Configuration Manual

MSc Research Project
Msc in Data Analytics

Sreevishnuvardhan Mandala
Student ID: x23197293

School of Computing
National College of Ireland

Supervisor: Cristina Hava Muntean

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Sreevishnuvardhan Mandala		
Student ID:	x23197293		
Programme:	MSc in Data Analytics	**Year:**	2024
Module:	MSc in Research Project		
Lecturer:	Cristina Hava Muntean		
Submission Due Date:	12/12/2024		
Project Title:	Methods and Challenges of Data Analytics to Combat Fake News		
Word Count:	398	**Page Count:**	9

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: Sreevishnuvardhan Mandala

Date: 12/12/2024

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

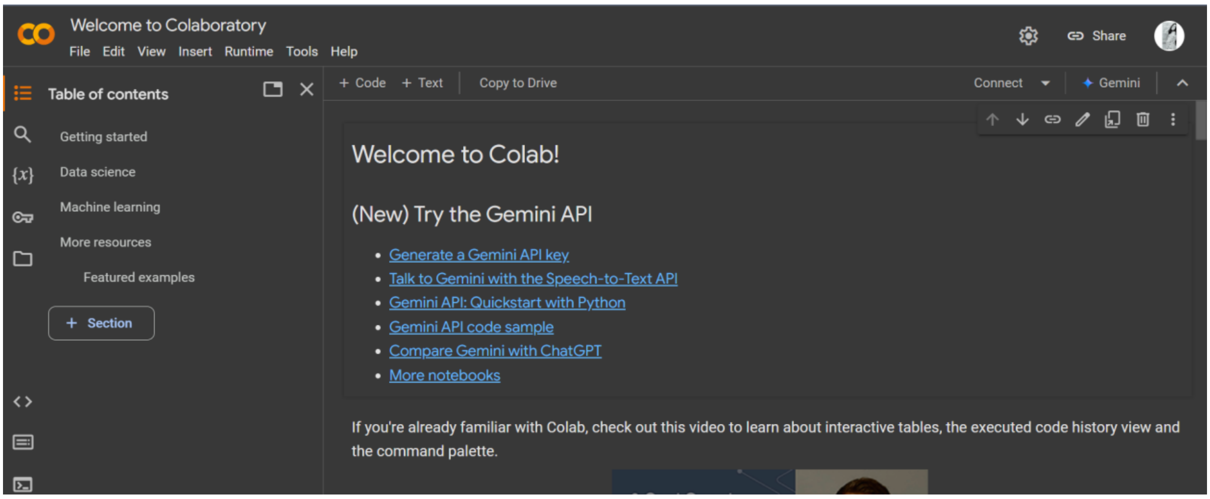
Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Configuration Manual

Sreevishnuvardhan Mandala
Student ID: x23197293

1 Tools and Libraries used

Tool	Version	Purpose
Python	3.12.5	Core language for implementing data preprocessing, analysis, and Model Building.
Jupyter Notebook	6.5.4	Interactive environment for coding and visualizing results.
pandas	2.1.1	Handles data manipulation, cleaning, and structuring for Model analysis.
numpy	1.25.3	Provides numerical operations required in Machine building algorithm modeling.
statsmodels	0.14.0	Implements machine learning algorithm models for analysis .
sklearn	1.4.0	Assists with preprocessing and evaluation metrics in machine learning workflows.



Welcome to Colab!

(New) Try the Gemini API

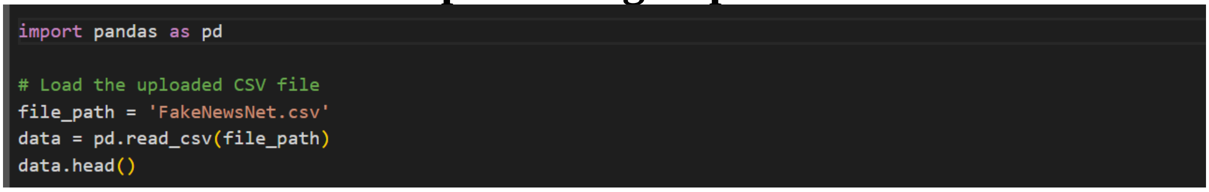
- Generate a Gemini API key
- Talk to Gemini with the Speech-to-Text API
- Gemini API: Quickstart with Python
- Gemini API code sample
- Compare Gemini with ChatGPT
- More notebooks

If you're already familiar with Colab, check out this video to learn about interactive tables, the executed code history view and the command palette.

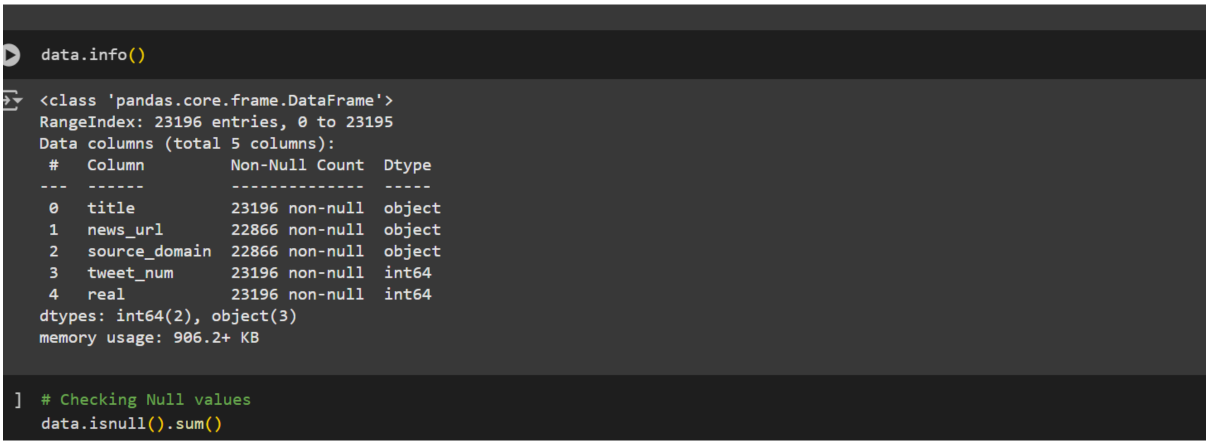
2 Dataset used

The dataset used in this file is related to fake or real news and the dataset name is FakeNewsNet.csv

3 Libraries and Pre-processing steps



```
import pandas as pd

# Load the uploaded CSV file
file_path = 'FakeNewsNet.csv'
data = pd.read_csv(file_path)
data.head()
```

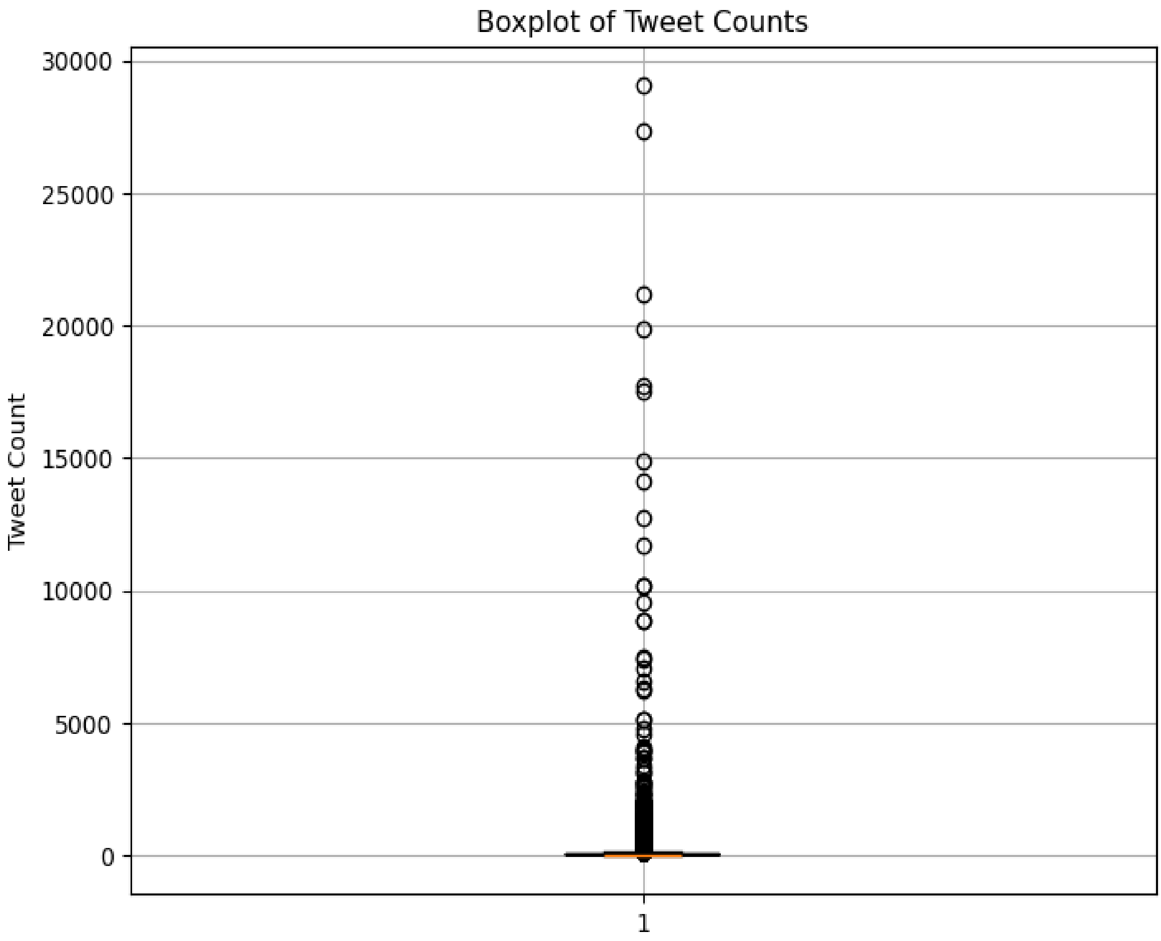


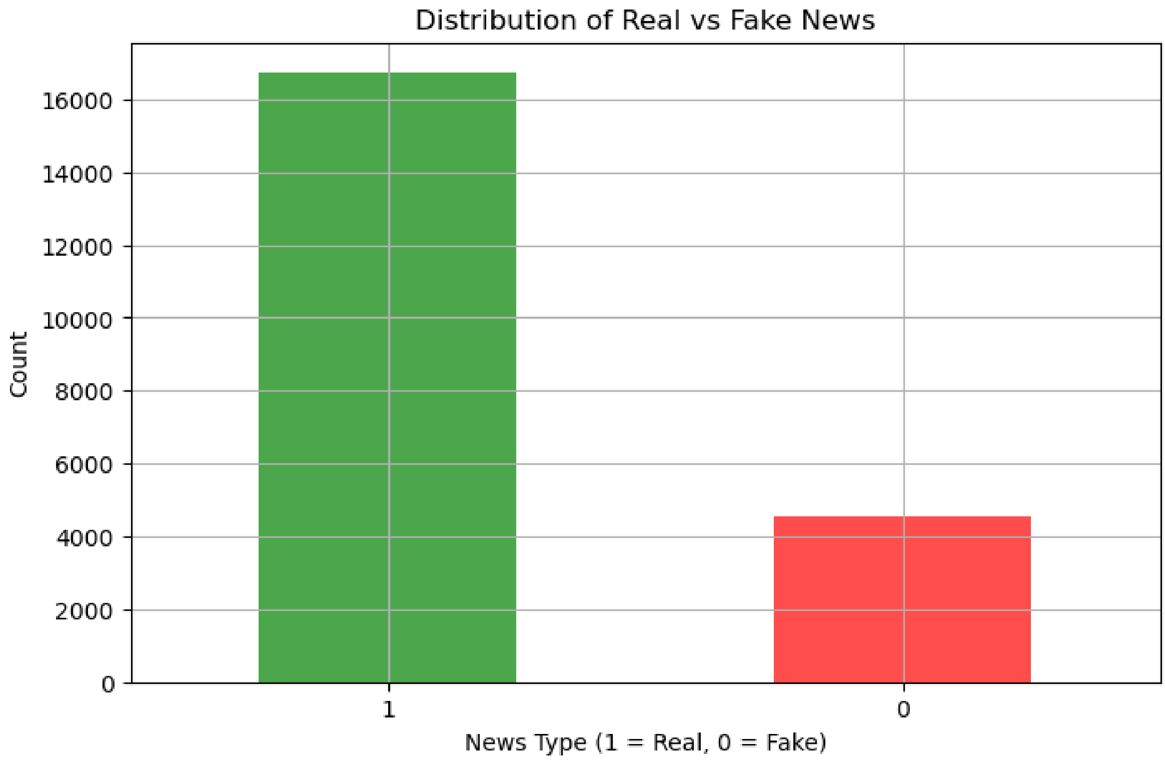
```
data.info()
```

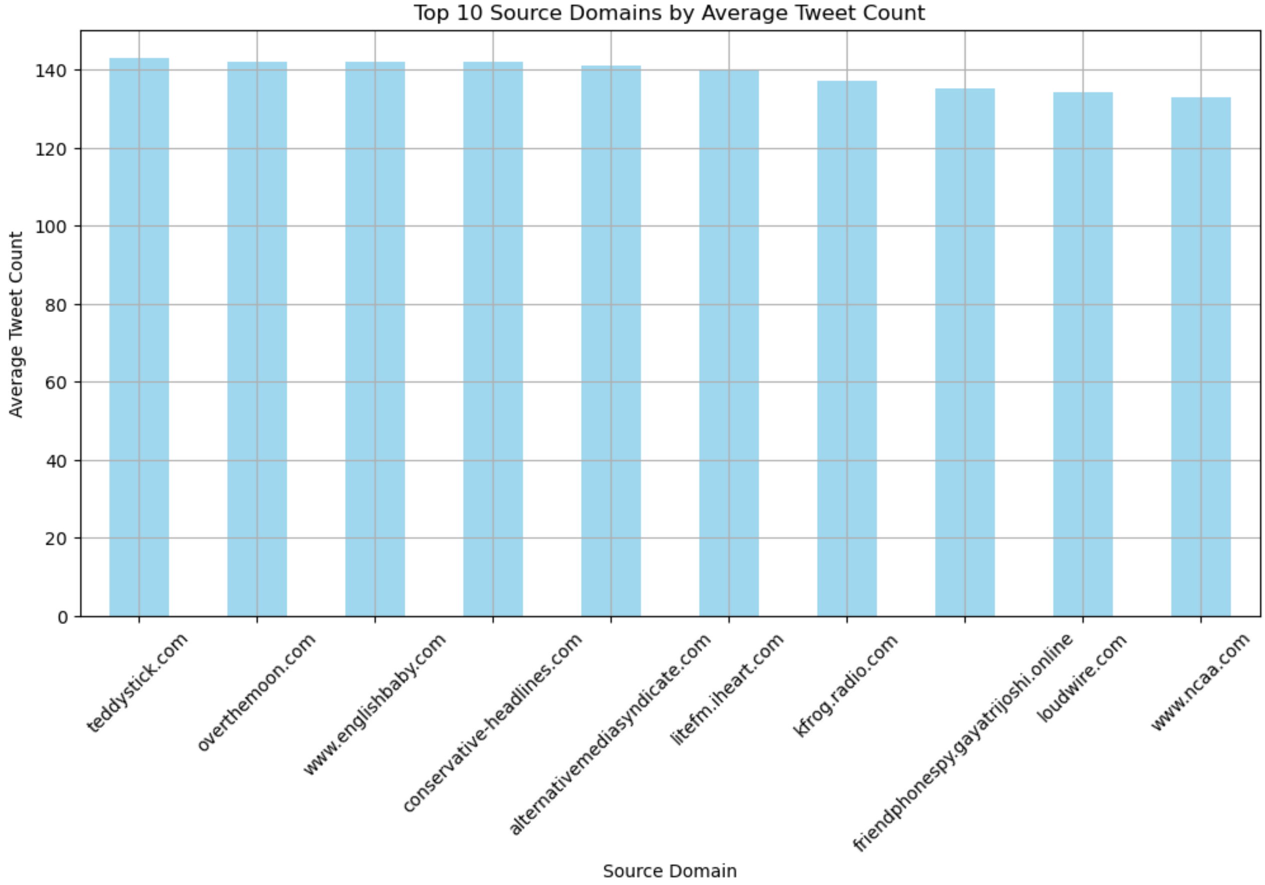
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 23196 entries, 0 to 23195
Data columns (total 5 columns):
 #   Column         Non-Null Count  Dtype 
---  ------         --------------  ----- 
 0   title          23196 non-null  object
 1   news_url       22866 non-null  object
 2   source_domain  22866 non-null  object
 3   tweet_num      23196 non-null  int64 
 4   real           23196 non-null  int64 
dtypes: int64(2), object(3)
memory usage: 906.2+ KB
```

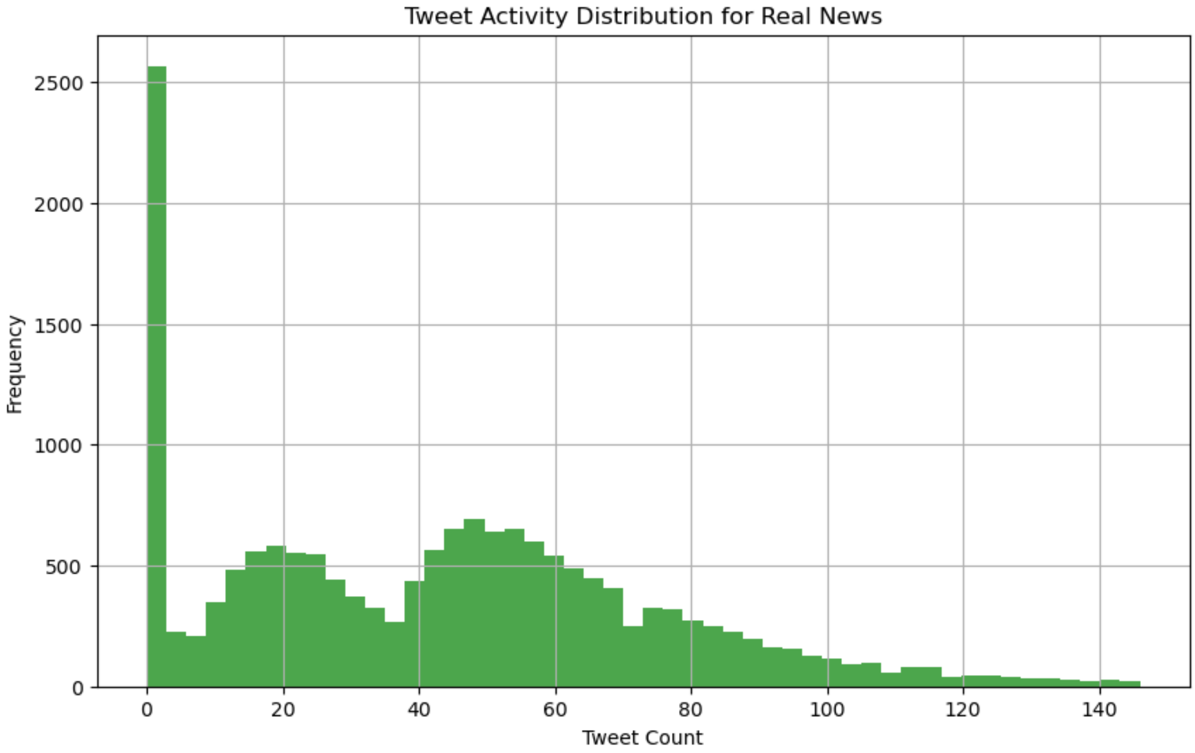
```
# Checking Null values
data.isnull().sum()
```

4 EDA



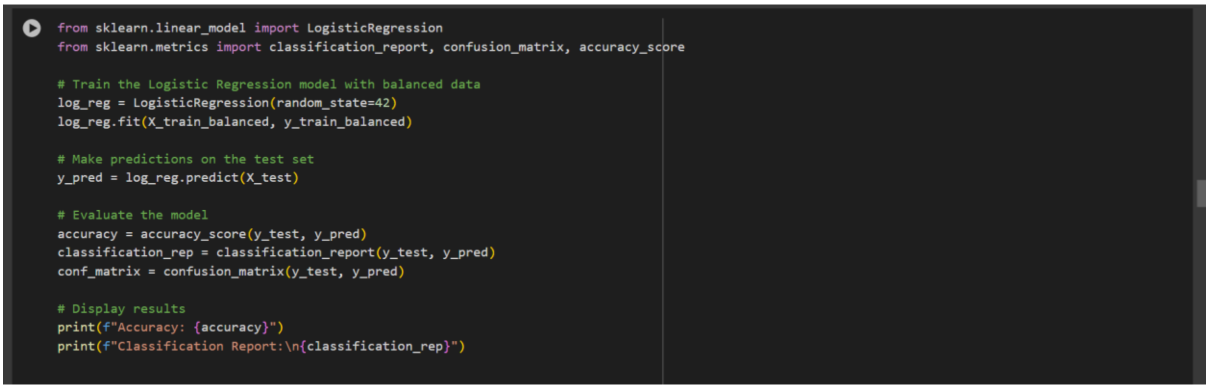

Distribution of Real vs Fake News
Count
0
2000
4000
6000
8000
10000
12000
14000
16000
1
0
News Type (1 = Real, 0 = Fake)


Top 10 Source Domains by Average Tweet Count
Average Tweet Count
0
20
40
60
80
100
120
140
teddystick.com
overthemoon.com
www.englishbaby.com
conservative-headlines.com
alternativemediasyndicate.com
litefm.iheart.com
kfrog.radio.com
friendphonespy.gayatrijoshi.online
loudwire.com
www.ncaa.com
Source Domain

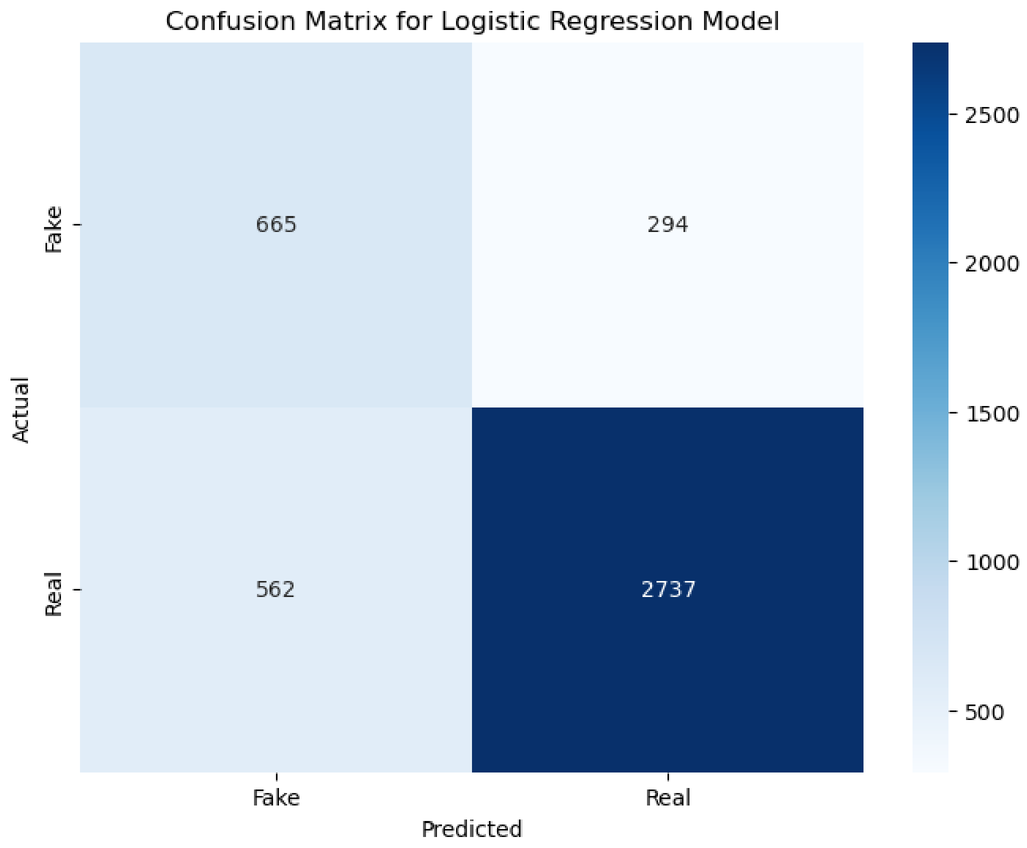


5 Model Implementation

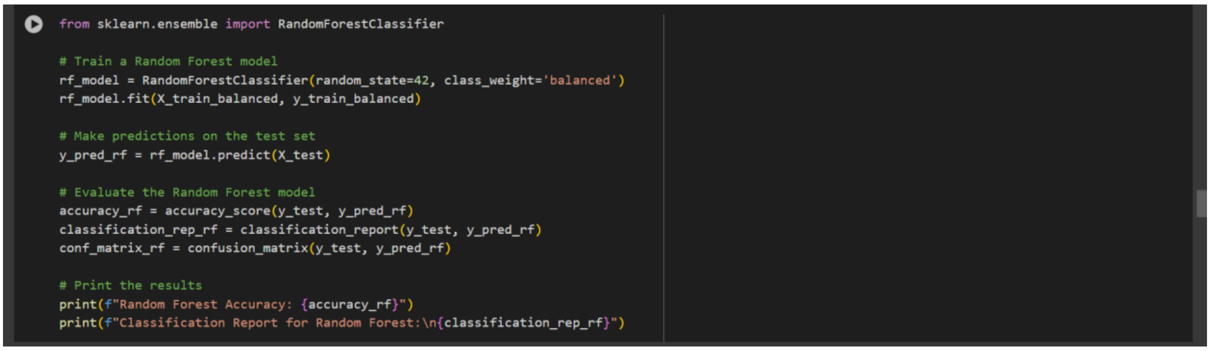
5.1 Logistic Regression



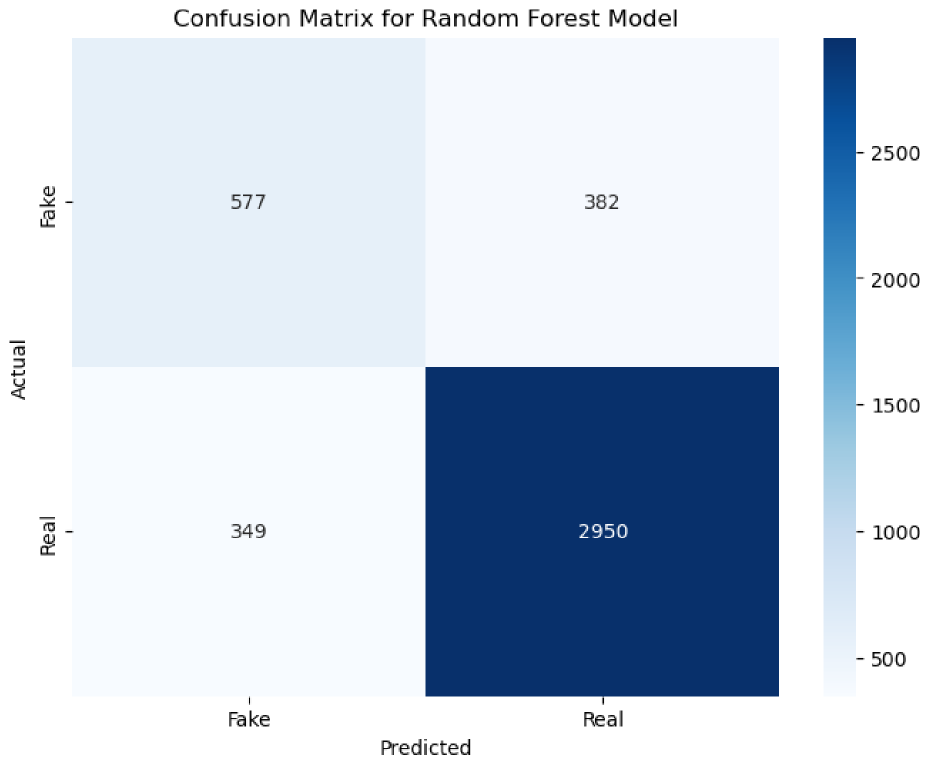
```
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import classification_report, confusion_matrix, accuracy_score

# Train the Logistic Regression model with balanced data
log_reg = LogisticRegression(random_state=42)
log_reg.fit(X_train_balanced, y_train_balanced)

# Make predictions on the test set
y_pred = log_reg.predict(X_test)

# Evaluate the model
accuracy = accuracy_score(y_test, y_pred)
classification_rep = classification_report(y_test, y_pred)
conf_matrix = confusion_matrix(y_test, y_pred)

# Display results
print(f"Accuracy: {accuracy}")
print(f"Classification Report:\n{classification_rep}")
```



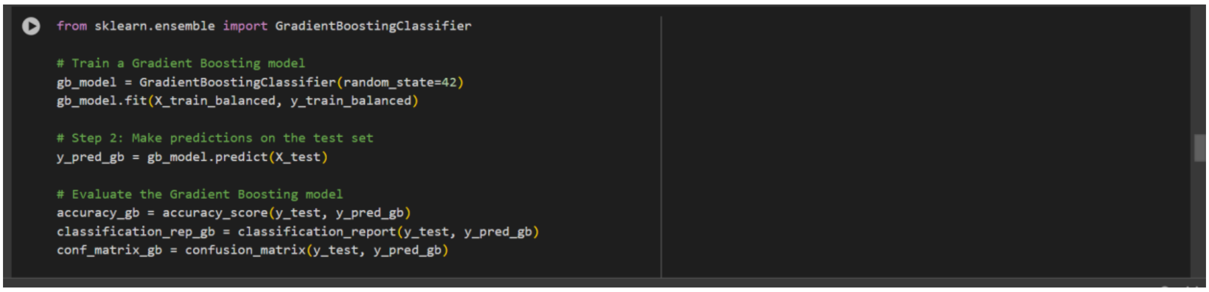
5.2 Random Forest Classifier



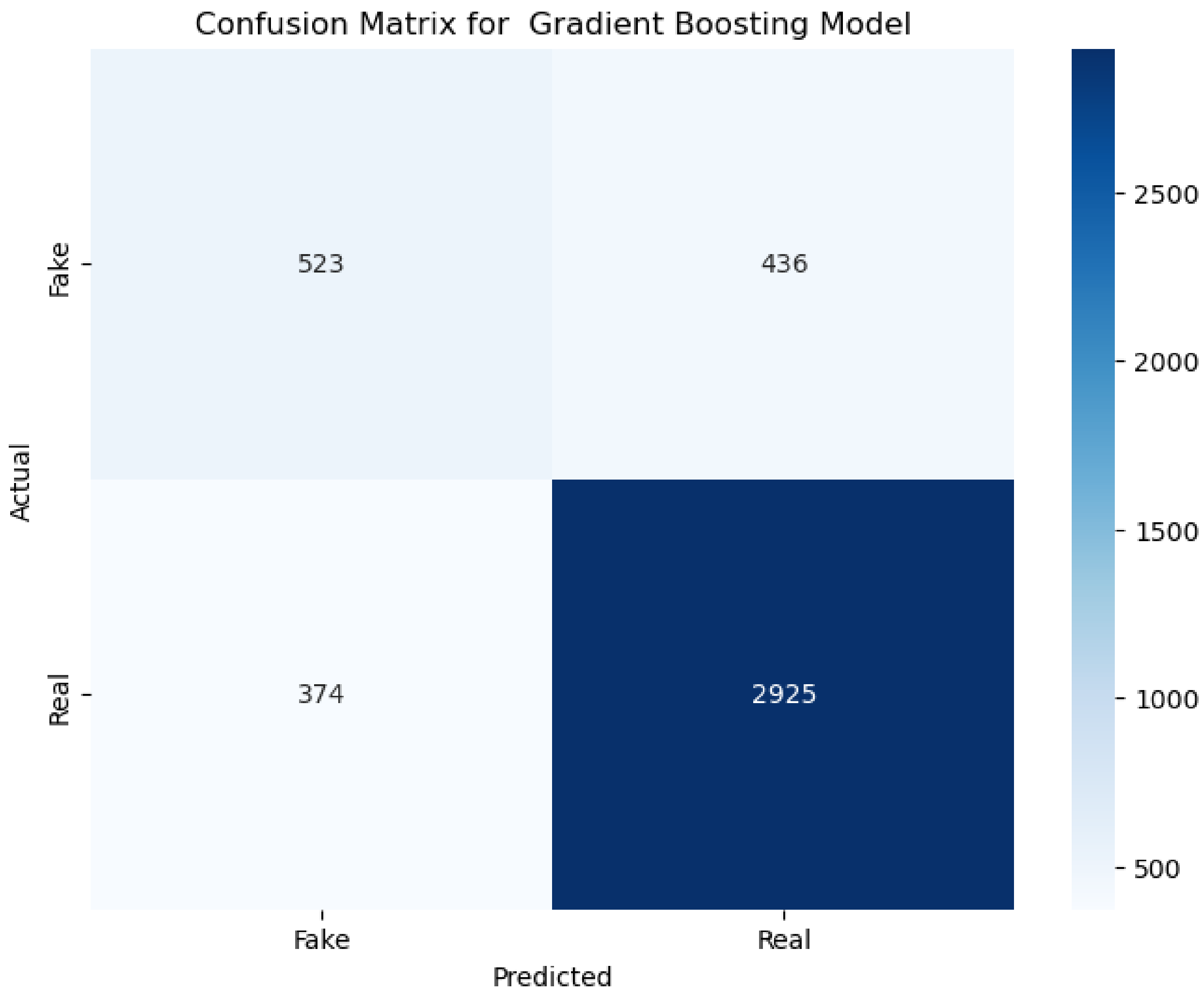
```
from sklearn.ensemble import RandomForestClassifier

# Train a Random Forest model
rf_model = RandomForestClassifier(random_state=42, class_weight='balanced')
rf_model.fit(X_train_balanced, y_train_balanced)

# Make predictions on the test set
y_pred_rf = rf_model.predict(X_test)

# Evaluate the Random Forest model
accuracy_rf = accuracy_score(y_test, y_pred_rf)
classification_rep_rf = classification_report(y_test, y_pred_rf)
conf_matrix_rf = confusion_matrix(y_test, y_pred_rf)

# Print the results
print(f"Random Forest Accuracy: {accuracy_rf}")
print(f"Classification Report for Random Forest:\n{classification_rep_rf}")
```



5.3 Gradient Boosting Classifier



```
from sklearn.ensemble import GradientBoostingClassifier

# Train a Gradient Boosting model
gb_model = GradientBoostingClassifier(random_state=42)
gb_model.fit(X_train_balanced, y_train_balanced)

# Step 2: Make predictions on the test set
y_pred_gb = gb_model.predict(X_test)

# Evaluate the Gradient Boosting model
accuracy_gb = accuracy_score(y_test, y_pred_gb)
classification_rep_gb = classification_report(y_test, y_pred_gb)
conf_matrix_gb = confusion_matrix(y_test, y_pred_gb)
```



References

https://colab.research.google.com/notebooks/intro.ipynb