REPORT

Lab Assignment 9

Vrushali Pandit - B20BB047

https://colab.research.google.com/drive/1gZzP6rvNPPpAlqBBO43pEPqT9KoGB4_F?usp=sharing

Data:

The dataset contains information about patients with or without diabetes.

The **goal** is to try to **predict from given features** of medical conditions and history of the patient **whether they have diabetes or no**t. Features given : number of pregnancies, glucose, blood pressure, skin thickness, insulin, BMI, DiabetesPedigreeFunction and their age.

Pre-processing:

1. Although there aren't any null values in the data, there are many values = 0 which don't make sense. Such as-
    a. Glucose
    b. Blood pressure
    c. BMI
    d. Skin thickness
2. Hence we impute these 0 values of the aforementioned features by the mean values.
3. And we must standardize the data to remove any possible bias before applying LDA, PCA, or KNN.

Q1) Comparing which is better PCA or LDA on the basis of accuracy in classification

- Because we are dealing with a Binary Classification problem, LDA (Linear Discriminant Analysis) can only project the original data onto 1 dimensional space.
- We import *LinearDiscriminantAnalysis* from *sklearn.discriminant_analysis*
- No matter what n_componeents we set for LDA, it will project the data onto a 1D space.
- We fit and transform the original data (supervised) to get the reduced feature space.

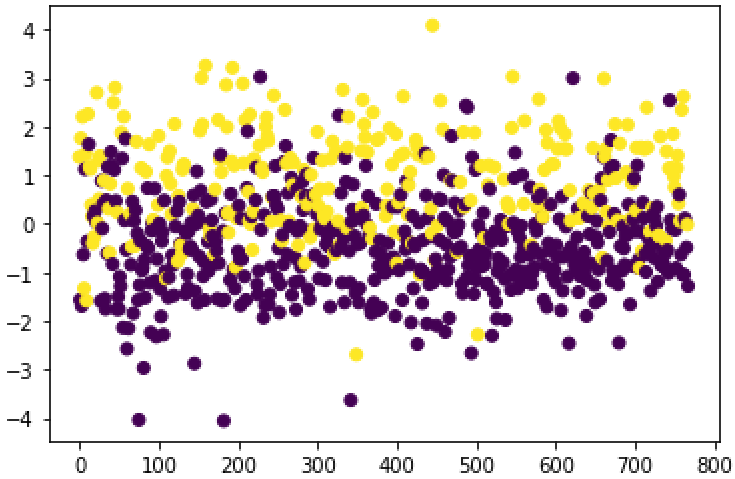
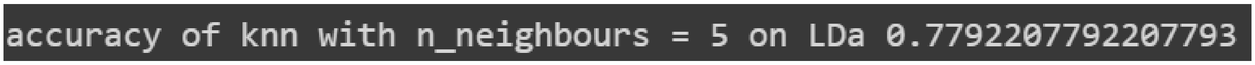


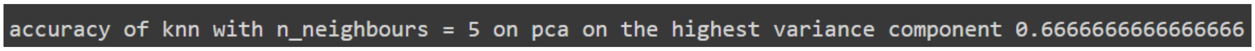
Figure 1 : Plot of features extracted by LDA

- I then applied PCA with a number of components = what would capture 95% of the variance.
- This returns n_components_ = 7.
- We can't compare the classification performance of 1 feature of LDA vs multiple features of PCA.
- Hence we take individual components of PCA (highest variance then 2nd highest variance and so on) to do the comparison.
- LDA accuracy:

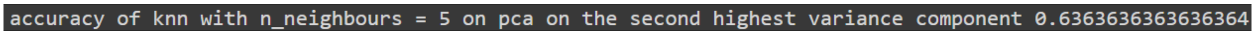


```
accuracy of knn with n_neighbours = 5 on LDa 0.7792207792207793
```

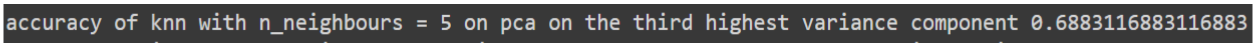
- PCA accuracy with different components:



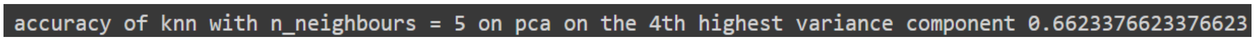
```
accuracy of knn with n_neighbours = 5 on pca on the highest variance component 0.6666666666666666
```



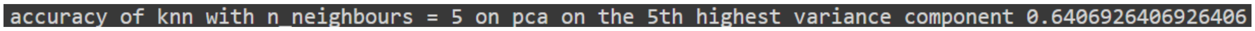
```
accuracy of knn with n_neighbours = 5 on pca on the second highest variance component 0.6363636363636364
```



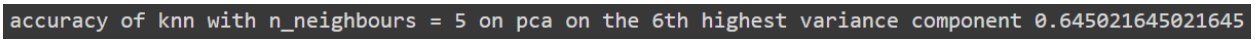
```
accuracy of knn with n_neighbours = 5 on pca on the third highest variance component 0.6883116883116883
```



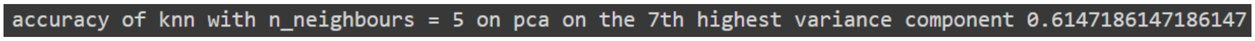
```
accuracy of knn with n_neighbours = 5 on pca on the 4th highest variance component 0.6623376623376623
```



```
accuracy of knn with n_neighbours = 5 on pca on the 5th highest variance component 0.6406926406926406
```

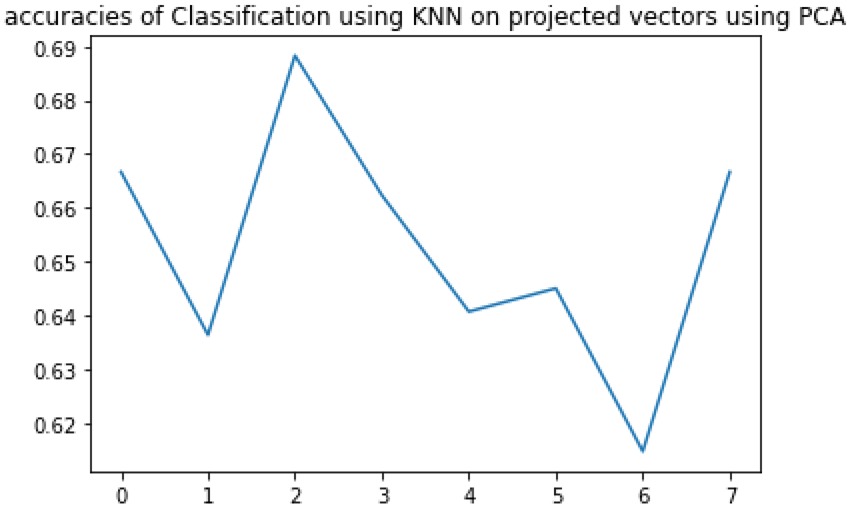


```
accuracy of knn with n_neighbours = 5 on pca on the 6th highest variance component 0.645021645021645
```



```
accuracy of knn with n_neighbours = 5 on pca on the 7th highest variance component 0.6147186147186147
```

A plot of the same:

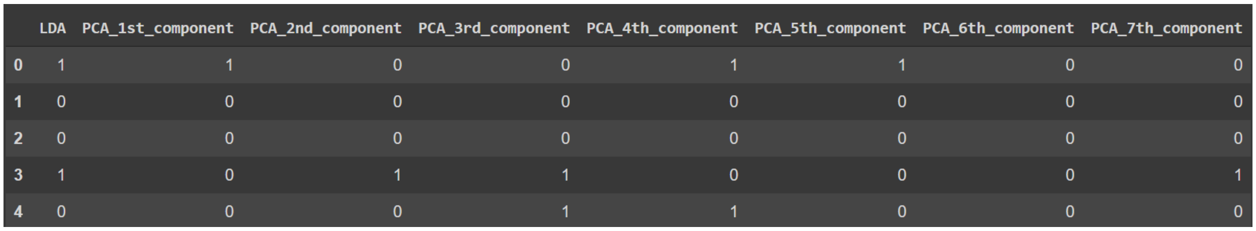


As we can see, classification on features extracted by LDA outperform PCA.

WHY?
- LDA is specifically designed to extract features in classification tasks.
- It maximizes the difference between the classes and minimizes the difference within classes.
- While PCA merely tries to maximize the variance of differences between all the points irrespective of labels.

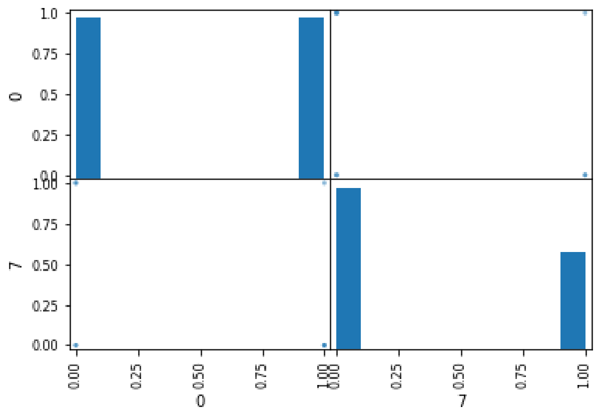
Q2) Plotting the scatter matrices for different comparisons (LDA the first 7 components of PCA)

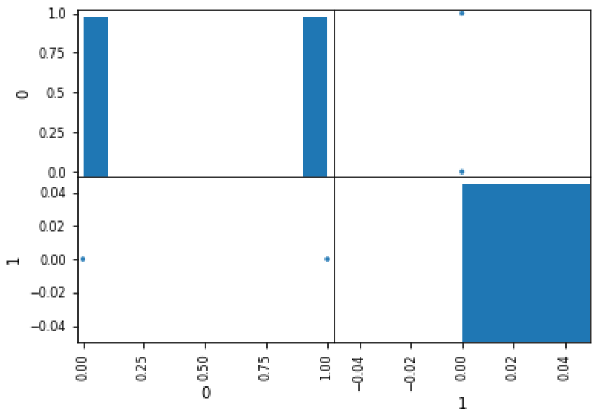
- To do so we first need to create a dataframe with rows and columns as shown: (class prediction as done in q1)

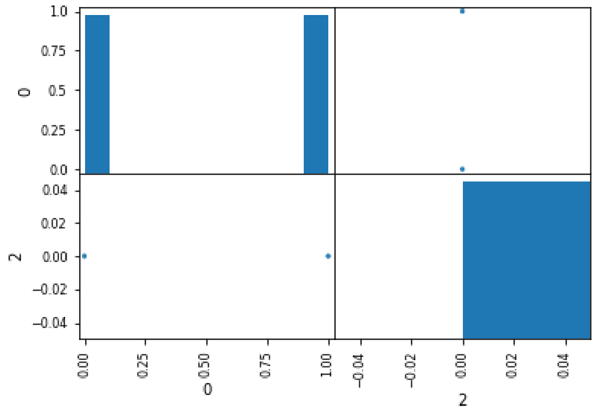


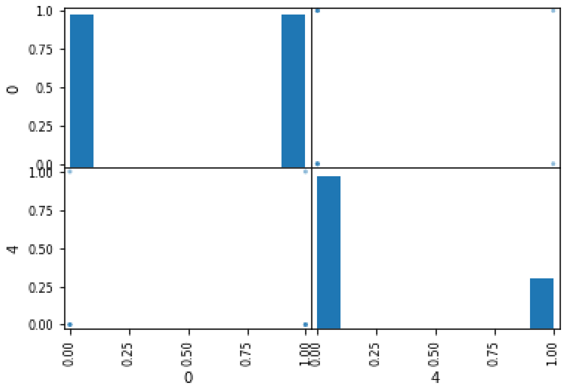
	LDA	PCA_1st_component	PCA_2nd_component	PCA_3rd_component	PCA_4th_component	PCA_5th_component	PCA_6th_component	PCA_7th_component
0	1	1	0	0	1	1	0	0
1	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0
3	1	0	1	1	0	0	0	1
4	0	0	0	1	1	0	0	0

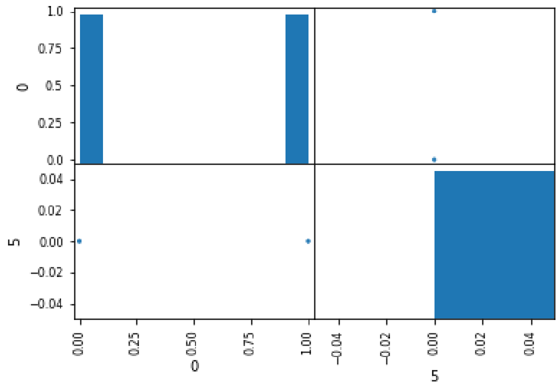
-

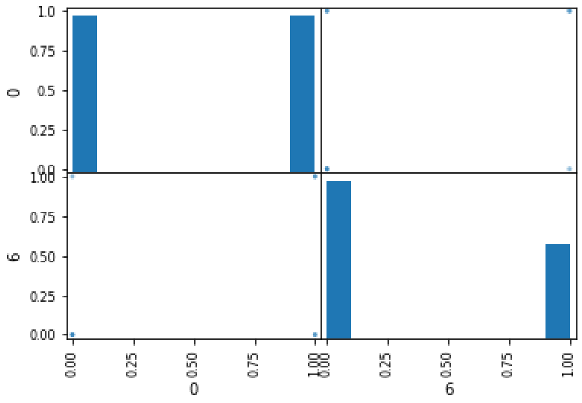












Q3) Learn a Logistic Regression Classifier and compare its results on original data and features extracted from LDA

- Accuracy when Logistic Regression Classifier is trained on entire data = **0.7965**
- Accuracy when Logistic Regression Classifier is trained on features extracted by LDA = **0.806**

- We see that fewer and more condensed clusters of data allows Logistic Regression to give a higher accuracy. Because less features means less need of regularization and hence it prevents overfitting.

Q4) Learn a MLP - Multi Layer Perceptron Classifier on the original dataset, and experiment with different parameters and HPs

- I used a single hidden layer (3 layer MLP) and experimented with nodes = 10,30,50,70,90
- And activation = 'identity’, ‘logistic’, ‘tanh’, ‘relu’
- The highest test accuracy was obtained with 10 nodes and 'identity' activation. = **0.8138**
- This is the highest accuracy we have achieved because MLPs can learn complex functions.
- Also MLPs with larger number of nodes have a tendency to learn or memorize the data. This leads to overfitting.
- Hence an MLP with fewest nodes is giving the highest accuracy on test data.