K- means, GMM and PCA

In this Assignment we have to perform K-means and Gaussian mixture models, also we have to visualize the clusters by dimensionality reduction using PCA.

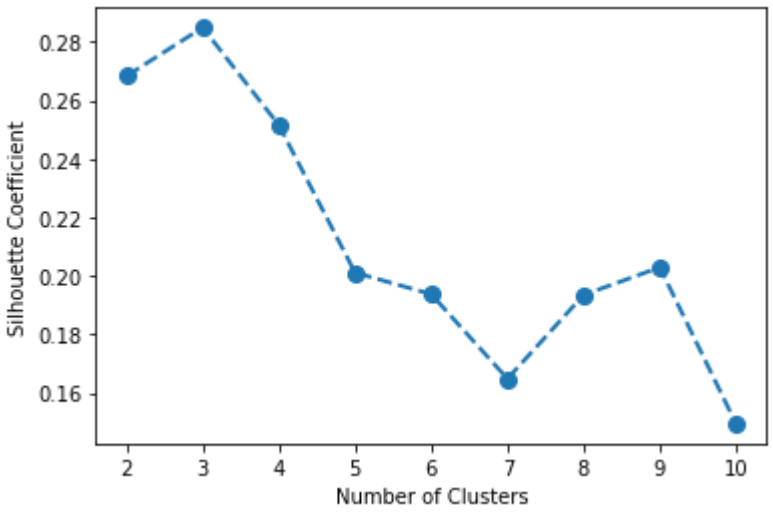
Datasets:

For this Assignment, I have used two datasets which are Wine and Iris Dataset can be found at Kaggle Links provided below:

1. Wine Dataset: https://www.kaggle.com/harrywang/wine-dataset-for-clustering.
2. Iris Dataset: https://www.kaggle.com/rutujavaidya/iris-dataset.
3. A small dataset was created using the make_blobs function provided by the sklearn library.

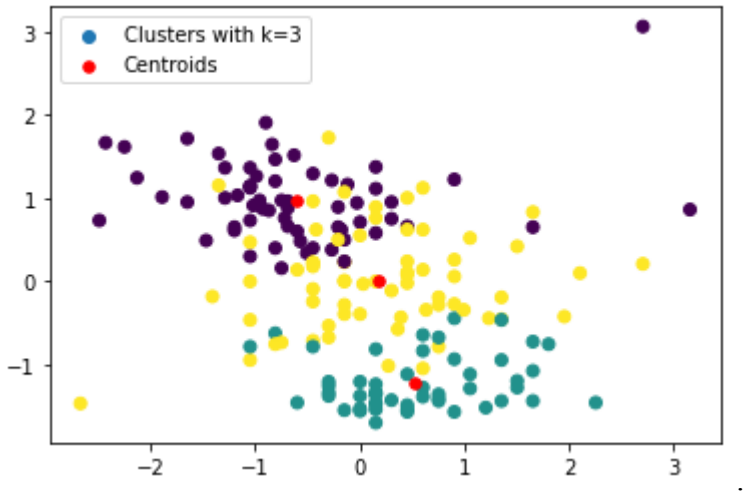
Functions created and used:

1. **For Kmeans finding the best cluster using silhouette score:** I have created a function `kmeans_clustering_find_clusters` which uses the `KMeans` function defined in sklearn and also uses the `silhouette_score` function to calculate the scores of different clusters and the best score is selected and the respective cluster is used.

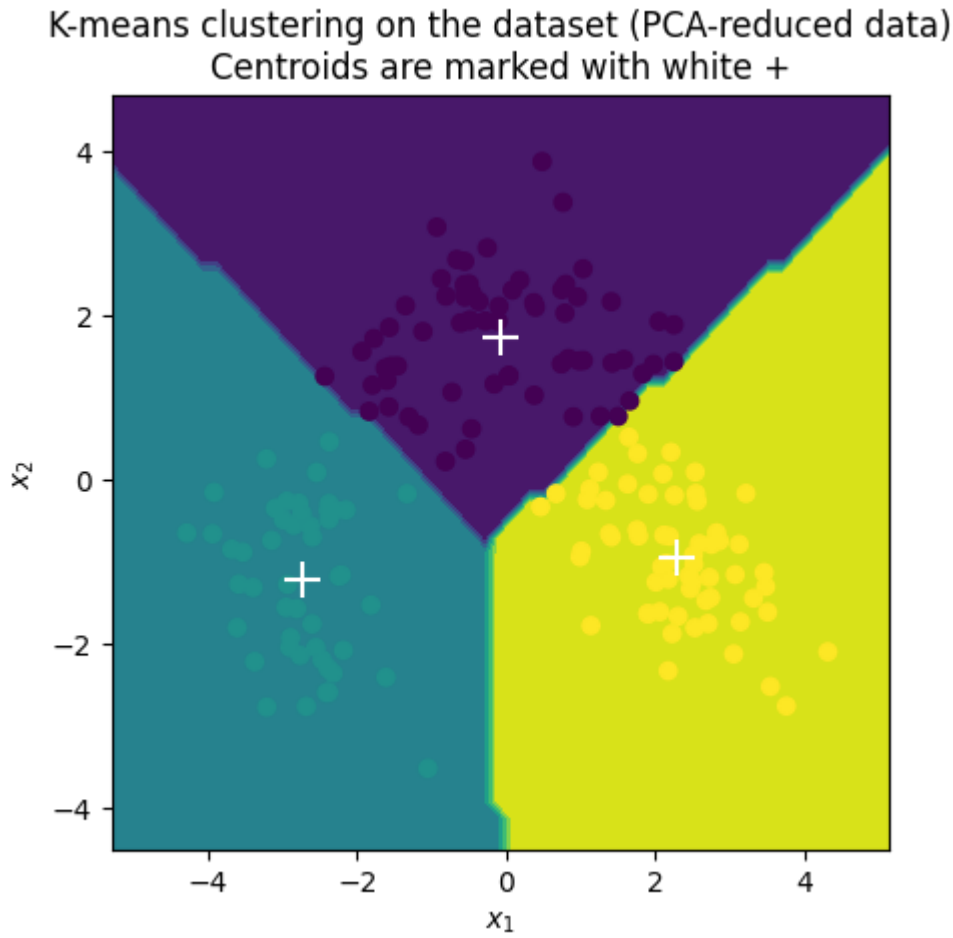


Like in the above-plotted graph best score is 0.28 for 3 clusters hence 3 clusters are used.

2. **Kmeans:** `final_kmeans(scaled_features,plot_x,plot_y,cluster=2)` is used to find the kmeans and do the plotting

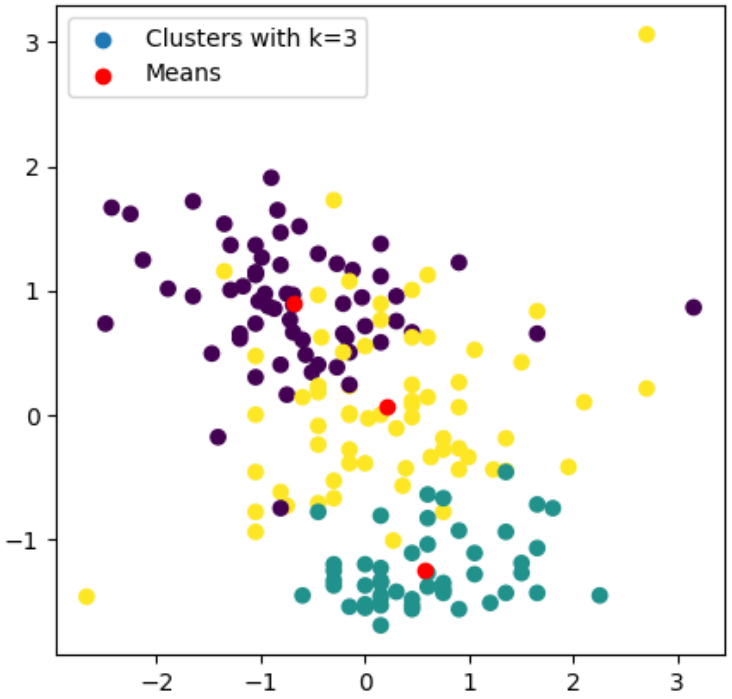


3. **Kmenas using PCA:** `kmeans_pca(scaled_features,cluster=2)` is used to find the kmeans after the data is reduced using PCA.

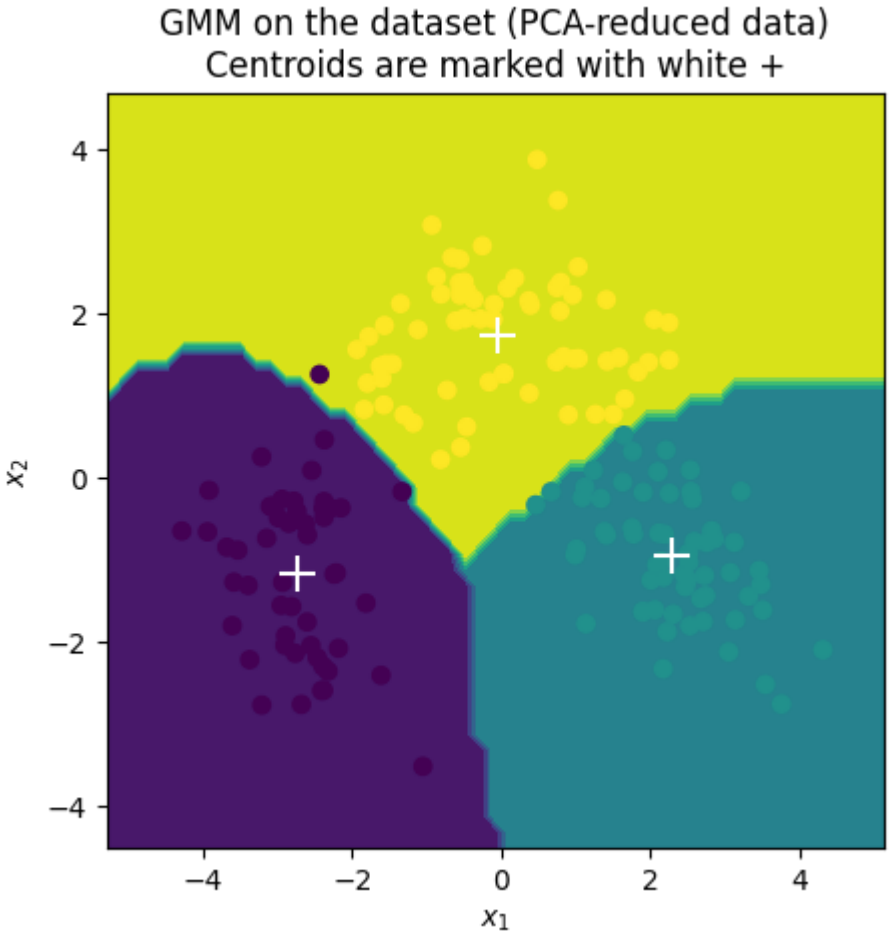


4. **GMM:**

`gausianmixturemodel(scaled_features,plot_x,plot_y,cluster=2)` is used for performing the GMM on the data it performs GMM using the means generated by kmenas and also uses different covariance matrices.



5. **GMM using PCA:** `gmm_pca(scaled_features,cluster=2)` is used to perform GMM on the data reduced by PCA then Plot the data.



Description:
In the above functions, cluster means the best cluster on a dataset, as the dataset is multi-dimensional except for PCA plot_x and plot_y are used to provide the features which you want to plot of this multidimensional dataset.

Experiments conducted:

1. **Full, Diagonal, and Identity covariance matrices:**
   In this, there is not much difference between the output of GMM for the three matrices except for the changes in the values of cluster centers.

2. **For Overlapping examples:** In the case of the Iris dataset the examples are overlapping this can find out easily since the target variable has 3 unique labels but according to silhouette scores 2 clusters are found for the best performance hence we can say there are overlapping examples.
   In this case for Iris dataset i have used 3 clusters.