```
from google.colab import files
files.upload()
```

Choose Files No file chosen Upload widget is only available when the cell has been executed in the current browser session. Please rerun this cell to enable.

```
Saving kaggle.json to kaggle.json
{'kaggle.json': b'{"username":"sanskarsingh003","key":"f522945b733ac1077107a31d9b3b90af"}'}
```

```
!mkdir -p ~/.kaggle
!cp kaggle.json ~/.kaggle/
!chmod 600 ~/.kaggle/kaggle.json
```

```
!kaggle datasets download -d clmentbisaillon/fake-and-real-news-dataset
```

```
Dataset URL: https://www.kaggle.com/datasets/clmentbisaillon/fake-and-real-news-dataset
License(s): CC-BY-NC-SA-4.0
Downloading fake-and-real-news-dataset.zip to /content
  0% 0.00/41.0M [00:00<?, ?B/s]
100% 41.0M/41.0M [00:00<00:00, 1.18GB/s]
```

```
!unzip '*.zip'
```

```
Archive:  fake-and-real-news-dataset.zip
  inflating: Fake.csv
  inflating: True.csv
```

```
import pandas as pd
import numpy as np
import string
import nltk
nltk.download('punkt')
nltk.download('wordnet')
nltk.download('stopwords')
```

```
[nltk_data] Downloading package punkt to /root/nltk_data...
[nltk_data]   Unzipping tokenizers/punkt.zip.
[nltk_data] Downloading package wordnet to /root/nltk_data...
[nltk_data] Downloading package stopwords to /root/nltk_data...
[nltk_data]   Unzipping corpora/stopwords.zip.
True
```

```
from nltk.tokenize import word_tokenize
from nltk.corpus import stopwords
from nltk.stem import WordNetLemmatizer
```

```
from sklearn.model_selection import StratifiedKFold, cross_val_score
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.tree import DecisionTreeClassifier
from sklearn.ensemble import RandomForestClassifier
from xgboost import XGBClassifier
```

```
stop_words = set(stopwords.words('english'))
lemmatizer = WordNetLemmatizer()
```

```
def clean_text(text):
    text = text.lower()
    tokens = word_tokenize(text)
    tokens = [word for word in tokens if word.isalpha()]  # remove punctuation/numbers
    tokens = [word for word in tokens if word not in stop_words]  # remove stopwords
    tokens = [lemmatizer.lemmatize(word) for word in tokens]  # lemmatization
    return ' '.join(tokens)
```

```
df_fake = pd.read_csv('Fake.csv')
df_true = pd.read_csv('True.csv')
```

```
df_fake.head()
```

	title	text	subject	date
0	Donald Trump Sends Out Embarrassing New Year'...	Donald Trump just couldn t wish all Americans ...	News	December 31, 2017
1	Drunk Bragging Trump Staffer Started Russian ...	House Intelligence Committee Chairman Devin Nu...	News	December 31, 2017
2	Sheriff David Clarke Becomes An Internet Joke...	On Friday, it was revealed that former Milwauk...	News	December 30, 2017
3	Trump Is So Obsessed He Even Has Obama's Name...	On Christmas day, Donald Trump announced that ...	News	December 29, 2017
4	Pope Francis Just Called Out Donald Trump Dur...	Pope Francis used his annual Christmas Day mes...	News	December 25, 2017

```
df_true.head()
```

	title	text	subject	date
0	As U.S. budget fight looms, Republicans flip t...	WASHINGTON (Reuters) - The head of a conservat...	politicsNews	December 31, 2017
1	U.S. military to accept transgender recruits o...	WASHINGTON (Reuters) - Transgender people will...	politicsNews	December 29, 2017
2	Senior U.S. Republican senator: 'Let Mr. Muell...	WASHINGTON (Reuters) - The special counsel inv...	politicsNews	December 31, 2017
3	FBI Russia probe helped by Australian diplomat...	WASHINGTON (Reuters) - Trump campaign adviser ...	politicsNews	December 30, 2017
4	Trump wants Postal Service to charge 'much mor...	SEATTLE/WASHINGTON (Reuters) - President Donal...	politicsNews	December 29, 2017

```
df_fake['label'] = 'fake'
df_true['label'] = 'true'
```

```
# Combine the two
df= pd.concat([df_fake, df_true], ignore_index=True)
```

```
df = df.sample(frac=1, random_state=42).reset_index(drop=True)
```

```
df.head()
```

	title	text	subject	date	label
0	Ben Stein Calls Out 9th Circuit Court: Committ...	21st Century Wire says Ben Stein, reputable pr...	US_News	February 13, 2017	fake
1	Trump drops Steve Bannon from National Securit...	WASHINGTON (Reuters) - U.S. President Donald T...	politicsNews	April 5, 2017	true
2	Puerto Rico expects U.S. to lift Jones Act shi...	(Reuters) - Puerto Rico Governor Ricardo Rosse...	politicsNews	September 27, 2017	true
3	OOPS: Trump Just Accidentally Confirmed He Le...	On Monday, Donald Trump once again embarrassed...	News	May 22, 2017	fake
4	Donald Trump heads for Scotland to reopen a go...	GLASGOW, Scotland (Reuters) - Most U.S. presid...	politicsNews	June 24, 2016	true

```
df['clean_text'] = df['text'].apply(clean_text)
```

```
X=df['clean_text']
y=df['label']
```

```
vectorizer = TfidfVectorizer(max_df=0.7)
X_tfidf = vectorizer.fit_transform(X)
```

```
cv = StratifiedKFold(n_splits=5, shuffle=True, random_state=42)
```

```
models = {
    "Decision Tree": DecisionTreeClassifier(random_state=42),
    "Random Forest": RandomForestClassifier(n_estimators=100, random_state=42),
    "XGBoost": XGBClassifier(use_label_encoder=False, eval_metric='logloss', random_state=42)
}
```

```
from sklearn.preprocessing import LabelEncoder

for name, model in models.items():
    scores = cross_val_score(model, X_tfidf, y, cv=cv, scoring='accuracy')
    print(f"{name} - Accuracy: {np.mean(scores) * 100:.2f}% (+/- {np.std(scores) * 100:.2f}%)")

# Label encoding
label_encoder = LabelEncoder()
y = label_encoder.fit_transform(y)
# Vectorization
X_tfidf = vectorizer.fit_transform(X)
```

```
Decision Tree - Accuracy: 99.53% (+/- 0.05%)
Random Forest - Accuracy: 98.70% (+/- 0.22%)
/usr/local/lib/python3.11/dist-packages/xgboost/core.py:158: UserWarning: [18:32:54] WARNING: /workspace/src/learner.cc:740:
Parameters: { "use_label_encoder" } are not used.

  warnings.warn(smsg, UserWarning)
/usr/local/lib/python3.11/dist-packages/xgboost/core.py:158: UserWarning: [18:34:20] WARNING: /workspace/src/learner.cc:740:
Parameters: { "use_label_encoder" } are not used.

  warnings.warn(smsg, UserWarning)
/usr/local/lib/python3.11/dist-packages/xgboost/core.py:158: UserWarning: [18:35:41] WARNING: /workspace/src/learner.cc:740:
Parameters: { "use_label_encoder" } are not used.

  warnings.warn(smsg, UserWarning)
/usr/local/lib/python3.11/dist-packages/xgboost/core.py:158: UserWarning: [18:37:01] WARNING: /workspace/src/learner.cc:740:
Parameters: { "use_label_encoder" } are not used.

  warnings.warn(smsg, UserWarning)
/usr/local/lib/python3.11/dist-packages/xgboost/core.py:158: UserWarning: [18:38:22] WARNING: /workspace/src/learner.cc:740:
Parameters: { "use_label_encoder" } are not used.

  warnings.warn(smsg, UserWarning)
XGBoost - Accuracy: 99.72% (+/- 0.03%)
```

```
# 1. Train the final model
final_model = XGBClassifier(use_label_encoder=False, eval_metric='logloss', random_state=42)
final_model.fit(X_tfidf, y)

# 2. Save model and vectorizer
import joblib
joblib.dump(final_model, 'xgb_fake_news_model.pkl')
joblib.dump(vectorizer, 'tfidf_vectorizer.pkl')
```

```
/usr/local/lib/python3.11/dist-packages/xgboost/core.py:158: UserWarning: [18:42:00] WARNING: /workspace/src/learner.cc:740:
Parameters: { "use_label_encoder" } are not used.

  warnings.warn(smsg, UserWarning)
['tfidf_vectorizer.pkl']
```

```
from google.colab import files

# Save and download model
files.download('xgb_fake_news_model.pkl')
files.download('tfidf_vectorizer.pkl')
```