Lab 10. Clustering using Kmeans and Gaussian Mixture Model(GMM)

You are given with Iris flower dataset file (Iris.scv). The data set consists of 50 samples from each of three species of Iris (Iris setosa, Iris virginica and Iris versicolor). Four features were measured from each sample: the length and the width of the sepals and petals, in centimeters.

1. Apply PCA and select first two directions to convert the data in to 2D. (Exclude the attribute “Species” for PCA)

2. Apply K-means (K=3) clustering on the reduced data. Plot the data points in these clusters (use different colors for each cluster). Obtain the sum of squared distances of samples to their closest cluster center.
(Use **kmeans.fit** to train the model and **kmeans.labels_** to obtaine the cluster labels).

3. Build a GMM with 3 components (use **GMM.fit**) on the reduced data. Use this GMM to cluster the data points (Use **GMM.predict**). Plot the points in these clusters.

4. Obtain the purity score for the Kmeans and GMM clustering (K=3).

5. Repeat part 2 and 3 for K= 2, 3, 4, 5, 6 and 7 and Find the optimum number of clusters (K)
using Elbow method for K-means and GMM.

##

```
# Purity score
# Get y_true using the original file

from sklearn import metrics
def purity_score(y_true, y_pred):
    # compute contingency matrix (also called confusion matrix)
    contingency_matrix = metrics.cluster.contingency_matrix(y_true, y_pred)
    # return purity
    return np.sum(np.amax(contingency_matrix, axis=0)) / np.sum(contingency_matrix)
```