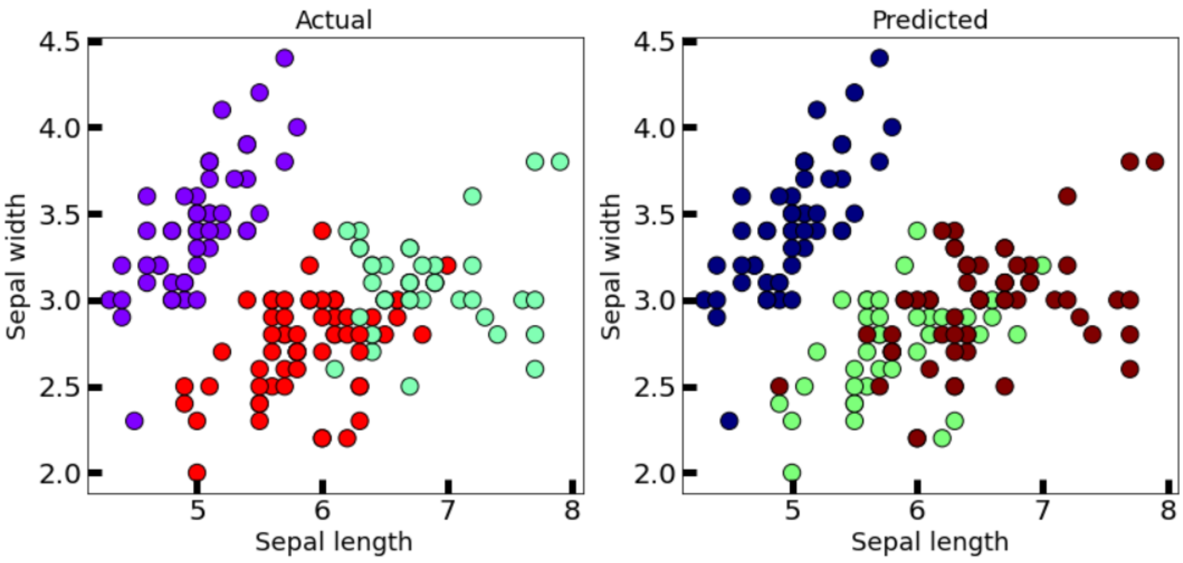
on the following UCI datasets (can be loaded from the package itself):
a. **Iris plants dataset**: https://archive.ics.uci.edu/ml/datasets/Iris/
b. **Wine Dataset**: https://archive.ics.uci.edu/ml/datasets/wine
Additionally, implement **K-means++** and **Bisecting K-means**.

> **Github Link of this Assignment :** https://github.com/Droyder7/ML-Lab-Assignments

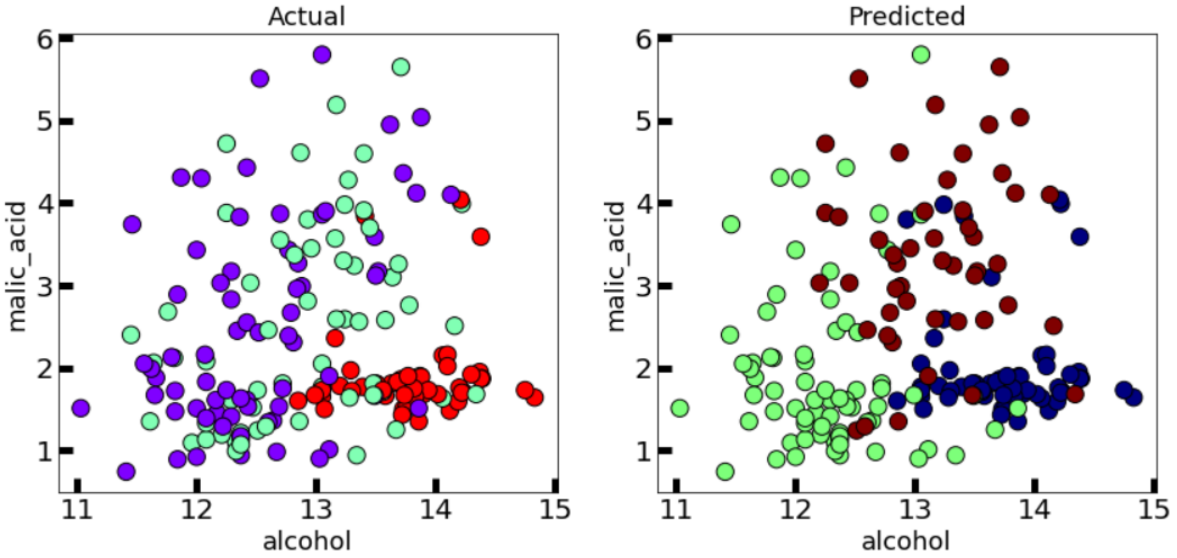
A. Partition Based

1. K-means

i. Iris plants dataset



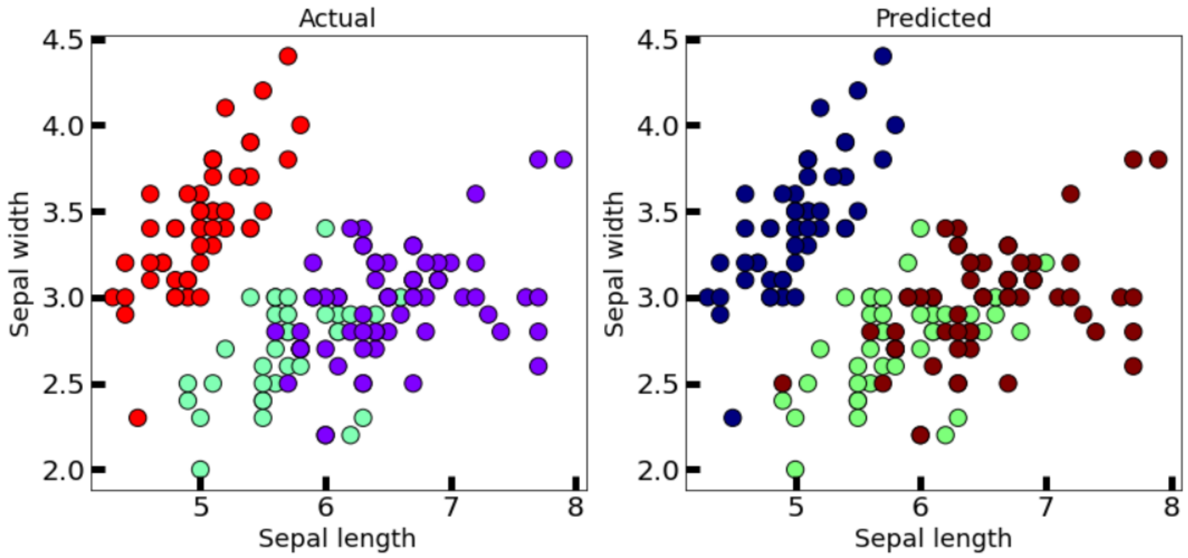
ii. Wine dataset



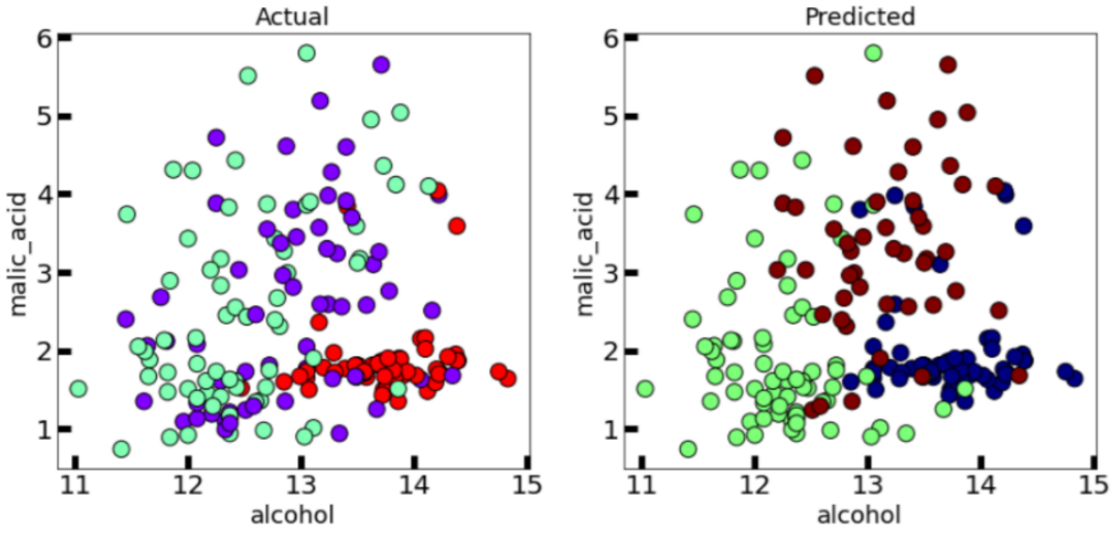
> **This algorithm generalizes to clusters of different shapes and sizes, such as elliptical clusters. The problem with it is that we need to manually choose the value of “k”.**

2. K-medoids

i. Iris plants dataset



ii. Wine dataset

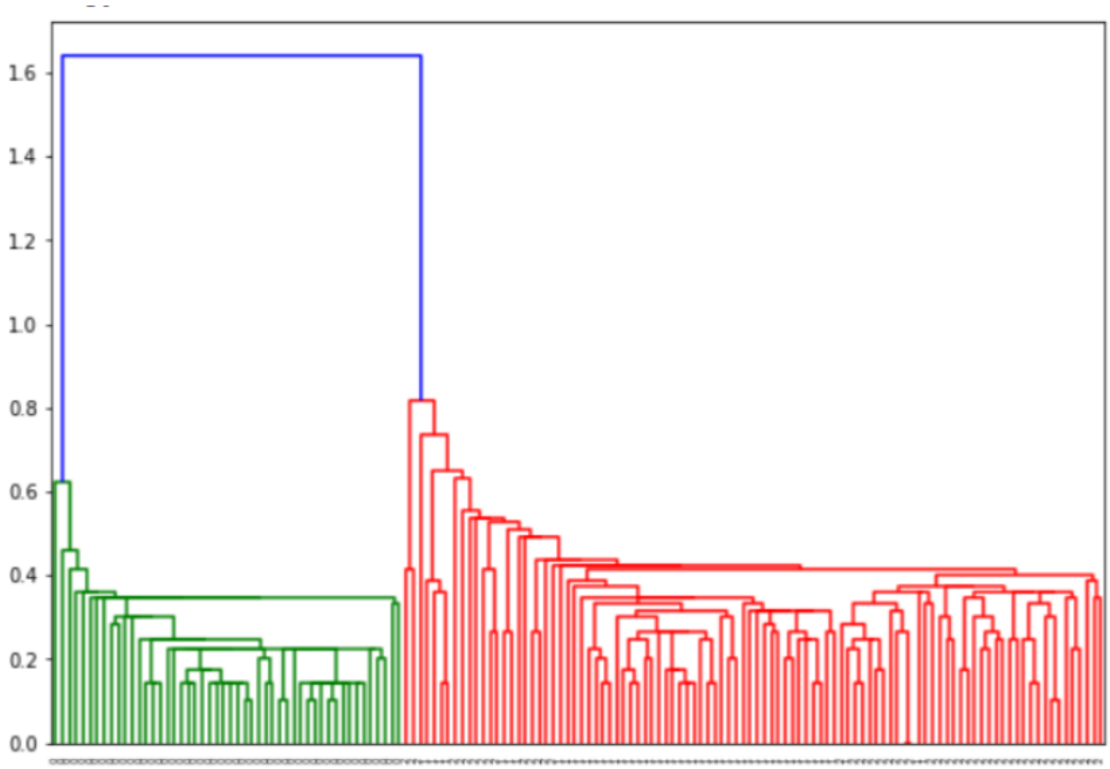


> **This algorithm solves the problem with the K-means algorithm.**
>
> **K-means attempts to minimize the total squared error, while k-medoids minimizes the sum of dissimilarities between points labeled to be in a cluster and a point designated as the center of that cluster.**

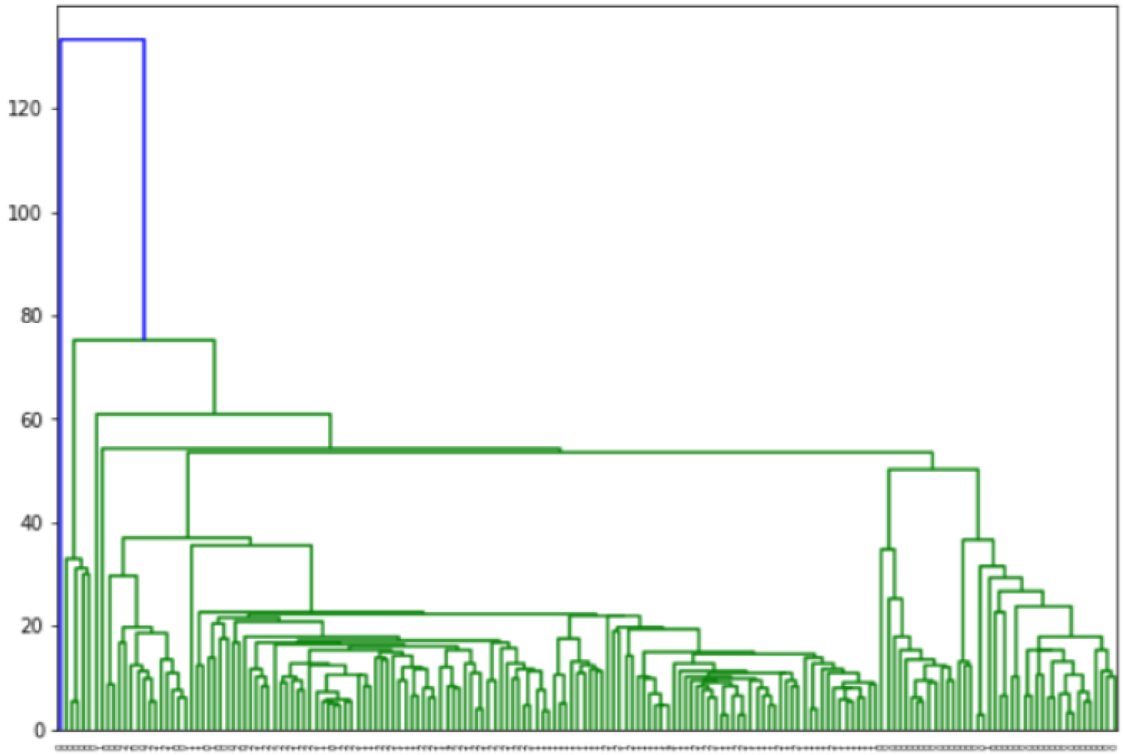
B. Hierarchical

1. Dendrogram

i. Iris plants dataset



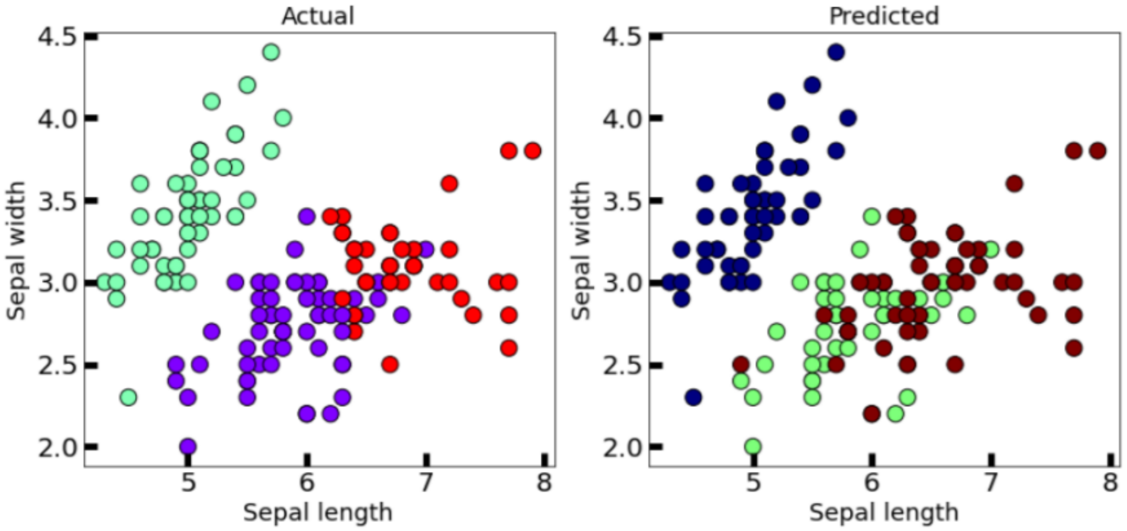
ii. Wine dataset



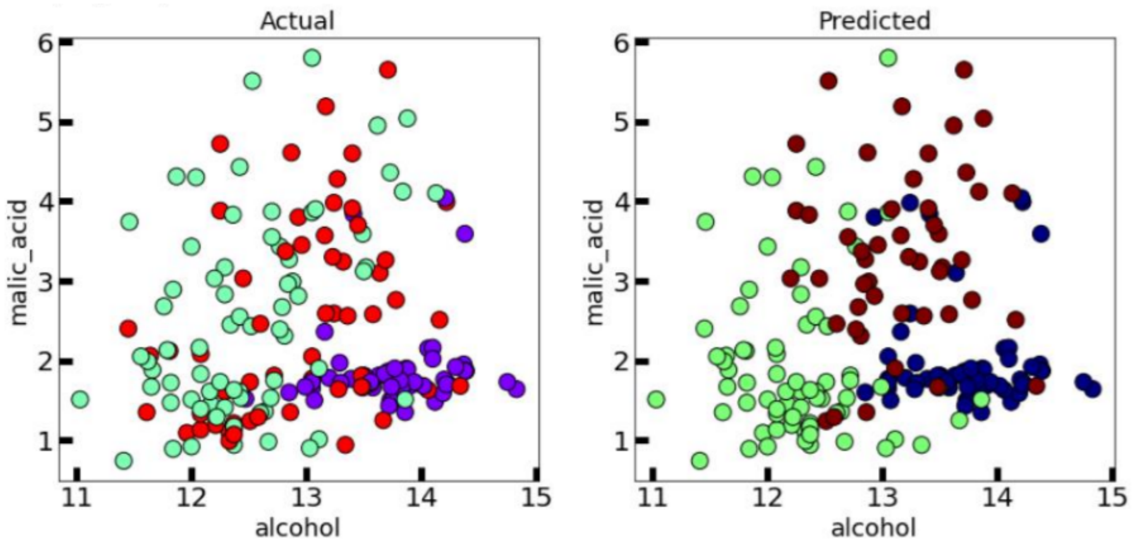
> **A dendrogram is a diagram that shows the hierarchical relationship between objects. It is most commonly created as an output from hierarchical clustering. The main use of a dendrogram is to work out the best way to allocate objects to clusters**

2. AGNES

i. Iris plants dataset



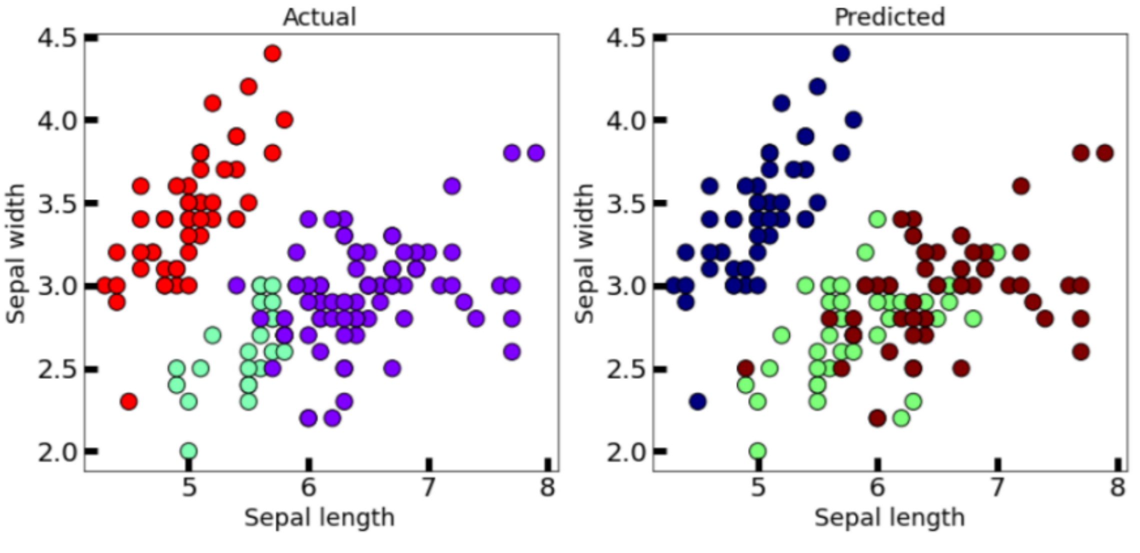
ii. Wine dataset



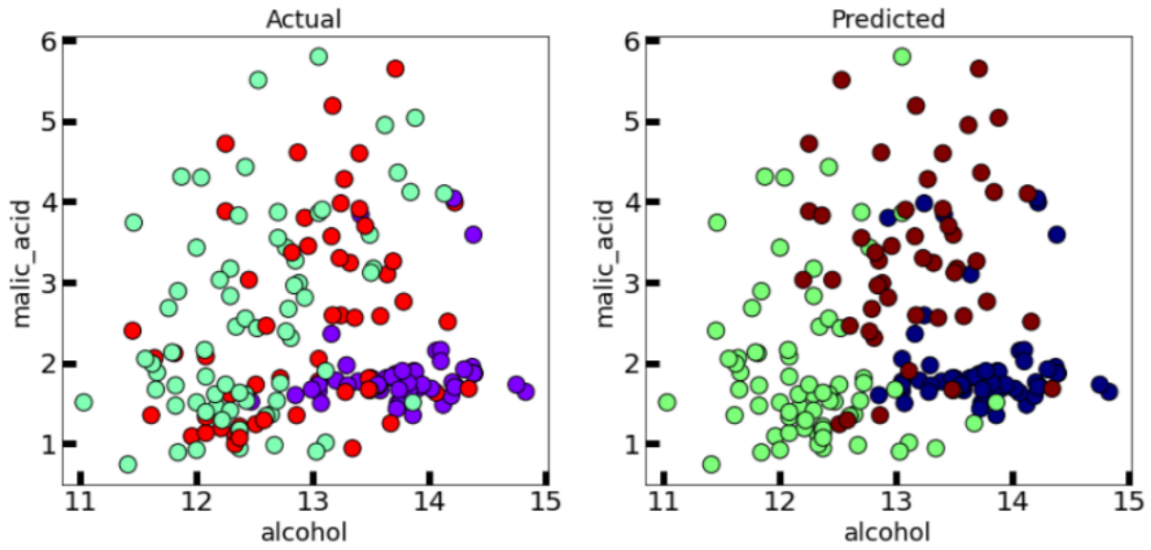
> The agglomerative clustering is the most common type of hierarchical clustering used to group objects in clusters based on their similarity. It's also known as AGNES (Agglomerative Nesting).
>
> The algorithm starts by treating each object as a singleton cluster. Next, pairs of clusters are successively merged until all clusters have been merged into one big cluster containing all objects.
>
> The result is a tree-based representation of the objects, named dendrogram.

3. BIRCH

i. Iris plants dataset



ii. Wine dataset

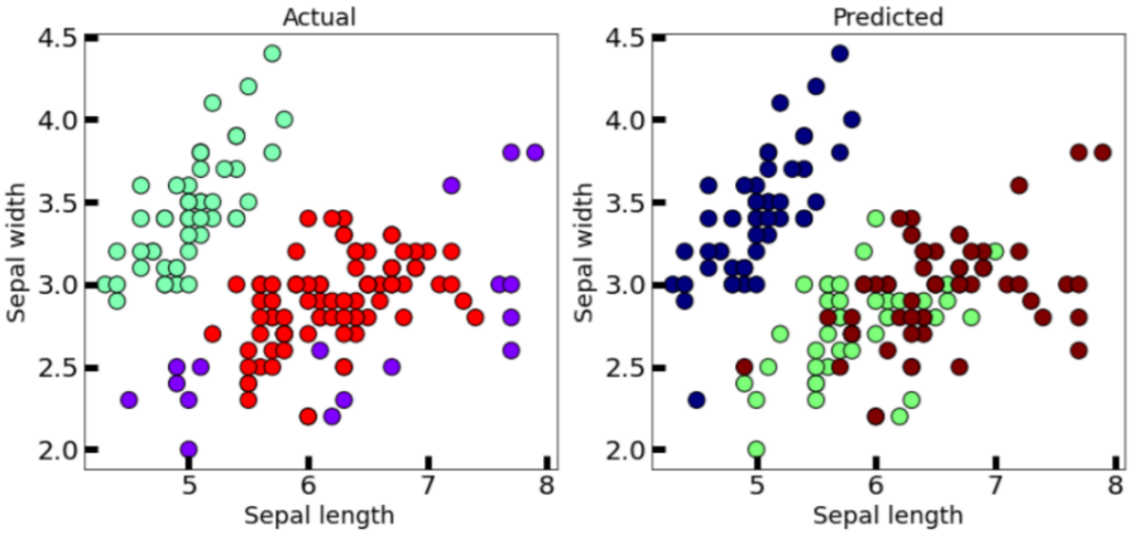


> **Balanced Iterative Reducing and Clustering using Hierarchies (BIRCH) is a clustering algorithm that can cluster large datasets by first generating a small and compact summary of the large dataset that retains as much information as possible.**
>
> **This smaller summary is then clustered instead of clustering the larger dataset**

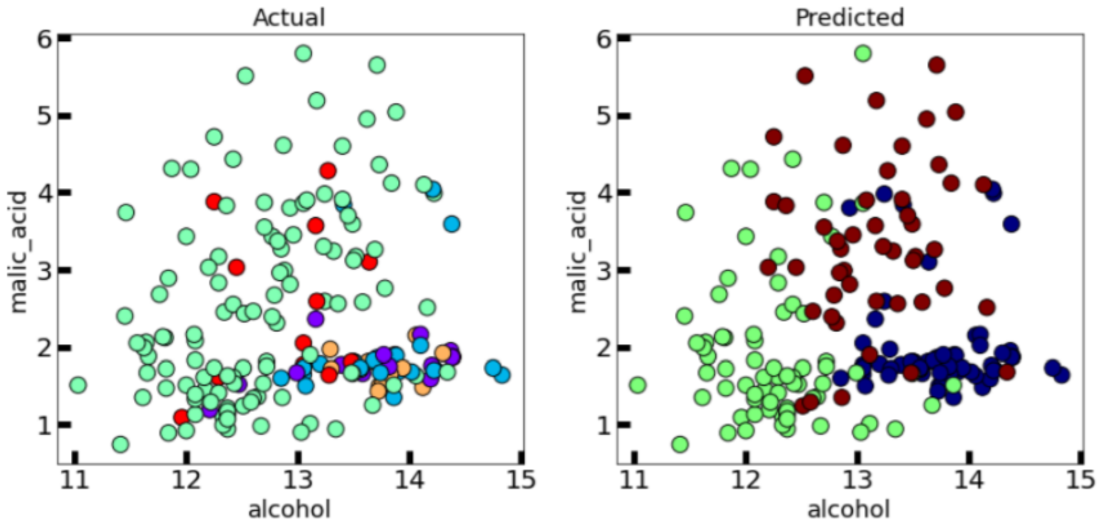
C. Density Based

1. DBSCAN

i. Iris plants dataset



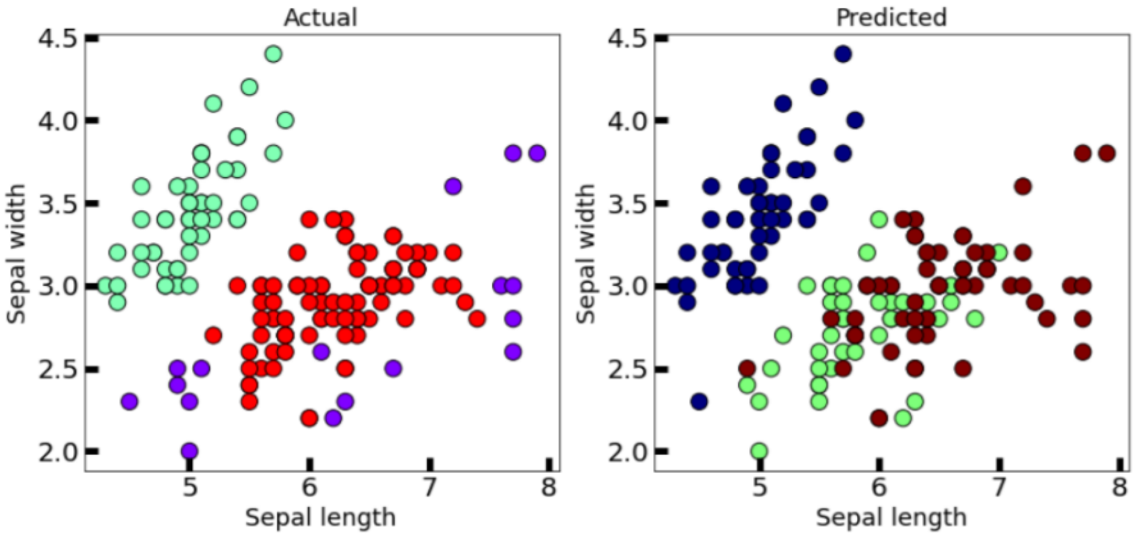
ii. Wine dataset



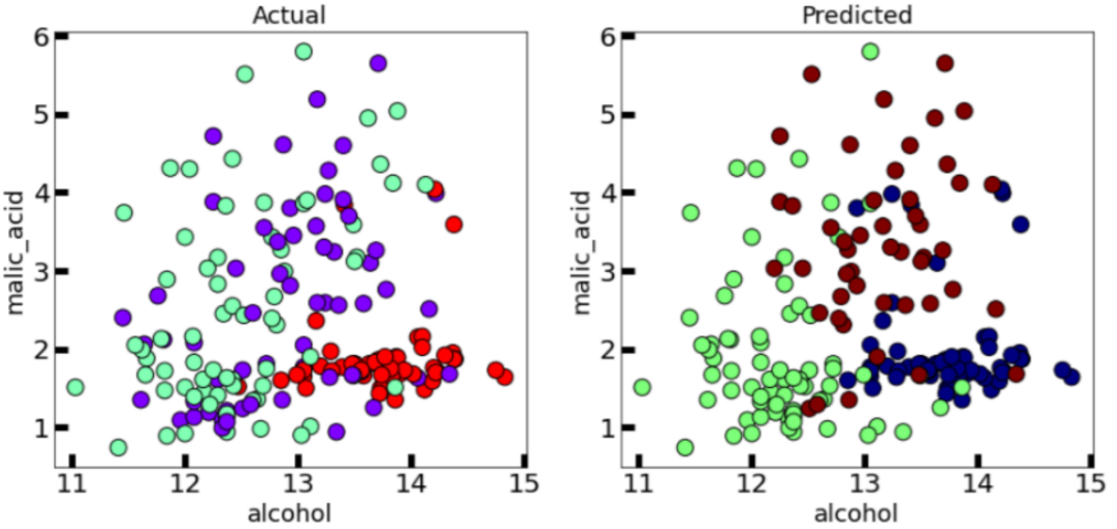
> **Clusters are dense regions in the data space, separated by regions of the lower density of points. The DBSCAN algorithm is based on this intuitive notion of “clusters” and “noise”.**
>
> **The key idea is that for each point of a cluster, the neighborhood of a given radius has to contain at least a minimum number of points**

2. OPTICS

i. Iris plants dataset



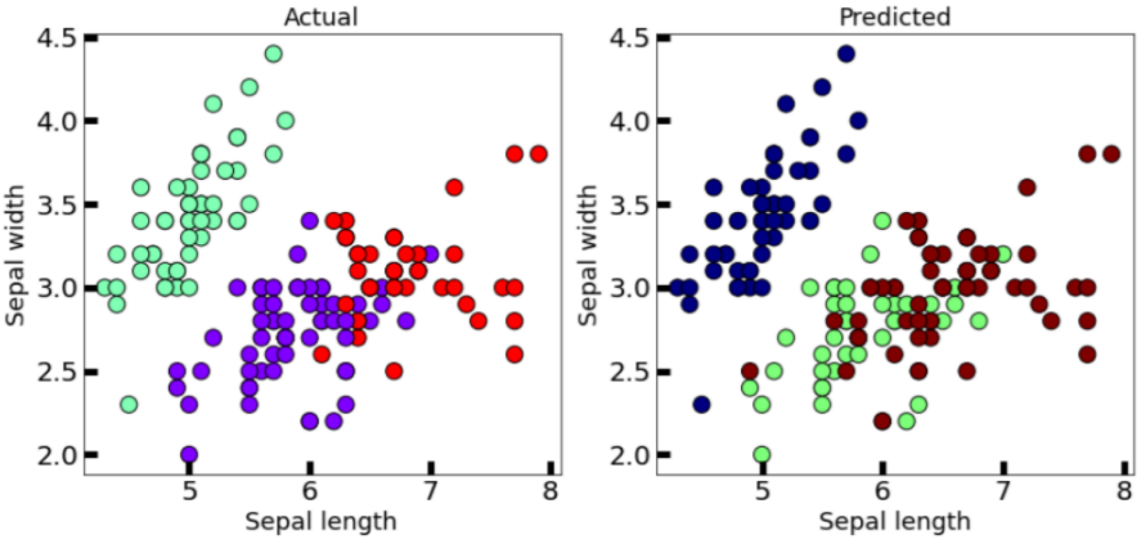
ii. Wine dataset



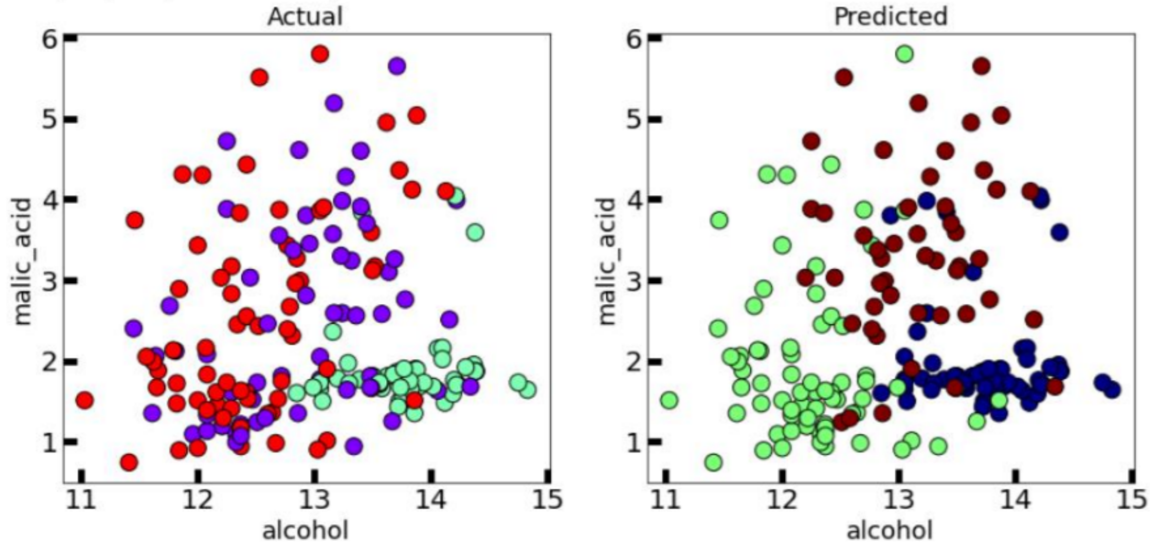
> **This clustering technique is different from other clustering techniques in the sense that this technique does not explicitly segment the data into clusters.**
>
> **Instead, it produces a visualization of Reachability distances and uses this visualization to cluster the data.**

D. K-means++

i. Iris plants dataset



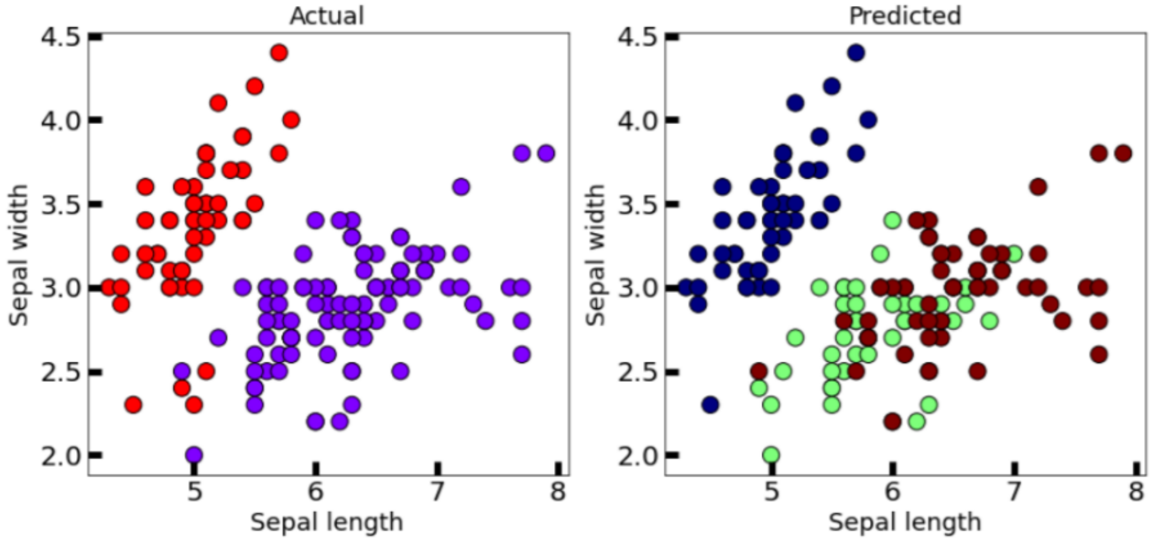
ii. Wine dataset



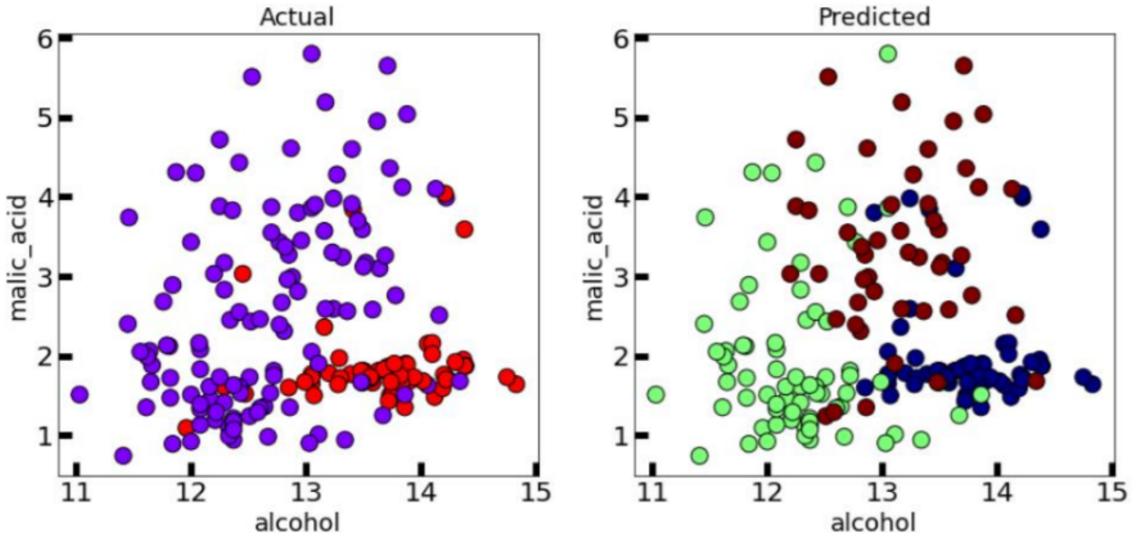
> **In the case of K-Means clustering, we were using randomization. The initial k-centroids were picked randomly from the data points.**
>
> **This randomization of picking k-centroids points results in the problem of initialization sensitivity. This problem tends to affect the final formed clusters. The final formed clusters depend on how initial centroids were picked.**
>
> **K-Means++ solves the above problem.**

E. Bisecting K-means

i. Iris plants dataset



ii. Wine dataset



> **Bisecting K-means clustering technique is a little modification to the regular K-Means algorithm, wherein we can fix the procedure of dividing the data into clusters.**
>
> **So, similar to K-means, we first initialize K centroids (You can either do this randomly or can have some prior).**
>
> **After which we apply regular K-means with K=2 (that's why the word bisecting). We keep repeating this bisection step until the desired number of clusters are reached.**

2. Evaluate and compare the performances of the algorithms for each type of clustering, based on the following metrics:
    1. **Silhouette Coefficient**
    2. **Calinski-Harabasz Index**
    3. **Davies-Bouldin Index**
3. Also determine the **Cohesion** and **Separation** performance scores using **Sum of Squared Error (SSE)** and **Sum of Squares Between groups (SSB)**.

    Show the performance comparison for each category of clustering algorithms in a tabular form.

Comparison table for clustering algorithms

Algorithm	Dataset	Silhouette Score	Calinski Harabasz Score	Davies Bouldin Score	SSE	SSB
K-means	IRIS PLANT DATASET	0.5528190124	561.6277566	0.6619715465	78.85144143	24.18035247
	WINE DATASET	0.5711381938	561.8156579	0.5342431775	2370689.687	-2370571.805
K-medoids	IRIS PLANT DATASET	0.5201984013	521.5609065	0.668624441	98.86857318	5.11476015
	WINE DATASET	0.5666480409	539.3792354	0.5292394126	16376.96932	-16243.64278
Dendrogram	IRIS PLANT DATASET	-	-	-	-	-
	WINE DATASET	-	-	-	-	-
AGNES	IRIS PLANT DATASET	0.5543236611	558.0580408	558.0580408	-	-
	WINE DATASET	0.5644796402	552.8517115	0.5357343074	-	-
BIRCH	IRIS PLANT DATASET	0.5019524848	458.4725106	0.6258305924	-	-
	WINE DATASET	0.5644796402	552.8517115	0.5357343074	-	-
DBSCAN	IRIS PLANT DATASET	0.486034197	220.297515	7.222448016	-	-
	WINE DATASET	0.4413295945	208.9449396	7.812129203	-	-
OPTICS	IRIS PLANT DATASET	0.486034197	220.297515	7.222448016	-	-
	WINE DATASET	0.4413295945	208.9449396	7.812129203	-	-
K-means++	IRIS PLANT DATASET	0.5528190124	561.6277566	0.6619715465	78.85144143	24.18035247
	WINE DATASET	0.5711381938	561.8156579	0.5342431775	2370689.687	-2370571.805
Bisecting K-means	IRIS PLANT DATASET	0.3093066205	61.17725176	1.099971025	152.3479518	-44.7653206
	WINE DATASET	0.00384025695	1.06619667	9.045634695	4543749.615	-4543626.929