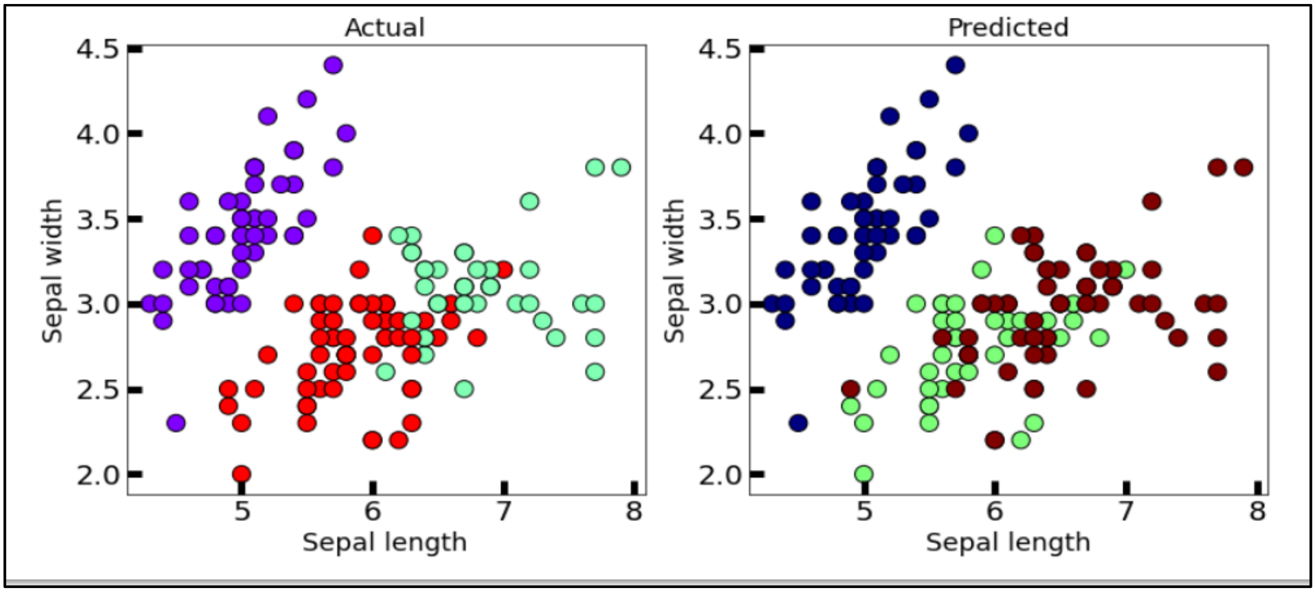
Machine Learning Assignment - 4

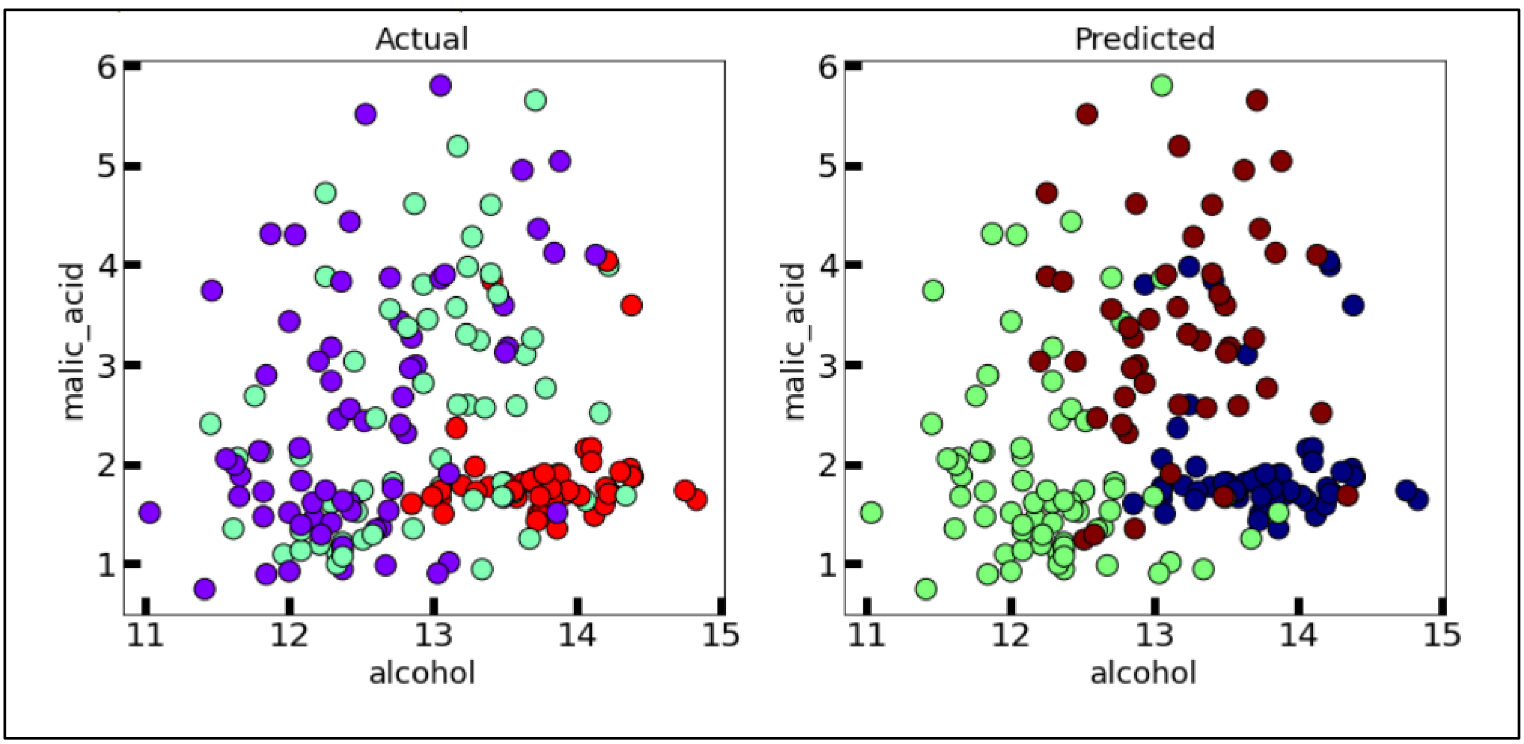
Name - Subhojit Ghoshal
Roll - 001811001048
Semester - 7
Year - 4
Department - Information Technology

1) *Partition based: K-means*

1.1) IRIS PLANT DATASET



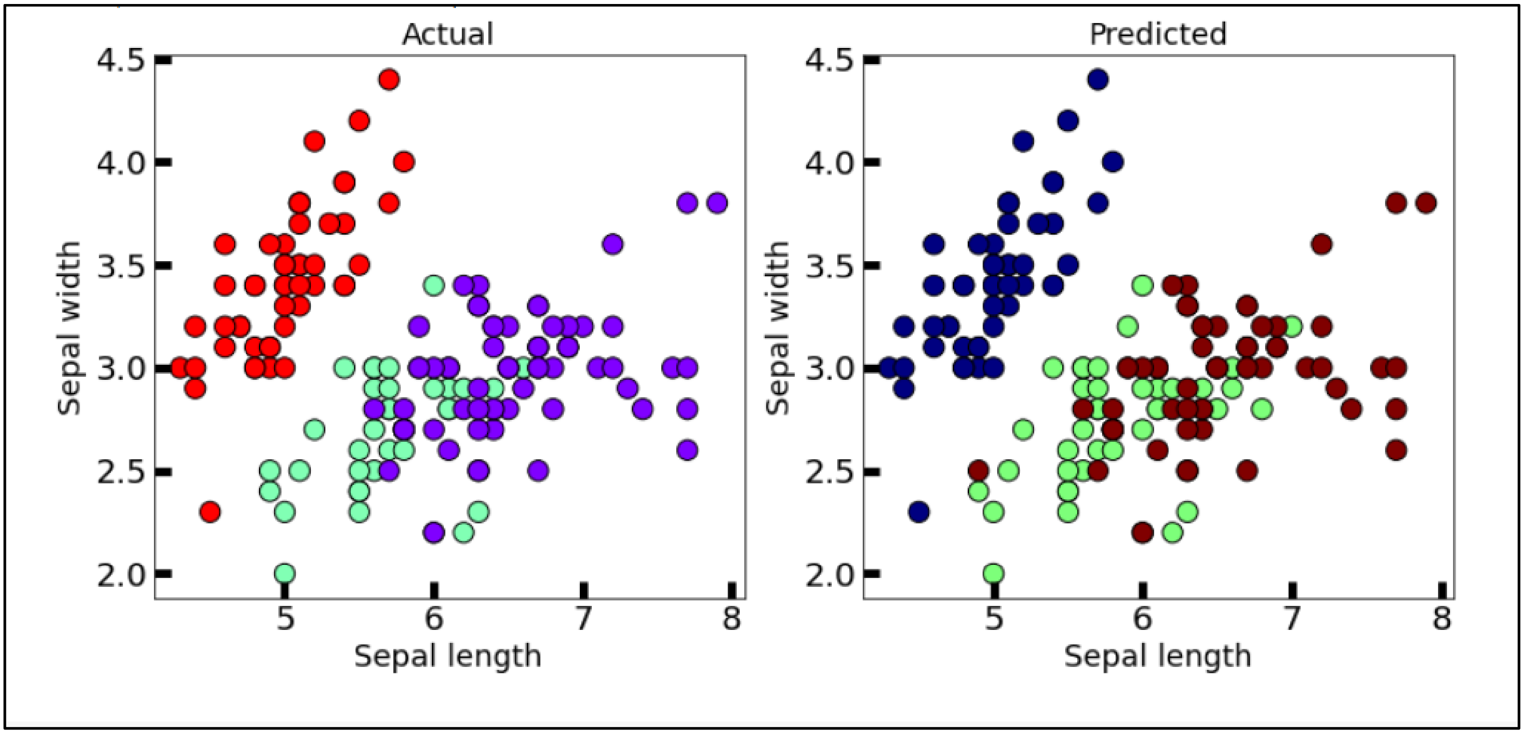
1.2) WINE DATASET



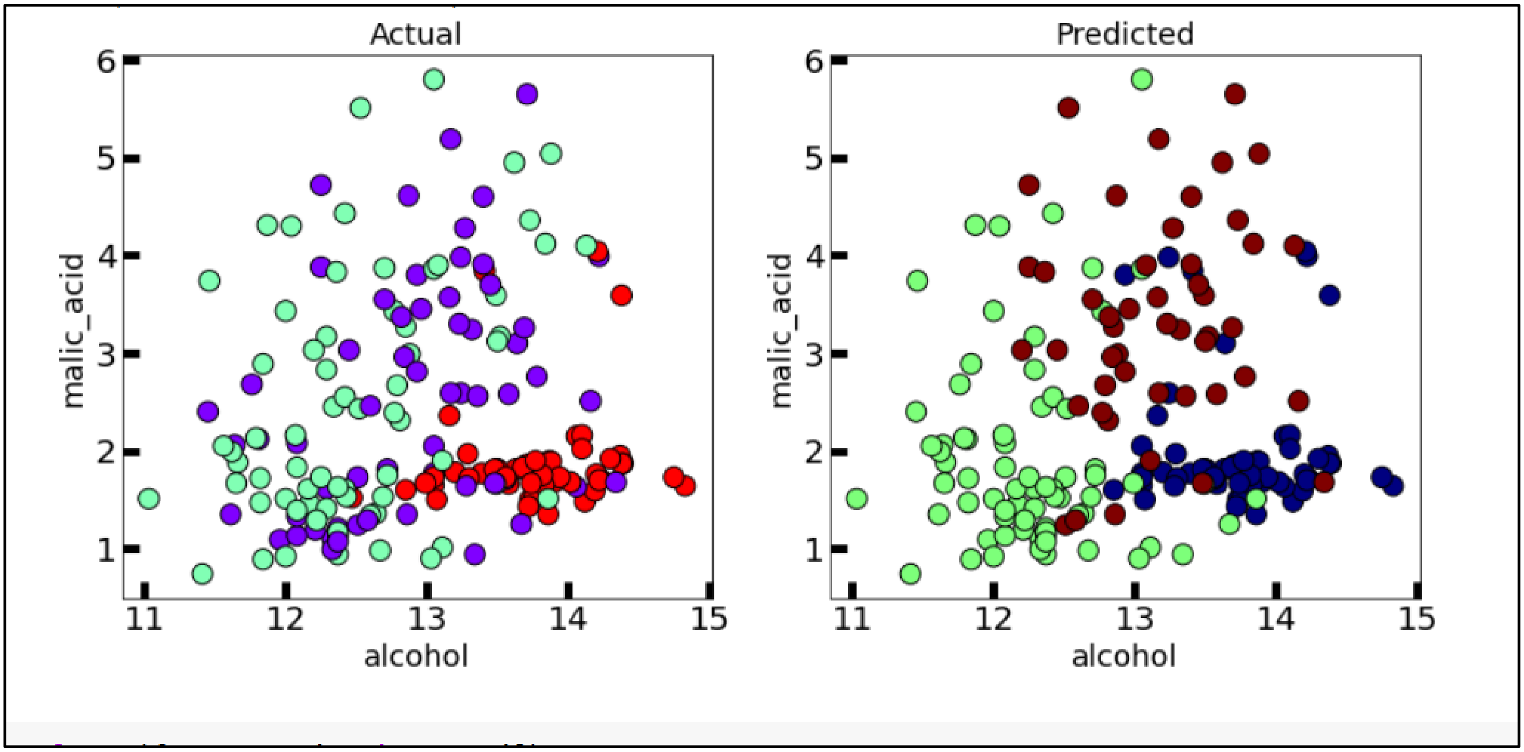
This algorithm generalizes to clusters of different shapes and sizes, such as elliptical clusters. The problem with it is that we need to manually choose the value of "k".

2) *Partition based: K-medoids*

2.1) IRIS PLANT DATASET



2.2) WINE DATASET

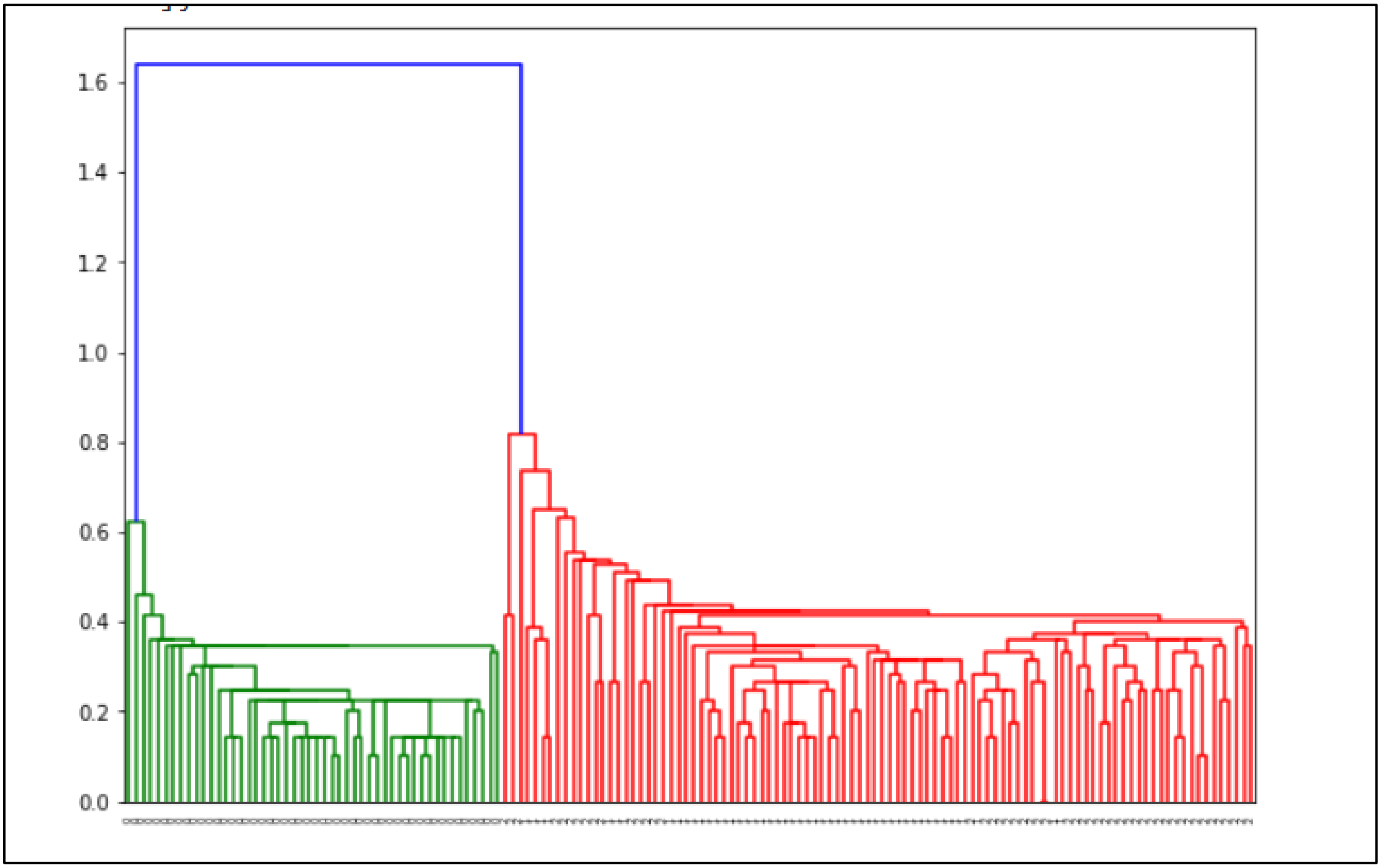


This algorithm solves the problem with the K-means algorithm.

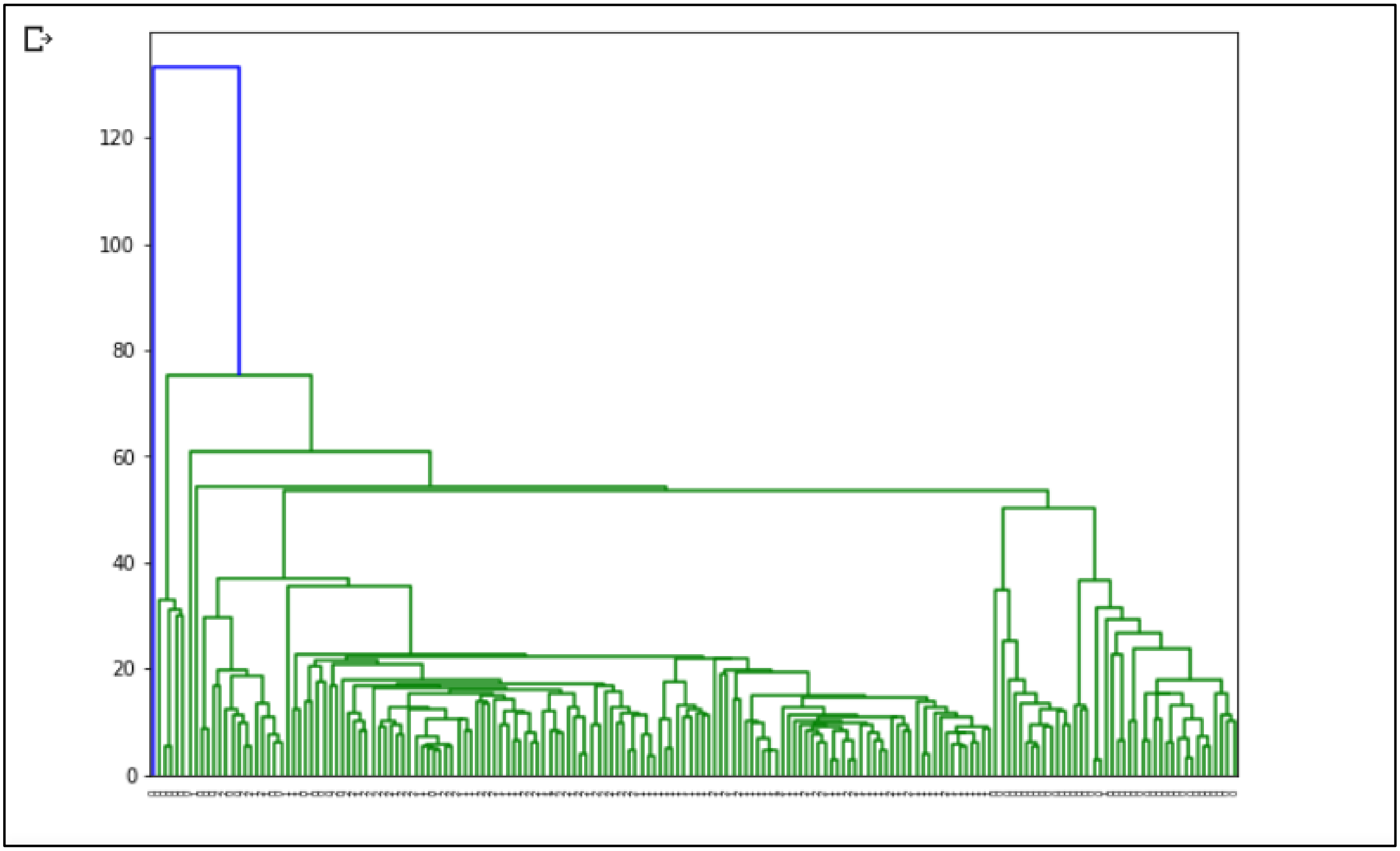
K-means attempts to minimize the total squared error, while k-medoids minimizes the sum of dissimilarities between points labeled to be in a cluster and a point designated as the center of that cluster..

3) *Hierarchical: Dendrogram*

3.1) IRIS PLANT DATASET



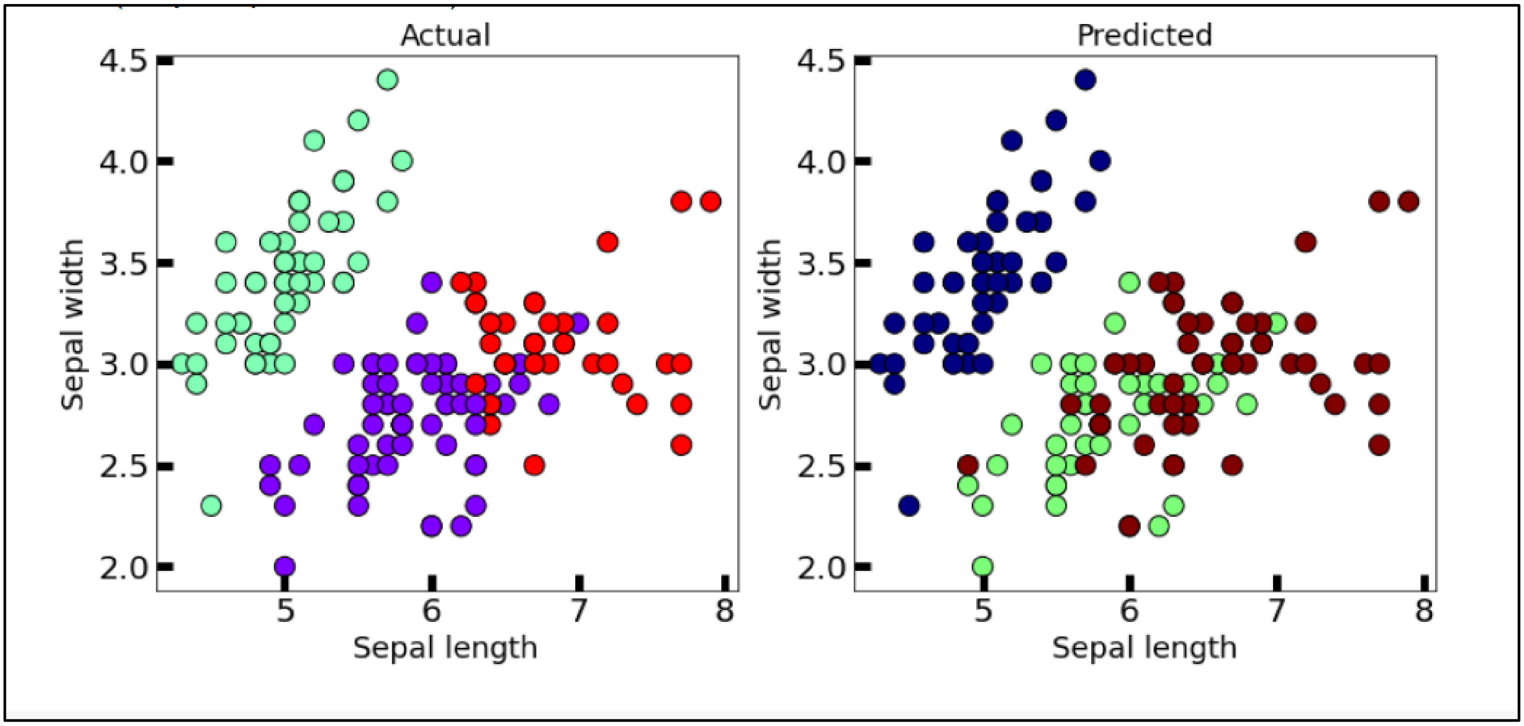
3.2) WINE DATASET



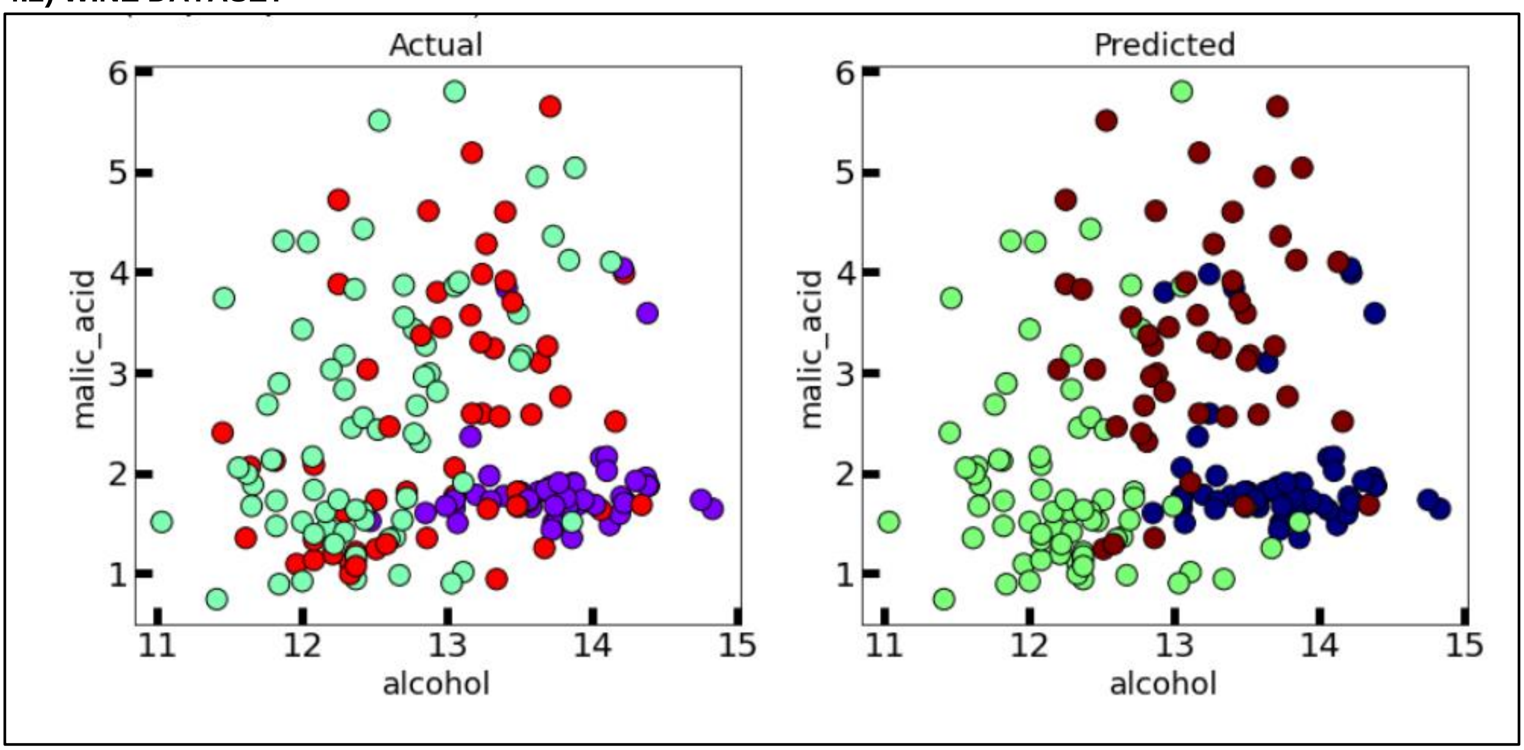
A dendrogram is a diagram that shows the hierarchical relationship between objects. It is most commonly created as an output from hierarchical clustering. The main use of a dendrogram is to work out the best way to allocate objects to clusters.

4) *Hierarchical: AGNES*

4.1) IRIS PLANT DATASET



4.2) WINE DATASET



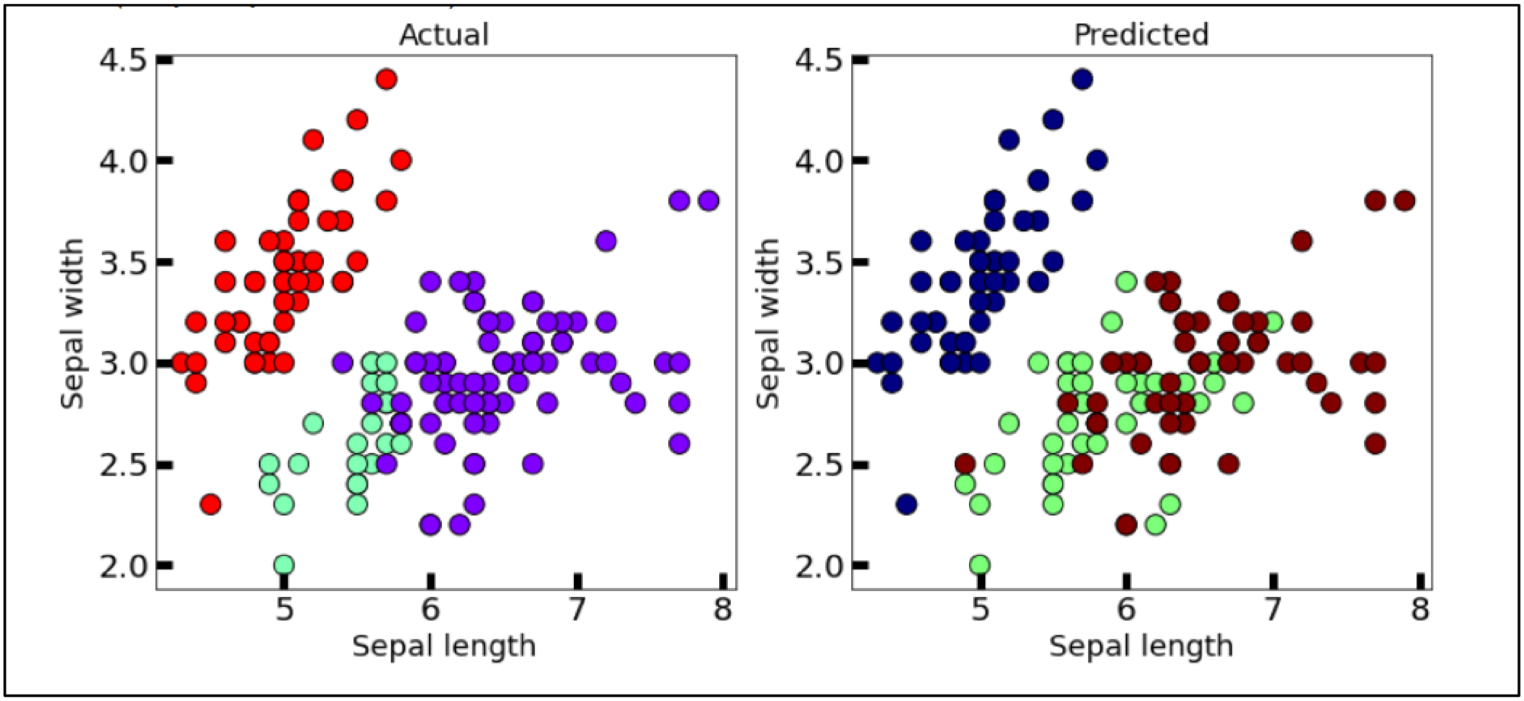
The agglomerative clustering is the most common type of hierarchical clustering used to group objects in clusters based on their similarity. It’s also known as AGNES (Agglomerative Nesting).

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.

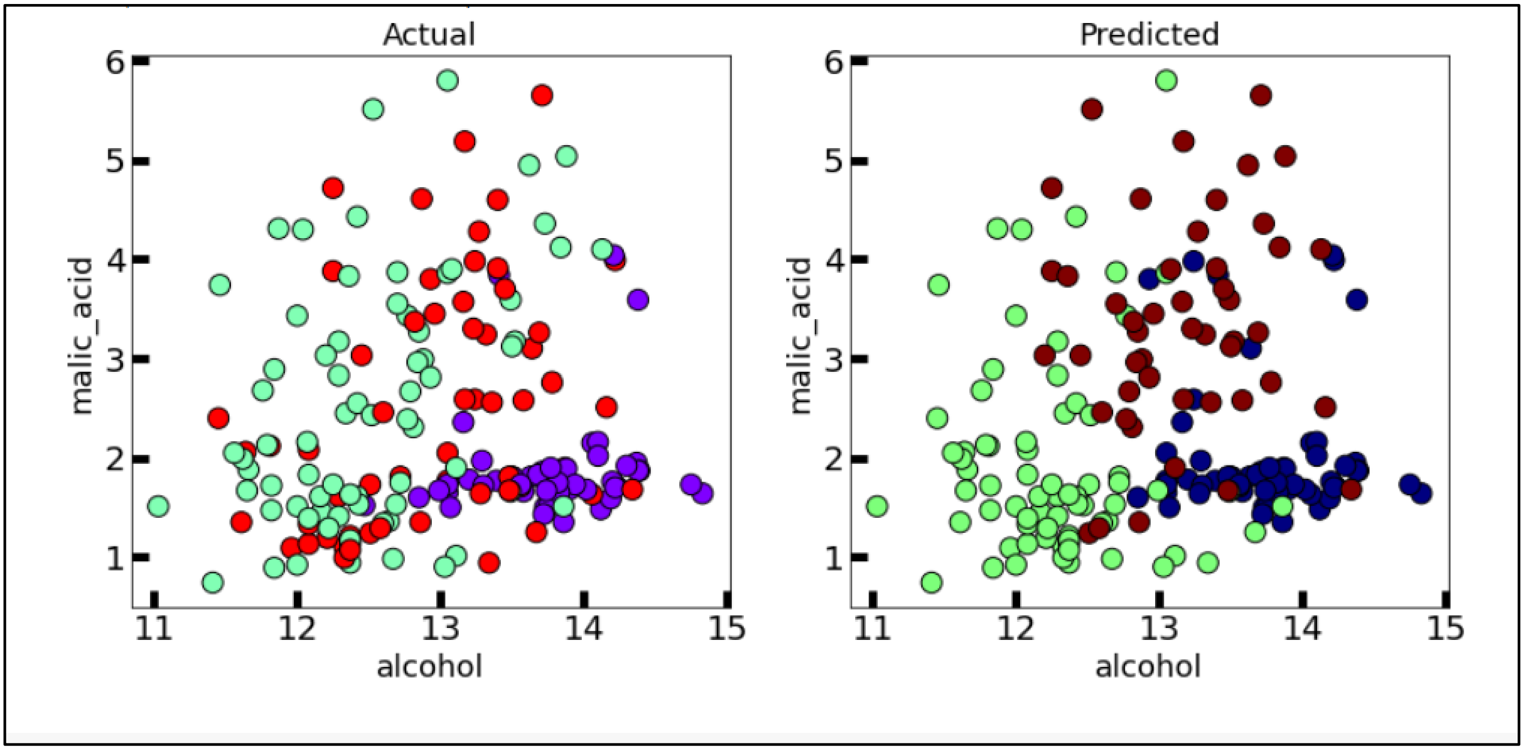
The result is a tree-based representation of the objects, named dendrogram.

5) *Hierarchical: BIRCH*

5.1) IRIS PLANT DATASET



5.2) WINE DATASET

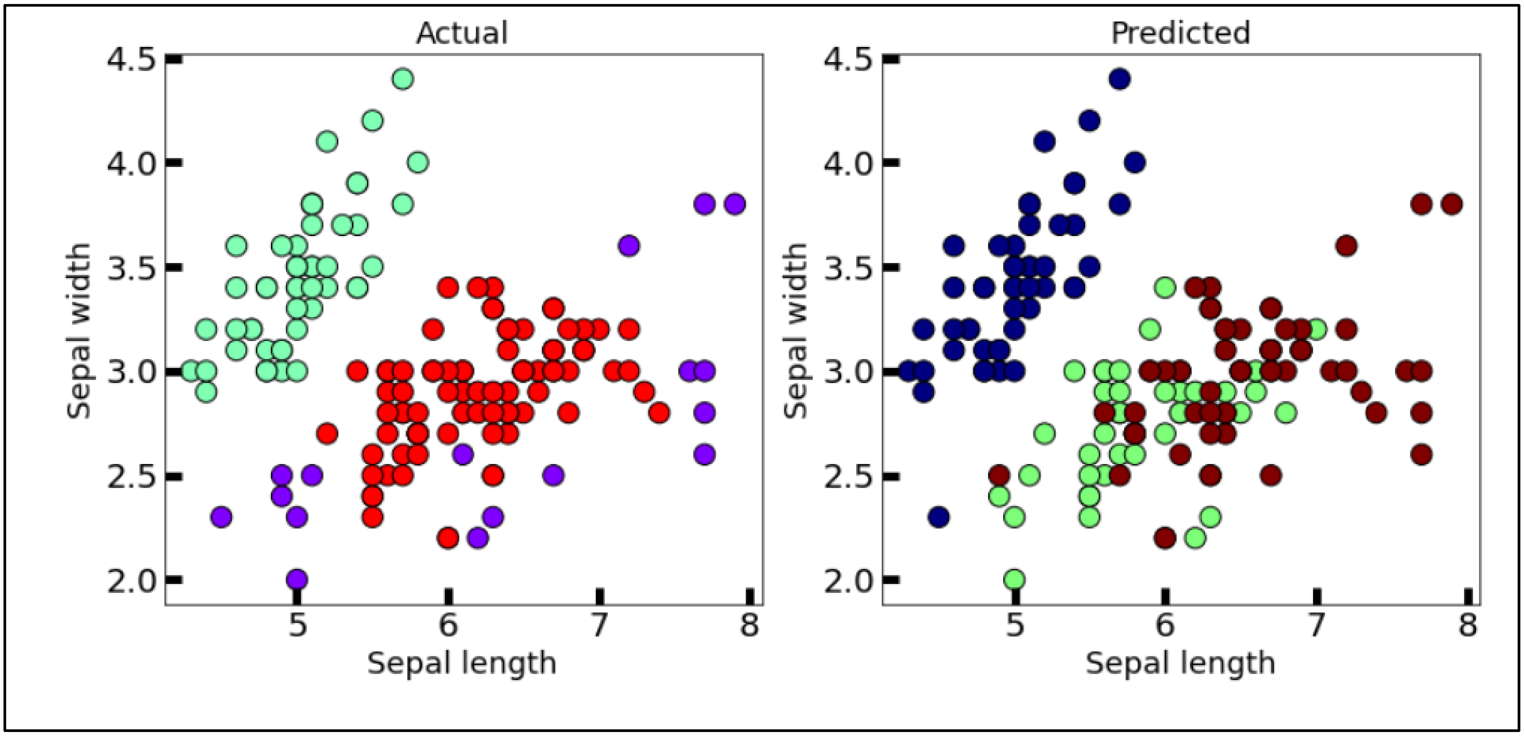


Balanced Iterative Reducing and Clustering using Hierarchies (BIRCH) is a clustering algorithm that can cluster large datasets by first generating a small and compact summary of the large dataset that retains as much information as possible.

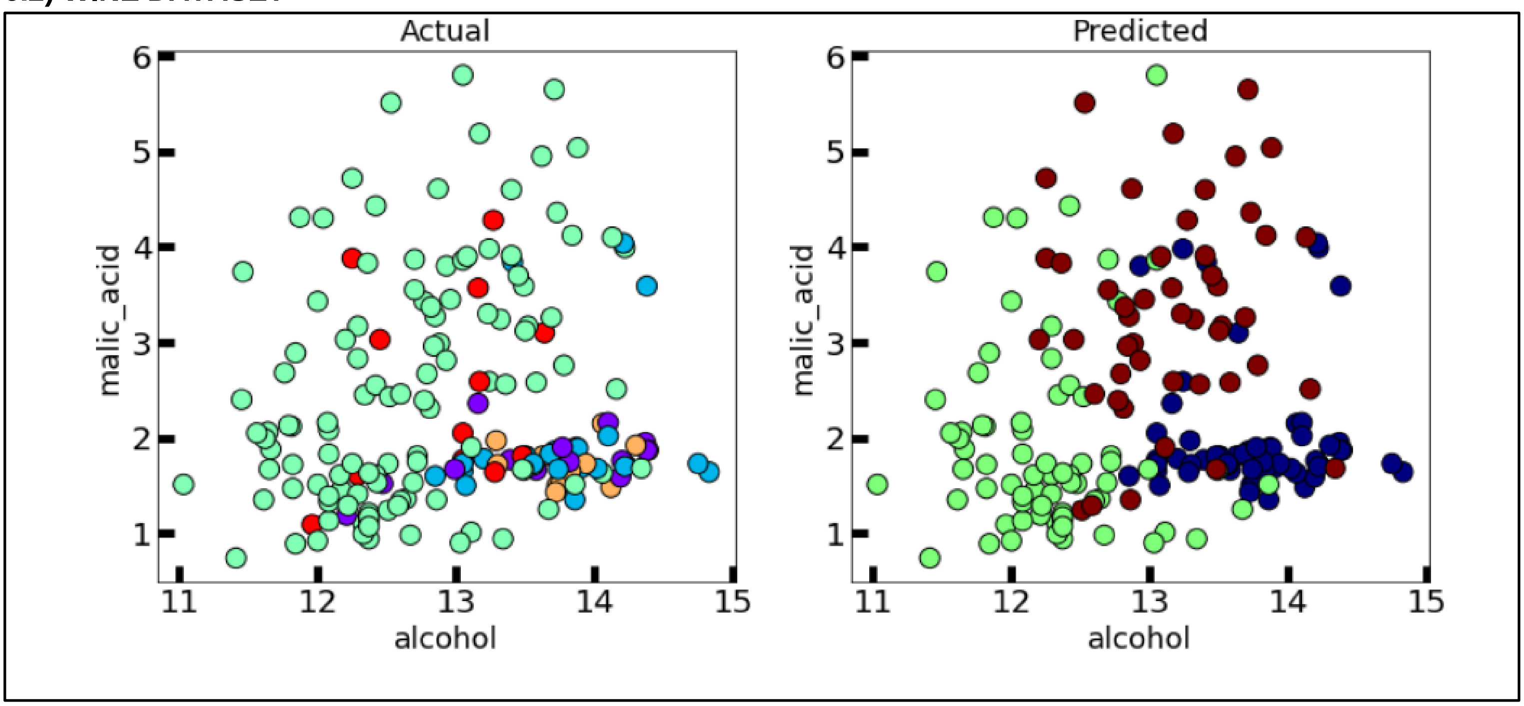
This smaller summary is then clustered instead of clustering the larger dataset

6) ***Density based: DBSCAN***

6.1) IRIS PLANT DATASET



6.2) WINE DATASET

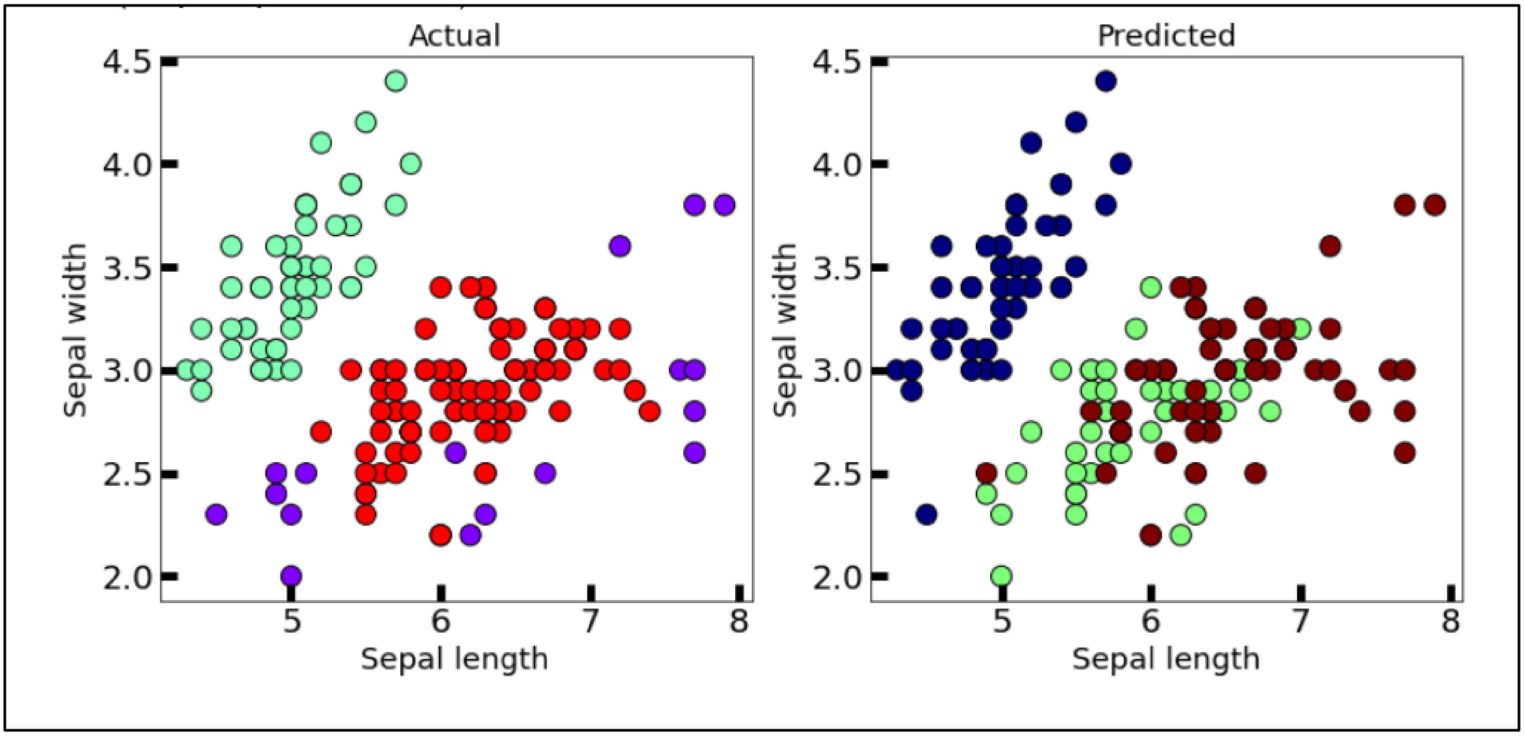


Clusters are dense regions in the data space, separated by regions of the lower density of points. The DBSCAN algorithm is based on this intuitive notion of "clusters" and "noise".

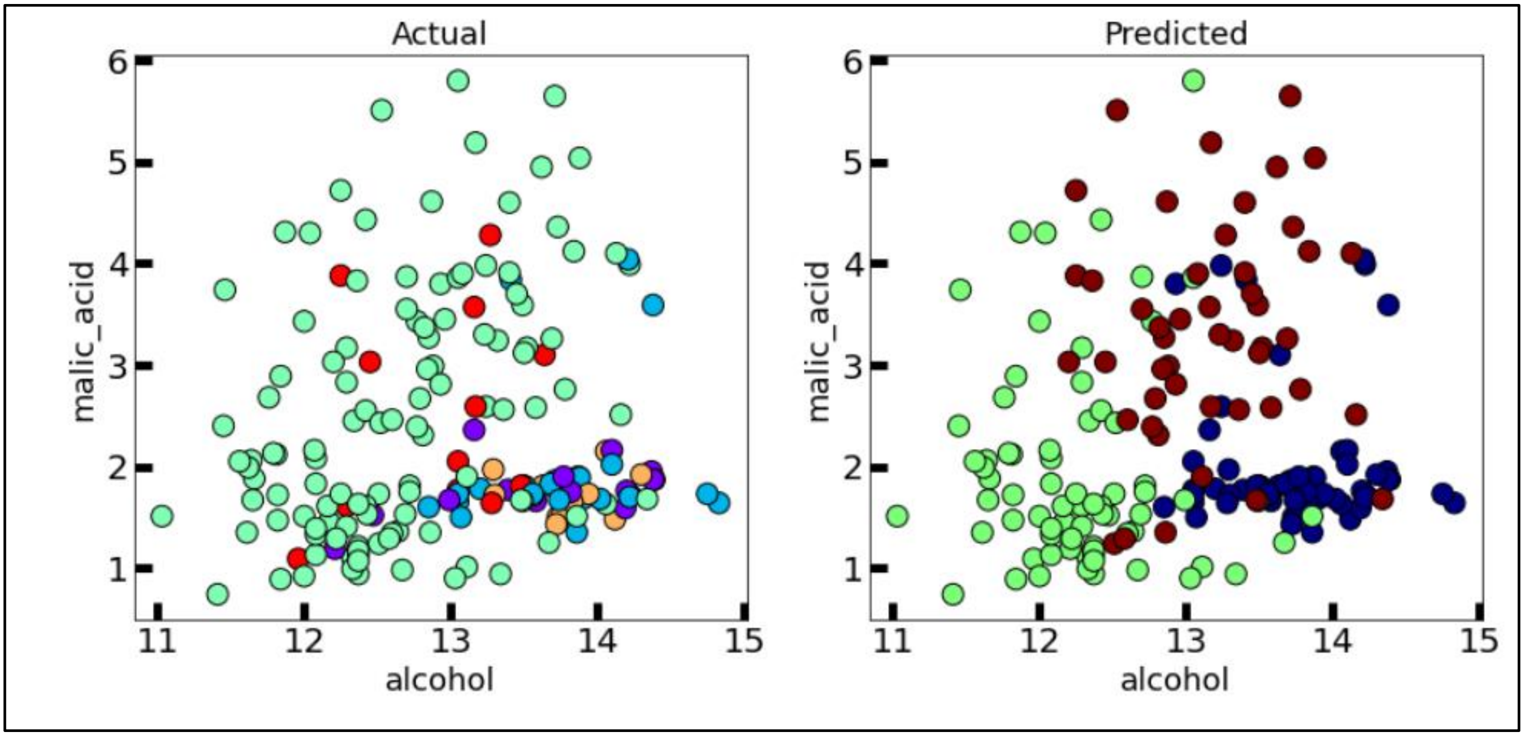
The key idea is that for each point of a cluster, the neighborhood of a given radius has to contain at least a minimum number of points.

7) *Density based: OPTICS*

7.1) IRIS PLANT DATASET



7.2) WINE DATASET

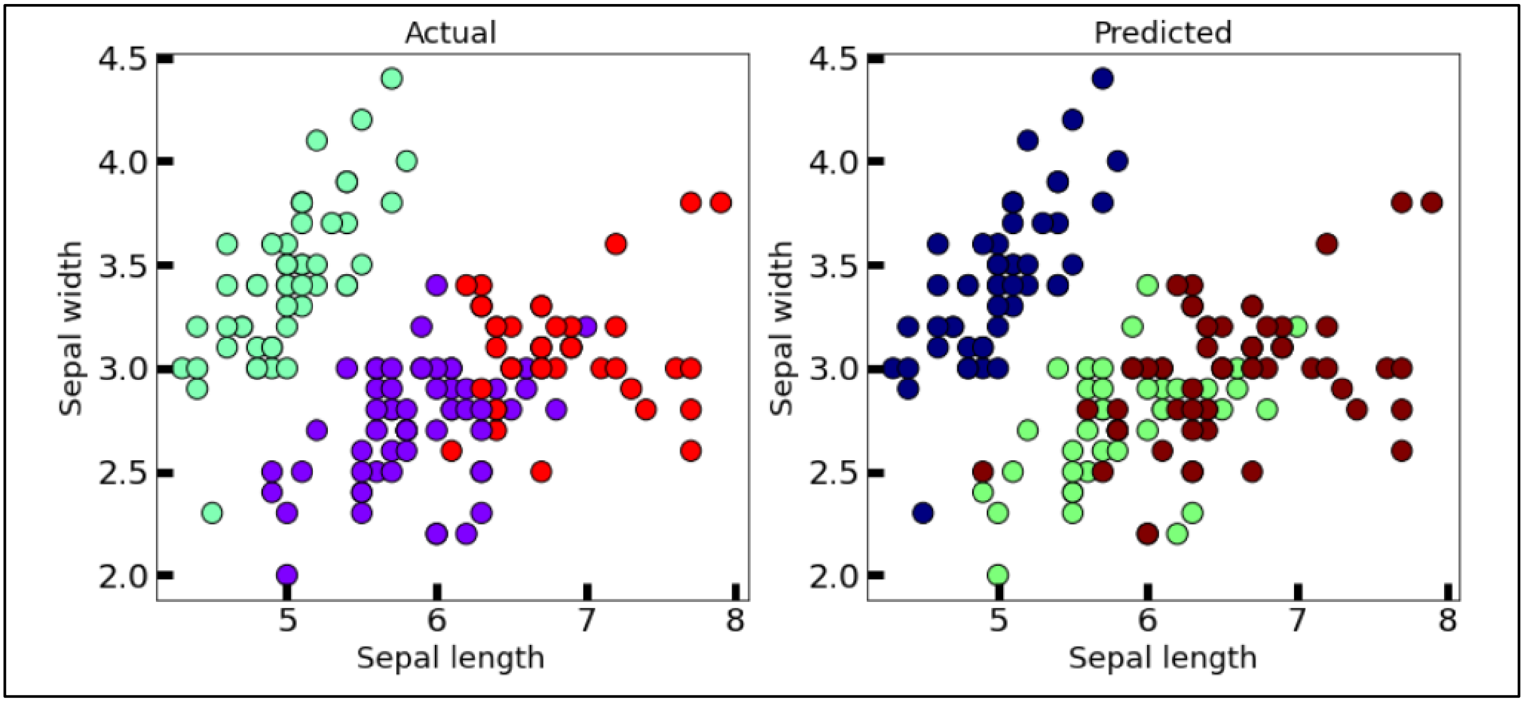


This clustering technique is different from other clustering techniques in the sense that this technique does not explicitly segment the data into clusters.

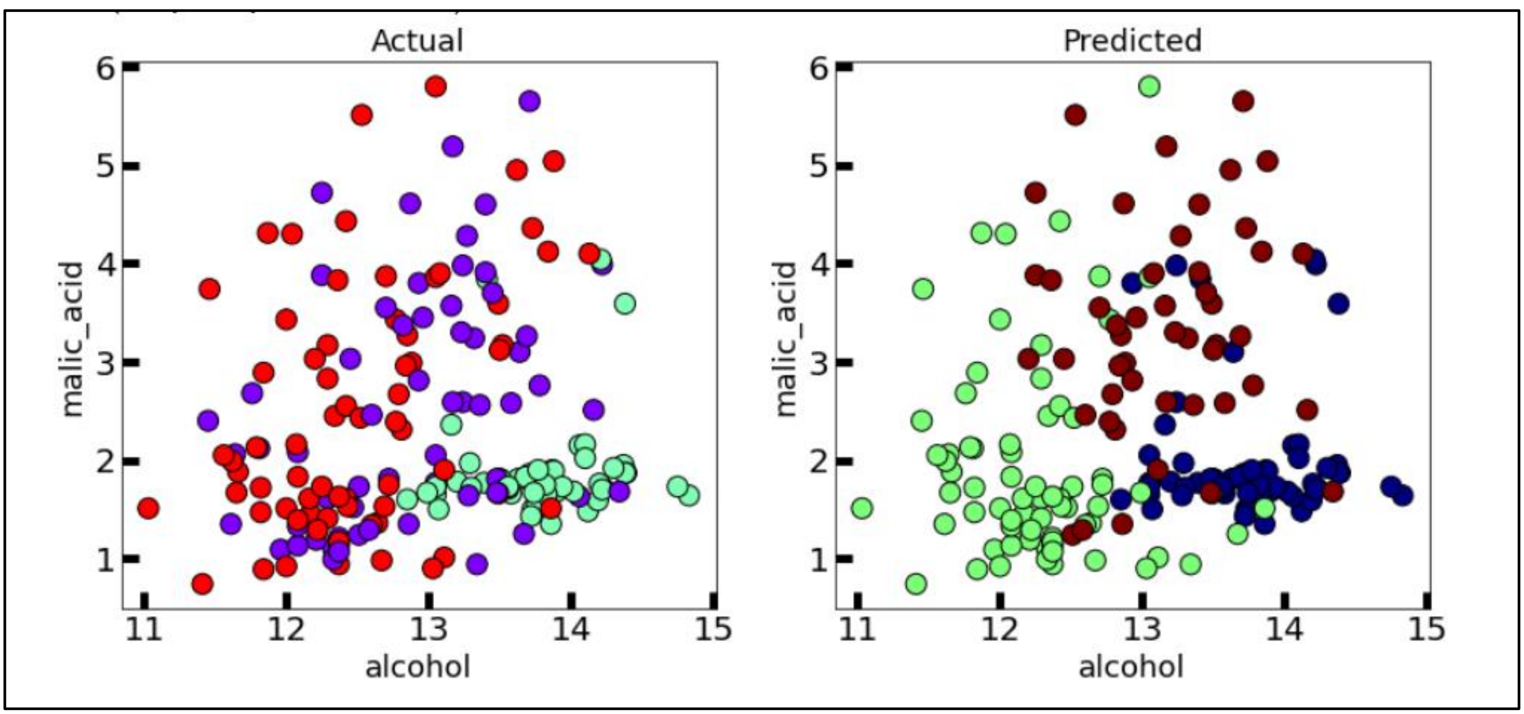
Instead, it produces a visualization of Reachability distances and uses this visualization to cluster the data.

8) *K-means++*

8.1) IRIS PLANT DATASET



8.2) WINE DATASET



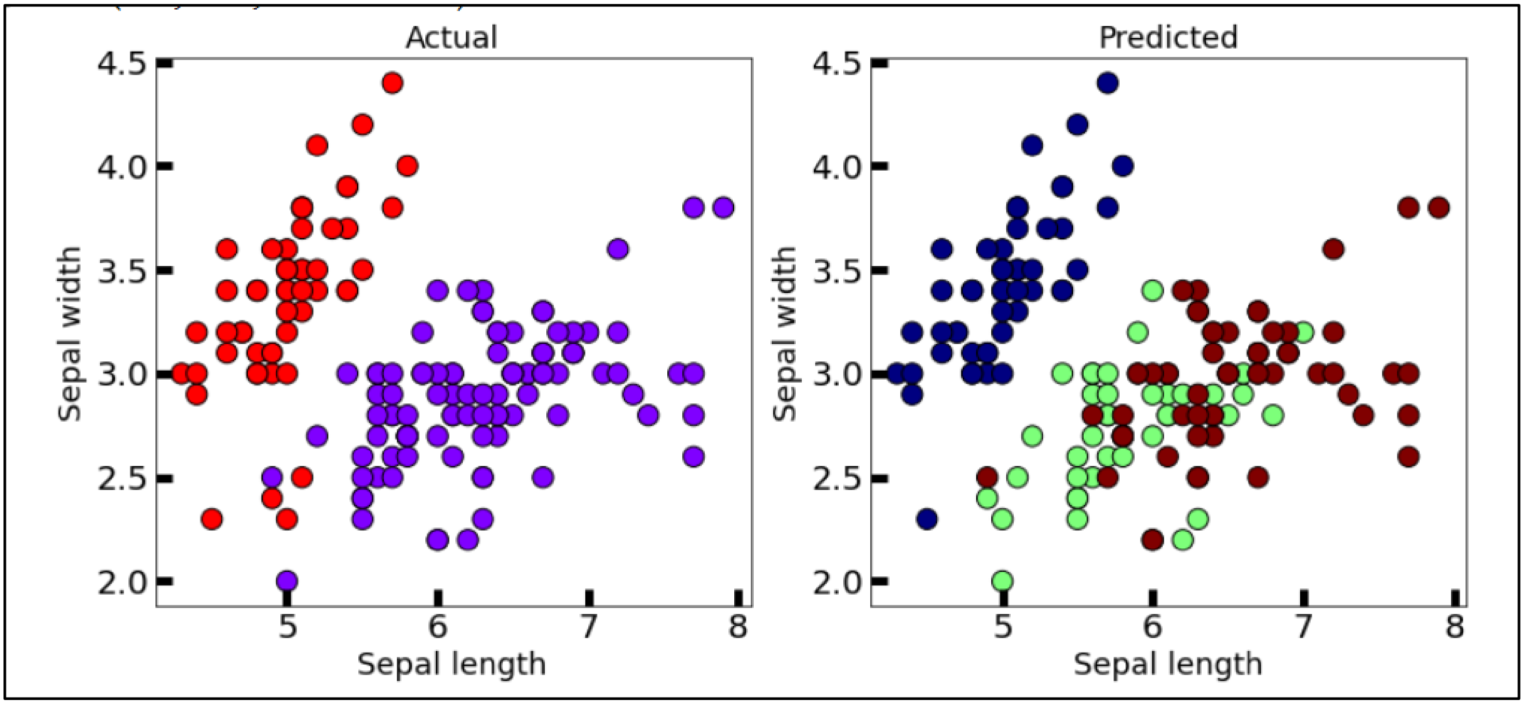
In the case of K-Means clustering, we were using randomization. The initial k-centroids were picked randomly from the data points.

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.

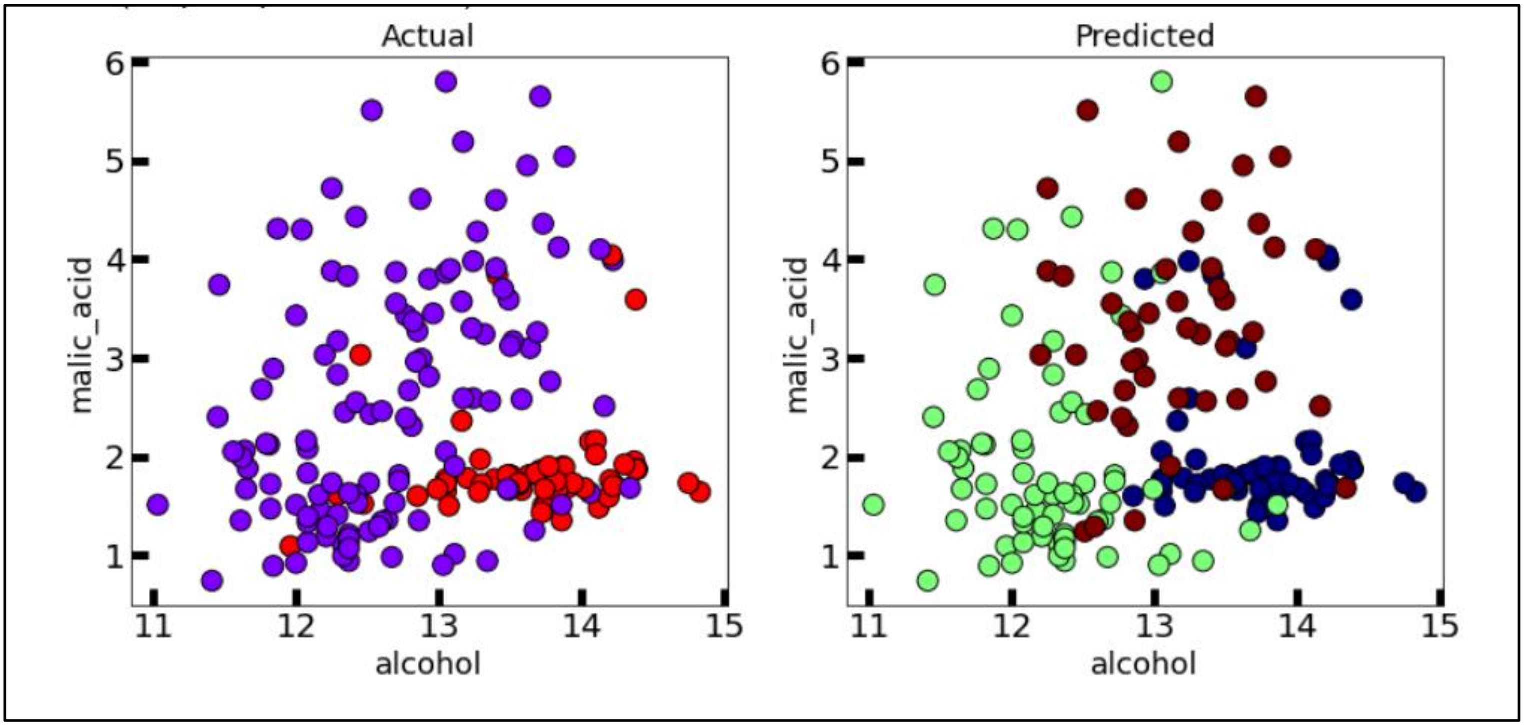
K-Means++ solves the above problem.

9) *Bisecting K-means*

9.1) IRIS PLANT DATASET



9.2) 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.