Lab7: Data classification using Bayes Classifier with Gaussian Mixture Model (GMM) and Effect of Dimension Reduction in Classification

You are given the **Pima Indians Diabetes Database** as a csv file (`pima-indians-diabetes.csv`). This dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases. It consists 768 tuples each having 9 attributes. The last attribute for every tuple signifies the class label (0 for non-diabetes and 1 diabetes). It is a two class problem. Other attributes are input features.

1. Show the performance of **K-nearest neighbor (KNN) classifier** for different values of *K* **(1, 3, 5, 7, 9, 11, 13, 15, 17, 19, 21)**

   A. Find **confusion matrix** (use 'confusion_matrix') for each *K*.

   B. Find the **classification accuracy** (You can use 'accuracy_score') for each *K*. Note the value of *K* for which the accuracy is high.

2. Build a **Bayes classifier** with Multi-modal Gaussian distribution (GMM) with *Q* components (modes) as class conditional density for each class. Show the performance for different values of *Q* **(2, 4, 8, 16)**. Estimate the parameters of the Gaussian Mixture Model (mixture coefficients, mean vectors and covariance matrices) using maximum likelihood method.

   A. Find **confusion matrix** (use 'confusion_matrix') for each *Q*.
   B. Find the **classification accuracy** (You can use 'accuracy_score') for each *Q*.
   C. Observe the values in the covariance matrix in each case and comment.
   D. Compare the results with that obtained using **Bayes classifier** with unimodal Gaussian distribution ($Q = 1$).

3. Reduce this multidimensional data into *l* dimensions using **principle component analysis (PCA)**. Now repeat Part 1 and 2 using reduced dimensional representation of each samples. Show the results for different values of *l* (1, 2, ..., *d*). Here *d* is the actual dimension of the data.

Observation:

I. Compare and comment on the accuracy for each classifiers.

II. Is there any improvement in the accuracy of the Bayes classifier after using GMM compared to Bayes classifier with unimodal Gaussian?

Notes:

Use the function "mixture.GaussianMixture" from scikit-learn to build GMM.

a) Standardize the data before building classifiers.

b) 70% of data from each class should be used for training and remaining for testing.

c) Results should be shown using confusion matrix and classification accuracy for all the assignment.
(Use inbuilt function 'confusion_matrix')