Fake News Detection - Approach

Nityanandan Paramasivam

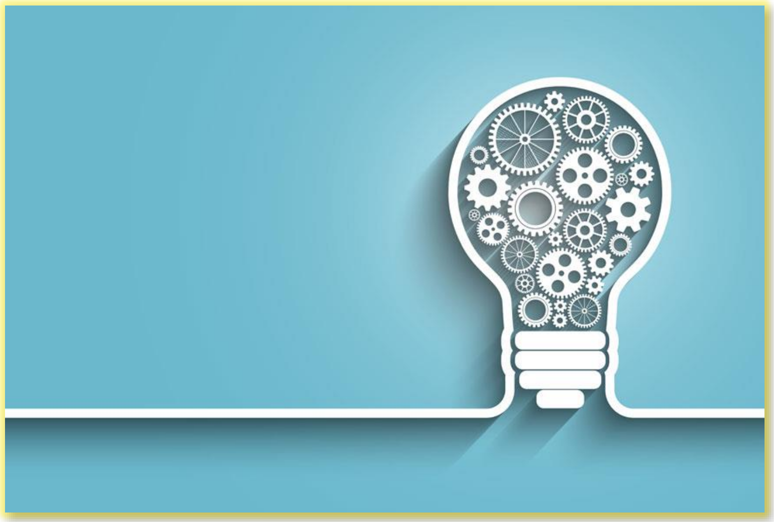
THE OBJECTIVE

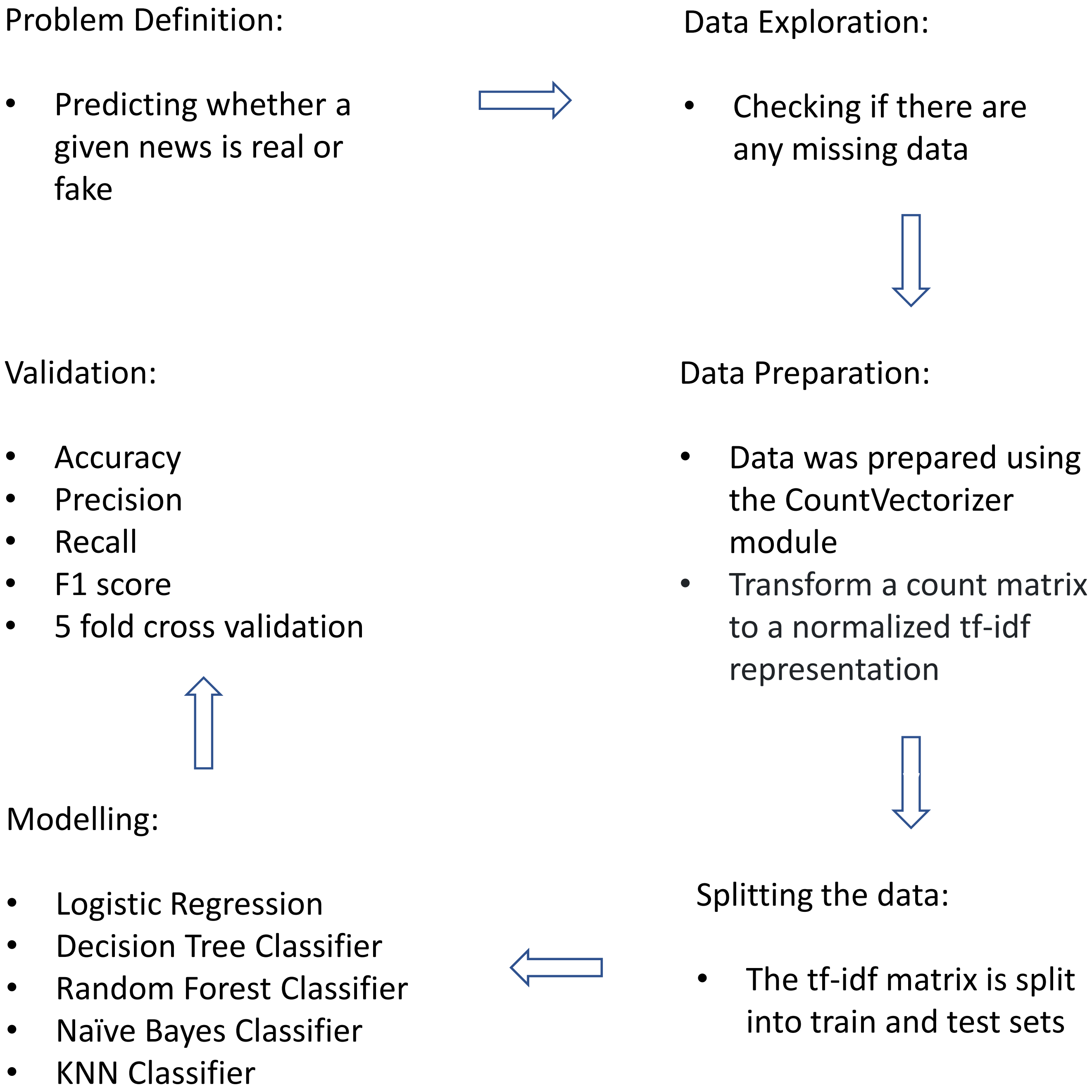
Develop a model to detect if a given news article is a fake one or a legitimate one.

Challenge: The data contains only text data and there are no numerical data except id, which is only used to identify records

I have handled the data here using text analytics modules from Scikit-Learn. It will actually convert that into numerical data and we can fit them into models

APPROACH





Problem Definition:

- Predicting whether a given news is real or fake

Data Exploration:

- Checking if there are any missing data

Data Preparation:

- Data was prepared using the CountVectorizer module
- Transform a count matrix to a normalized tf-idf representation

Splitting the data:

- The tf-idf matrix is split into train and test sets

Modelling:

- Logistic Regression
- Decision Tree Classifier
- Random Forest Classifier
- Naïve Bayes Classifier
- KNN Classifier

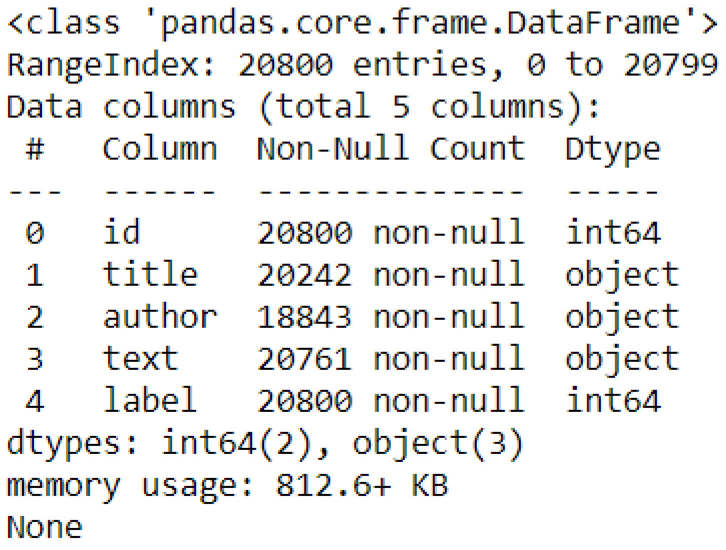
Validation:

- Accuracy
- Precision
- Recall
- F1 score
- 5 fold cross validation

DATA EXPLORATION

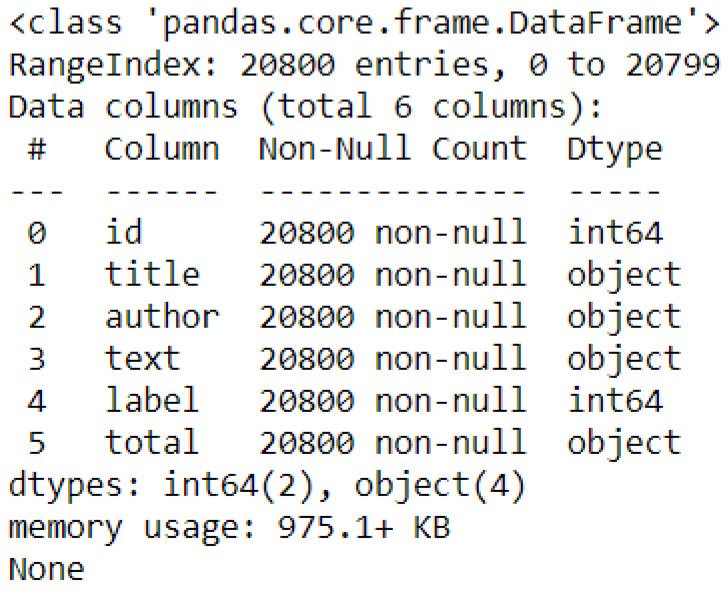
- There were a lot of missing data in the dataset
- They are all unique news, headlines and names of authors
- These cannot be imputed
- So I filled all of them with empty values

Before Filling:



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 20800 entries, 0 to 20799
Data columns (total 5 columns):
 #   Column  Non-Null Count  Dtype 
---  ------  --------------  ----- 
 0   id      20800 non-null  int64 
 1   title   20242 non-null  object
 2   author  18843 non-null  object
 3   text    20761 non-null  object
 4   label   20800 non-null  int64 
dtypes: int64(2), object(3)
memory usage: 812.6+ KB
None
```

After Filling and adding a new feature:

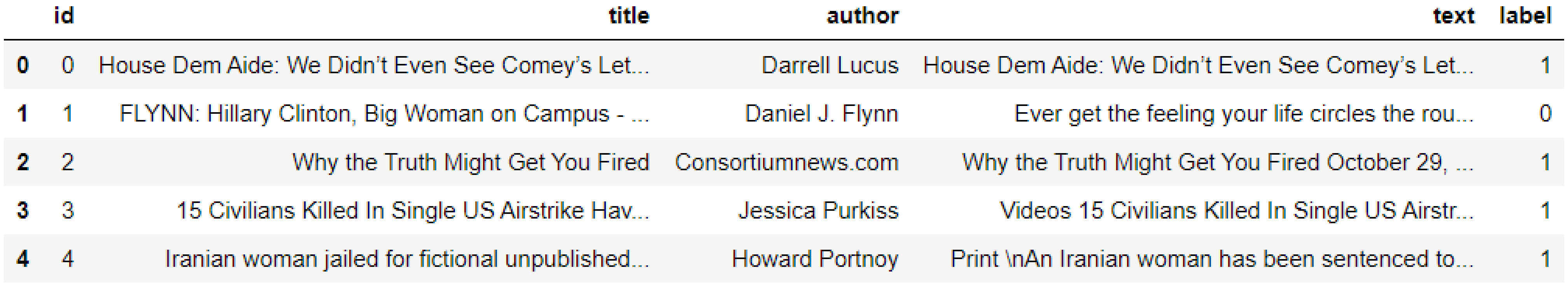


```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 20800 entries, 0 to 20799
Data columns (total 6 columns):
 #   Column  Non-Null Count  Dtype 
---  ------  --------------  ----- 
 0   id      20800 non-null  int64 
 1   title   20800 non-null  object
 2   author  20800 non-null  object
 3   text    20800 non-null  object
 4   label   20800 non-null  int64 
 5   total   20800 non-null  object
dtypes: int64(2), object(4)
memory usage: 975.1+ KB
None
```

DATA PREPARATION

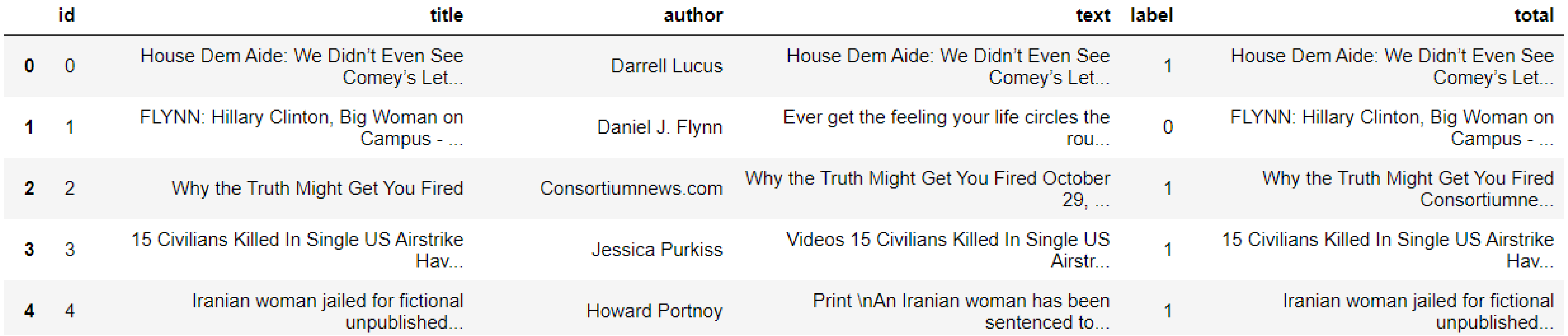
- We create a new column in the dataset to add all the previous columns.
- This adds a new feature which contains the title, author and the news
- Now we use CountVectorizer from Scikit-learn to filter and tokenize the stopwords and also help in text pre-processing.
- This builds a dictionary of features and transforms documents to feature vectors.
- Then we use the TfidfTransformer from Scikit-learn to transform a count matrix to a normalized tf-idf representation.

Data before adding new feature



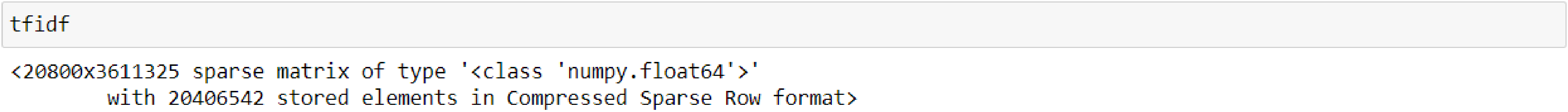
	id	title	author	text	label
0	0	House Dem Aide: We Didn’t Even See Comey’s Let...	Darrell Lucus	House Dem Aide: We Didn’t Even See Comey’s Let...	1
1	1	FLYNN: Hillary Clinton, Big Woman on Campus - ...	Daniel J. Flynn	Ever get the feeling your life circles the rou...	0
2	2	Why the Truth Might Get You Fired	Consortiumnews.com	Why the Truth Might Get You Fired October 29, ...	1
3	3	15 Civilians Killed In Single US Airstrike Hav...	Jessica Purkiss	Videos 15 Civilians Killed In Single US Airstr...	1
4	4	Iranian woman jailed for fictional unpublished...	Howard Portnoy	Print \nAn Iranian woman has been sentenced to...	1

Data after adding new feature



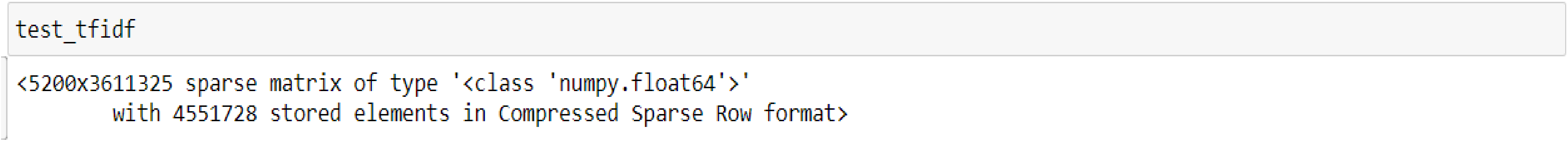
	id	title	author	text	label	total
0	0	House Dem Aide: We Didn’t Even See Comey’s Let...	Darrell Lucus	House Dem Aide: We Didn’t Even See Comey’s Let...	1	House Dem Aide: We Didn’t Even See Comey’s Let...
1	1	FLYNN: Hillary Clinton, Big Woman on Campus - ...	Daniel J. Flynn	Ever get the feeling your life circles the rou...	0	FLYNN: Hillary Clinton, Big Woman on Campus - ...
2	2	Why the Truth Might Get You Fired	Consortiumnews.com	Why the Truth Might Get You Fired October 29, ...	1	Why the Truth Might Get You Fired Consortiumne...
3	3	15 Civilians Killed In Single US Airstrike Hav...	Jessica Purkiss	Videos 15 Civilians Killed In Single US Airstr...	1	15 Civilians Killed In Single US Airstrike Hav...
4	4	Iranian woman jailed for fictional unpublished...	Howard Portnoy	Print \nAn Iranian woman has been sentenced to...	1	Iranian woman jailed for fictional unpublished...

After fitting data into TfIdfTransformer:



```
tfidf
```

```
<20800x3611325 sparse matrix of type '<class 'numpy.float64'>'
	with 20406542 stored elements in Compressed Sparse Row format>
```

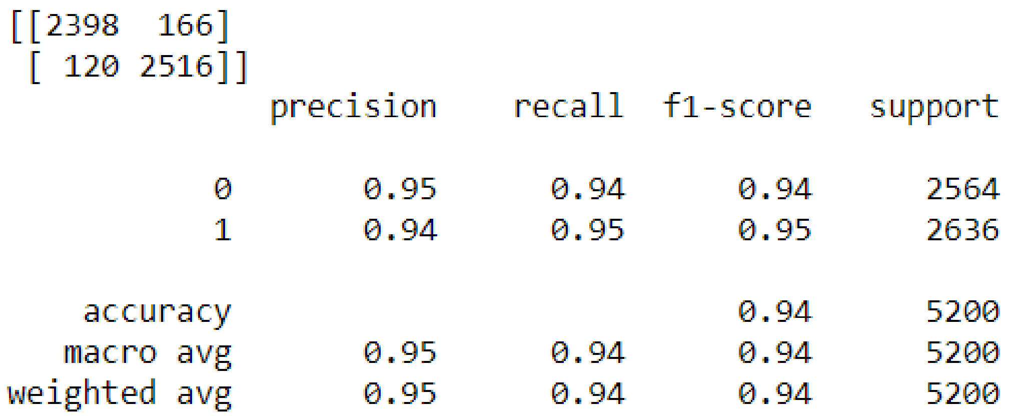


```
test_tfidf
```

```
<5200x3611325 sparse matrix of type '<class 'numpy.float64'>'
	with 4551728 stored elements in Compressed Sparse Row format>
```

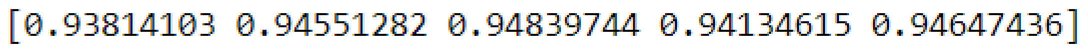
MODELLING AND VALIDATION

- Logistic Regression:



```
[[2398  166]
 [ 120 2516]]
              precision    recall  f1-score   support

           0       0.95      0.94      0.94      2564
           1       0.94      0.95      0.95      2636

    accuracy                           0.94      5200
   macro avg       0.95      0.94      0.94      5200
weighted avg       0.95      0.94      0.94      5200
```

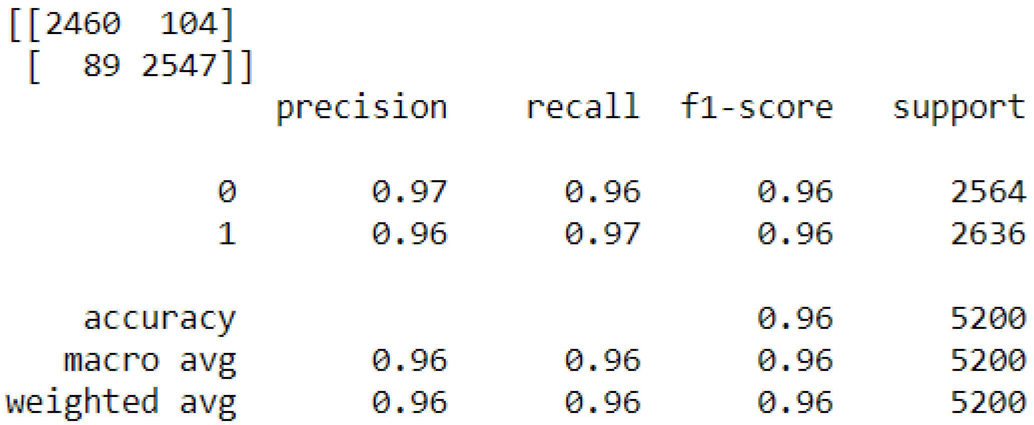
- Cross Validation:



```
[0.93814103 0.94551282 0.94839744 0.94134615 0.94647436]
```

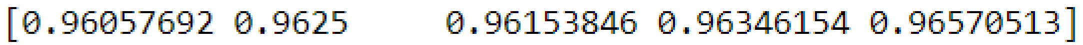
MODELLING AND VALIDATION

- Decision Tree Classifier:



```
[[2460  104]
 [  89 2547]]
              precision    recall  f1-score   support

           0       0.97      0.96      0.96      2564
           1       0.96      0.97      0.96      2636

    accuracy                           0.96      5200
   macro avg       0.96      0.96      0.96      5200
weighted avg       0.96      0.96      0.96      5200
```

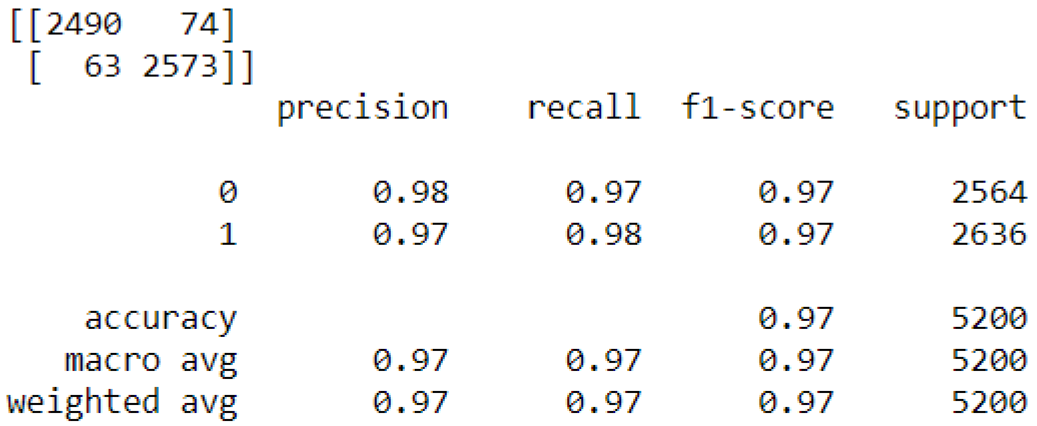
- Cross Validation:



```
[0.96057692 0.9625     0.96153846 0.96346154 0.96570513]
```

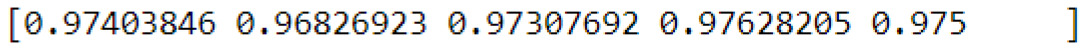
MODELLING AND VALIDATION

- Random Forest Classifier:



```
[[2490   74]
 [  63 2573]]
              precision    recall  f1-score   support

           0       0.98      0.97      0.97      2564
           1       0.97      0.98      0.97      2636

    accuracy                           0.97      5200
   macro avg       0.97      0.97      0.97      5200
weighted avg       0.97      0.97      0.97      5200
```

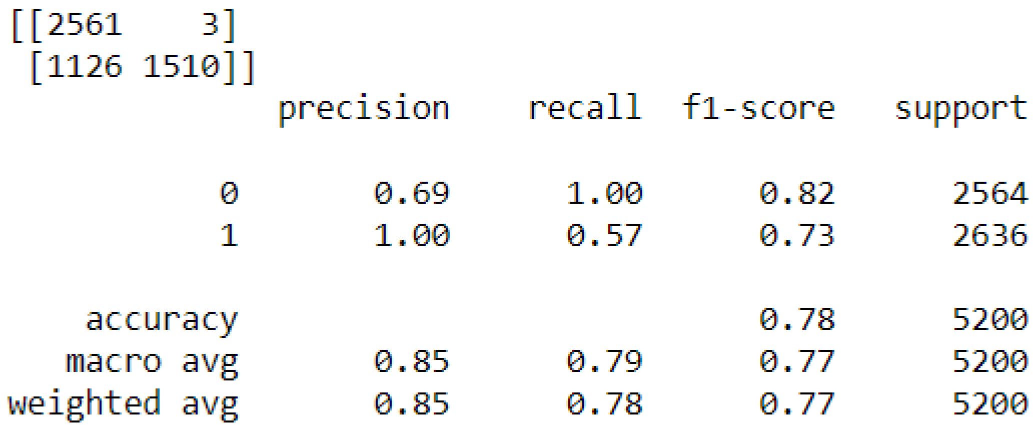
- Cross Validation:



```
[0.97403846 0.96826923 0.97307692 0.97628205 0.975     ]
```

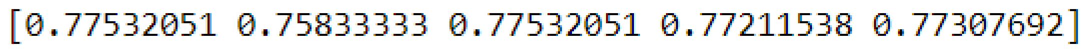
MODELLING AND VALIDATION

- Naïve Bayes Classifier:



```
[[2561    3]
 [1126 1510]]
              precision    recall  f1-score   support

           0       0.69      1.00      0.82      2564
           1       1.00      0.57      0.73      2636

    accuracy                           0.78      5200
   macro avg       0.85      0.79      0.77      5200
weighted avg       0.85      0.78      0.77      5200
```

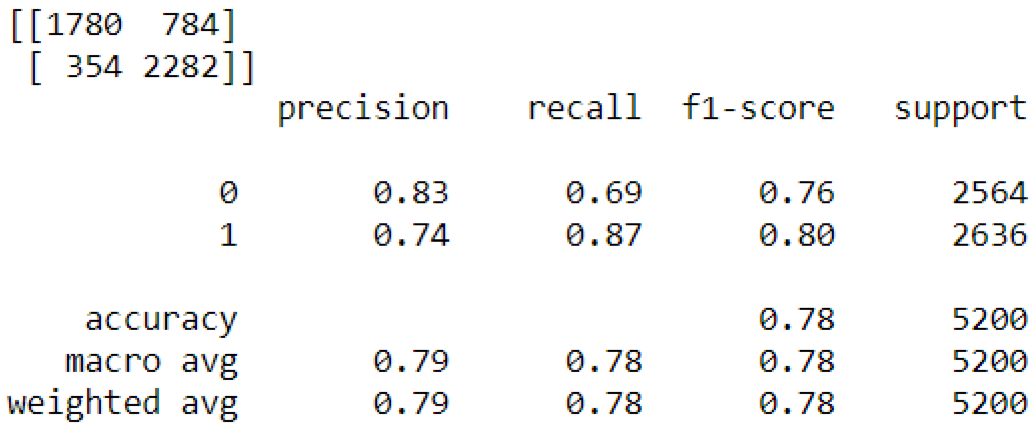
- Cross Validation:



```
[0.77532051 0.75833333 0.77532051 0.77211538 0.77307692]
```

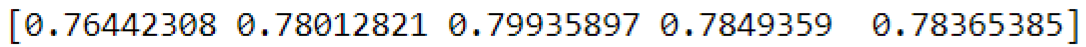
MODELLING AND VALIDATION

- K Nearest Neighbor Classifier:



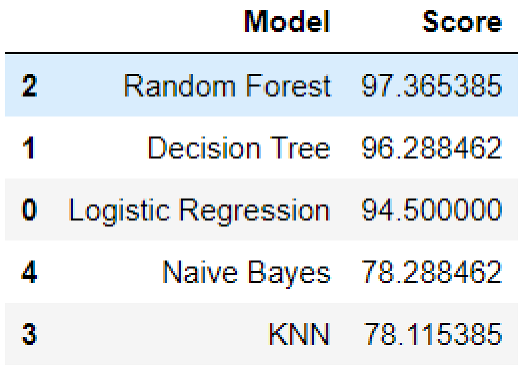
```
[[1780  784]
 [ 354 2282]]
              precision    recall  f1-score   support

           0       0.83      0.69      0.76      2564
           1       0.74      0.87      0.80      2636

    accuracy                           0.78      5200
   macro avg       0.79      0.78      0.78      5200
weighted avg       0.79      0.78      0.78      5200
```

- Cross Validation:



```
[0.76442308 0.78012821 0.79935897 0.7849359  0.78365385]
```

SORTING MODELS AS PER THE ACCURACY



	Model	Score
2	Random Forest	97.365385
1	Decision Tree	96.288462
0	Logistic Regression	94.500000
4	Naive Bayes	78.288462
3	KNN	78.115385

- We can see that the accuracy of Random Forest, Decision Tree and Logistic Regression is better than Naïve Bayes and KNN.

- This is because Random forest is better at handling high dimensional spaces and large training samples than KNN and Naïve bayes.

- Decision Tree and Logistic regression are better here compared to KNN because there is no distance metric. KNN determines neighborhoods, so there must be a distance metric. This implies that all the features must be numeric.

THANK YOU