LAB Report:

Name:Devyani Gorkar
Roll no:B20ME027

About Code:

df is the data frame consisting of the diabetes.csv file.X is the input array and Y is the output or classes containing column.X_lda_new is the new array with reduced dimension.This new input is divided into train and test sets with X_train_lda,X_test_lda,y_train_lda,y_test_lda.We feed the same input X into PCA model to reduce its dimension and then split it into. X_train_pca, X_test_pca ,y_train_pca, y_test_pca.In LDA the reduced no. of features are less than or equal to min(n_classes-1,n_features),so in our case the number of features after applying LDA is less than or equal to 1.To make the comparison effective we keep n_components in PCA equal to 1.We use KNN(k=5) for PCA reduced X and LDA reduced X.When we use LDA to reduce the dimensions of X and after applying KNN on X_lda_new set we the accuracy score in test dataset as 0.753246 but when we use PCA instead of LDA we get the accuracy score as 0.636363 .Further we print the confusion matrix for both techniques and also plot the graphs. Later we use decision tree classifiers on input without reduced features and input with reduced features using LDA.Atlast we use multilayer perceptron classifier to classify the data.

Observation.
- Comparison of KNN score for both cases.
  On fitting both type of training data in the KNN model we get the accuracy score 0.753246 for LDA reduced X and for PCA reduced X we get the accuracy score as 0.636363.

So we can see that LDA creates a better separation between two classes while reducing the dimensions.
- Comparison of confusion matrix and plots
  Confusion matrix of LDA testdata:

```
confusion_matrix(y_test_lda,y_lda_new)
[46]
...  array([[120,  27],
            [ 30,  54]], dtype=int64)
```

Confusion matrix of PCA test data:
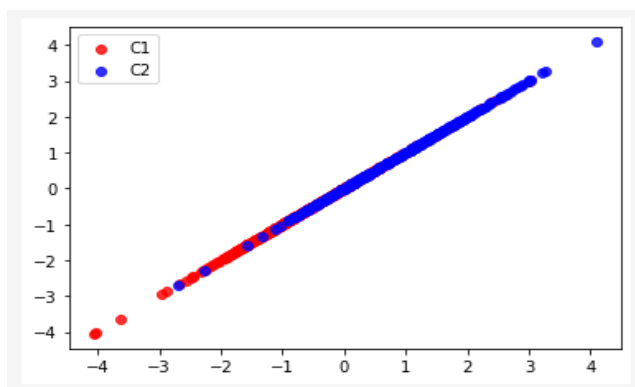
```
    ▷ ∨        confusion_matrix(y_test_pca,y_pca_new)

[47]

  ...   array([[109,  47],
               [ 41,  34]], dtype=int64)
```
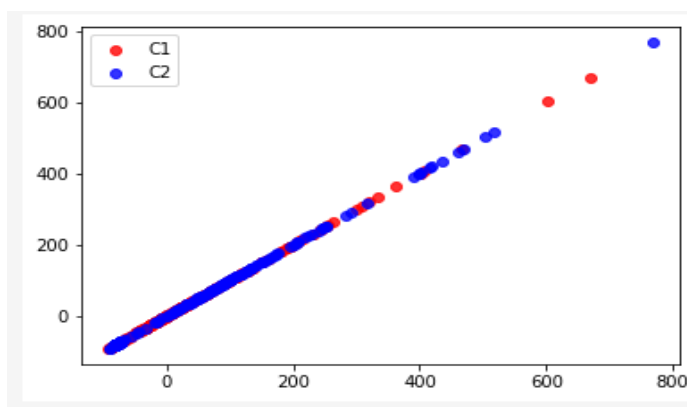
We can see that the true values (TT) and FF values for LDA reduced data is more compared to PCA reduced data .This shows that LDA classifies test data more accurately compared to PCA.
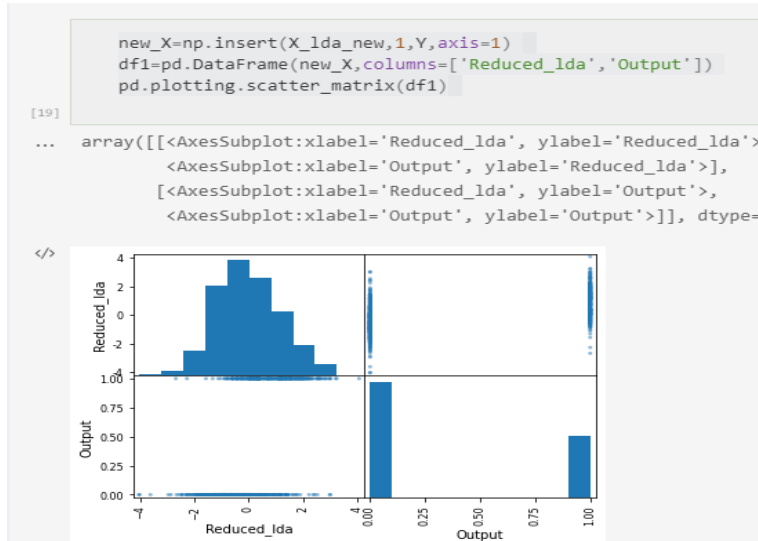
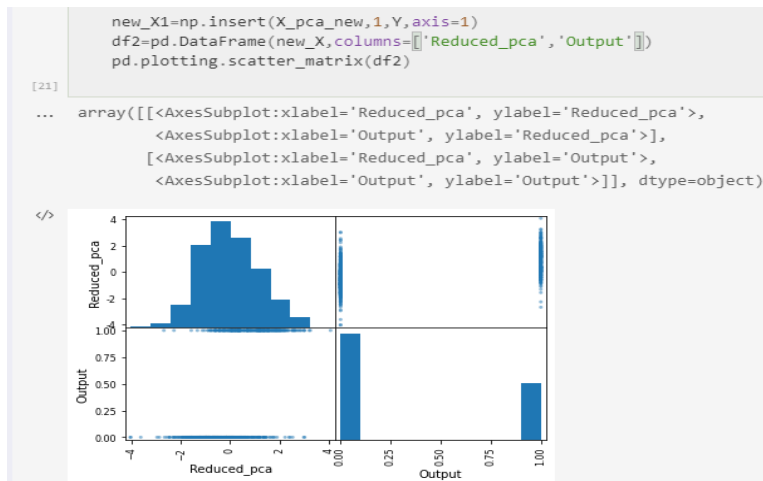Plot of LDA reduced X:

Plot of PCA reduced X:

We can see that in X_lda_new plot the red and blue points are somehow separable whereas in the X_pca_new plot the data points (blue and red) are overlapping most of the times and hence the separation becomes difficult in this case

- Scatter Matrix:
  For LDA:

```
new_X=np.insert(X_lda_new,1,Y,axis=1)
df1=pd.DataFrame(new_X,columns=['Reduced_lda','Output'])
pd.plotting.scatter_matrix(df1)
```
```
[19]
...    array([[<AxesSubplot:xlabel='Reduced_lda', ylabel='Reduced_lda'>
              <AxesSubplot:xlabel='Output', ylabel='Reduced_lda'>],
             [<AxesSubplot:xlabel='Reduced_lda', ylabel='Output'>,
              <AxesSubplot:xlabel='Output', ylabel='Output'>]], dtype=
```


  For PCA:

```
new_X1=np.insert(X_pca_new,1,Y,axis=1)
df2=pd.DataFrame(new_X,columns=['Reduced_pca','Output'])
pd.plotting.scatter_matrix(df2)
```
```
[21]
...    array([[<AxesSubplot:xlabel='Reduced_pca', ylabel='Reduced_pca'>,
              <AxesSubplot:xlabel='Output', ylabel='Reduced_pca'>],
             [<AxesSubplot:xlabel='Reduced_pca', ylabel='Output'>,
              <AxesSubplot:xlabel='Output', ylabel='Output'>]], dtype=object)
```


- Comparison of decision tree results with and without lda.
  The accuracy score of decision tree on original data 0.696969696969697 and the accuracy of reduced data using LDA is 0.6753246753246753 respectively.This is because we have lost some information while reducing dimensions of the data therefore the accuracy slightly reduces.
- Accuracy score for different hidden layers of MLPclassifier.
  The accuracy of MLP classifier on using it over X_train_lda is way higher than the accuracy of decision tree.The highest accuracy we obtain on using MLP classifier is 0.8000 for hidden_layer_size=300.

Results:

Hence we can conclude that it is better to use LDA than PCA to reduce the number of features while doing classification.The MLPclassifier gives better accuracy compared to other classifiers like decision trees.

Colab link:

https://colab.research.google.com/drive/12WiUaW-LzNe9CQS0vkGTtVj299NzUSx-?usp=sharing