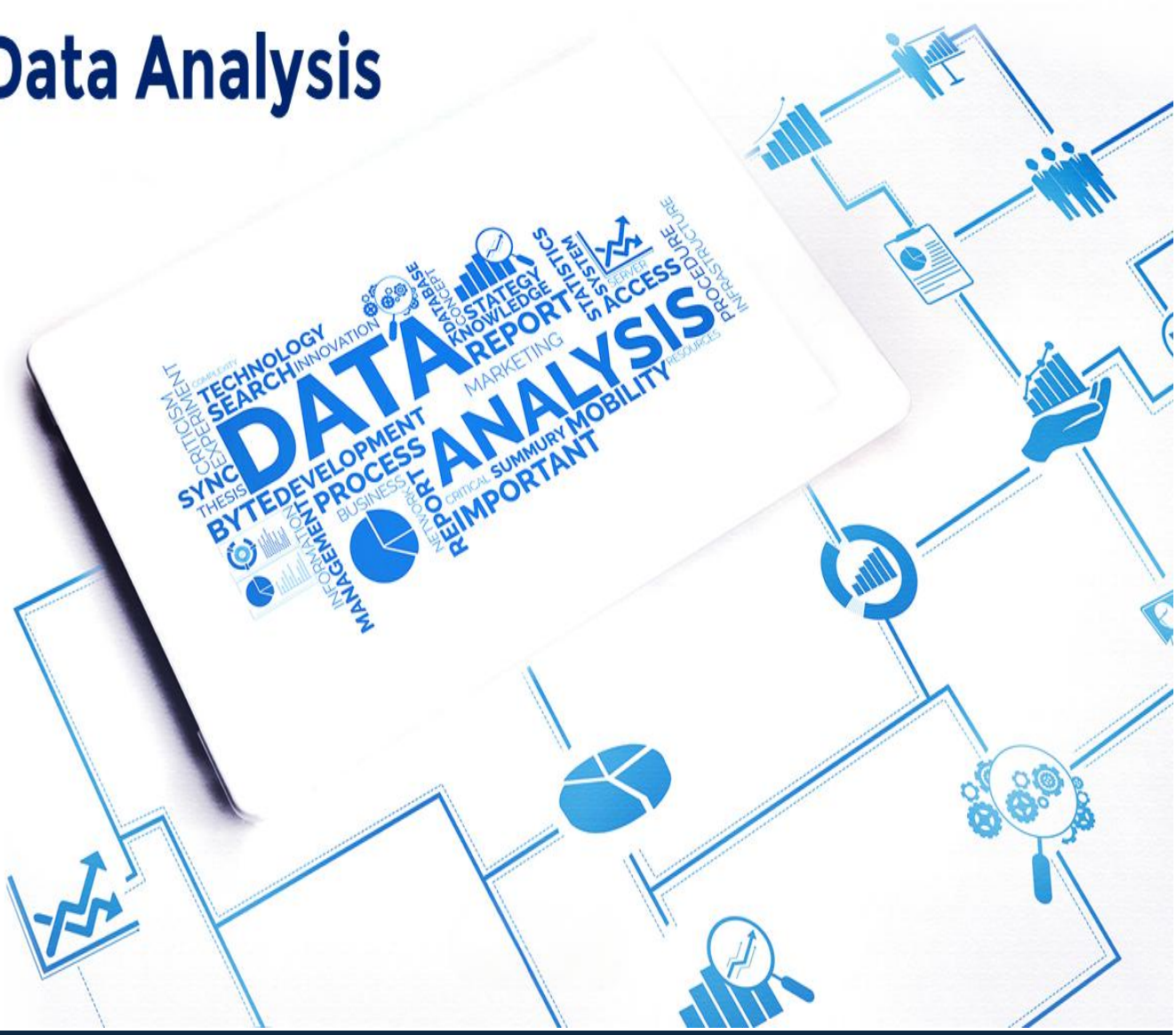
Data Analysis



Group Data Analysis Report

DURHAM COLLEGE
SUCCESS MATTERS

Table of Contents

Abstract

In the era of digital technology, fake news travels fast and it is becoming hard to distinguish between the two, real and fake news. The rising influence of fake information yielded to our client the need to find a machine learning solution that will be able to detect fake news articles basing on the content of those articles. The objective of the project is to build and test classification models using the Fake and Real News Dataset available on Kaggle and build a model that will identify news as fake or real with high accuracy. Due to natural language processing (NLP) techniques, the data have been cleaned, prepared, and analyzed, with useful characteristics being extracted, including the length of the article and the term importance. Three machine learning algorithms were trained and compared: Logistic Regression, Naive Bayes, and Random Forest, using such important indicators as Accuracy, Precision, Recall, and F1-Score. We were aiming to have a reliable model with a reliable percentage F1-score of at least 85. The working model, the one that has been chosen, will be built into an API to verify fake news in real-time and shown to the representatives of the stakeholders to implement the model currently.

Rational Statement

In today's digital age it is very difficult to differentiate between fake and real news in the media articles. Our clients want to develop a machine learning model to predict if the news is "Fake" or "Real" based on the content of the news. This model will help us in identifying fake news so that it can be fact-checked and help to mitigate the impact caused by fake news.

We will use "fake-and-real-news-dataset" from Kaggle and implement classification algorithms which will help us to accurately classify news articles as real or fake. We will use metrices like Accuracy, Precision, Recall, and F1-Score to evaluate model effectiveness. Our model's success will be defined by F1-score of 85% or above.

Data Description

Our data set consist of two CSV spreadsheets with the data of True news and Fake news. If we combine them, they include 44,898 entries across 5 variables. These variables are described as below.

Variable	Description
Title	This is the headline or title given to the news article. It gives you a quick idea of the news content, often holding keywords helpful for classification.
Text	Here's the full content/body of the news article. This is the most informative aspect used for figuring out if the news is authentic or not.
Subject	This shows a wide topic category the article belongs to (e.g., News, politics News). It might give a useful background, helping with classification.
Date	This is the publication date shown for the article. While identifying time-related patterns is possible, the date isn't always key for making classifications.
Label	This is the binary target variable we're aiming for: 0 stands for Fake news (coming from Fake.csv) and 1 stands for Real news (taken from True.csv). It's what our machine learning model needs to learn how to predict.

Data Preparation

We performed data preparation to apply it for a natural language processing (NLP) task through the following main stages:

1. Missing Value Treatment:

   Title and any text data without values were imputed with blank strings because we did not want to lose valuable records and still protect the structure of the article in the process. This prevented loss of data and consistency in the whole data.

2. Label Encoding:

   A column to state the binary label was included with a fake news as 0 labeling and a real news as 1 labeling. This made controlled classification modelling possible.

3. Text Consolidation:

   The column title and the column text were combined into one column of contents such that it captured the complete context of every article which was to be represented fully as a feature when vectorized.

4. Text Cleaning:

   Articles were changed to small letters. We dropped URLs, punctuations, special characters and non-alphabetic tokens as they added noise to our data and there would be no uniformity in the output format.

5. Feature Engineering Content Length:

   Another new column content length was generated by the counting of words of every cleaned article. This was applied to the set of exploratory data analysis to find out stylistic distinctions between fake and real news for example fake news is generally shorter.

6. Visual Analysis:

   We created histograms, boxplots, count plots to check the balance of the classes, and the length of the article's distribution. These graphics assisted in ensuring the data is balanced and obtained useful trends that were worthy to be modelled.

7. TF-IDF Vectorization:

   TF-IDF was used to convert text to a numerical feature that would represent the significance of text in a document. The noises and dimensionality were reduced by removing stop words and eliminating excessively repetitive terms.

All the above steps were done diligently because we wanted our data not just to be standardized, clean, but also add some value to it in the form of the relevant patterns that can be applied in the process of identifying fake news.

Exploratory Data Analysis (EDA)

We conducted thorough data analysis for the purpose of finding patterns and relations in the data, especially between the text of articles and chances of their being fake or real. Below are the analysis we have done on our data.

1. Class Distribution Analysis
   - A countplot has been used to compare the frequency of fake (0) and real (1) news articles. The classes proved to be relatively balanced, there was no significant resampling and thus it enhances generalization of the model.
2. Text Length Exploration
   - The explorations of the text length were conducted to determine how long web-based text should be presented. A new feature content_length (word count) was defined to perform the analysis of the verbosity of articles.

Boxplot and histogram depict that real news items are longer and mainly diverse in their length in comparison with false ones. This showed that the length of articles can be a significant attribute to classification.

3. Visualizations used
    - **Histograms** - It was shown that fake news is characterised by its shorter length with histograms demonstrating most common lengths per class.
    - **Boxplots -** The boxplots provided an overview of the range and distribution of the length of the contents with a finding that fake articles are concentrated in a narrower range than real news.
    - **Bar charts** - Bar charts will give an overview of the number of classes and optionally the top 10 most frequently occurring news subjects.

These observations substantiated our main assumptions about the text layout and helped in designing our model, such as the significance of the length of an article, patterns of vocabulary, and the ratio of labels between classes.

Data Analysis Solution

We used three different model to classify whether news articles are Fake or True. Three models and its results are as below.

1. **Logistic Regression**

The Logistic Regression model provided an excellent 98% accuracy with weighted F1-score of 0.98 which indicates balanced performance across both classes. It had an accuracy of 0.99 on fake news and 0.98 on real news and a recall of 0.98 on fake and 0.99 on real which means that it was able to accurately identify both the types of news. As seen in the confusion matrix, the model had only 137 misclassifications when the total number of predictions is 8,980, making it extremely reliable. Also, Logistic Regression has a clear interpretability, and it can be used to easily determine the features or words that contribute to the label of fake news the most and is thus useful in both performance and explaining ability.

2. **Naive Bayes**

Naive Bayes model provided a decent overall performance of 94 percent inaccuracy and a weighted F1-score of 0.94 even though the performance is not as high as Logistic Regression. It upheld same accuracy of 0.94 in fake and real news with recall in fake as 0.94 and in real as 0.93. The confusion matrix showed that 565 samples were misclassified, which is largely explained by the duplication of the vocabulary that fake and real articles utilized, and which could lead to mixing up the probabilistic assumptions

of the model. However, the fact that Naive Bayes is very fast and uses a minimal number of computations makes it a good option when the detection is supposed to be in real-time or used on a very large scale and efficiency is the main point.

3. **Random Forest Classifier**

When comparing the performance of the models, the Random Forest model outperform the rest as it installed an 99% accuracy rate against a weighted F1-scores of 0.99. It showed excellent accuracy and recall of 0.99 on the fake and real articles, thus providing a good indication that it can differentiate between the two classes accurately. Random Forest successfully extracted non-linear, complicated conjugations in the text with only 92 misclassifications on the total amount of data. Nonetheless, the superior level of generalization does not make the model as interpretable as the Logistic Regression, which can be a problematic feature in explaining its decisions to stakeholders or situations where the transparency of models is key.

We are including the results of our model and ROC curve of them in the appendix.

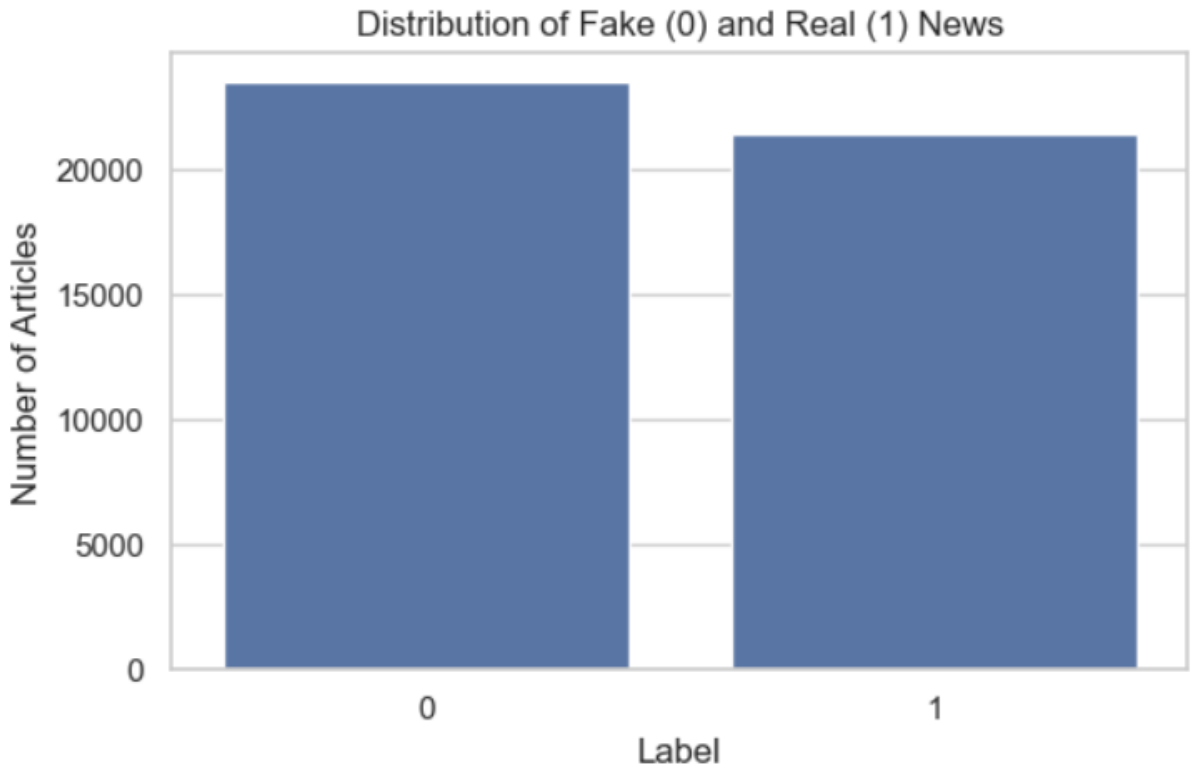
Conclusion

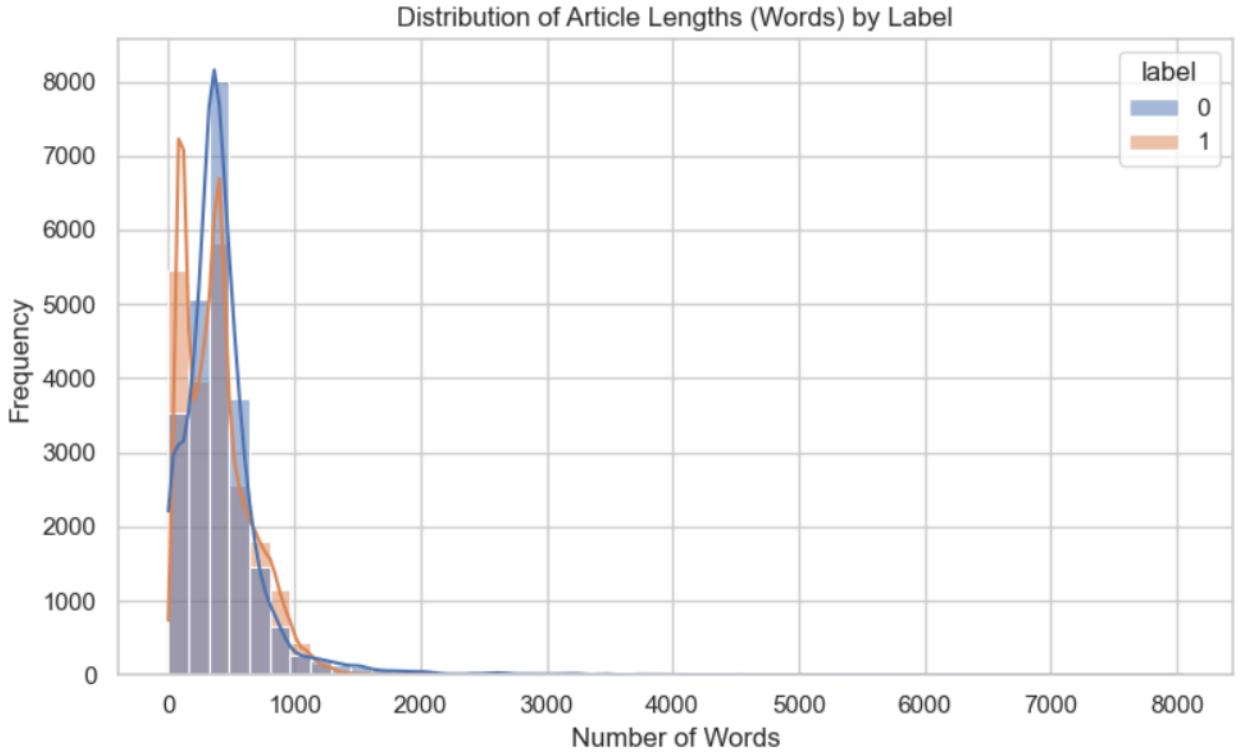
We have created three models which are Logistic Regression, Naive Bayes, and Random Forest which showed good categorization performance, with Random Forest showing the best result in terms of accuracy 99% and the lowest misclassification rate. Logistic Regression was of great use in terms of interpretability, which is useful in terms of stakeholder communication and transparency of a model, whereas Naive Bayes was significantly more efficient, which would be useful in real-time applications and large-scale usage. It was also found that the length of the text in an article and the vocabulary used was also influential in determining the fake news and the real text.

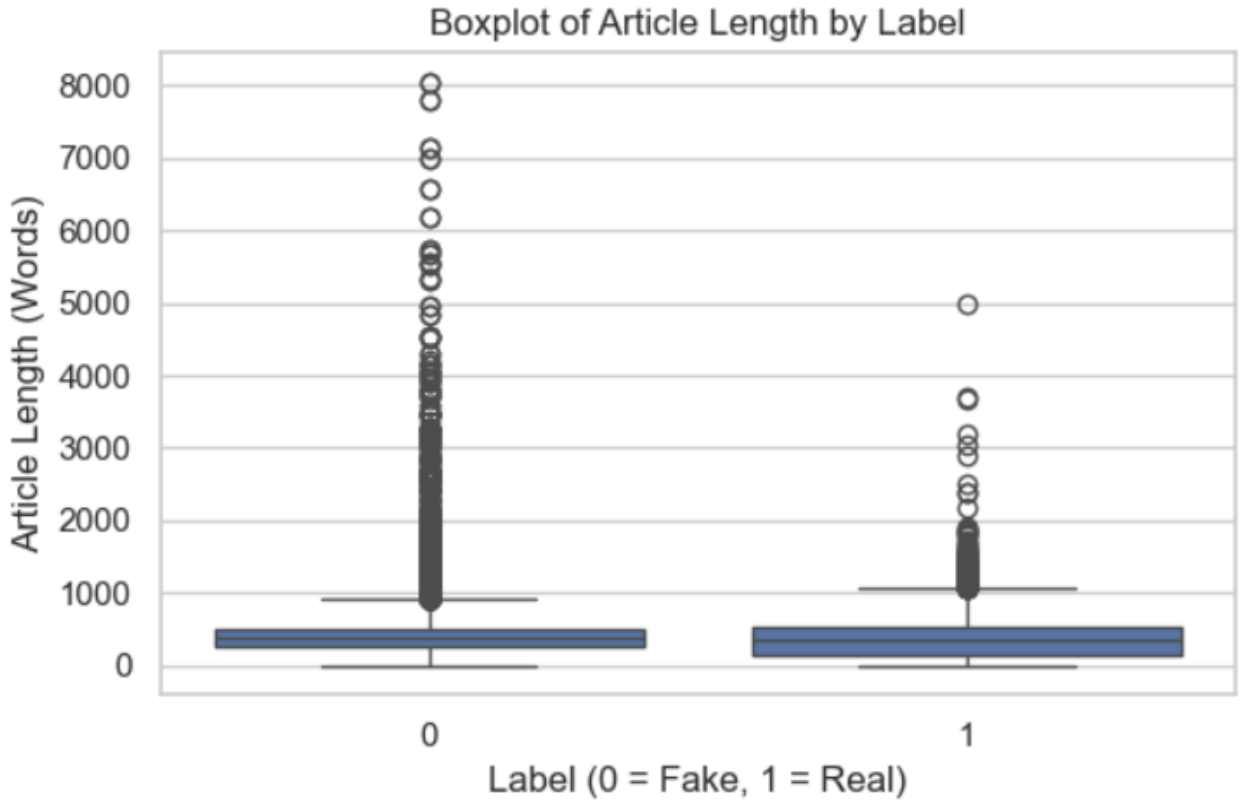
We will decide one of the models to be integrated through an API and on account of accuracy, computational efficiency, and explainability. This API will be presented in the final stakeholder presentation where its prospects of integrating easily in platforms that need them to detect fake news on a real-time basis will come out.

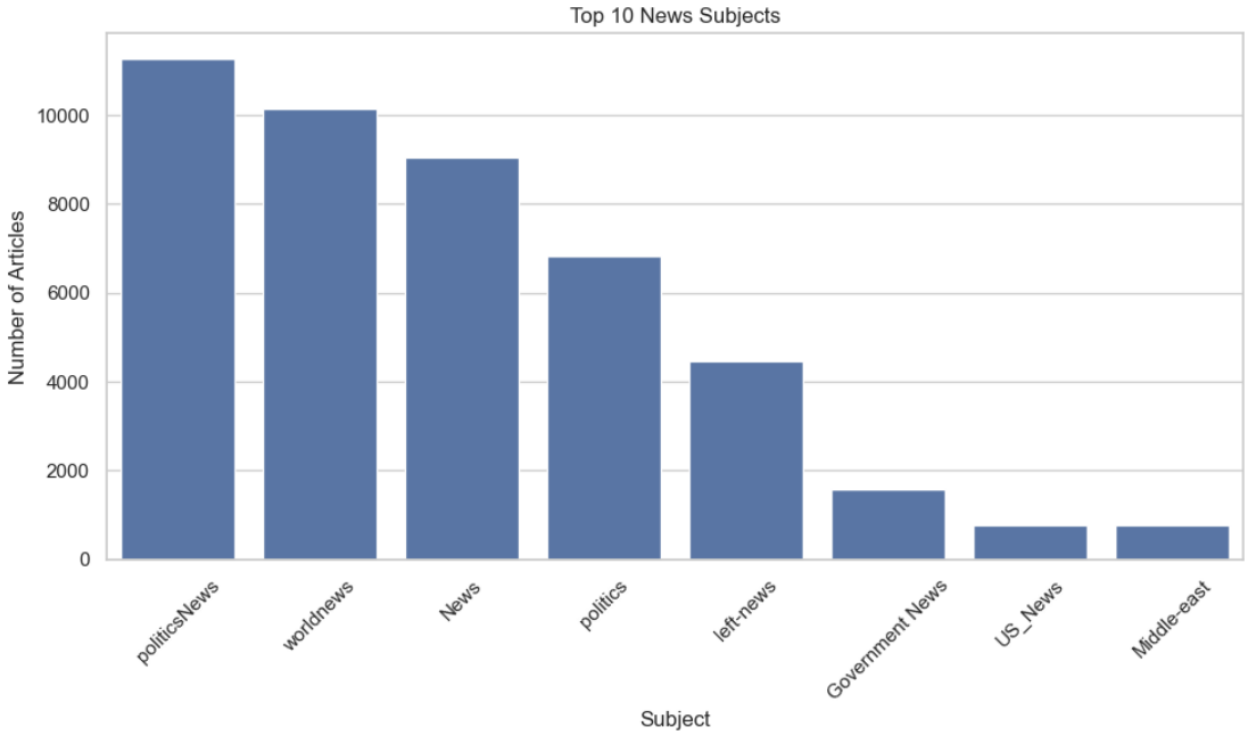
Appendix

Charts




Distribution of Article Lengths (Words) by Label
label
0
1
Frequency
Number of Words
0
1000
2000
3000
4000
5000
6000
7000
8000


Boxplot of Article Length by Label
Article Length (Words)
Label (0 = Fake, 1 = Real)
0
1000
2000
3000
4000
5000
6000
7000
8000



Performance metrics for models

```
Logistic Regression Results:

[[4621   75]
 [  62 4222]]
              precision    recall  f1-score   support

           0       0.99      0.98      0.99      4696
           1       0.98      0.99      0.98      4284

    accuracy                           0.98      8980
   macro avg       0.98      0.98      0.98      8980
weighted avg       0.98      0.98      0.98      8980
```

```
Naive Bayes Results:

[[4405  291]
 [ 274 4010]]
              precision    recall  f1-score   support

           0       0.94      0.94      0.94      4696
           1       0.93      0.94      0.93      4284

    accuracy                           0.94      8980
   macro avg       0.94      0.94      0.94      8980
weighted avg       0.94      0.94      0.94      8980
```

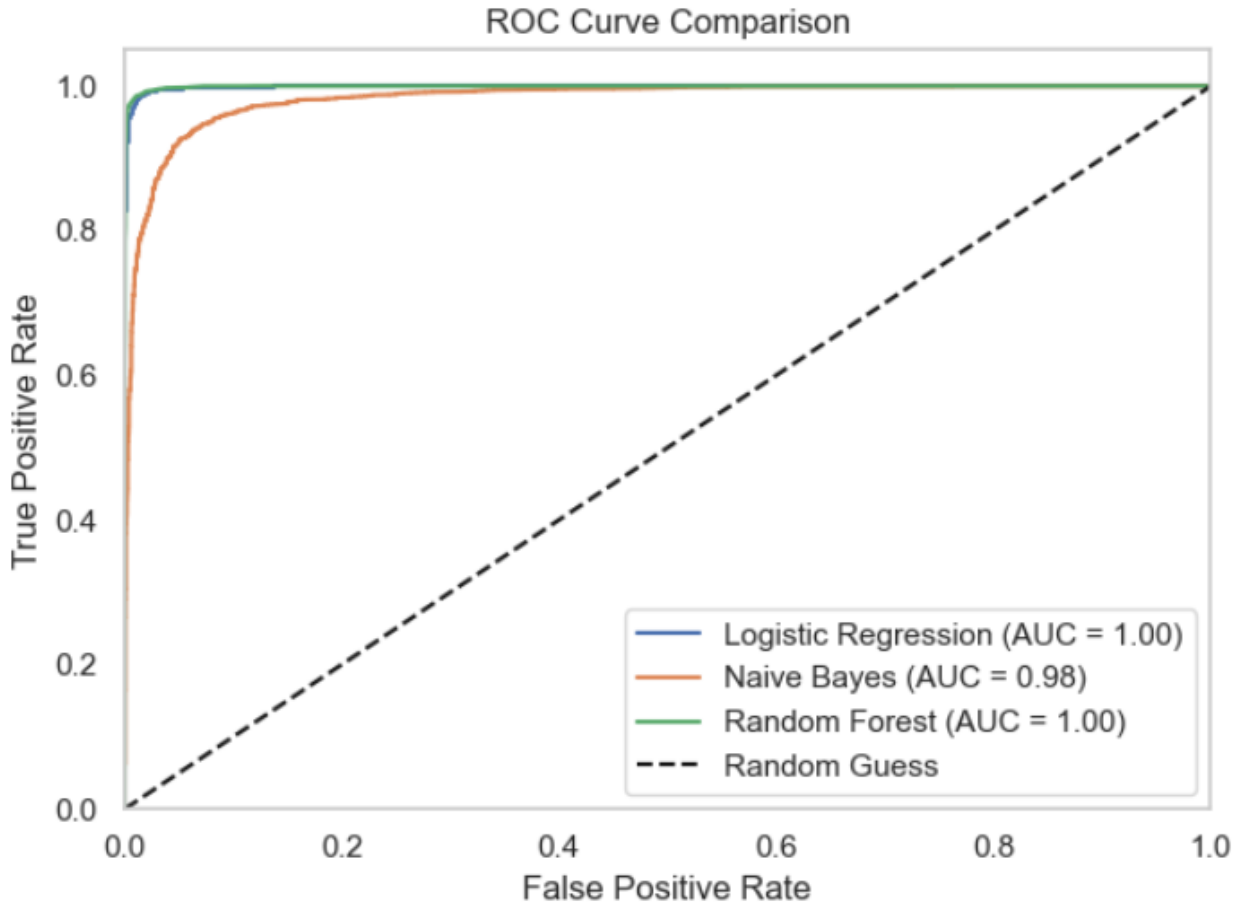
```
Random Forest Results:

[[4645   51]
 [  59 4225]]
              precision    recall  f1-score   support

           0       0.99      0.99      0.99      4696
           1       0.99      0.99      0.99      4284

    accuracy                           0.99      8980
   macro avg       0.99      0.99      0.99      8980
weighted avg       0.99      0.99      0.99      8980
```

ROC Curve



(Tetikoglu, Week9 - Data2204-Notes, 2025) (Tetikoglu, Week3 - Data2204-Notes, 2025) (Tetikoglu F. , 2025)

References

Tetikoglu, F. (2025). *Week3 - Data2204-Notes*. Retrieved August 1, 2025, from DC Connect Durham College: https://durhamcollege.desire2learn.com/d2l/le/content/620732/viewContent/8605747/View

Tetikoglu, F. (2025). *Week4 - Data2204-Notes*. Retrieved August 1, 2025, from DC Connect Durham College: https://durhamcollege.desire2learn.com/d2l/le/content/620732/viewContent/8605762/View

Tetikoglu, F. (2025). *Week9 - Data2204-Notes*. Retrieved August 1, 2025, from DC Connect Durham College: https://durhamcollege.desire2learn.com/d2l/le/content/620732/Home?itemIdentifier=D2L.LE.Content.ContentObject.ModuleCO-8605682