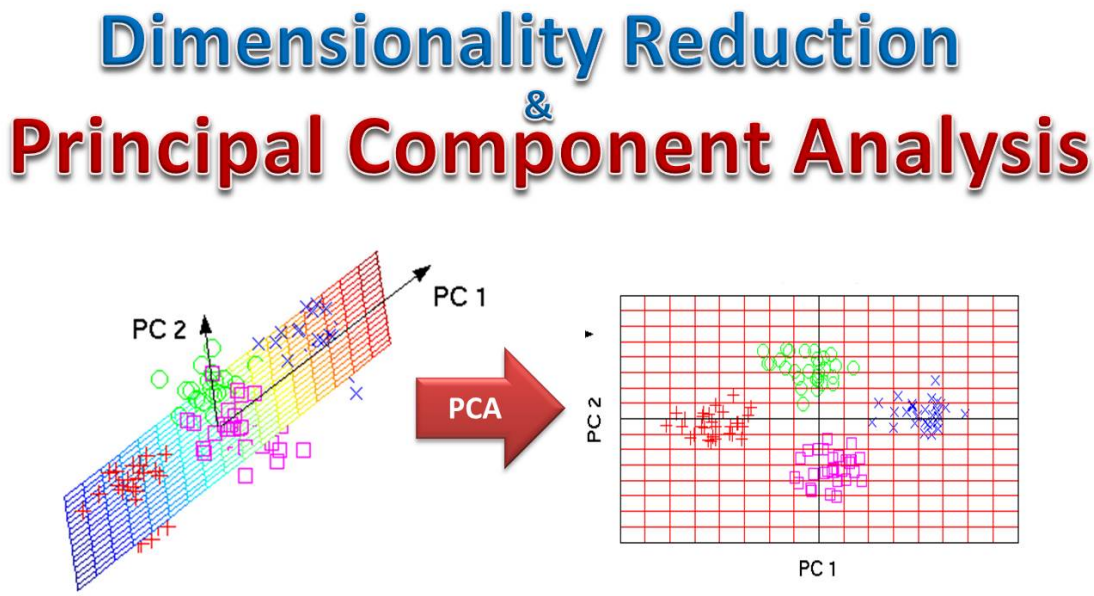
Pattern Recognition and Machine Learning

Lab 7 Report Kartik Choudhary [B20CS025]

Dimensionality Reduction

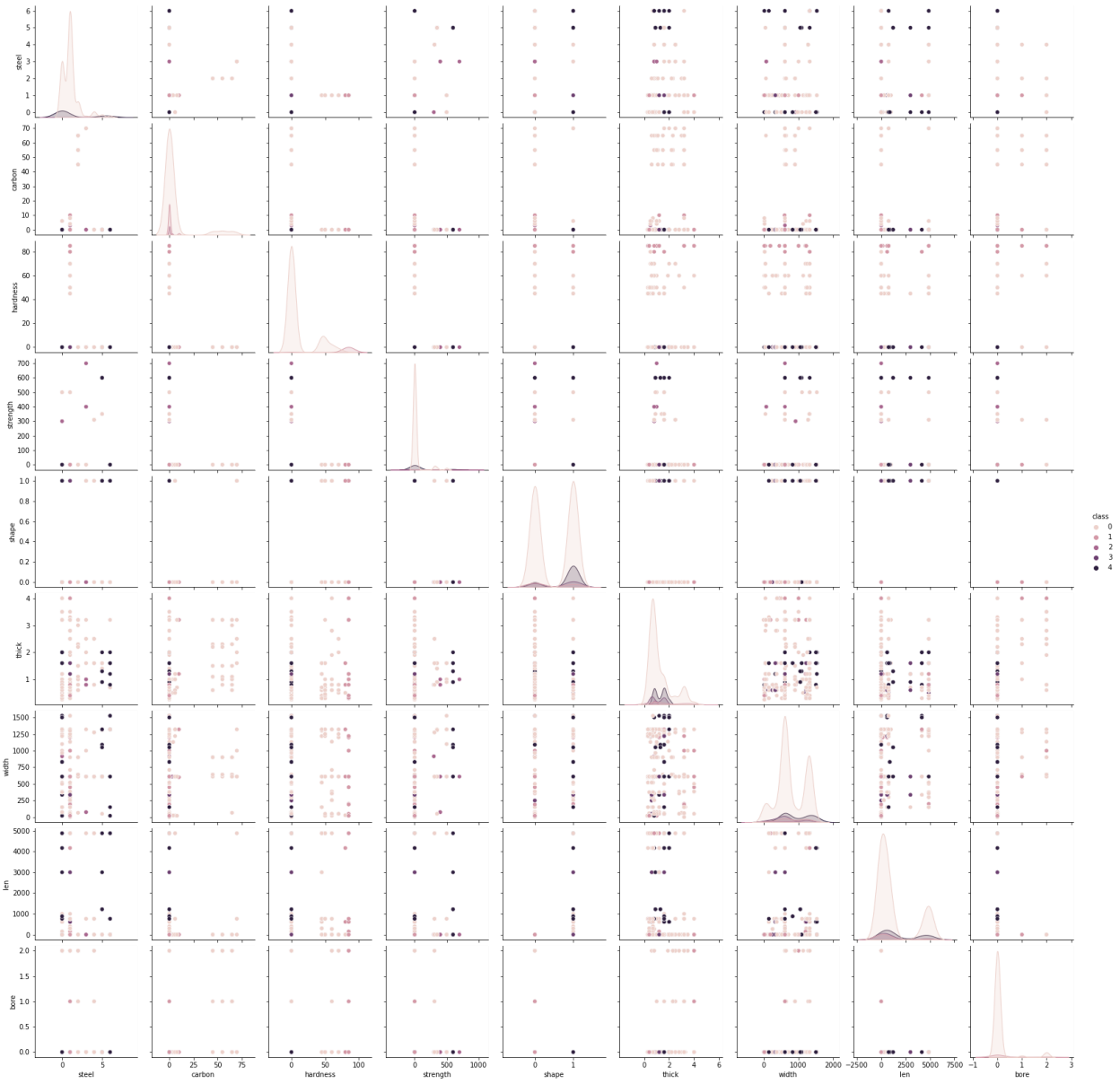


Question 1

1. The data was converted into a readable format using the pd.read_csv.
2. After pre-processing, the data obtained by us comprised of 9 different features namely :
    a. Steel
    b. Carbon
    c. Hardness
    d. Strength
    e. Shape
    f. Thick
    g. Width
    h. Len

i. Bore

Plotting the data gives us the following pair plot.



3. As asked in the question, we had to choose 2-3 classification methods. The methods chosen by me for the particular classification task were :

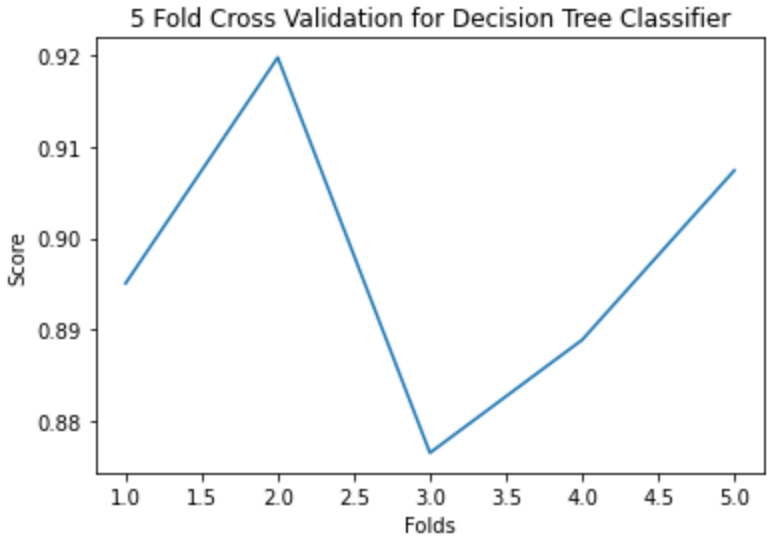
   Decision Tree Classifier

   Gaussian Naive Bayes Classifier

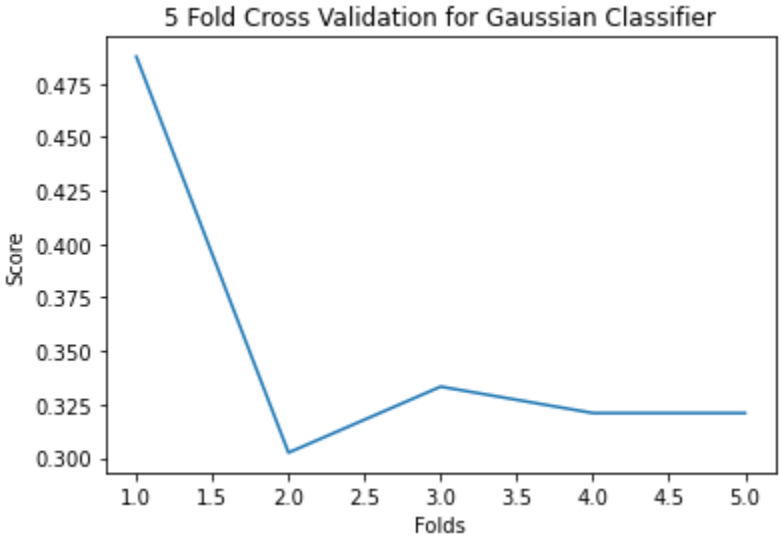
Multinomial Naive Bayes Classifier

CrossVal Score Obtained without PCA are as follows :

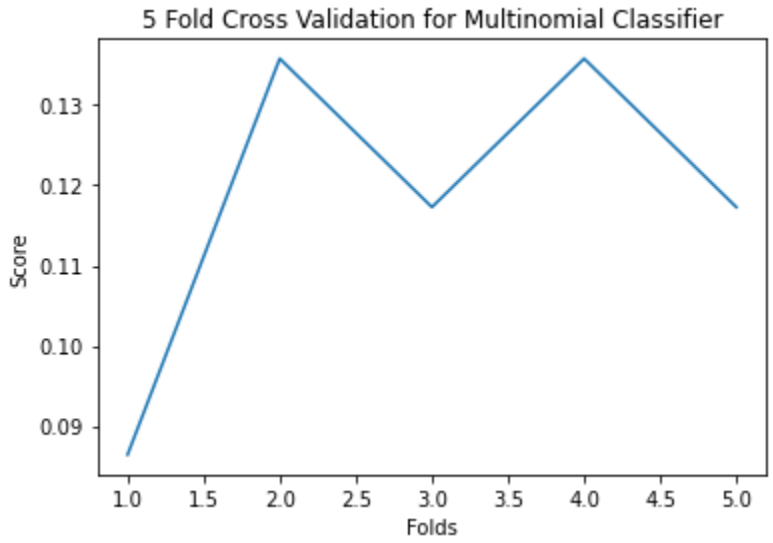
For Decision Tree Classifier:



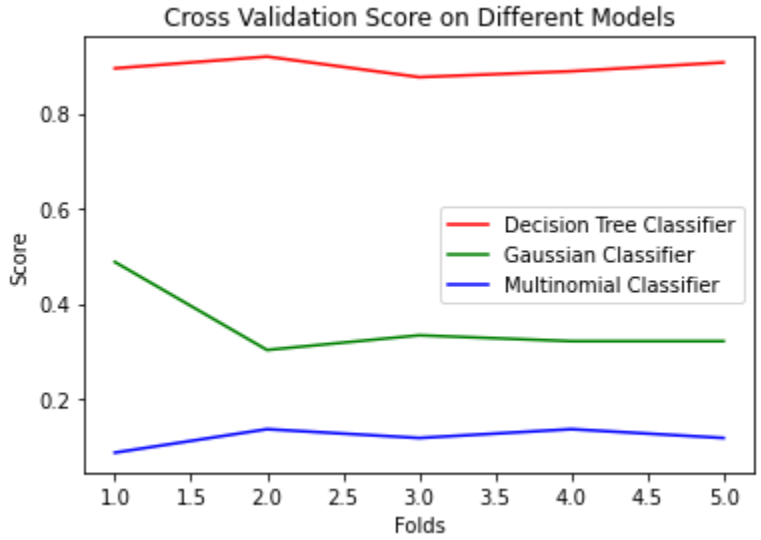
For Gaussian Classifier



For Multinomial Classifier



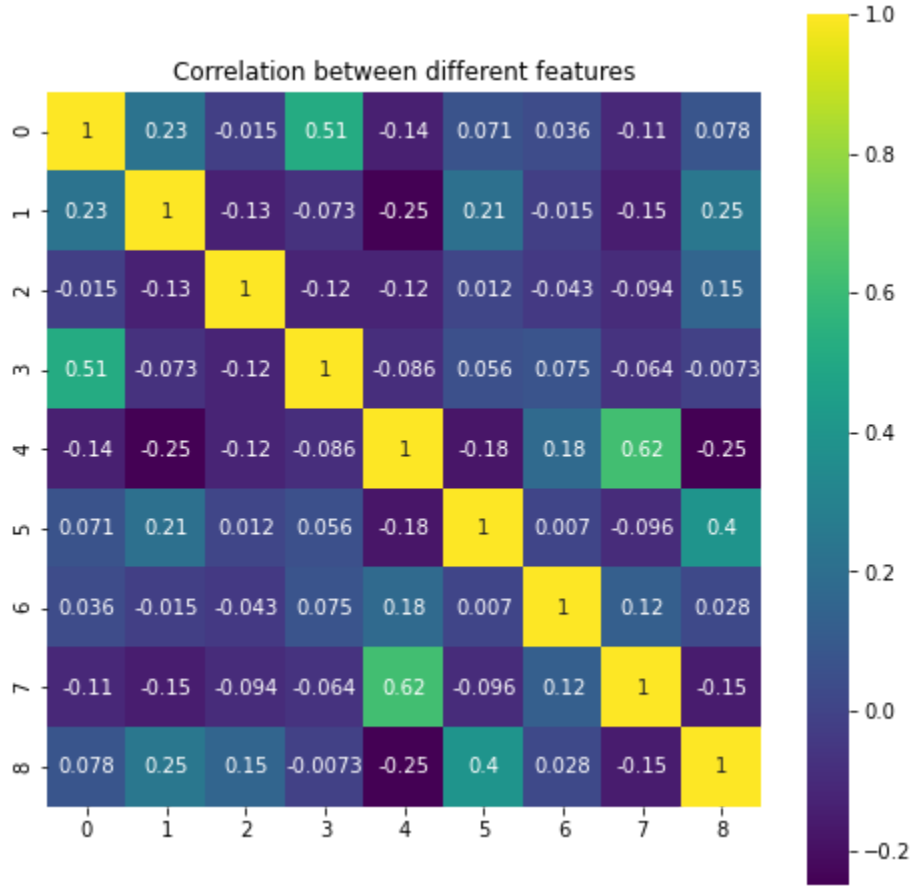
Five Fold Cross-Validation of the above models

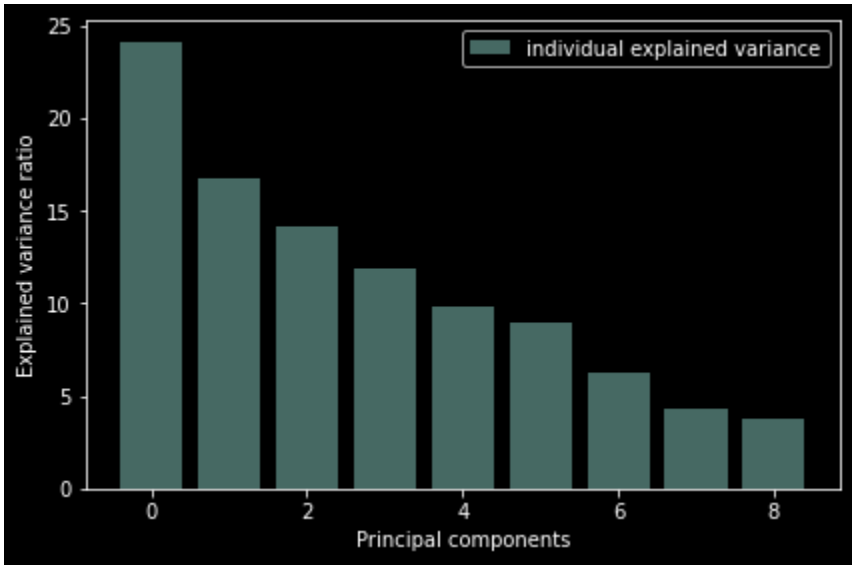


4. **Principal Component Analysis**

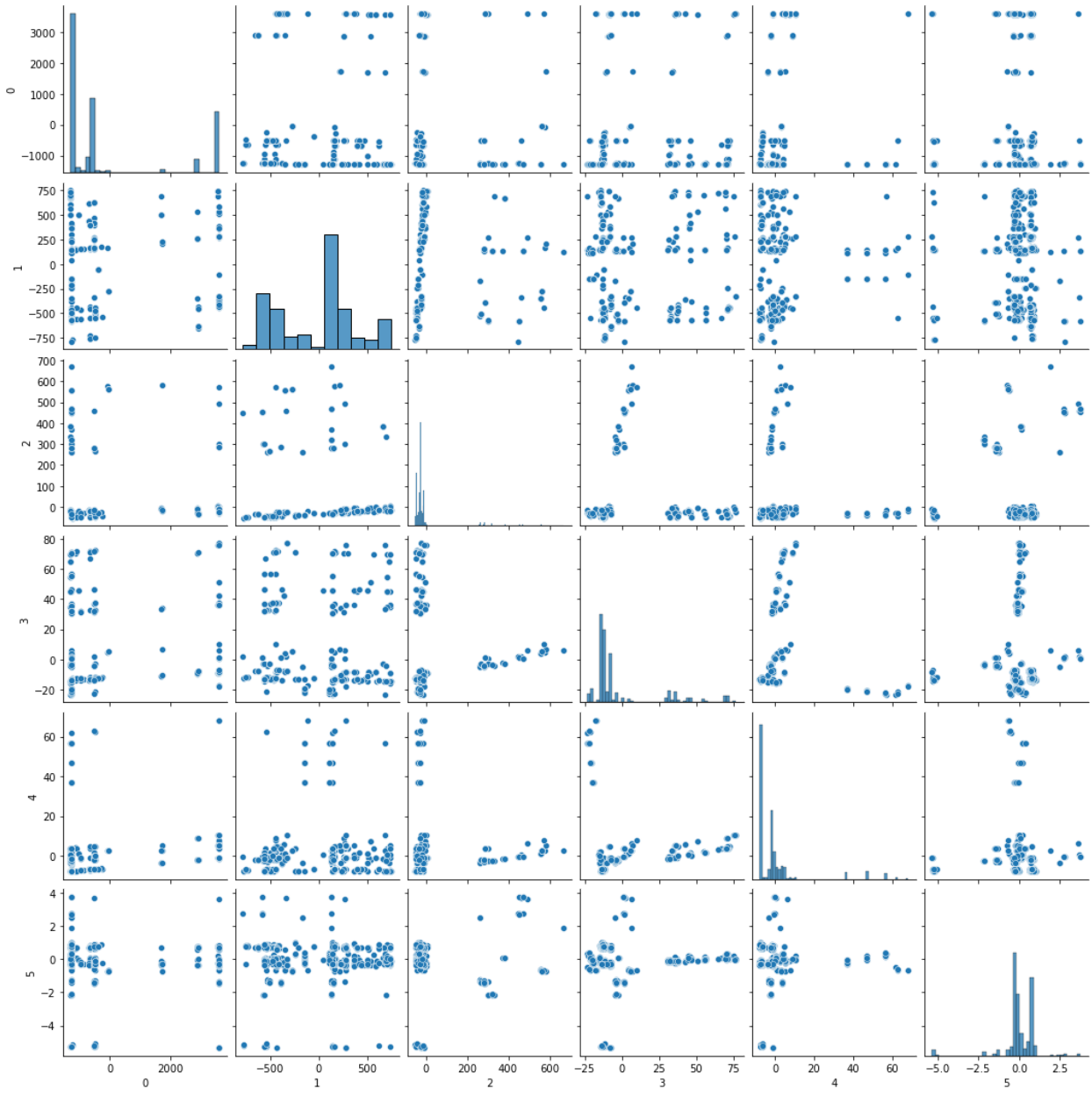
For Principal Component Analysis (PCA), I have used feature-wise-mean centralization. Then the original feature space was projected to PCA space, which comprised 6 components.

While projecting the feature space to PCA space, **covariance matrix** obtained was as follows :





Plot after doing the PCA



Accuracy and F1 Score obtained after PCA are as follows :

FOR DECISION TREE CLASSIFIER:

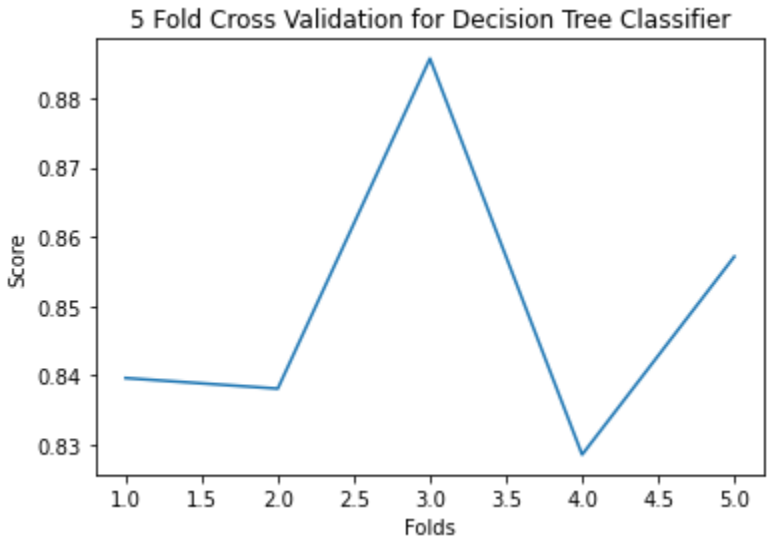
Cross-Validation Score: `[0.83962264 0.83809524 0.88571429 0.82857143 0.85714286]`

Accuracy Score: `86.97183098591549 %`

```
[0.91928251 0.93333333 0.8        0.46153846 0.6557377 ]
```

F1 Score:

5 Fold Cross-Validation for Decision Tree Classifier:



FOR GAUSSIAN CLASSIFIER:

Cross-Validation Score:

```
[0.63207547 0.56190476 0.51428571 0.61904762 0.58095238]
```

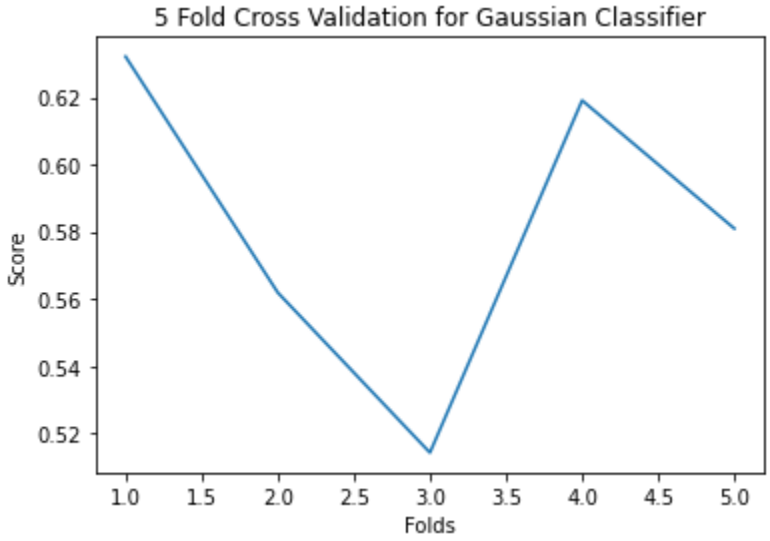
```
57.3943661971831 %
```

Accuracy Score:

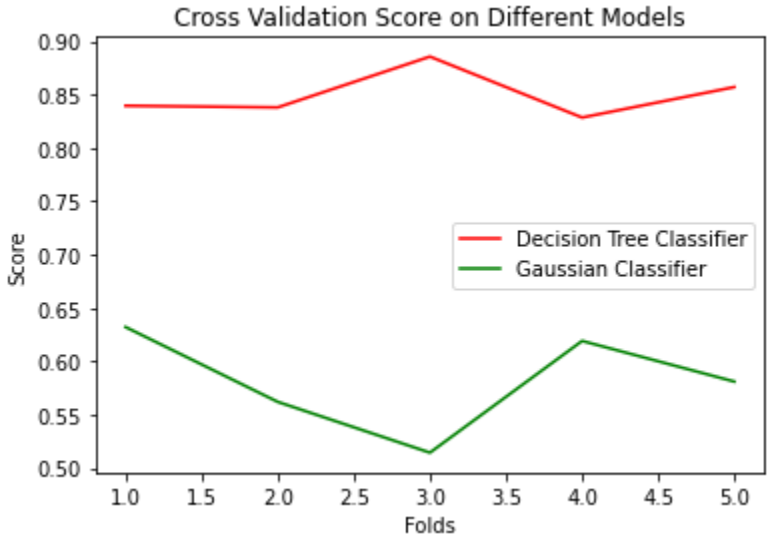
```
[0.68586387 0.63829787 0.28571429 0.28915663 0.14285714]
```

F1_Score:

5 Fold Cross-Validation for Gaussian Classifier:



Cross-Validation Score on Different ModelsCross Validation Score on Different Models:



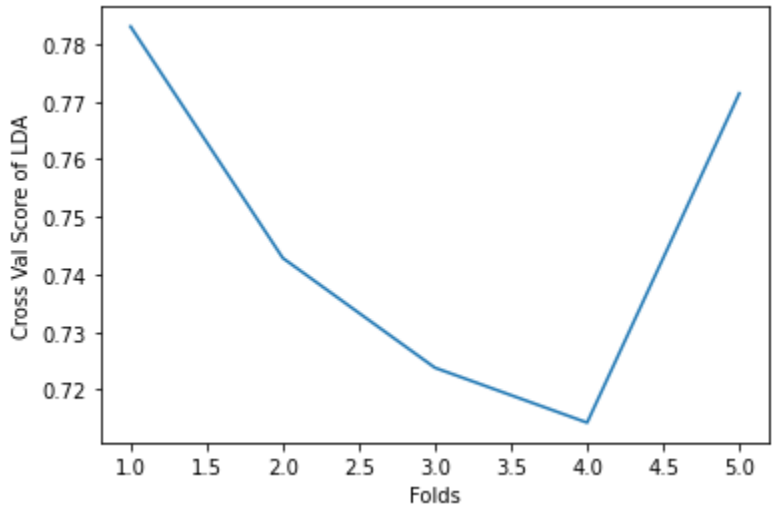
After applying PCA, we can observe that since there was a low correlation between the original feature space, therefore the effect of PCA on accuracy was not much prominent. Whereas, if the correlation would have been slightly higher, then the PCA would have contributed towards increasing the accuracy of the classification model.

Effect of PCA on the distribution of the dataset

As observed in the plots shown above, PCA has tried to capture as much variance in the dataset as possible, whereas in the original dataset distribution variance was not much visible.

Question 2: Linear Discriminant Analysis:

Plot after doing Linear Discriminant Analysis (LDA) :



As observed above the feature with high impact is thick and width

Classification 2*2 table is as follows :

	Decision Tree classifier	Quadratic Discriminant Analysis
PCA	83.098592	54.225352
LDA	87.323944	54.225352

Obtained ROC and AUC after 5-fold cross-validation is as follows :

Plotting ROC and AUC is not possible as for plotting ROC and AUC, we need the no of output classes to be equal to 2, whereas here we have 5.