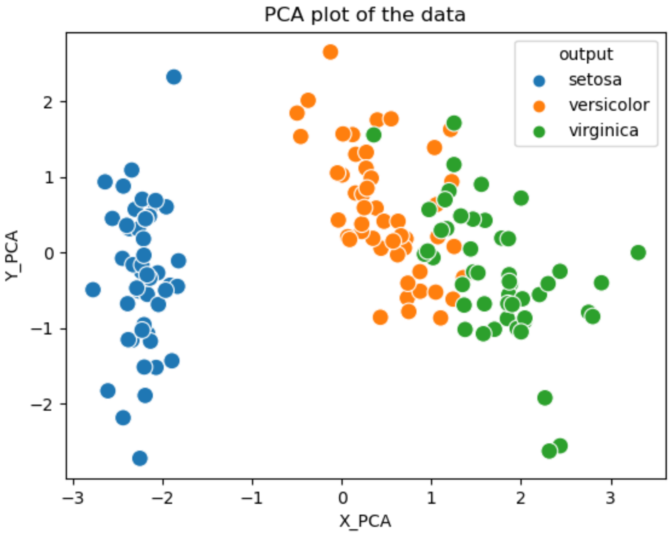
K-means and GMM

Saurabh Burewar (B18CSE050)

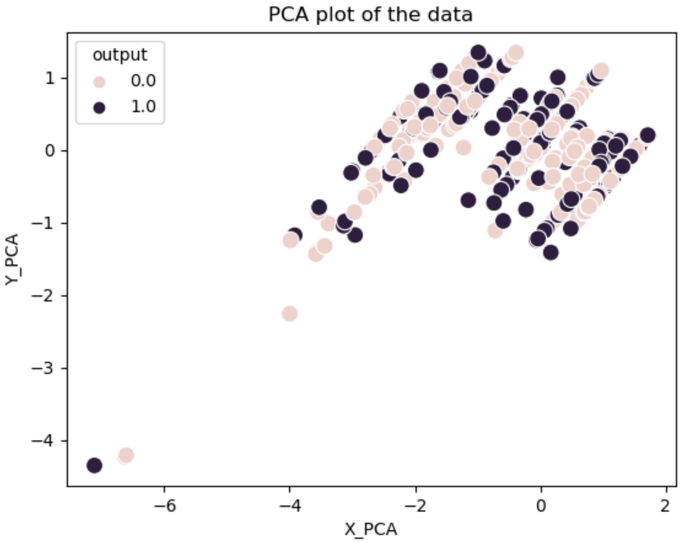
Dataset

I have used 3 datasets to get a good performance grasp of the models being tested.

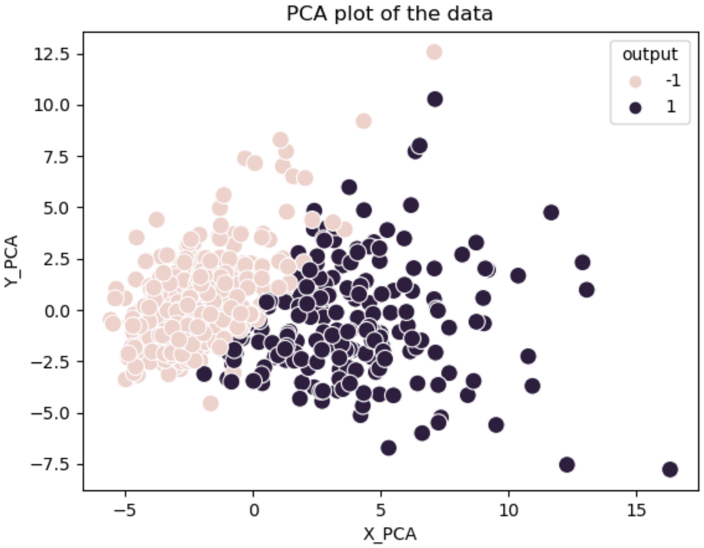
- The first dataset is the Iris dataset with 3 classes and 4 features. This dataset has an interesting distribution where 1 cluster is far apart from the other two but the other two are mixed together which makes it harder to classify the mixed clusters. Plotting these using PCA we get the following plot -



- The second dataset is the Titanic dataset with 2 classes and 5 features (I have only taken 5 features here). This dataset is a very mixed dataset originally and is not a good one for predicting through clustering algorithms. I included this just to see how these algorithms perform on such a dataset. Plotting these using PCA, we get -



- The third dataset is the Breast Cancer dataset with 2 classes and 10 features. This is a very clear dataset which makes it a more ideal one for clustering algorithms. Plotting these using PCA, we get -



PCA

Since the data is in higher dimensions, PCA is used to project the data in two dimensions by taking features with high variances.

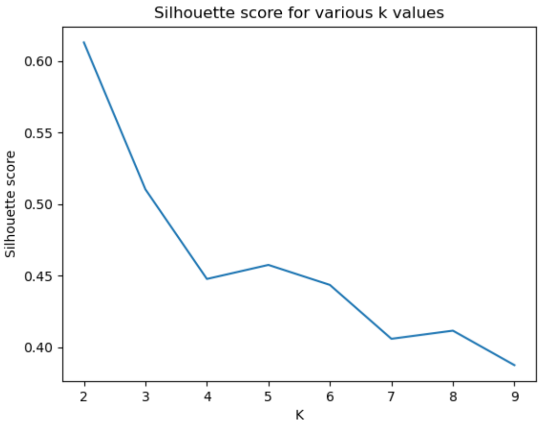
K-means

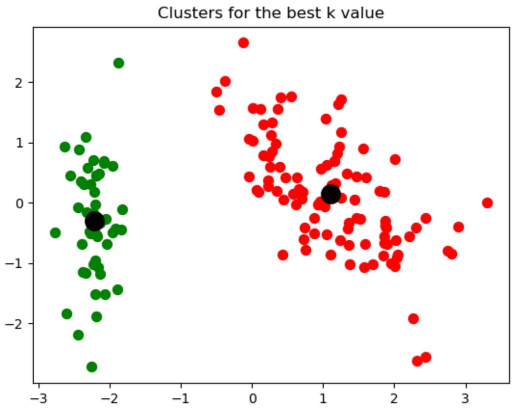
I have implemented the K-means clustering algorithm by taking the first k examples as the starting centroids. All the examples are assigned their clusters and then we find new centroids from the mean of clusters. These two steps are repeating for around 300 iterations for all 3 datasets.

K-means is run for different values of k ranging from 2 to 10 and the Silhouette score for all is calculated. Then, the value of k which gives the best result is chosen which is shown below.

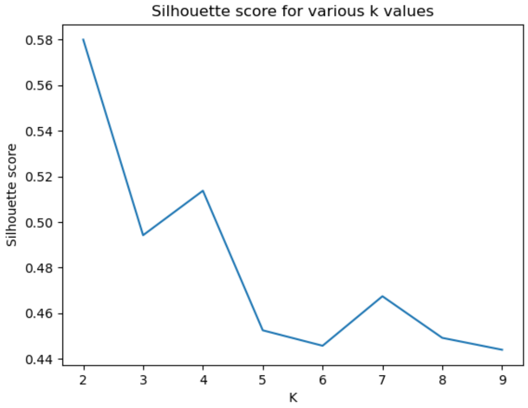
[Note: The black points are the centroids of the clusters]

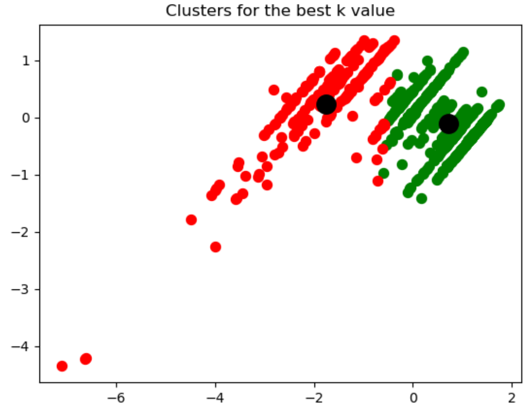
For Iris dataset - Interesting thing to note here is that k-means gives the best results for k = 2 since the mixed clusters are considered as a single cluster.



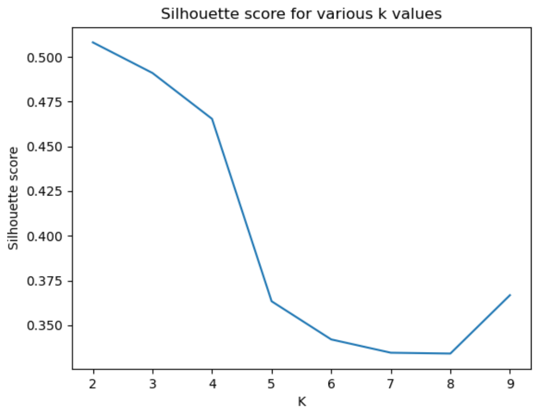


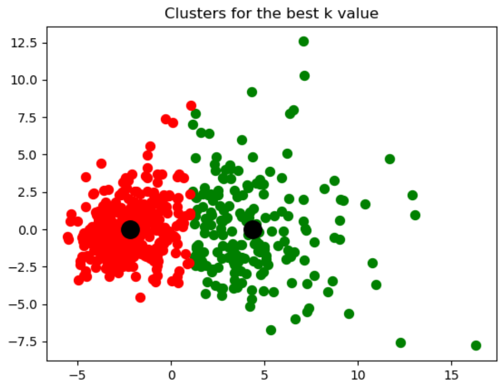
For Titanic dataset - There is a big difference between scores of different k values here and the appropriate number is 2.





For Breast cancer dataset - Here, there isn't a big difference between lower values of k but the best results are for k = 2 and the original classification also has 2 clusters.



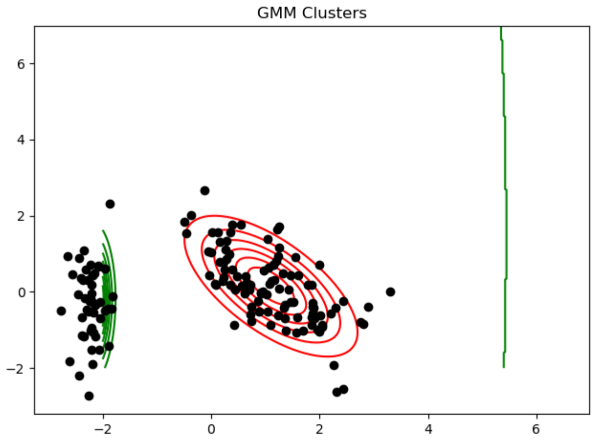


GMM using K-means result as starting centroids

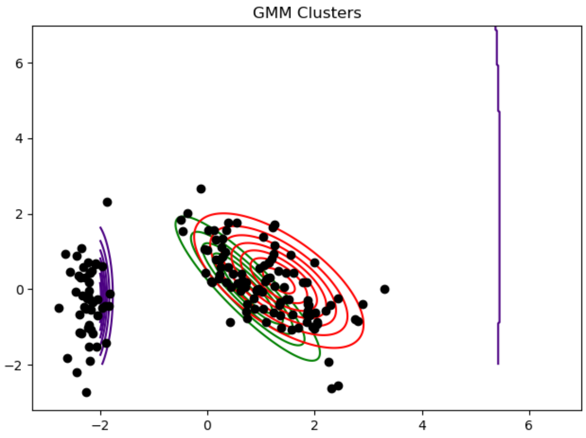
GMM is implemented with an extra argument in the fit function that allows custom starting centroids. With this, GMM can be started using the K-means result as the starting centroids. If this argument is empty, random centroids will be chosen. The GMM gives the following clusters for the 3 datasets -

For Iris dataset - Since the Iris dataset originally has 3 classes, I have run GMM with number of clusters as 2 and 3. There is not much optimization to be done to the K-means results which makes the result for k = 2 similar to k-means. The k = 3 case shows improvement in the clustering and the clusters are now very similar to the original clusters.

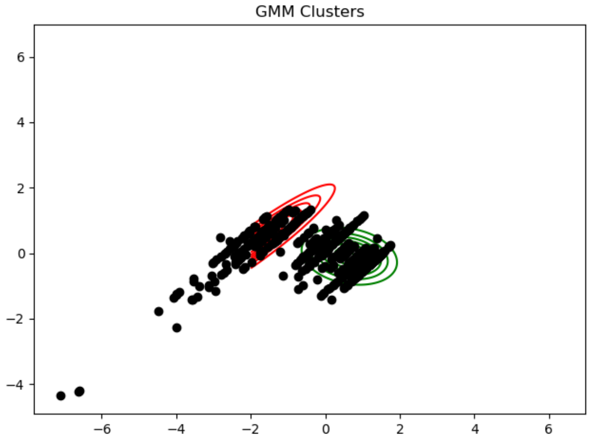
k = 2



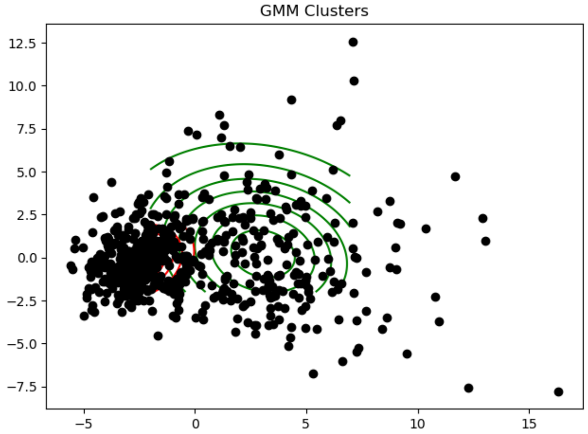
k = 3



For Titanic dataset - Even though not the same as the original clusters, the clusters made here are distinct.



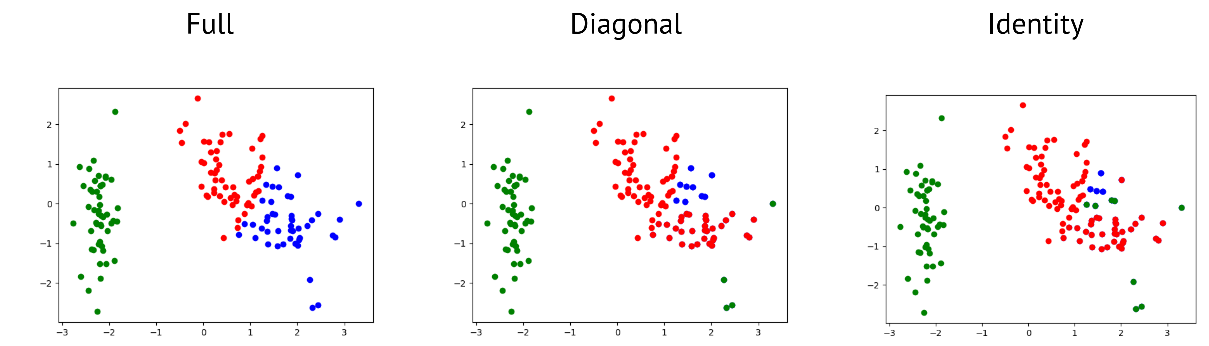
For Breast cancer dataset - The clustering here is similar to the result given by k-means which is again very similar to the original clusters.



GMM with different covariance types

Here, GMM is run with different types of covariance matrices. The following results are seen for the 3 datasets -

For Iris dataset - The number of clusters is kept as 3 here. There is a big difference in the clusters formed by different covariance types. The ‘Full’ covariance gives a good result with distinct clusters similar to original clusters because they can independently adopt any shape. The ‘Diagonal’ covariance doesn't give as good results because the contours are oriented along the axes. The ‘Identity’ gives a very different result because of spherical contours.

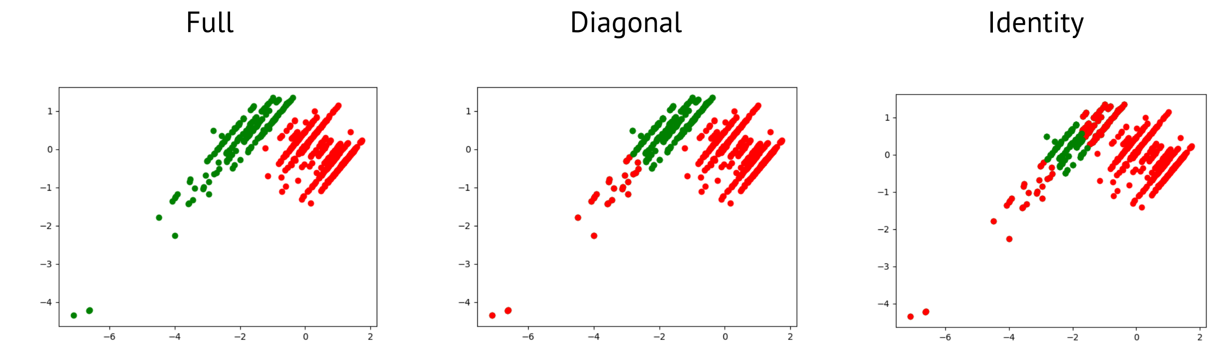


Full

Diagonal

Identity

For Titanic dataset - Full covariance gives a clear clustering result but since diagonal is oriented along axes, the cluster points below y = -1 are not taken in the green cluster. Similar case is seen in Identity covariance due to spherical contours.

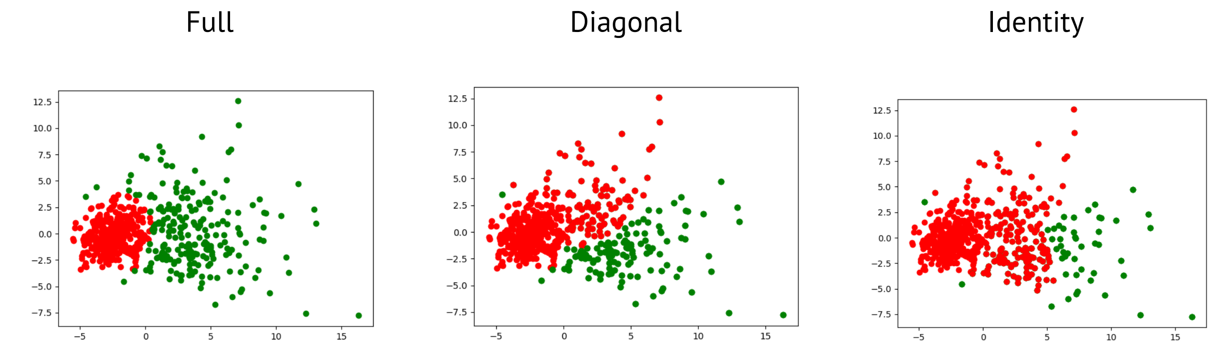


Full

Diagonal

Identity

For Breast cancer dataset - Again, full covariance gives good results, but here diagonal also gives good results because even though contours are oriented along axes, the data points can still be classified similarly. Identity due to spherical contours makes overlapping contours which gives completely different clusters.



Full

Diagonal

Identity