Homework 3 (ML Practice)

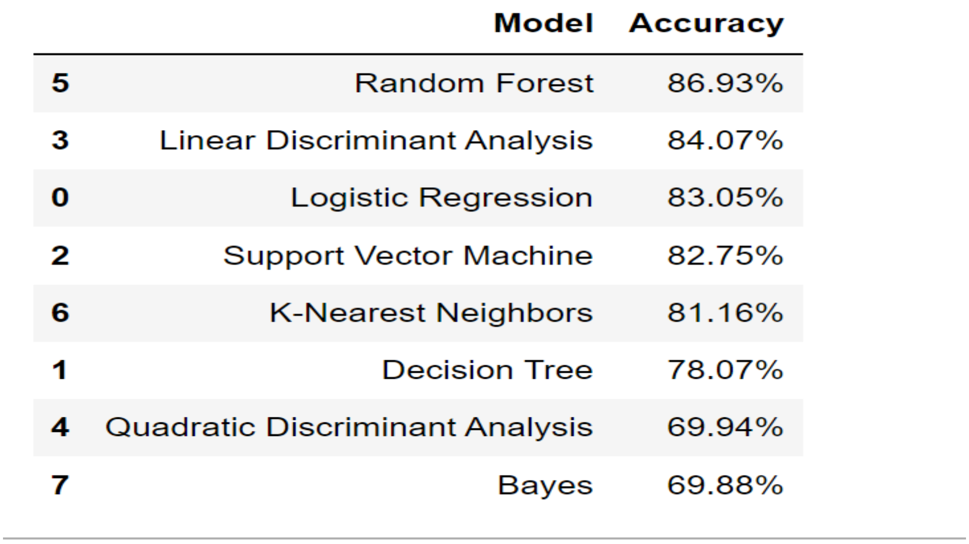
Name: Nasir Khan

Student Number: 0075244

1. We first import the data file cses4 cut.csv"

2. Using ***train_test_split*** from sklearn.model_selection, we split the data so that 70% data is used for training 30% is utilized for testing.

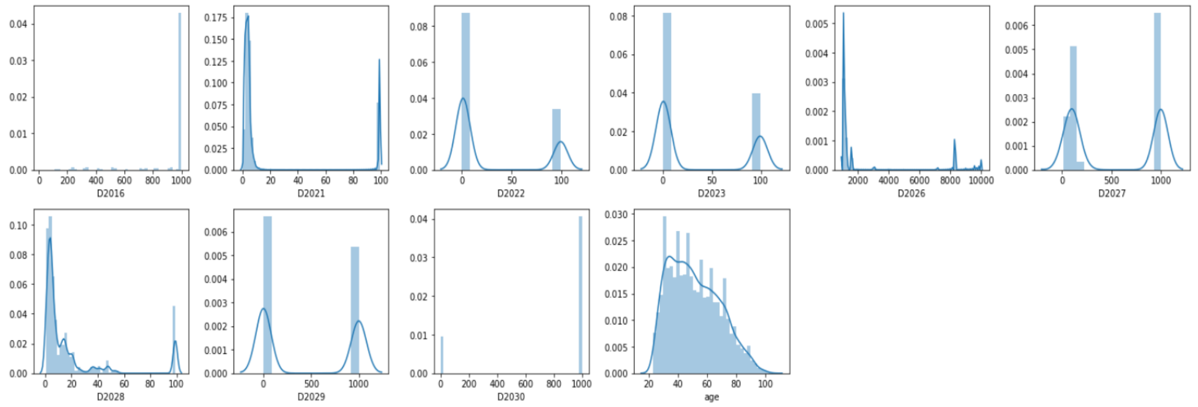
3. We compare the 'Logistic Regression', 'Decision Tree', 'Support Vector Machine', 'Linear Discriminant Analysis', 'Quadratic Discriminant Analysis', 'Random Forest', 'K-Nearest Neighbors', and 'Bayes' classifier and compare their accuracy.



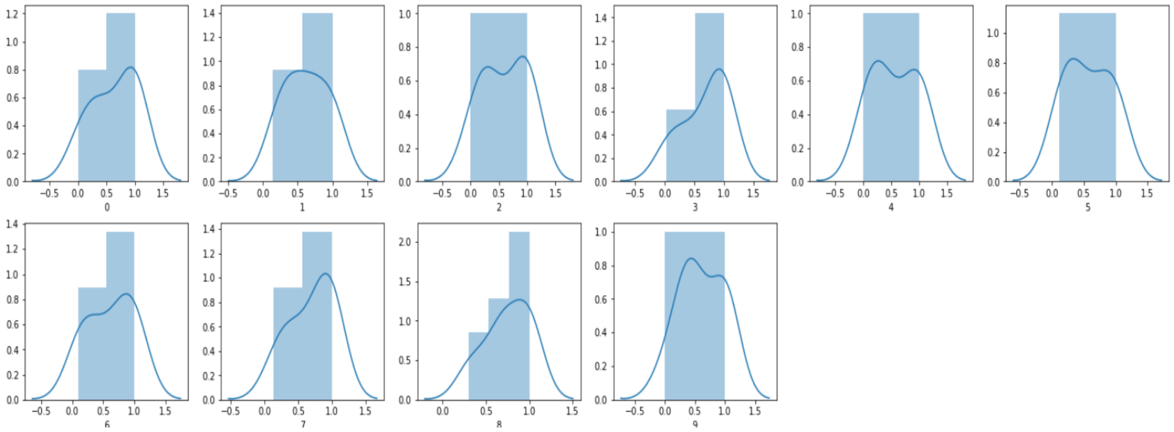
	Model	Accuracy
5	Random Forest	86.93%
3	Linear Discriminant Analysis	84.07%
0	Logistic Regression	83.05%
2	Support Vector Machine	82.75%
6	K-Nearest Neighbors	81.16%
1	Decision Tree	78.07%
4	Quadratic Discriminant Analysis	69.94%
7	Bayes	69.88%

4. From the figure above we see that Random Forest performs the best which is due to the fact that in each iteration the number of decision nodes get reduced.
5.
6. We analyze the feature selection and dimensionality-reduction by utilizing ***sklearn.feature_selection*** to compare the top 20 features with the highest score.

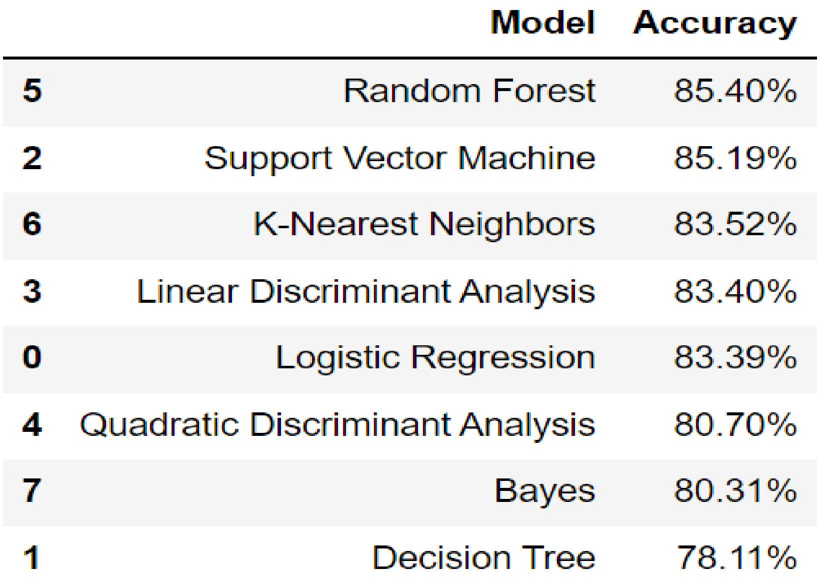
7. We then select subset of 10 feature selection starting from 2016 to visualize them and use seaborn library for this purpose. The figure shows the trends for the different years:



8. From the above figure, we see that the data is not normal Gaussian, so we preprocess the data distribution to make it in Gaussian form. The results of transformation are as follows:



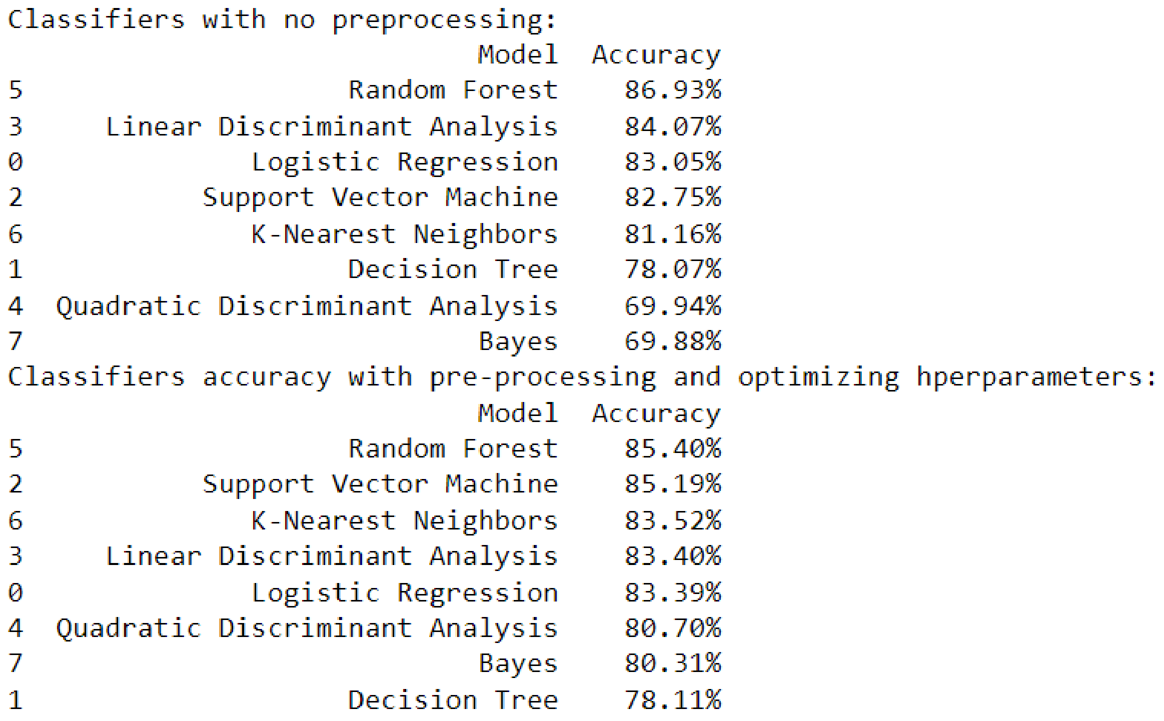
9. Now, we perform and check accuracy of the different classifiers again on this transformed and pre-processed subset of the data i-e., for classifiers with dimensionality-reduction and pre-processing.



	Model	Accuracy
5	Random Forest	85.40%
2	Support Vector Machine	85.19%
6	K-Nearest Neighbors	83.52%
3	Linear Discriminant Analysis	83.40%
0	Logistic Regression	83.39%
4	Quadratic Discriminant Analysis	80.70%
7	Bayes	80.31%
1	Decision Tree	78.11%

10. The overall accuracy performance is improved but the trend remains the same i-e, Random forest still performs the best.

    Conclusion:
    The overall performance comparison of different classifiers before and after preprocessing and optimization is shown below:



```
Classifiers with no preprocessing:
                             Model  Accuracy
5                    Random Forest    86.93%
3     Linear Discriminant Analysis    84.07%
0              Logistic Regression    83.05%
2           Support Vector Machine    82.75%
6              K-Nearest Neighbors    81.16%
1                    Decision Tree    78.07%
4  Quadratic Discriminant Analysis    69.94%
7                            Bayes    69.88%
Classifiers accuracy with pre-processing and optimizing hperparameters:
                             Model  Accuracy
5                    Random Forest    85.40%
2           Support Vector Machine    85.19%
6              K-Nearest Neighbors    83.52%
3     Linear Discriminant Analysis    83.40%
0              Logistic Regression    83.39%
4  Quadratic Discriminant Analysis    80.70%
7                            Bayes    80.31%
1                    Decision Tree    78.11%
```