Internship Report

Fake News Detection

Name: Ankita Gupta

Course: ML1119

Duration: 2 Weeks

Problem Statement: The main objective is to detect the fake news, build a Machine Learning model to differentiate between “Real” news and “Fake” news.

Github link: https://github.com/Ankita30-ui/Fake_news_detection

Prerequisites

What things you need to install the software and how to install them:

Python 3.6 This setup requires that your machine has python 3.6 installed on it. you can refer to this url https://www.python.org/downloads/ to download python. Once you have python downloaded and installed, you will need to setup PATH variables (if you want to run python program directly, detail instructions are below in how to run software section). To do that check this: https://www.pythoncentral.io/add-python-to-path-python-is-not-recognized-as-an-internal-or-external-command/. Setting up PATH variable is optional as you can also run program without it and more instruction are given below on this topic. Second and easier option is to download anaconda and use its anaconda prompt to run the commands. To install anaconda check this url https://www.anaconda.com/download/ You will also need to download and install below 3 packages after you install either python or anaconda from the steps above Sklearn (scikit-learn) numpy scipy if you have chosen to install python 3.6 then run below commands in command prompt/terminal to install these packages pip install -U scikit-learn pip install numpy pip install scipy if you have chosen to install anaconda then run below commands in anaconda prompt to install these packages conda install -c scikit-learn conda install -c anaconda numpy conda install -c anaconda scipy

Dataset used

The data source used for this project is news.csv The news.csv Data Set contains attributes like: title, text and labels (fake, real).

Link: https://drive.google.com/drive/folders/1Dzj0gD6irtFA97BkBlpd3lnwWRPzHBHp?usp=sharing

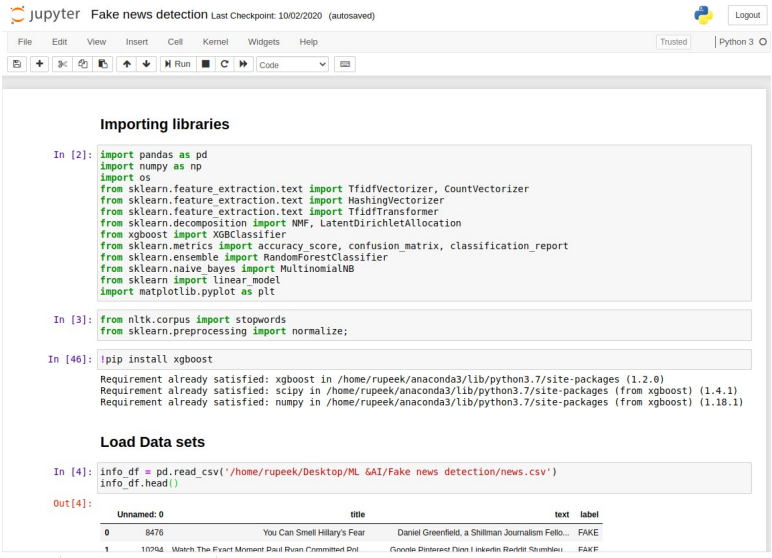
Applied algorithms

XGB, Random Forest Classifier, Multinomial Naive Bayes.

Accuracy comparision

According to the accuracy , Multinomial NB works the best.

Importing the libraries:



jupyter Fake news detection Last Checkpoint: 10/02/2020 (autosaved)

File Edit View Insert Cell Kernel Widgets Help

Importing libraries

In [2]:
```
import pandas as pd
import numpy as np
import os
from sklearn.feature_extraction.text import TfidfVectorizer, CountVectorizer
from sklearn.feature_extraction.text import HashingVectorizer
from sklearn.feature_extraction.text import TfidfTransformer
from sklearn.decomposition import NMF, LatentDirichletAllocation
from xgboost import XGBClassifier
from sklearn.metrics import accuracy_score, confusion_matrix, classification_report
from sklearn.ensemble import RandomForestClassifier
from sklearn.naive_bayes import MultinomialNB
from sklearn import linear_model
import matplotlib.pyplot as plt
```

In [3]:
```
from nltk.corpus import stopwords
from sklearn.preprocessing import normalize;
```

In [46]:
```
!pip install xgboost
```

```
Requirement already satisfied: xgboost in /home/rupeek/anaconda3/lib/python3.7/site-packages (1.2.0)
Requirement already satisfied: scipy in /home/rupeek/anaconda3/lib/python3.7/site-packages (from xgboost) (1.4.1)
Requirement already satisfied: numpy in /home/rupeek/anaconda3/lib/python3.7/site-packages (from xgboost) (1.18.1)
```

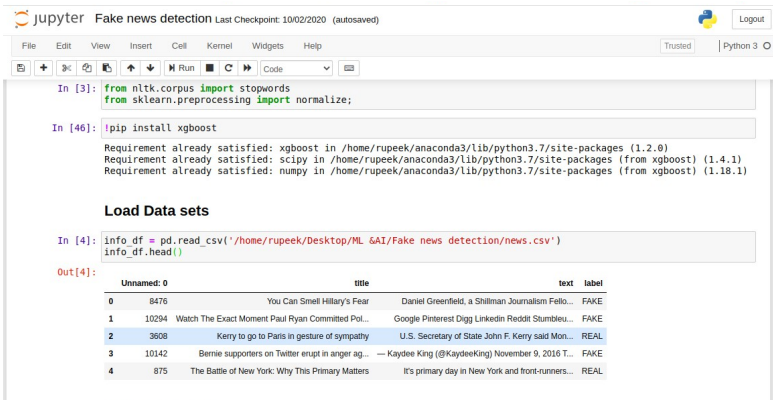
Load Data sets

In [4]:
```
info_df = pd.read_csv('/home/rupeek/Desktop/ML &AI/Fake news detection/news.csv')
info_df.head()
```

Out[4]:

	Unnamed: 0	title	text	label
0	8476	You Can Smell Hillary's Fear	Daniel Greenfield, a Shillman Journalism Fello...	FAKE
1	10294	Watch The Exact Moment Paul Ryan Committed Pol...	Google Pinterest Digg Linkedin Reddit Stumbleu...	FAKE

2.Loading the dataset:



jupyter Fake news detection Last Checkpoint: 10/02/2020 (autosaved)

File Edit View Insert Cell Kernel Widgets Help

In [3]:
```
from nltk.corpus import stopwords
from sklearn.preprocessing import normalize;
```

In [46]:
```
!pip install xgboost
```

```
Requirement already satisfied: xgboost in /home/rupeek/anaconda3/lib/python3.7/site-packages (1.2.0)
Requirement already satisfied: scipy in /home/rupeek/anaconda3/lib/python3.7/site-packages (from xgboost) (1.4.1)
Requirement already satisfied: numpy in /home/rupeek/anaconda3/lib/python3.7/site-packages (from xgboost) (1.18.1)
```

Load Data sets

In [4]:
```
info_df = pd.read_csv('/home/rupeek/Desktop/ML &AI/Fake news detection/news.csv')
info_df.head()
```

Out[4]:

	Unnamed: 0	title	text	label
0	8476	You Can Smell Hillary's Fear	Daniel Greenfield, a Shillman Journalism Fello...	FAKE
1	10294	Watch The Exact Moment Paul Ryan Committed Pol...	Google Pinterest Digg Linkedin Reddit Stumbleu...	FAKE
2	3608	Kerry to go to Paris in gesture of sympathy	U.S. Secretary of State John F. Kerry said Mon...	REAL
3	10142	Bernie supporters on Twitter erupt in anger ag...	— Kaydee King (@KaydeeKing) November 9, 2016 T...	FAKE
4	875	The Battle of New York: Why This Primary Matters	It's primary day in New York and front-runners...	REAL

3. Exploring the data:

jupyter Fake news detection Last Checkpoint: 10/02/2020 (autosaved)

File Edit View Insert Cell Kernel Widgets Help

Explore the Data

In [7]:
```
#instantiate CountVectorizer()
cv=CountVectorizer()#create matrix dXm(doc and words)

# this steps generates word counts for the words in your docs
word_count_vector=cv.fit_transform(info_df["text"])
```

In [8]:
```
word_count_vector.shape
```

Out[8]: (6335, 67659)

In [9]:
```
tfidf_transformer = TfidfTransformer(smooth_idf=True,use_idf=True)
tf_idf_vector = tfidf_transformer.fit_transform(word_count_vector)
```

In [10]:
```
cv.get_feature_names()
```

```
'22m3',
'22map',
'22mapposition',
'22nd',
'22newest',
'22overlays',
'22plates',
'22pm',
'22restrictlisttomap',
'22search',
'22sort',
'22th',
'22timezone',
'22us1000778i',
'22utc',
'22viewmodes',
'23',
'230',
'231',
```

In [11]:
```
feature_names = cv.get_feature_names()
```

Activities Google Chrome Oct 9 15:47

Eckovation Course | Career Eckovation | Fake news detection - Jup | Ankita30-ui/Fake_news_d

localhost:8888/notebooks/Desktop/ML%20%26AI/Fake%20news%20detection.ipynb#

Apps Jennifer Lope... Fmovies - Wat... Watch Friends... Download Fri... TorHD - Down... Watch Where... YouTube to M... Drug Result P... Torasemide -... Mirzapur 1x3... Hypothyroidis...

jupyter Fake news detection Last Checkpoint: 10/02/2020 (autosaved)

File Edit View Insert Cell Kernel Widgets Help

In [22]:
```
stopw=stopwords.words('english')
stopw
def stopwords_remove(x):
    terms=x.split()#get the words
    terms=[w for w in terms if w not in stopw]#if word isnt possible in stopw
    sentence=' '.join(terms)#joining words after stopword removal
    return sentence
data_text['Refined_text']=data_text['text'].apply(lambda x: stopwords_remove(x))
```

In [23]:
```
def word_count(x):#no. of terms in data
    terms = x.split()
    return len(terms)
data_text['Word_Count']=data_text['Refined_text'].apply(lambda x: word_count(x))
```

In [24]:
```
data_text.head()
```

Out[24]:

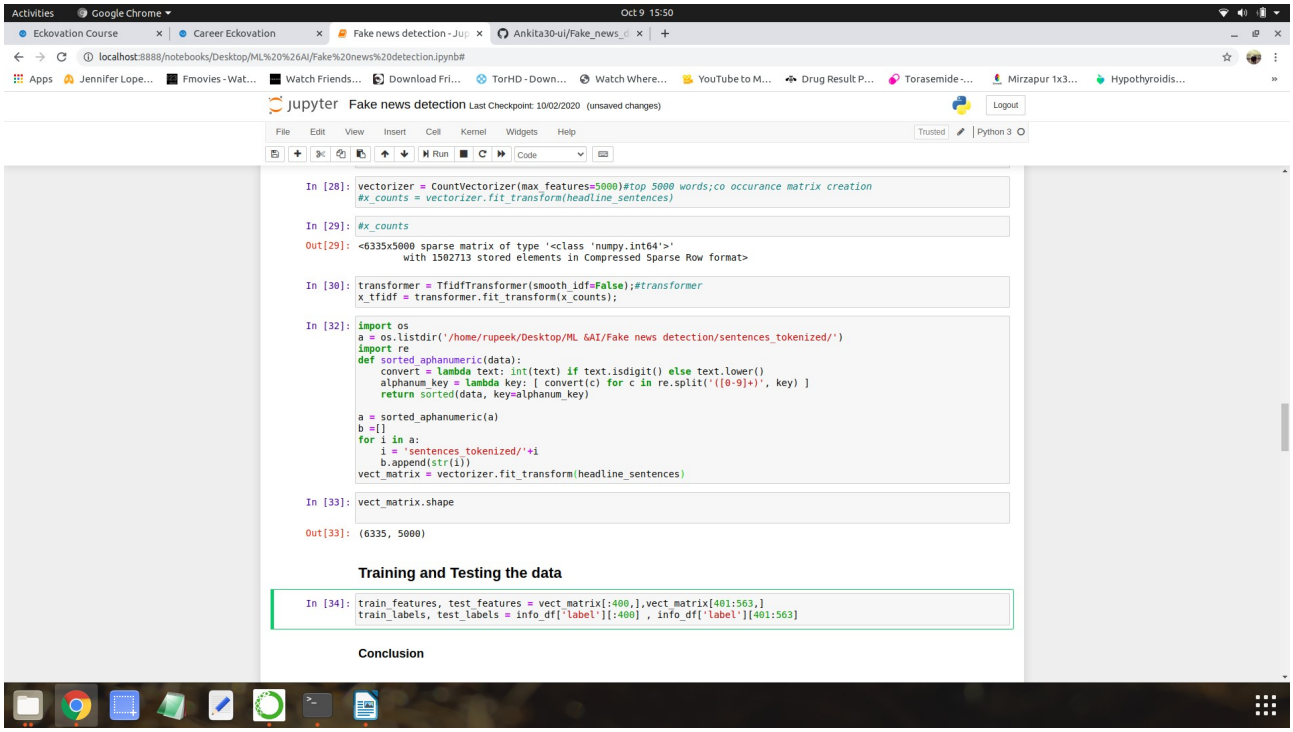
	text	Refined_text	Word_Count
0	Daniel Greenfield, a Shillman Journalism Fello...	Daniel Greenfield, Shillman Journalism Fellow ...	765
1	Google Pinterest Digg LinkedIn Reddit Stumbleu...	Google Pinterest Digg LinkedIn Reddit Stumbleu...	272
2	U.S. Secretary of State John F. Kerry said Mon...	U.S. Secretary State John F. Kerry said Monday...	259
3	— Kaydee King (@KaydeeKing) November 9, 2016 T...	— Kaydee King (@KaydeeKing) November 9, 2016 T...	277
4	It's primary day in New York and front-runners...	It's primary day New York front-runners Hillar...	197

In [25]:
```
data_text['Word_Count'].describe()
```

Out[25]:
```
count     6335.000000
mean       480.641515
std        519.722037
min          0.000000
25%        182.000000
50%        373.000000
75%        628.000000
max      12212.000000
Name: Word_Count, dtype: float64
```

Fake news....ipynb | Car_sales_....ipynb | new | Show all

4. Training and Testing the data



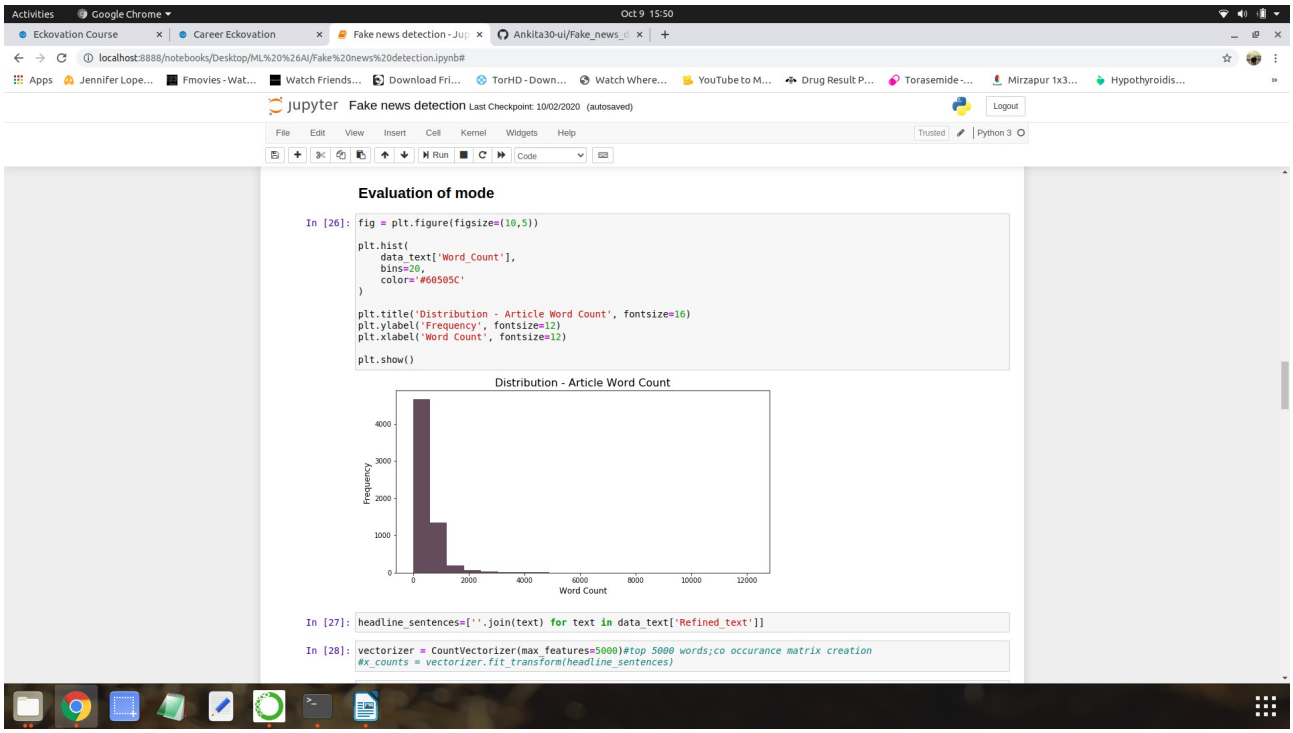
```
In [28]: vectorizer = CountVectorizer(max_features=5000)#top 5000 words;co occurance matrix creation
         #x_counts = vectorizer.fit_transform(headline_sentences)

In [29]: #x_counts

Out[29]: <6335x5000 sparse matrix of type '<class 'numpy.int64'>'
                 with 1502713 stored elements in Compressed Sparse Row format>

In [30]: transformer = TfidfTransformer(smooth_idf=False);#transformer
         x_tfidf = transformer.fit_transform(x_counts);

In [32]: import os
         a = os.listdir('/home/rupeek/Desktop/ML &AI/Fake news detection/sentences_tokenized/')
         import re
         def sorted_aphanumeric(data):
             convert = lambda text: int(text) if text.isdigit() else text.lower()
             alphanum_key = lambda key: [ convert(c) for c in re.split('([0-9]+)', key) ]
             return sorted(data, key=alphanum_key)

         a = sorted_aphanumeric(a)
         b =[]
         for i in a:
             i = 'sentences_tokenized/'+i
             b.append(str(i))
         vect_matrix = vectorizer.fit_transform(headline_sentences)

In [33]: vect_matrix.shape

Out[33]: (6335, 5000)
```

Training and Testing the data

```
In [34]: train_features, test_features = vect_matrix[:400,],vect_matrix[401:563,]
         train_labels, test_labels = info_df['label'][:400] , info_df['label'][401:563]
```

Conclusion

5. Evaluation of mode



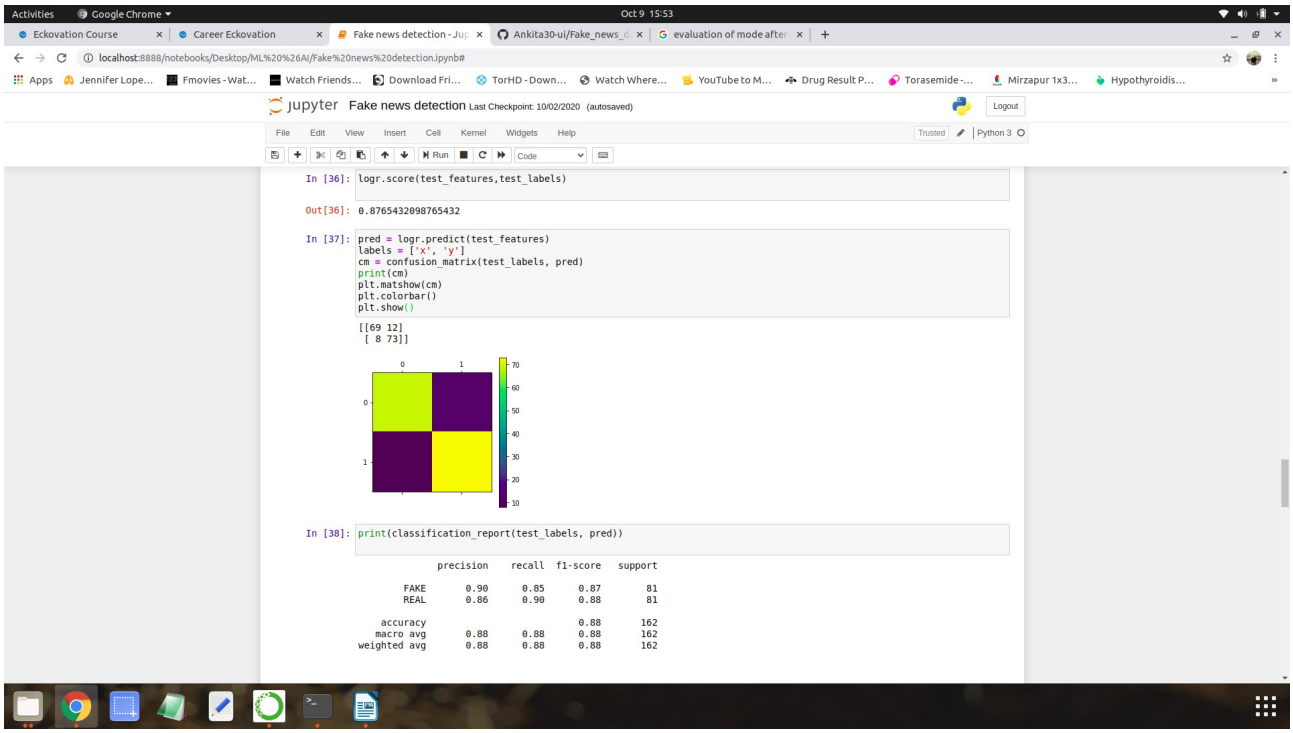
Evaluation of mode

```
In [26]: fig = plt.figure(figsize=(10,5))

         plt.hist(
             data_text['Word_Count'],
             bins=20,
             color='#60505C'
         )

         plt.title('Distribution - Article Word Count', fontsize=16)
         plt.ylabel('Frequency', fontsize=12)
         plt.xlabel('Word Count', fontsize=12)

         plt.show()
```

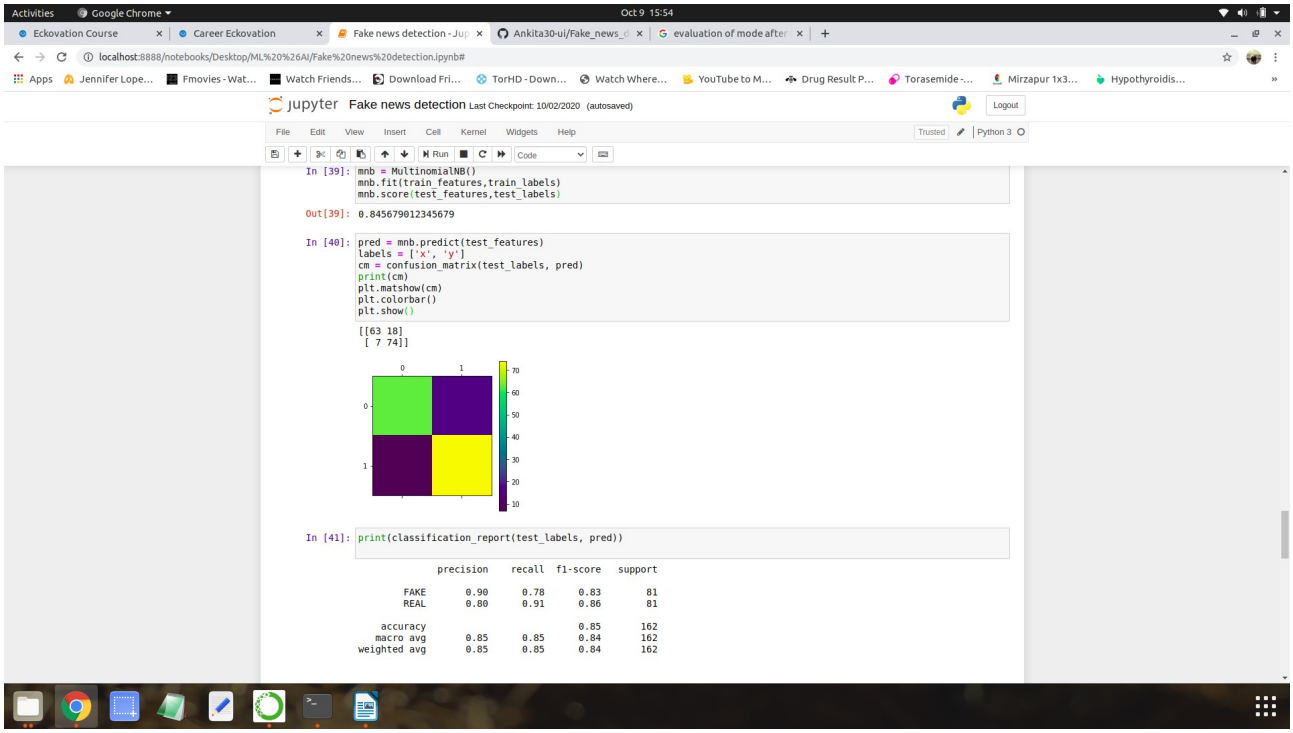


```
In [27]: headline_sentences=[''.join(text) for text in data_text['Refined_text']]

In [28]: vectorizer = CountVectorizer(max_features=5000)#top 5000 words;co occurance matrix creation
         #x_counts = vectorizer.fit_transform(headline_sentences)
```

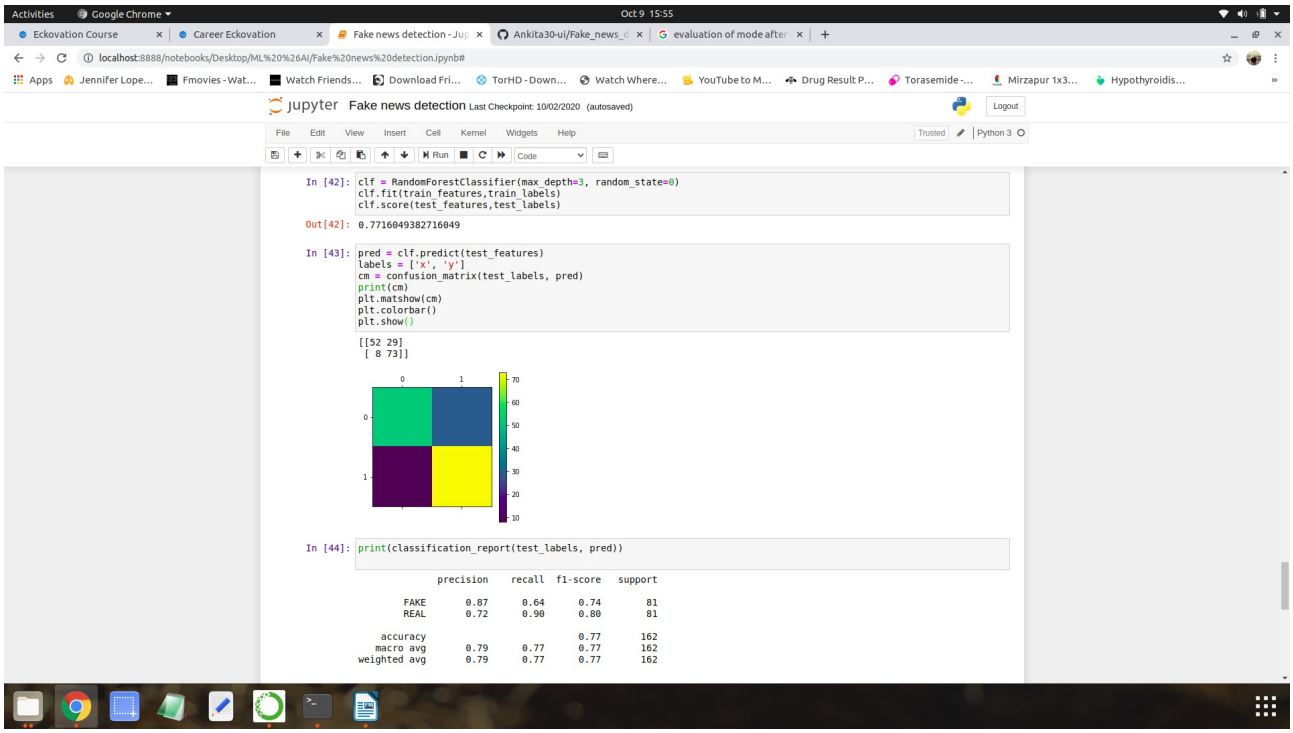
6. Logistic regression



7. MultinomialNB:

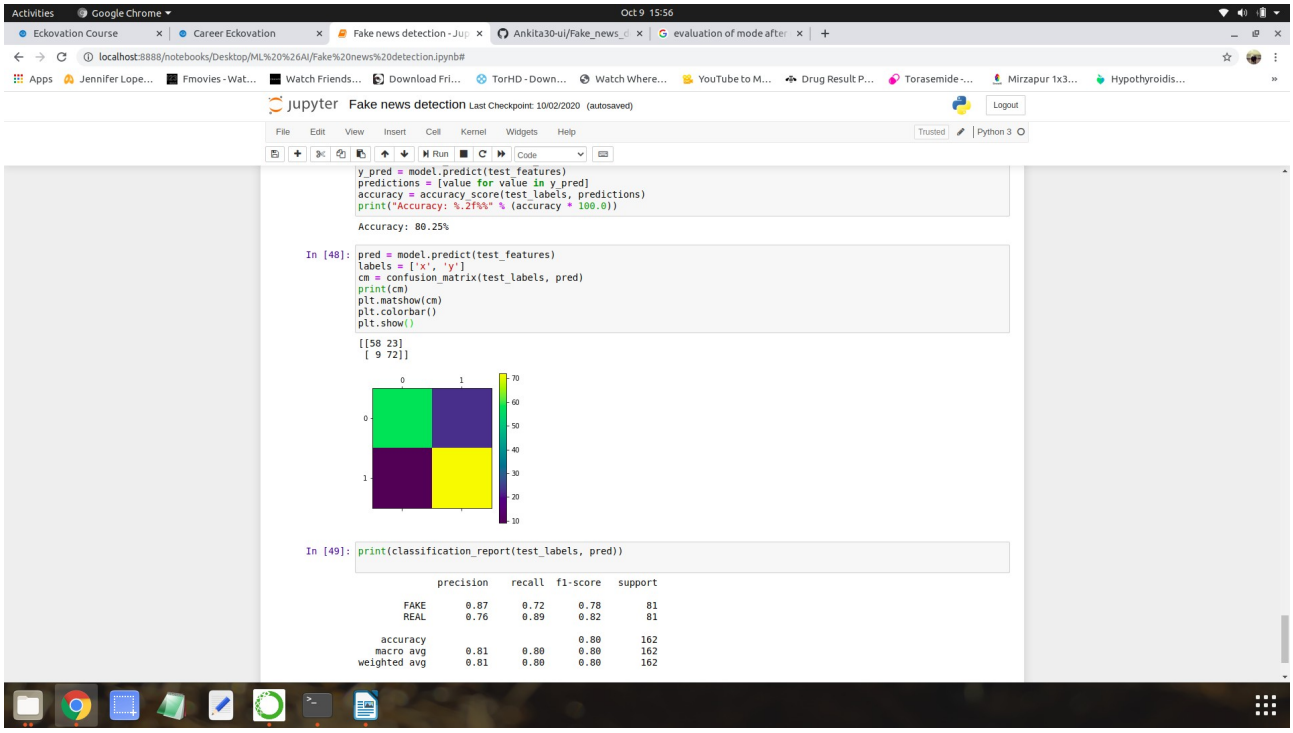


8. Random forest:



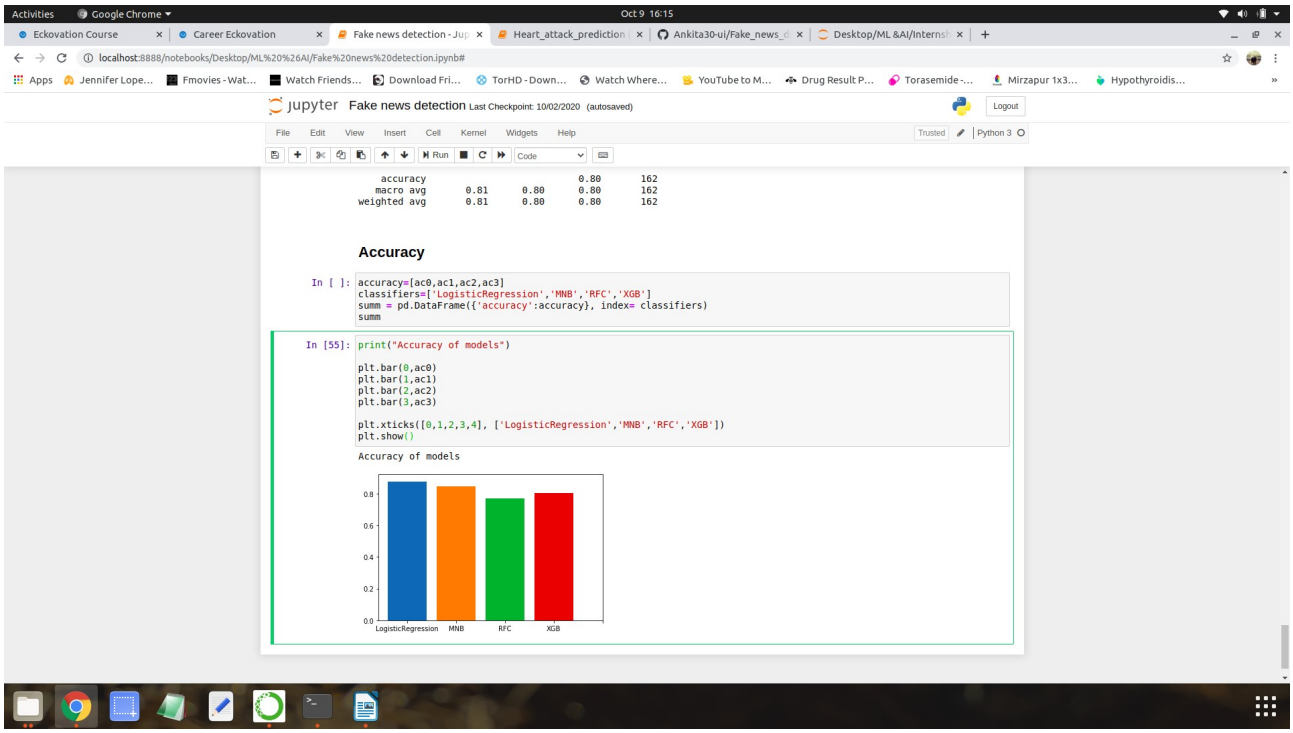
```
In [42]: clf = RandomForestClassifier(max_depth=3, random_state=0)
         clf.fit(train_features,train_labels)
         clf.score(test_features,test_labels)

Out[42]: 0.7716049382716049

In [43]: pred = clf.predict(test_features)
         labels = ['x', 'y']
         cm = confusion_matrix(test_labels, pred)
         print(cm)
         plt.matshow(cm)
         plt.colorbar()
         plt.show()

         [[52 29]
          [ 8 73]]

In [44]: print(classification_report(test_labels, pred))

                    precision    recall  f1-score   support

              FAKE       0.87      0.64      0.74        81
              REAL       0.72      0.90      0.80        81

          accuracy                           0.77       162
         macro avg       0.79      0.77      0.77       162
      weighted avg       0.79      0.77      0.77       162
```

9. XGB:



```
         y_pred = model.predict(test_features)
         predictions = [value for value in y_pred]
         accuracy = accuracy_score(test_labels, predictions)
         print("Accuracy: %.2f%%" % (accuracy * 100.0))

         Accuracy: 80.25%

In [48]: pred = model.predict(test_features)
         labels = ['x', 'y']
         cm = confusion_matrix(test_labels, pred)
         print(cm)
         plt.matshow(cm)
         plt.colorbar()
         plt.show()

         [[58 23]
          [ 9 72]]

In [49]: print(classification_report(test_labels, pred))

                    precision    recall  f1-score   support

              FAKE       0.87      0.72      0.78        81
              REAL       0.76      0.89      0.82        81

          accuracy                           0.80       162
         macro avg       0.81      0.80      0.80       162
      weighted avg       0.81      0.80      0.80       162
```

10. Accuracy


Activities
Google Chrome
Oct 9 16:15
Eckovation Course
Career Eckovation
Fake news detection - Jup
Heart_attack_prediction
Ankita30-ui/Fake_news_d
Desktop/ML &AI/Internsh
localhost:8888/notebooks/Desktop/ML%20%26AI/Fake%20news%20detection.ipynb#
Apps
Jennifer Lope...
Fmovies - Wat...
Watch Friends...
Download Fri...
TorHD - Down...
Watch Where...
YouTube to M...
Drug Result P...
Torasemide -...
Mirzapur 1x3...
Hypothyroidis...
jupyter
Fake news detection Last Checkpoint: 10/02/2020 (autosaved)
Logout
File
Edit
View
Insert
Cell
Kernel
Widgets
Help
Trusted
Python 3
Run
Code
accuracy 0.80 162
macro avg 0.81 0.80 0.80 162
weighted avg 0.81 0.80 0.80 162
Accuracy
In [ ]:
accuracy=[ac0,ac1,ac2,ac3]
classifiers=['LogisticRegression','MNB','RFC','XGB']
summ = pd.DataFrame({'accuracy':accuracy}, index= classifiers)
summ
In [55]:
print("Accuracy of models")
plt.bar(0,ac0)
plt.bar(1,ac1)
plt.bar(2,ac2)
plt.bar(3,ac3)
plt.xticks([0,1,2,3,4], ['LogisticRegression','MNB','RFC','XGB'])
plt.show()
Accuracy of models
0.8
0.6
0.4
0.2
0.0
LogisticRegression
MNB
RFC
XGB