DTSC 620 - STATISTICS FOR DATA SCIENCE PROJECT ASSIGNMENT 2

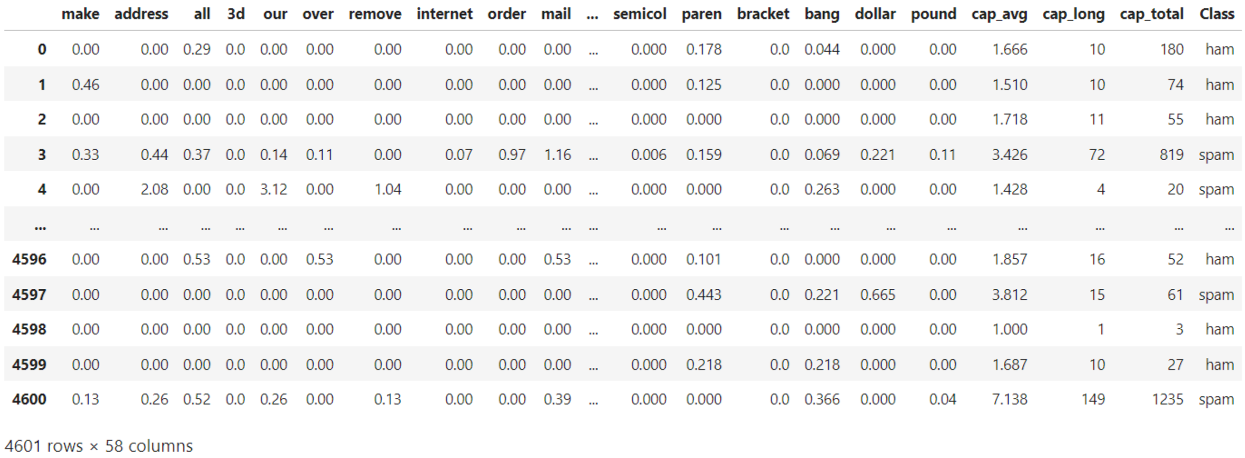
BAHADURI PRACHITI JAGDISH
NYIT ID: 1317686
Manhattan Campus (M01)

Project Report

Project objective: To fuse three classifiers using the majority voting rule: (1) Decision Tree, (2) Gaussian Naïve Bayes, and (3) Logistic Regression. Then compare the accuracy of the fused model with: (4)AdaBoost Ensemble with Decision Trees as the base learner, and (5) Random Forests

Project requirements: Well-written report and compatible code

About the data: 57 features constitute the number of times a particular word or character occurred in an email message with a total of 4601 instances. This data has to be classified into 2 classes viz ‘ham’ or ‘spam’.



	make	address	all	3d	our	over	remove	internet	order	mail	...	semicol	paren	bracket	bang	dollar	pound	cap_avg	cap_long	cap_total	Class
0	0.00	0.00	0.29	0.0	0.00	0.00	0.00	0.00	0.00	0.00	...	0.000	0.178	0.0	0.044	0.000	0.00	1.666	10	180	ham
1	0.46	0.00	0.00	0.0	0.00	0.00	0.00	0.00	0.00	0.00	...	0.000	0.125	0.0	0.000	0.000	0.00	1.510	10	74	ham
2	0.00	0.00	0.00	0.0	0.00	0.00	0.00	0.00	0.00	0.00	...	0.000	0.000	0.0	0.000	0.000	0.00	1.718	11	55	ham
3	0.33	0.44	0.37	0.0	0.14	0.11	0.00	0.07	0.97	1.16	...	0.006	0.159	0.0	0.069	0.221	0.11	3.426	72	819	spam
4	0.00	2.08	0.00	0.0	3.12	0.00	1.04	0.00	0.00	0.00	...	0.000	0.000	0.0	0.263	0.000	0.00	1.428	4	20	spam
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
4596	0.00	0.00	0.53	0.0	0.00	0.53	0.00	0.00	0.00	0.53	...	0.000	0.101	0.0	0.000	0.000	0.00	1.857	16	52	ham
4597	0.00	0.00	0.00	0.0	0.00	0.00	0.00	0.00	0.00	0.00	...	0.000	0.443	0.0	0.221	0.665	0.00	3.812	15	61	spam
4598	0.00	0.00	0.00	0.0	0.00	0.00	0.00	0.00	0.00	0.00	...	0.000	0.000	0.0	0.000	0.000	0.00	1.000	1	3	ham
4599	0.00	0.00	0.00	0.0	0.00	0.00	0.00	0.00	0.00	0.00	...	0.000	0.218	0.0	0.218	0.000	0.00	1.687	10	27	ham
4600	0.13	0.26	0.52	0.0	0.26	0.00	0.13	0.00	0.00	0.39	...	0.000	0.000	0.0	0.366	0.000	0.04	7.138	149	1235	spam

4601 rows × 58 columns

Libraries Imported:

I have used the pandas and NumPy libraries for basic operations on the dataset. As stated in the previous project report, the warnings library is used to eliminate any kind of warnings while running a particular snippet of code. We have included 3 classifier models, a Decision tree classifier, a Gaussian Naive Bayes, and a Logistic Regression. We have incorporated Ensemble Learning into our project by fusing the 3 classifiers via the Majority voting technique. The metrics over which we will be evaluating the test data are the confusion matrix, accuracy score, and recall score per class.

Exploratory Data Analysis:

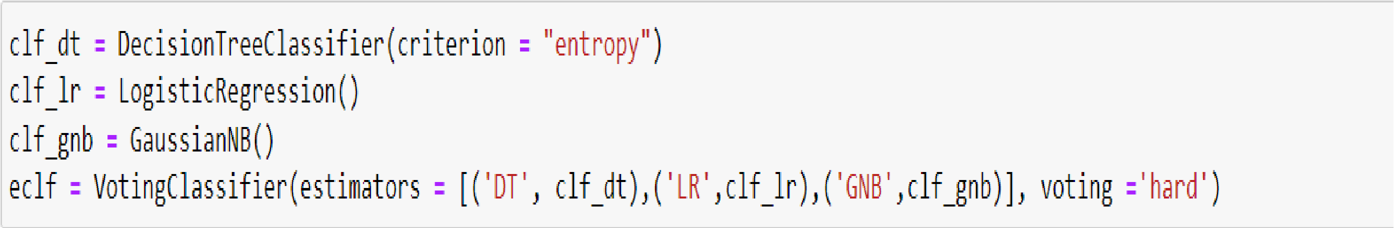
We have found out that the dataset has no null values in it with the help of isna() function. The describe() function is giving the count, mean, standard deviation, minimum value, maximum value, 25%, 50%, and 75% of each particular feature. The value_count() function gives the total number of ham and spam words or characters. Using the LabelEncoder library, we have labeled the class variables ham and spam as 0 and 1 respectively.

Splitting of dataset:

The dataset will be split into train and test data using the train_test_split function with a test size of 0.78 for 1000 training instances.

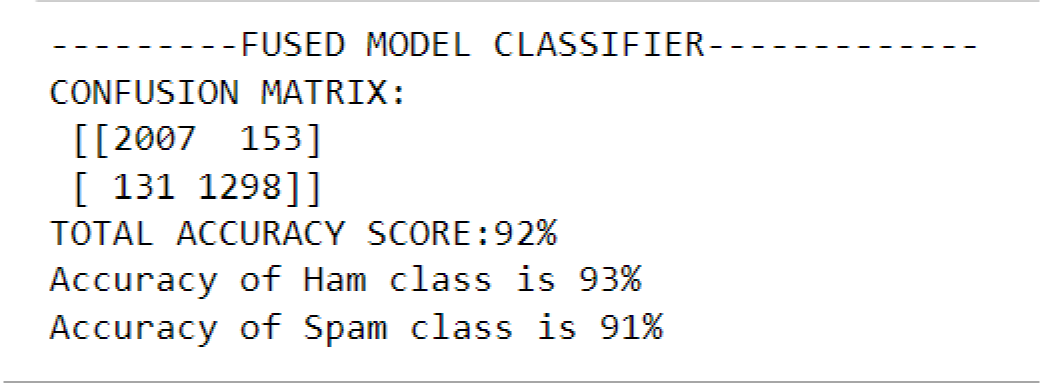
Fusing the 3 classifiers via Majoring voting technique:

Majority voting is an ensemble machine-learning model which combines different machine-learning models for better predictions and improves the performance of the model overall. In this project, we are ideally using the Decision Tree classifier, Logistic Regression, and the Gaussian Naive Bayes classifier.



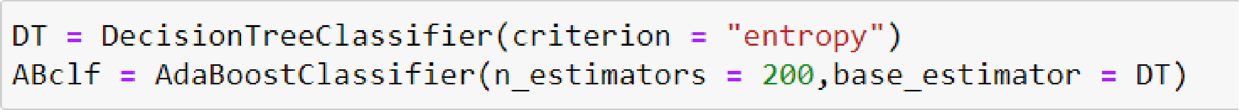
```
clf_dt = DecisionTreeClassifier(criterion = "entropy")
clf_lr = LogisticRegression()
clf_gnb = GaussianNB()
eclf = VotingClassifier(estimators = [('DT', clf_dt),('LR',clf_lr),('GNB',clf_gnb)], voting ='hard')
```

Then we found out the accuracy of the fused model and of the individual classes as well.

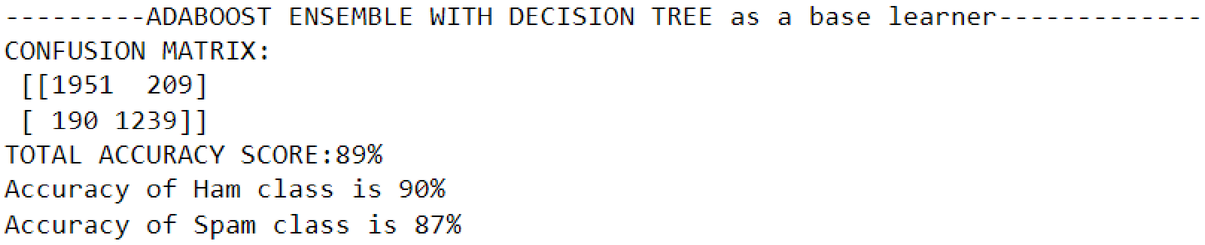


```
---------FUSED MODEL CLASSIFIER-------------
CONFUSION MATRIX:
 [[2007  153]
 [ 131 1298]]
TOTAL ACCURACY SCORE:92%
Accuracy of Ham class is 93%
Accuracy of Spam class is 91%
```

Now we need to determine the AdaBoost Ensemble's accuracy with the Decision Tree as the base estimator.



```
DT = DecisionTreeClassifier(criterion = "entropy")
ABclf = AdaBoostClassifier(n_estimators = 200,base_estimator = DT)
```

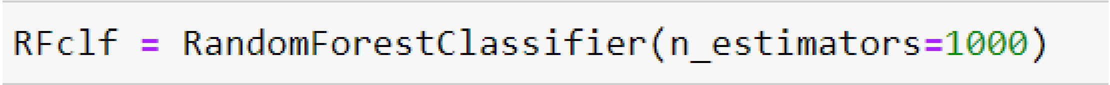


```
---------ADABOOST ENSEMBLE WITH DECISION TREE as a base learner-------------
CONFUSION MATRIX:
 [[1951  209]
 [ 190 1239]]
TOTAL ACCURACY SCORE:89%
Accuracy of Ham class is 90%
Accuracy of Spam class is 87%
```

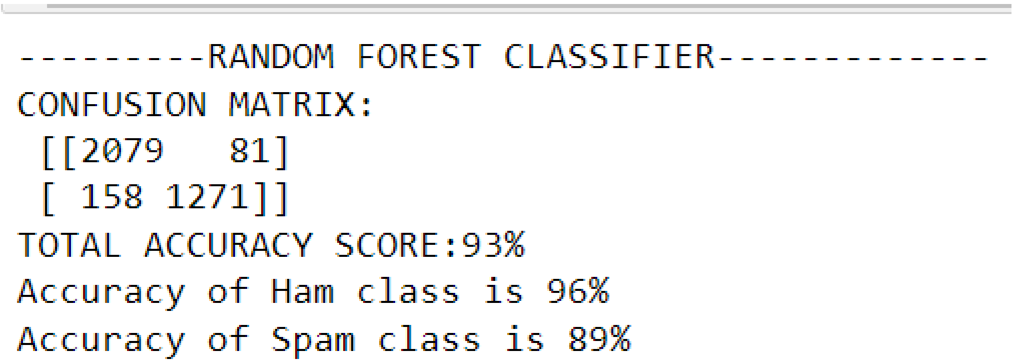
The comparison of the accuracies of the fused model and the AdaBoost Ensemble is constituted in the table below:

Models	Total Accuracy	ham	spam
Fused	92%	93%	91%
AdaBoost Ensemble with Decision Tree as a base estimator	89%	90%	87%

Now we further evaluate the accuracies for the Random Forest classifier with 1000 base learners.



```
RFclf = RandomForestClassifier(n_estimators=1000)
```



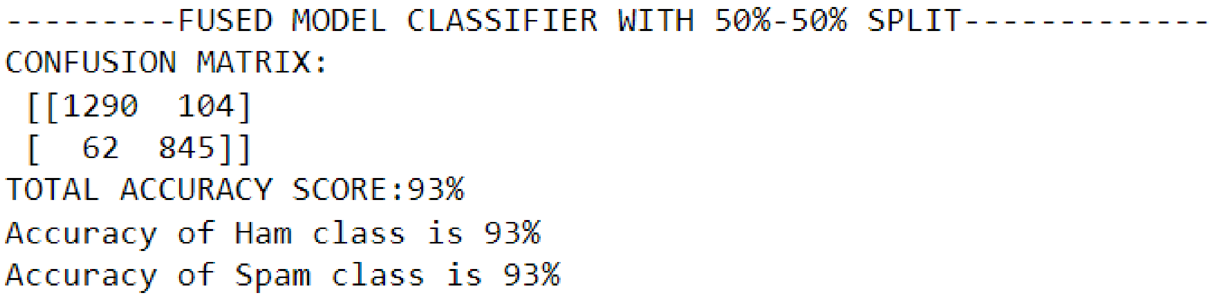
```
---------RANDOM FOREST CLASSIFIER-------------
CONFUSION MATRIX:
 [[2079   81]
 [ 158 1271]]
TOTAL ACCURACY SCORE:93%
Accuracy of Ham class is 96%
Accuracy of Spam class is 89%
```

The comparison of the accuracies of the fused model and the Random Forest classifier is constituted in the table below:

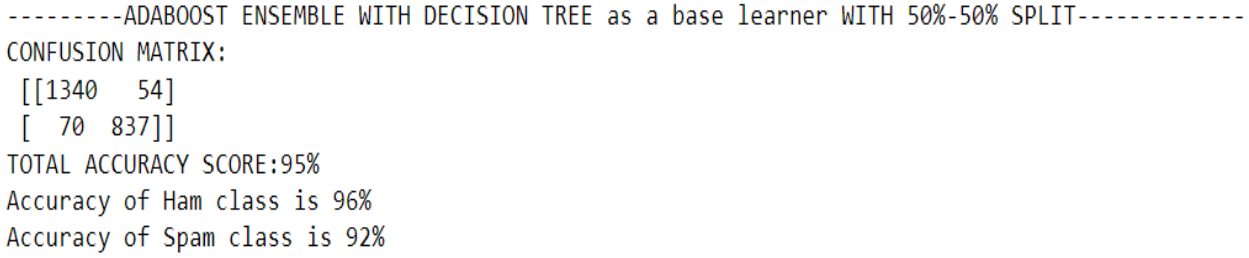
Models	Total Accuracy	ham	spam
Fused	92%	93%	91%
Random Forest classifier	93%	96%	89%

Impact of different training-testing splits on the accuracies of the models:

- When the training-testing splits are **50%-50%** then the accuracies of the fused model and the AdaBoost ensemble are varied as follows.



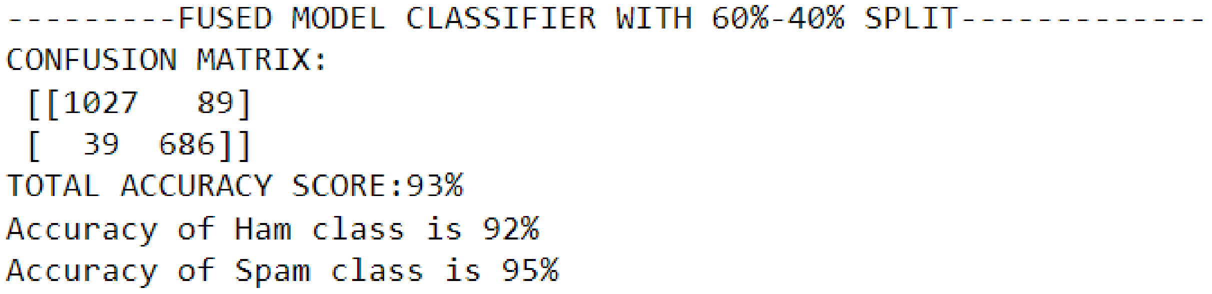
```
---------FUSED MODEL CLASSIFIER WITH 50%-50% SPLIT-------------
CONFUSION MATRIX:
 [[1290  104]
 [  62  845]]
TOTAL ACCURACY SCORE:93%
Accuracy of Ham class is 93%
Accuracy of Spam class is 93%
```



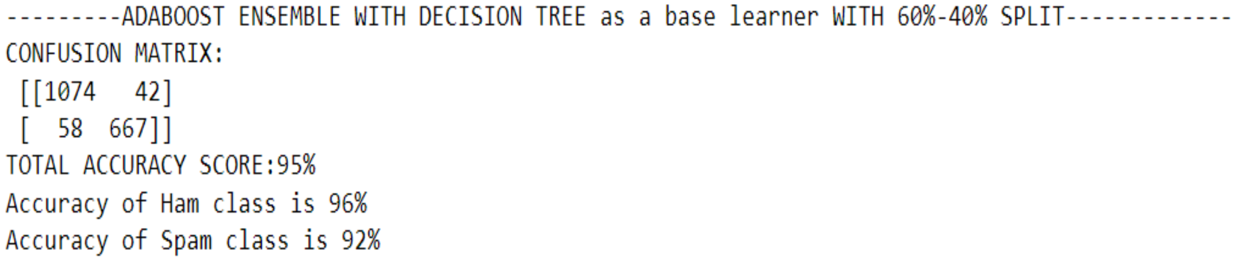
```
---------ADABOOST ENSEMBLE WITH DECISION TREE as a base learner WITH 50%-50% SPLIT------------
CONFUSION MATRIX:
 [[1340   54]
 [  70  837]]
TOTAL ACCURACY SCORE:95%
Accuracy of Ham class is 96%
Accuracy of Spam class is 92%
```

Models	Total Accuracy	ham	spam
Fused	93%	93%	93%
AdaBoost Ensemble	95%	96%	92%

- When the training-testing splits are **60%-40%** then the accuracies of the fused model and the AdaBoost ensemble are varied as follows.



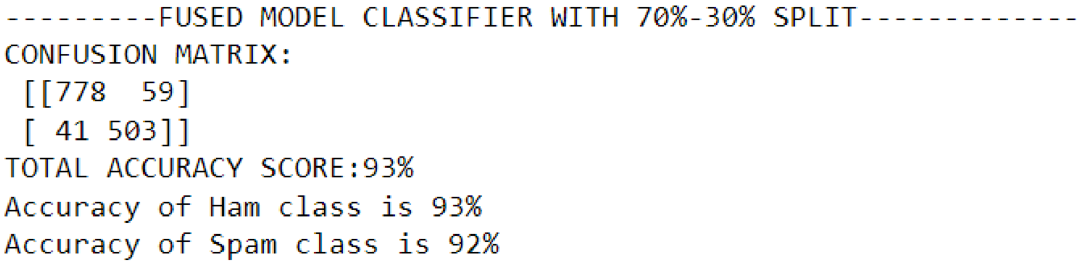
```
---------FUSED MODEL CLASSIFIER WITH 60%-40% SPLIT-------------
CONFUSION MATRIX:
 [[1027   89]
 [  39  686]]
TOTAL ACCURACY SCORE:93%
Accuracy of Ham class is 92%
Accuracy of Spam class is 95%
```



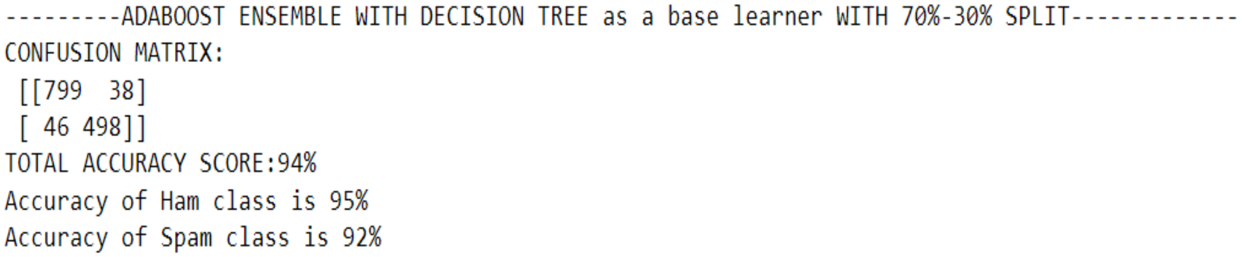
```
---------ADABOOST ENSEMBLE WITH DECISION TREE as a base learner WITH 60%-40% SPLIT------------
CONFUSION MATRIX:
 [[1074   42]
 [  58  667]]
TOTAL ACCURACY SCORE:95%
Accuracy of Ham class is 96%
Accuracy of Spam class is 92%
```

Models	Total Accuracy	ham	spam
Fused	93%	92%	95%
AdaBoost Ensemble	95%	96%	92%

- When the training-testing splits are 70%-30% then the accuracies of the fused model and the AdaBoost ensemble are varied as follows.



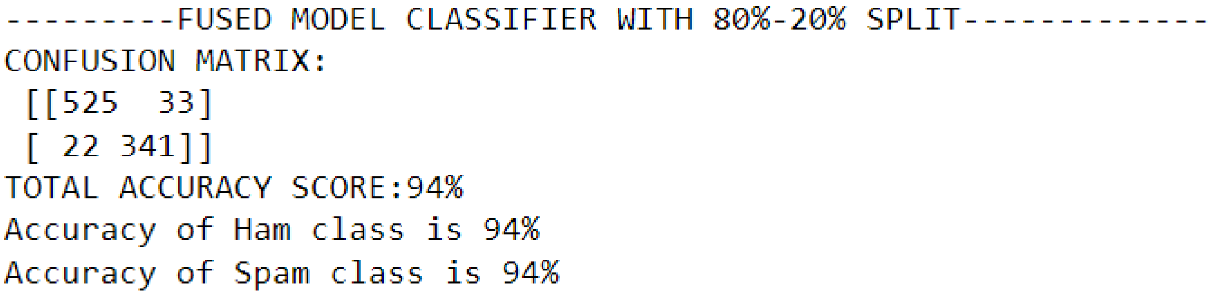
```
---------FUSED MODEL CLASSIFIER WITH 70%-30% SPLIT-------------
CONFUSION MATRIX:
 [[778  59]
 [ 41 503]]
TOTAL ACCURACY SCORE:93%
Accuracy of Ham class is 93%
Accuracy of Spam class is 92%
```



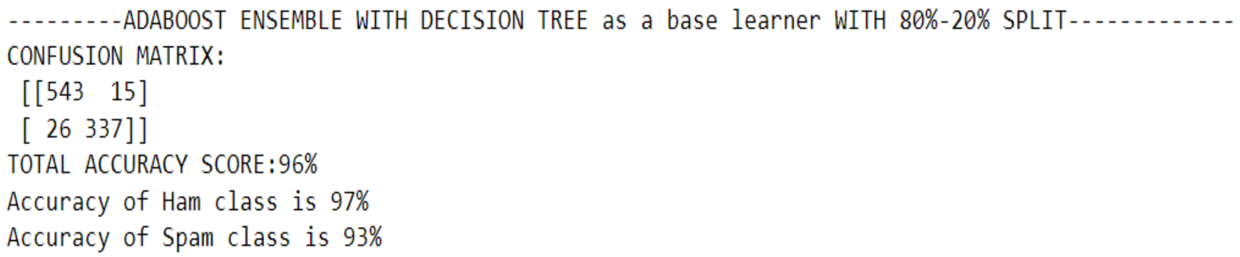
```
---------ADABOOST ENSEMBLE WITH DECISION TREE as a base learner WITH 70%-30% SPLIT------------
CONFUSION MATRIX:
 [[799  38]
 [ 46 498]]
TOTAL ACCURACY SCORE:94%
Accuracy of Ham class is 95%
Accuracy of Spam class is 92%
```

Models	Total Accuracy	ham	spam
Fused	93%	93%	92%
AdaBoost Ensemble	94%	95%	92%

- When the training-testing splits are 80%-20% then the accuracies of the fused model and the AdaBoost ensemble are varied as follows.



```
---------FUSED MODEL CLASSIFIER WITH 80%-20% SPLIT-------------
CONFUSION MATRIX:
 [[525  33]
 [ 22 341]]
TOTAL ACCURACY SCORE:94%
Accuracy of Ham class is 94%
Accuracy of Spam class is 94%
```



```
---------ADABOOST ENSEMBLE WITH DECISION TREE as a base learner WITH 80%-20% SPLIT------------
CONFUSION MATRIX:
 [[543  15]
 [ 26 337]]
TOTAL ACCURACY SCORE:96%
Accuracy of Ham class is 97%
Accuracy of Spam class is 93%
```

Models	Total Accuracy	ham	spam
Fused	94%	94%	94%
AdaBoost Ensemble	96%	97%	93%

Conclusion:

From this project, we determined that the Fused model gave better accuracies than the AdaBoost ensemble with Decision Tree as a base estimator. When it came to the Random Forest classifier, it offered better overall accuracies than the Fused model.

The impacts from all of the training-testing split viz 50-50%, 60-40%, 70-30%, 80-20% on the dataset also resulted in higher accuracies for the AdaBoost Ensemble with decision tree as a base learner than the Fused model, giving us clarity as to which model performed well. It is evident that when the dataset is split into 80-20% for training-testing, the accuracy is higher than the other splits.