Assignment-4
Clustering Techniques

Name: Harika Nalam
SJSU ID:015939963

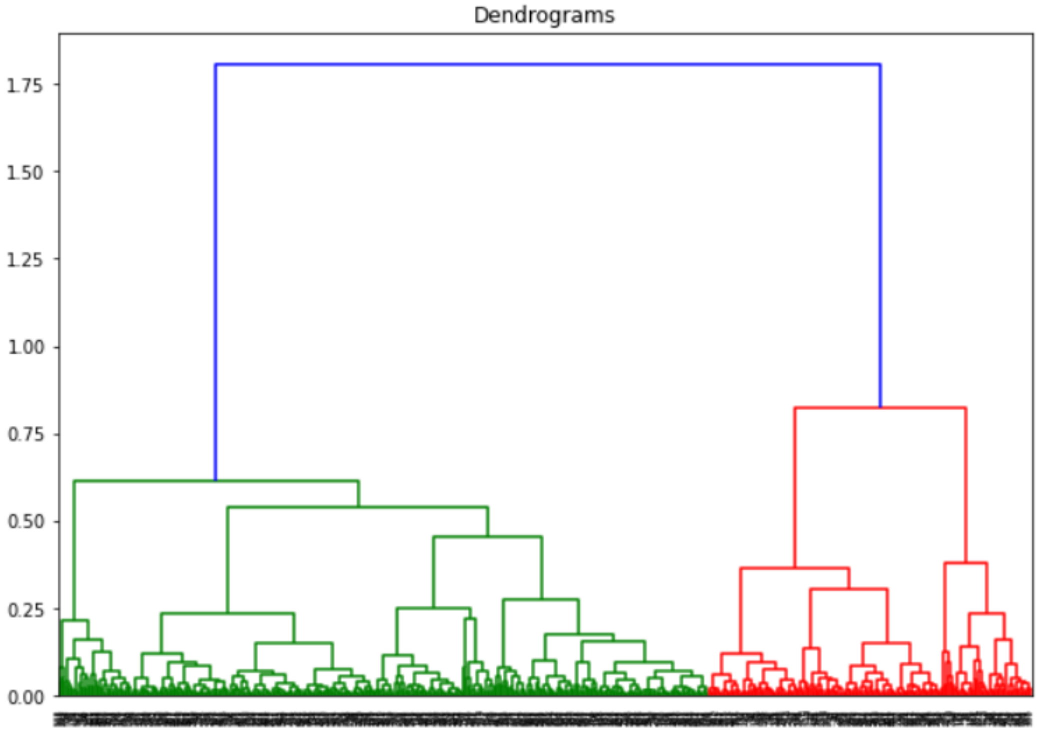
There are many types of clustering algorithms

1. Connectivity based clustering (Hierarchical clustering)
2. Centroid based clustering (KMeans)
3. Density-based clustering (DBSCAN)
4. Distributed based clustering (GMM)

Connectivity based clustering- Hierarchical clustering

Hierarchical Clustering is a method of unsupervised machine learning clustering where it begins with a pre-defined top to the bottom hierarchy of clusters. It then proceeds to perform a decomposition of the data objects based on this hierarchy.

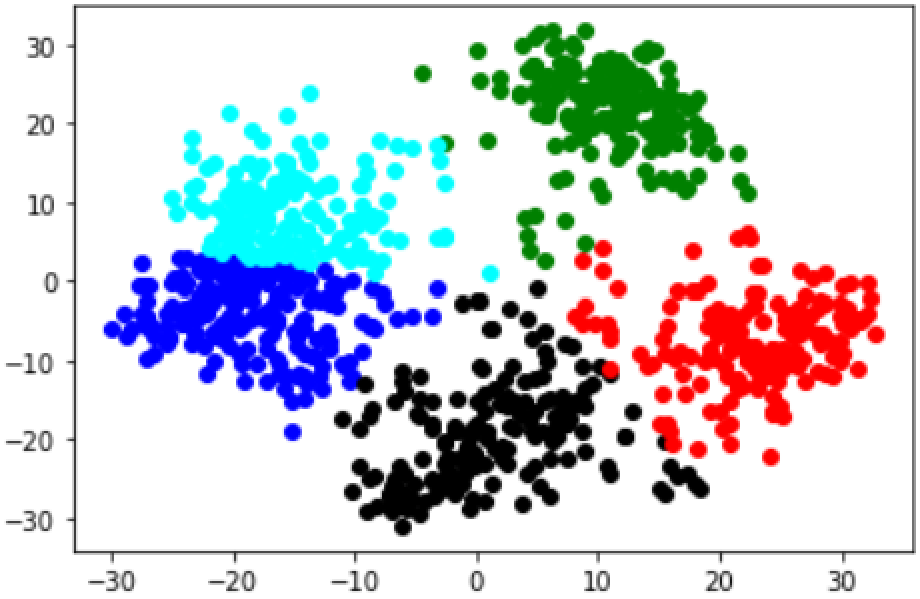
The breast cancer dataset is used to perform Hierarchical clustering. The dataset is classified into two classes.



Centroid based clustering - KMeans Clustering from scratch
Steps followed:

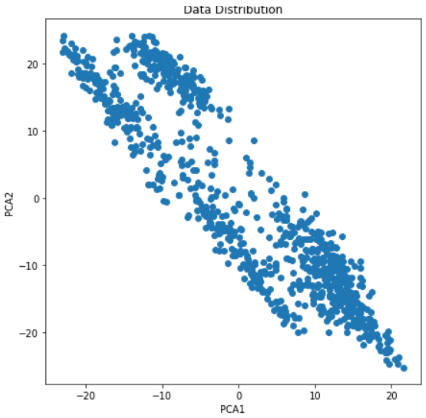
1. From the input data points select centroid randomly
2. Find the distance between the data points and randomly selected clusters.
3. Choose the nearest points to the centroids and cluster those points and find the new centroid from all the nearest points
4. These steps are repeated till the clusters formed in the previous iterations is the same as the clusters formed in this step.

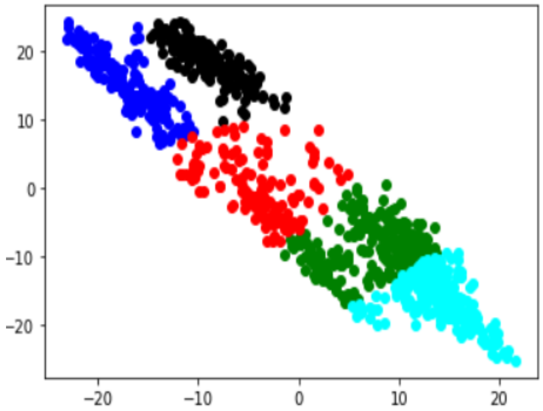
The dataset used is digits data. I have used 5 different class data (0to4)



But Kmeans clustering is not recommended for the clusters which are in non-circular shape. This technique forms clusters in circular boundaries.

This can be seen when data is stretched randomly





Density-based clustering - DBSCAN

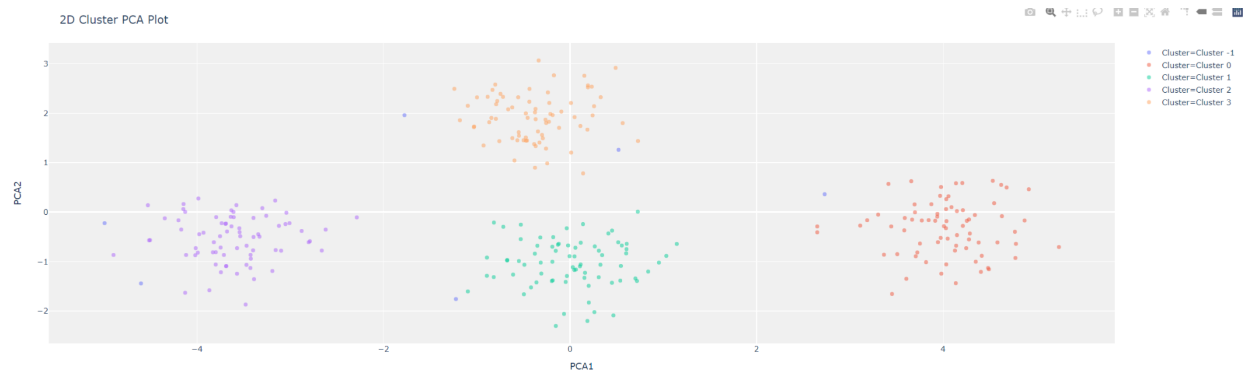
Density-based clustering methods take density into consideration instead of distances. Clusters are considered the densest region in a data space.

DBSCAN can get clusters with

1. Arbitrary shape
2. Without limitation in cluster size

DBSCAN is implemented using Pycaret.

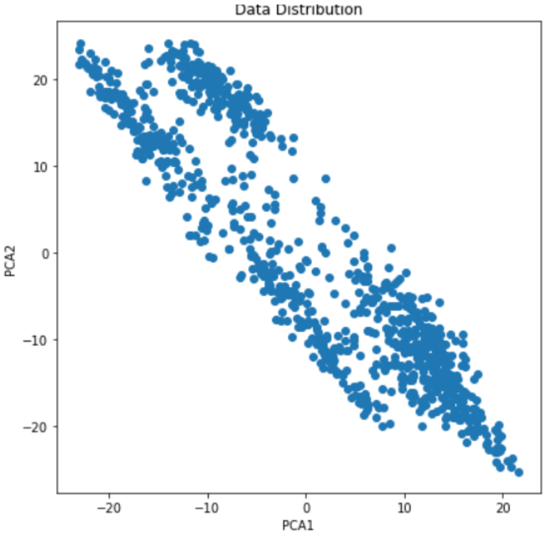
Used the “make_blobs” dataset.

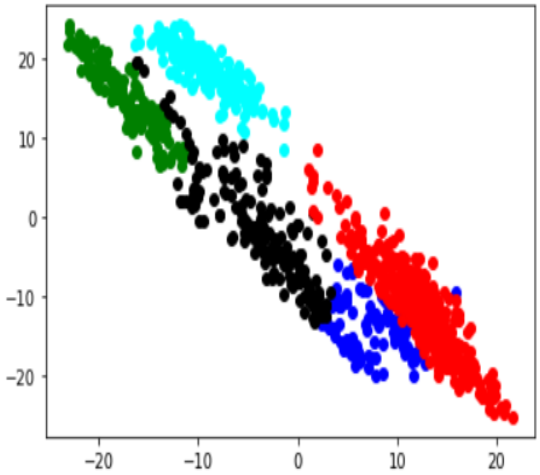


Distributed based clustering (GMM)

Distribution-based clustering creates and groups data points based on the same probability distribution (Gaussian, Binomial etc.) in the data.

Used digits dataset. I have applied random stretch on dataset to see if the GMM flows elliptical clusters.


Data Distribution
PCA2
PCA1


20
10
0
−10
−20
−20
−10
0
10
20