TECHNEX

**ENVISION**

Documentation

**TOPIC: FAKE NEWS DETECTION**

TEAM NAME: **TENSOR TITANS**

TEAM MEMBERS:

1. Aditya Bhalerao
2. Aryan Karmore
3. Raveena Basarimarad
4. Tushar Singh

**DATA SCRAPING:**

The rapid increase in the fake news possesses a significant threat to the society impacting public opinion, decision making and trust in media. This project aims to develop our robust machine learning model for the fake news detection starting from the ground up by collection of a diverse dataset of news from the sources like reputable websites such as PolitiFact and Reuters Global Fact Check platforms curating and analysing a custom dataset.. BeautifulSoup is used to extract key information including 5 columns viz sources, target, author, date and statement. This curated dataset serves as a foundation for future machine learning applications highlighting the critical role of web scraping in building a reliable pathway for preventing the spread of misinformation and advancing the fake news detection research.

**DATA PREPROCESSING:**

Steps performed in preprocessing-

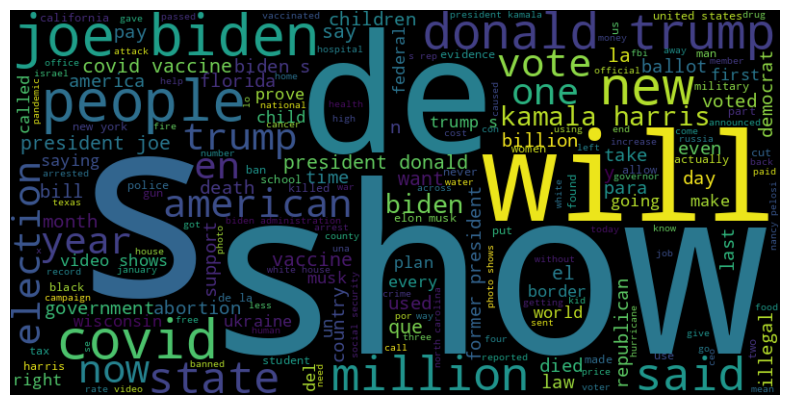
The dataset extracted from PolitiFact consisted of 5790 rows and 5 records, these are the various steps we had used for preprocessing

* The target variable consisted of 8 different classes which are True,False,Pants on Fire,Mostly True,Barely True, Half Flip,Half True, Full Flop. Most of these classes are just names for False, so we converted it into a 2-class problem mainly True and False using Python.
* After converting the target variable into a 2 variable problem we then performed one hot encoding using python
* As most of the data is textual, data was further preprocessed by
  + Converting the text into lowercase
  + Removing punctuation
  + Removing HTML tags
  + Removing /@ and other symbols
  + Removing Latex Equations
  + Removing further white spaces
  + Replacing hyphens with white spaces
  + Tokenising the data

**Data Visualisation**

Microsoft Power BI is used for data visualisation. A pie chart is introduced with “sources” field as Legend and “No. of Targets” (data count) as values. Also 2 interactive list slicers are introduced which handle Authors and Dates separately. The dashboard allows users to view the distribution of data among different sources. User can slice the data according to a particular Date or Author.

WordCloud was also used to find out which were the words that were used the most.



After performing all of these data processing methods we applied

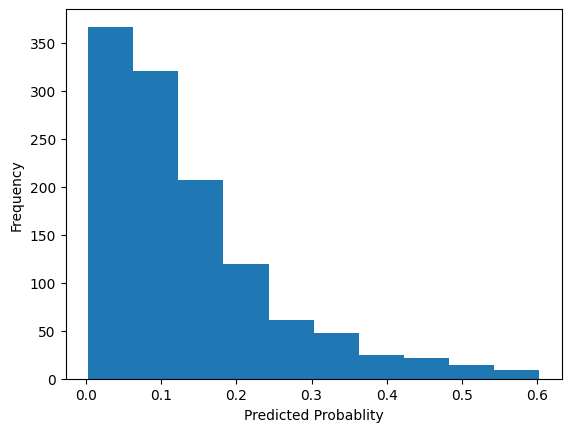
vectorisation to convert the text into vectors so that we can apply ML models

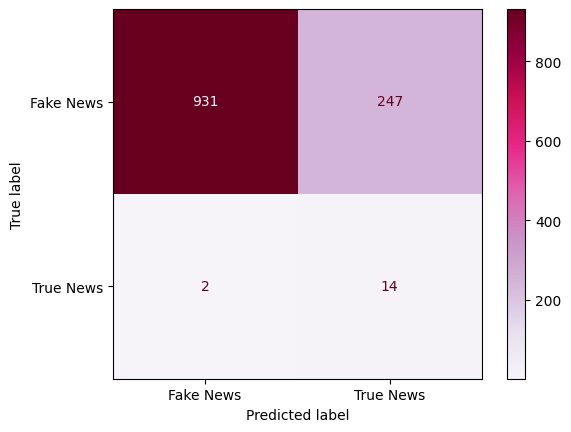
**ML models to apply**

**All of these models have been applied because the dataset is in 1 and 0**

**Naive Bayes:**

Naive Bayes is a ML model that is very useful in binary classification. It assumes that all of the features are independent of each other thus making the prediction using the Bayes equation. As the data has been converted into numerical format, -+Naive Bayes is ideal for predicting fake news as it can ignore all other features and focus on one.

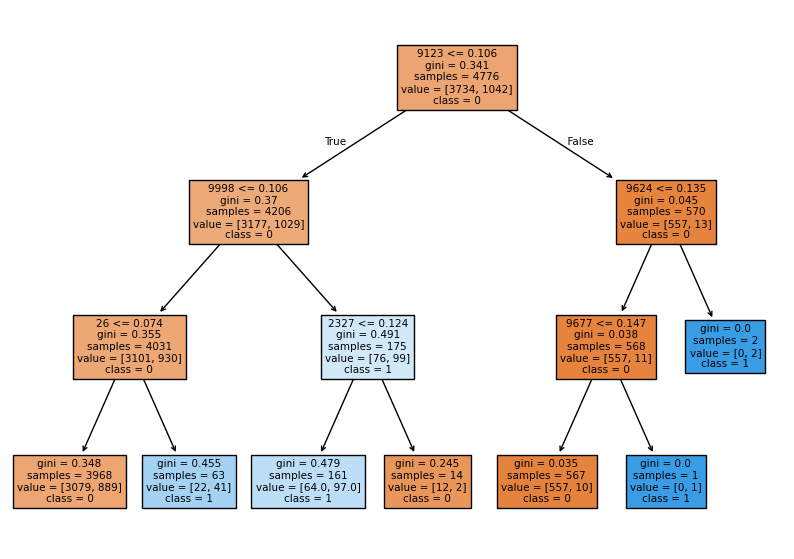
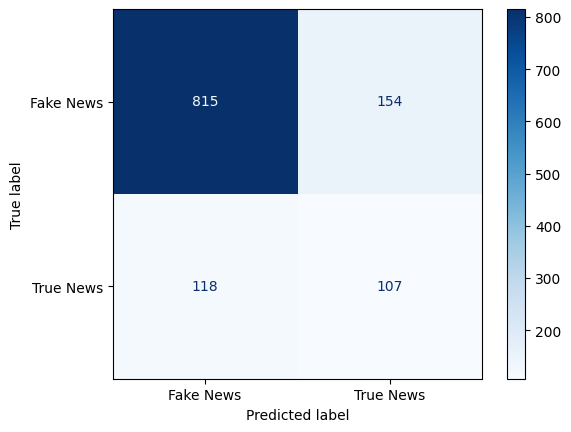
****

****

**Decision Trees**

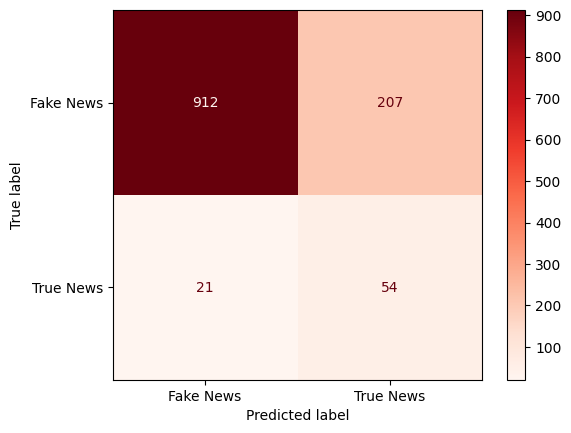
Decision Trees is a Machine Learning model that can be used in both classification and regression problems, this model predicts by splitting the data in trees and nodes by minimising the gini or the entropy function. Decision Trees can split the data based on the semantic values of the vectors.

Following are the images of the Confusion matrix and decision tree that the algorithm has followed



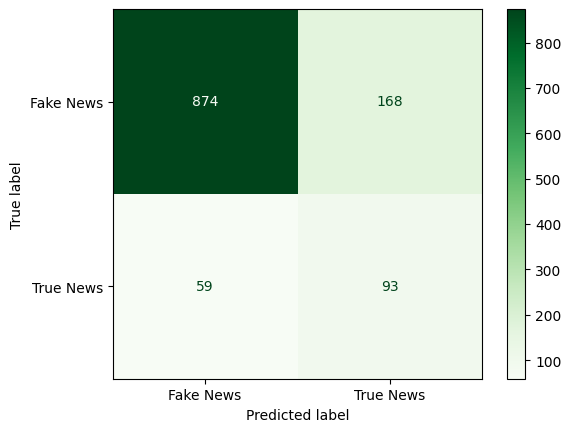
**Random Forest**

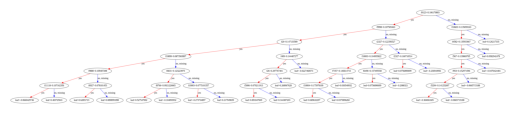
Random Forest is an ensemble ML model that collects various random decision trees. As decision trees is useful in classification, Random Forest will be a better classification technique as compared to decision trees.

****

**XGBoost**

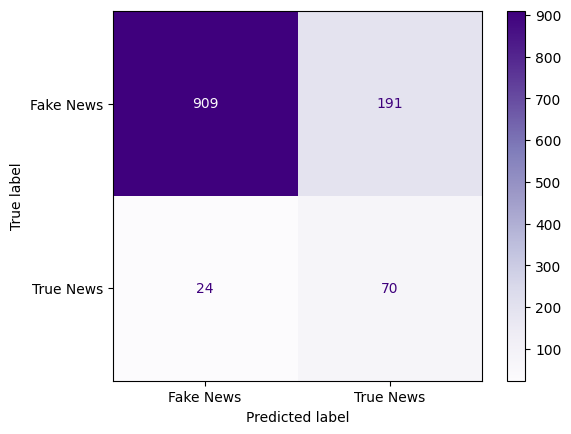
XGBoost is a type of decision tree model where each node splits equally. As decision tree was used,giving XGBoost may improve the overall accuracy

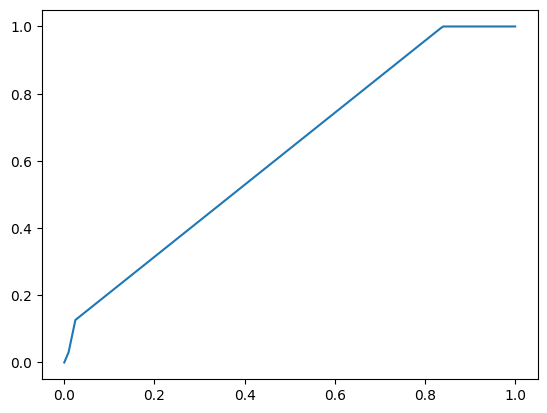
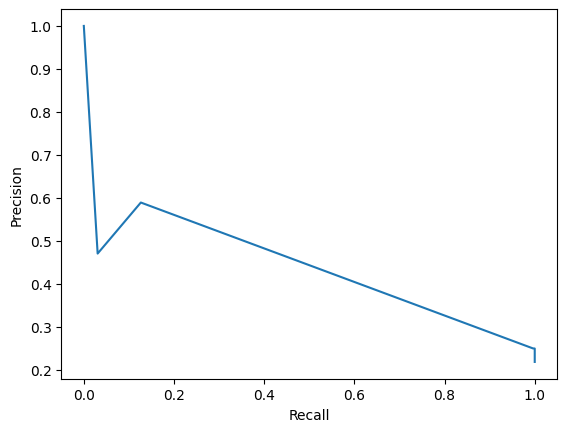
****

****

**Logistic Regression**

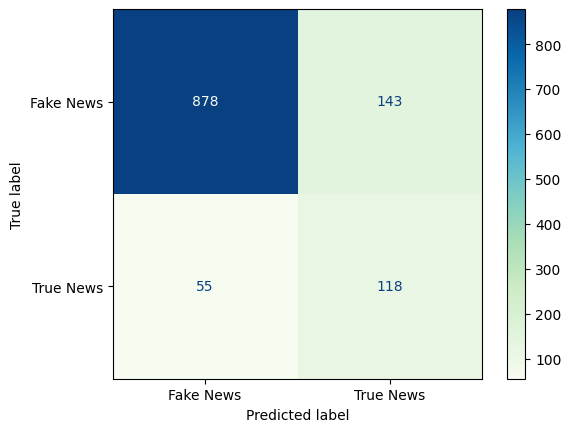
Logistic Regression is a type of ML model that classifies predictions based on a sigmoid function. If the value of the function is greater than the threshold value the output will be 1 if not 0





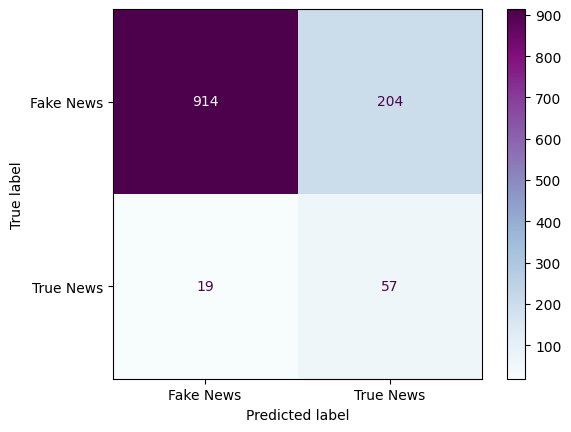
**Stacking**

Stacking is an ensemble model where various models are together stacked to come to a final prediction, it combines the prediction of various classifiers to come to a prediction. This is useful in increasing the accuracy and other evaluation metrics.

****

**Voting**

Voting is an ensemble model where various models are combined and weights are assigned to the models based on accuracy and other metrics. Based on the various votes and weights a final prediction is made. This combines many models to give us a better performance

****

**RNN:**

RNN or Recurrent Neural Network is a deep learning model that is applied to a sequential data. This type of neural network is particularly useful as it transveres forward and leftward at the same time and hence it can truly understand the semantic definition of the vectors and eventually come to a better performance

**Evaluation Metrics to be used:**

**Accuracy:** Accuracy is the percentage of correct predictions. A high accuracy does not mean a better prediction, data imbalance can be the problem. To address this issue accuracy might not be the best parameter.

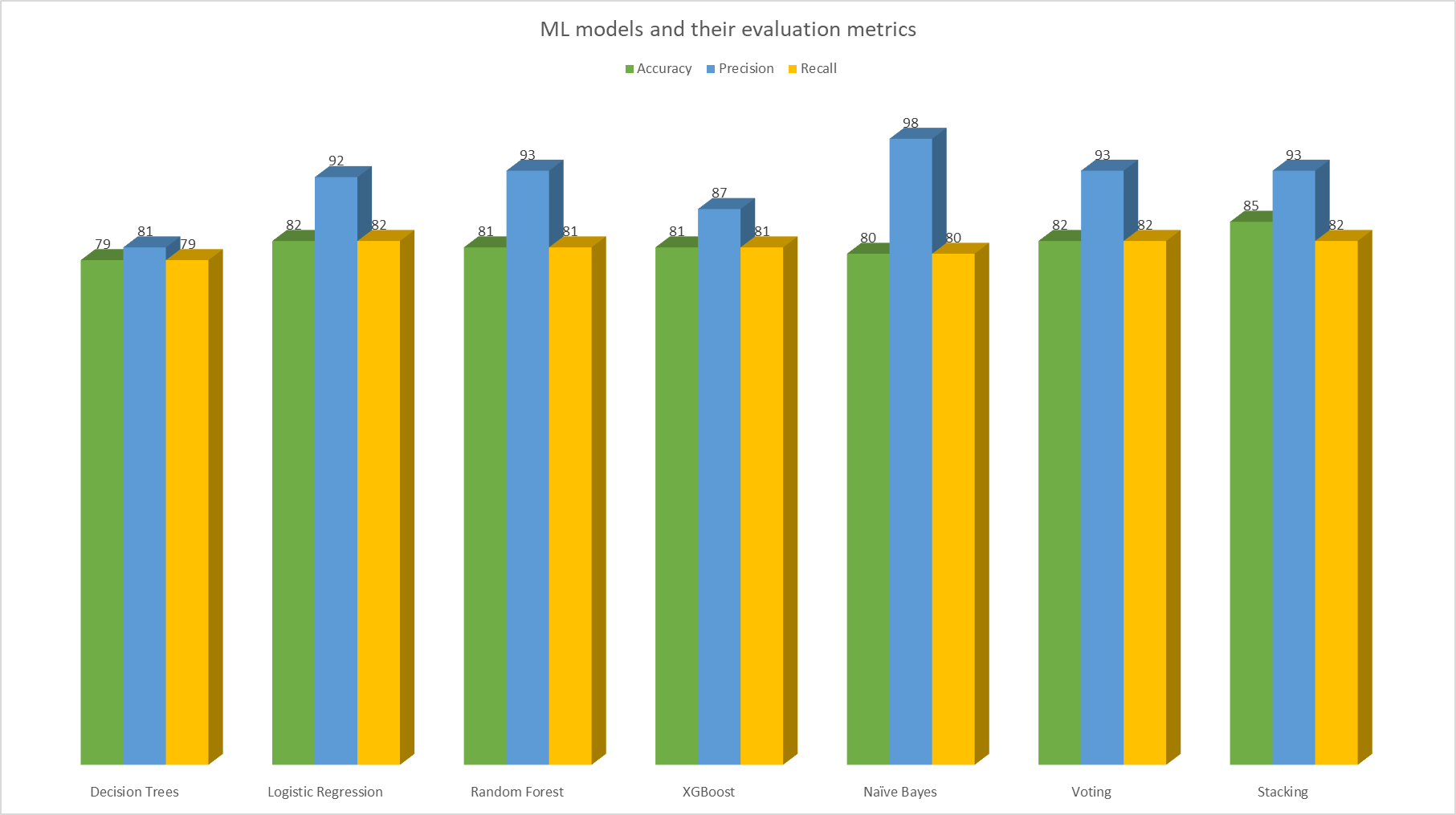
**Precession:** Precision is the percentage of positive predictions by the predicted positives

**Recall:** Recall is the percentage of positive predictions by the actual number of predictions.

Both Recall and Precision are important to find out how many values are actually true or false

**Confusion Matrix:** Classification Matrix is a matrix that shows True Positives,True Negatives, False Positives and False Negatives. A high number of True Positives/True Negatives will show that there is data imbalance

Applied all of the ML models as stated above and got the following evaluation metrics



Some of the images while performing ML models

ROC AUC curve

Precision-Recall Curve

**Technologies Used**

1. Python
2. Numpy
3. Pandas
4. Wordcloud
5. Nltk
6. CountVectorizer
7. Sklearn
8. Matplotlib
9. Contractions
10. Re
11. Microsoft PowerBi
12. Beautiful Soup
13. Requests
14. Time
15. Streamlit
16. Sckit Learn
17. XG boost