```
! pip install kaggle
! mkdir ~/.kaggle
```

```
Looking in indexes: https://pypi.org/simple, https://us-python.pkg.dev/colab-wheels/public/simple/
Requirement already satisfied: kaggle in /usr/local/lib/python3.7/dist-packages (1.5.12)
Requirement already satisfied: urllib3 in /usr/local/lib/python3.7/dist-packages (from kaggle) (1
Requirement already satisfied: requests in /usr/local/lib/python3.7/dist-packages (from kaggle) (
Requirement already satisfied: python-slugify in /usr/local/lib/python3.7/dist-packages (from kag
Requirement already satisfied: six>=1.10 in /usr/local/lib/python3.7/dist-packages (from kaggle)
Requirement already satisfied: python-dateutil in /usr/local/lib/python3.7/dist-packages (from ka
Requirement already satisfied: certifi in /usr/local/lib/python3.7/dist-packages (from kaggle) (20
Requirement already satisfied: tqdm in /usr/local/lib/python3.7/dist-packages (from kaggle) (4.64
Requirement already satisfied: text-unidecode>=1.3 in /usr/local/lib/python3.7/dist-packages (from
Requirement already satisfied: idna<3,>=2.5 in /usr/local/lib/python3.7/dist-packages (from reque
Requirement already satisfied: chardet<4,>=3.0.2 in /usr/local/lib/python3.7/dist-packages (from
```

```
! cp kaggle.json ~/.kaggle/
! chmod 600 ~/.kaggle/kaggle.json
! kaggle datasets download -d ashwithanoble/phishing-sites-url
```

```
cp: cannot stat 'kaggle.json': No such file or directory
chmod: cannot access '/root/.kaggle/kaggle.json': No such file or directory
Traceback (most recent call last):
  File "/usr/local/bin/kaggle", line 5, in <module>
    from kaggle.cli import main
  File "/usr/local/lib/python3.7/dist-packages/kaggle/__init__.py", line 23, in <module>
    api.authenticate()
  File "/usr/local/lib/python3.7/dist-packages/kaggle/api/kaggle_api_extended.py", line 166, in au
    self.config_file, self.config_dir))
OSError: Could not find kaggle.json. Make sure it's located in /root/.kaggle. Or use the environme
```

```
!cp /content/drive/MyDrive/kaggle.json  ~/.kaggle/kaggle.json
```

```
#downloading dataset
! kaggle datasets download -d ashwithanoble/phishing-sites-url
```

```
Downloading phishing-sites-url.zip to /content
 59% 5.00M/8.52M [00:00<00:00, 50.9MB/s]
100% 8.52M/8.52M [00:00<00:00, 30.4MB/s]
```

```
#unzipping the file
!unzip phishing-sites-url.zip
```

```
Archive:  phishing-sites-url.zip
  inflating: urls_for_phishing.csv
```

```
! pip install selenium
```

```
Looking in indexes: https://pypi.org/simple, https://us-python.pkg.dev/colab-wheels/public/simple/
Collecting selenium
  Downloading selenium-4.4.3-py3-none-any.whl (985 kB)
     |████████████████████████████████| 985 kB 33.5 MB/s
Requirement already satisfied: certifi>=2021.10.8 in /usr/local/lib/python3.7/dist-packages (from
Collecting trio-websocket~=0.9
  Downloading trio_websocket-0.9.2-py3-none-any.whl (16 kB)
Collecting trio~=0.17
  Downloading trio-0.21.0-py3-none-any.whl (358 kB)
     |████████████████████████████████| 358 kB 68.0 MB/s
Collecting urllib3[socks]~=1.26
  Downloading urllib3-1.26.12-py2.py3-none-any.whl (140 kB)
     |████████████████████████████████| 140 kB 72.6 MB/s
Collecting outcome
  Downloading outcome-1.2.0-py2.py3-none-any.whl (9.7 kB)
```

```
Requirement already satisfied: sortedcontainers in /usr/local/lib/python3.7/dist-packages (from t
Collecting async-generator>=1.9
  Downloading async_generator-1.10-py3-none-any.whl (18 kB)
Collecting sniffio
  Downloading sniffio-1.3.0-py3-none-any.whl (10 kB)
Requirement already satisfied: idna in /usr/local/lib/python3.7/dist-packages (from trio~=0.17->se
Requirement already satisfied: attrs>=19.2.0 in /usr/local/lib/python3.7/dist-packages (from trio
Collecting wsproto>=0.14
  Downloading wsproto-1.2.0-py3-none-any.whl (24 kB)
Requirement already satisfied: PySocks!=1.5.7,<2.0,>=1.5.6 in /usr/local/lib/python3.7/dist-packa
Collecting h11<1,>=0.9.0
  Downloading h11-0.13.0-py3-none-any.whl (58 kB)
     |████████████████████████████████| 58 kB 6.4 MB/s
Requirement already satisfied: typing-extensions in /usr/local/lib/python3.7/dist-packages (from l
Installing collected packages: sniffio, outcome, h11, async-generator, wsproto, urllib3, trio, tr
  Attempting uninstall: urllib3
    Found existing installation: urllib3 1.24.3
    Uninstalling urllib3-1.24.3:
      Successfully uninstalled urllib3-1.24.3
ERROR: pip's dependency resolver does not currently take into account all the packages that are i
requests 2.23.0 requires urllib3!=1.25.0,!=1.25.1,<1.26,>=1.21.1, but you have urllib3 1.26.12 wh
Successfully installed async-generator-1.10 h11-0.13.0 outcome-1.2.0 selenium-4.4.3 sniffio-1.3.0
```

```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
%matplotlib inline
import time

from sklearn.linear_model import LogisticRegression
from sklearn.naive_bayes import MultinomialNB
from sklearn.model_selection import train_test_split
from sklearn.metrics import classification_report
from sklearn.metrics import confusion_matrix
from nltk.tokenize import RegexpTokenizer
from nltk.stem.snowball import SnowballStemmer
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.pipeline import make_pipeline
from PIL import Image
from bs4 import BeautifulSoup
from selenium import webdriver
import networkx as nx
import pickle
import warnings
warnings.filterwarnings('ignore')
```

```
data=pd.read_csv('urls_for_phishing.csv')#loading the dataset
data.head(5)
```

	URL	Label
0	nobell.it/70ffb52d079109dca5664cce6f317373782/...	bad
1	www.dghjdgf.com/paypal.co.uk/cycgi-bin/webscrc...	bad
2	serviciosbys.com/paypal.cgi.bin.get-into.herf....	bad
3	mail.printakid.com/www.online.americanexpress....	bad
4	thewhiskeydregs.com/wp-content/themes/widescre...	bad

```
data.tail(5)
```

	URL	Label
507107	23.227.196.215/	bad
507108	apple-checker.org/	bad
507109	apple-iclods.org/	bad

```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 507112 entries, 0 to 507111
Data columns (total 2 columns):
 #   Column  Non-Null Count   Dtype 
---  ------  --------------   ----- 
 0   URL     507112 non-null  object
 1   Label   507112 non-null  object
dtypes: object(2)
memory usage: 7.7+ MB
```

```
data.shape
```

```
(507112, 2)
```

```
data.isnull().sum()
```

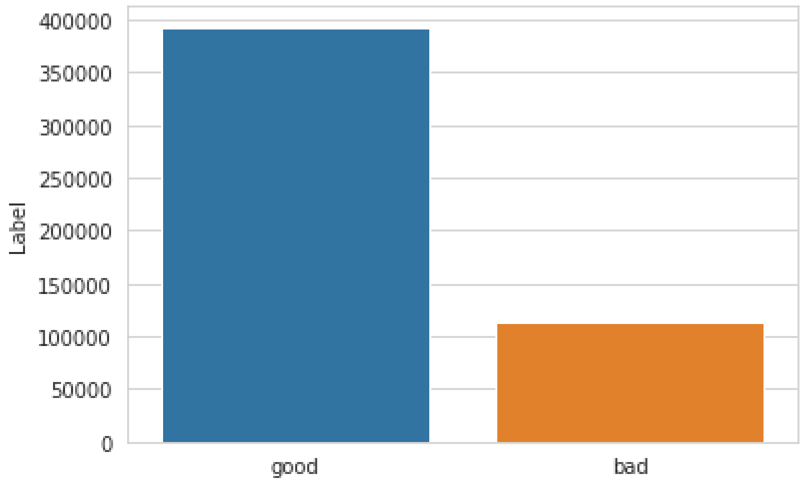
```
URL      0
Label    0
dtype: int64
```

```
data.duplicated().sum()
```

```
0
```

```
label_counts = pd.DataFrame(data.Label.value_counts())
sns.set_style('whitegrid')
sns.barplot(label_counts.index,label_counts.Label)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f9402384ed0>
```



TOKENIZING

```
tokenizer = RegexpTokenizer(r'[A-Za-z]+')
data.URL[0]
```

```
'nobell.it/70ffb52d079109dca5664cce6f317373782/login.SkyPe.com/en/cgi-bin/verification/login/70ff
b52d079109dca5664cce6f317373/index.php?cmd=_profile-ach&outdated_page_tmpl=p/gen/failed-to-load&n
av=0.5.1&login_access=1322408526'
```

```
# this will be pull letter which matches to expression
tokenizer.tokenize(data.URL[0]) # using first row
```

```
['nobell',
 'it',
 'ffb',
 'd',
 'dca',
 'cce',
 'f',
 'login',
 'SkyPe',
 'com',
 'en',
 'cgi',
 'bin',
 'verification',
 'login',
 'ffb',
 'd',
 'dca',
 'cce',
 'f',
 'index',
 'php',
 'cmd',
 'profile',
 'ach',
 'outdated',
 'page',
 'tmpl',
 'p',
 'gen',
 'failed',
 'to',
 'load',
 'nav',
 'login',
 'access']
```

```
print('Getting words tokenized ...')
t0= time.perf_counter()
data['text_tokenized'] = data.URL.map(lambda t: tokenizer.tokenize(t)) # doing with all rows
t1 = time.perf_counter() - t0
print('Time taken',t1 ,'sec')
```

```
Getting words tokenized ...
Time taken 2.502414115999997 sec
```

```
data.sample(5)
```

	URL	Label	text_tokenized
222314	pilotnewsmag.com/?p=11684	good	[pilotnewsmag, com, p]
186621	genforum.genealogy.com/bethune/	good	[genforum, genealogy, com, bethune]
480656	mxp2094.com	bad	[mxp, com]
23595	paypal.com.cgi.bin.webscr.cmd.flow.session.loh...	bad	[paypal, com, cgi, bin, webscr, cmd, flow, ses...
96109	stormpages.com/script/ping.txt	bad	[stormpages, com, script, ping, txt]

STEMMING

```
stemmer = SnowballStemmer("english")
print('Getting words stemmed ...')
t0= time.perf_counter()
data['text_stemmed'] = data['text_tokenized'].map(lambda l: [stemmer.stem(word) for word in l])
t1= time.perf_counter() - t0
print('Time taken',t1 ,'sec')
```

```
Getting words stemmed ...
Time taken 54.80282458100001 sec
```

```
data.sample(5)
```

	URL	Label	text_tokenized	text_stemmed
338696	gavinphotography.com/locations/los-angeles/los...	good	[gavinphotography, com, locations, los, angele...	[gavinphotographi, com, locat, los, angel, los...
74751	www.pitt.edu/~csna/Milligan/readme.html	good	[www, pitt, edu, csna, Milligan, readme, html]	[www, pitt, edu, csna, milligan, readm, html]
279850	auctionscc.com/auction/baseball-cards/	good	[auctionscc, com, auction, baseball, cards]	[auctionscc, com, auction, basebal, card]
58308	www.obis.com/agent/	good	[www, obis, com, agent]	[www, obi, com, agent]

```
print('Getting joiningwords ...')
t0= time.perf_counter()
data['text_sent'] = data['text_stemmed'].map(lambda l: ' '.join(l))
t1= time.perf_counter() - t0
print('Time taken',t1 ,'sec')
```

```
Getting joiningwords ...
Time taken 0.19705132799998637 sec
```

```
data.sample(5)
```

	URL	Label	text_tokenized	text_stemmed	text_se
122779	nbawallpaper.org/sess/edaeeaef161d13abf3a14adf...	bad	[nbawallpaper, org, sess, edaeeaef, d, abf, a,...	[nbawallpap, org, sess, edaeeaef, d, abf, a, a...	nbawallp org se edaeeae abf a ad
439120	thefullwiki.org/List_of_icebreakers	good	[thefullwiki, org, List, of, icebreakers]	[thefullwiki, org, list, of, icebreak]	thefullw org list icebre

VISUALIZATION

```
bad_sites = data[data.Label == 'bad']
good_sites =data[data.Label == 'good']
bad_sites.head(5)
```

	URL	Label	text_tokenized	text_stemmed	text_s
0	nobell.it/70ffb52d079109dca5664cce6f317373782/...	bad	[nobell, it, ffb, d, dca, cce, f, login, SkyPe...	[nobel, it, ffb, d, dca, cce, f, login, skype,...	nobel it ffb d cce f login sl com en
1	www.dghjdgf.com/paypal.co.uk/cycgi-bin/webscrc...	bad	[www, dghjdgf, com, paypal, co, uk, cycgi, bin...	[www, dghjdgf, com, paypal, co, uk, cycgi, bin...	www dgh com payp uk cycg websc
2	serviciosbys.com/paypal.cgi.bin.get-into.herf....	bad	[serviciosbys, com, paypal, cgi, bin, get, int...	[serviciosbi, com, paypal, cgi, bin, get, into...	serviciosbi paypal cg get into her

```
good_sites.head(5)
```

	URL	Label	text_tokenized	text_stemmed	text_s
18227	esxcc.com/js/index.htm?us.battle.net/noghn/en/...	good	[esxcc, com, js, index, htm, us, battle, net, ...	[esxcc, com, js, index, htm, us, battl, net, n...	esxcc js ir htr batt noghı

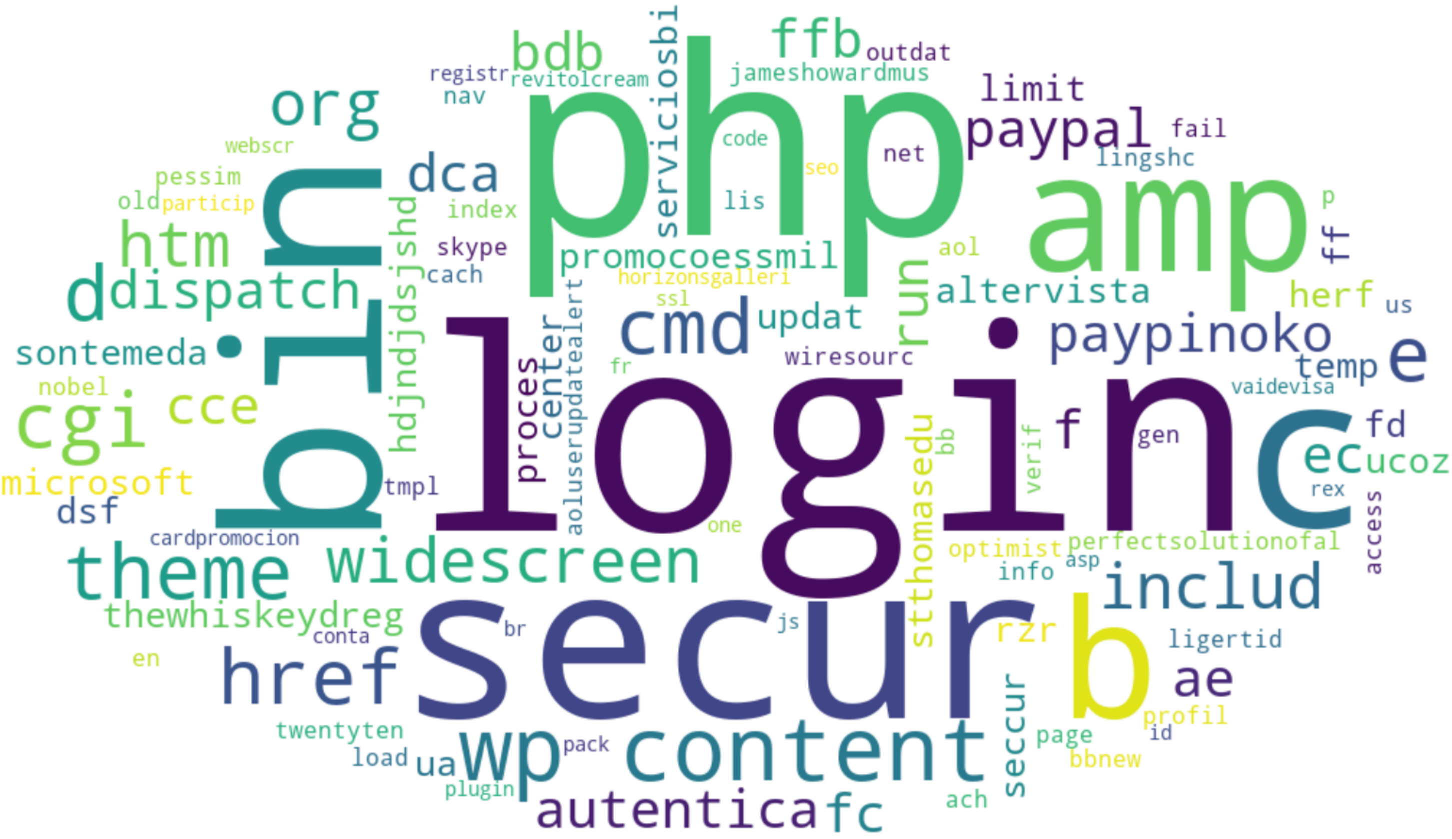
```
from os import path
from wordcloud import WordCloud, STOPWORDS
def google_authenticate():
  # Authenticate first so the Google Drive library can detect your credentials.
  from google.colab import auth
  auth.authenticate_user()

  from googleapiclient.discovery import build
  drive_service = build('drive', 'v3')
  return drive_service
drive_service = google_authenticate()
```

```
def read_file(file_id):
  file_id = file_id
  import io
  from googleapiclient.http import MediaIoBaseDownload
  request = drive_service.files().get_media(fileId=file_id)
  downloaded = io.BytesIO()
  downloader = MediaIoBaseDownload(downloaded, request)
  done = False
  while done is False:
    _, done = downloader.next_chunk()
  downloaded.seek(0)
  return downloaded
```

```
text_file = read_file("1SvLFtrpbxWgP7OTh5USrYQPebQRcCSEk")
document = text_file.read().decode('utf-8')
print(len(document))
print(document[0:100])
```

```
1037
thewhiskeydreg com wp content theme widescreen includ temp promocoessmil hdjndjdsjshd stthomasedu
```

```


#image_file = read_file("1egaK6EKgqPnYzswtS679-6NpZKRASSNQ")
# create mask
image_file = read_file("1jXEEsqgQ5M4nehxBUeCM3Qedi4LitdIM")
alice_mask = np.array(Image.open(image_file))

# remove stopwords
stopwords = set(STOPWORDS)
stopwords.add("said")

# generate word cloud
wc = WordCloud(background_color="white", max_words=2000, mask=alice_mask,stopwords=stopwords)
wc.generate(document)
# plot the word cloud
plt.figure(figsize=(20,10), dpi=120)
plt.imshow(wc, interpolation='bilinear')
plt.axis("off")
plt.show()
```



SCRAPE WEBSITE

```
!apt-get update # to update ubuntu to correctly run apt install
!apt install chromium-chromedriver
!cp /usr/local/bin/chromedriver.exe /usr/bin
```

```
Hit:10 http://ppa.launchpad.net/cran/libgit2/ubuntu bionic InRelease
Hit:11 http://ppa.launchpad.net/deadsnakes/ppa/ubuntu bionic InRelease
Get:12 http://archive.ubuntu.com/ubuntu bionic-updates/universe amd64 Packages [2,318 kB]
Hit:13 http://ppa.launchpad.net/graphics-drivers/ppa/ubuntu bionic InRelease
Get:14 http://archive.ubuntu.com/ubuntu bionic-updates/main amd64 Packages [3,397 kB]
Get:15 http://ppa.launchpad.net/c2d4u.team/c2d4u4.0+/ubuntu bionic/main Sources [2,105 kB]
Get:17 http://security.ubuntu.com/ubuntu bionic-security/universe amd64 Packages [1,540 kB]
Get:18 http://ppa.launchpad.net/c2d4u.team/c2d4u4.0+/ubuntu bionic/main amd64 Packages [1,079 k
Get:19 http://security.ubuntu.com/ubuntu bionic-security/main amd64 Packages [2,965 kB]
Fetched 13.7 MB in 3s (5,350 kB/s)
Reading package lists... Done
Reading package lists... Done
Building dependency tree
Reading state information... Done
The following package was automatically installed and is no longer required:

  libnvidia-common-460
Use 'apt autoremove' to remove it.
The following additional packages will be installed:
  chromium-browser chromium-browser-l10n chromium-codecs-ffmpeg-extra
Suggested packages:
  webaccounts-chromium-extension unity-chromium-extension
The following NEW packages will be installed:
  chromium-browser chromium-browser-l10n chromium-chromedriver
  chromium-codecs-ffmpeg-extra
```

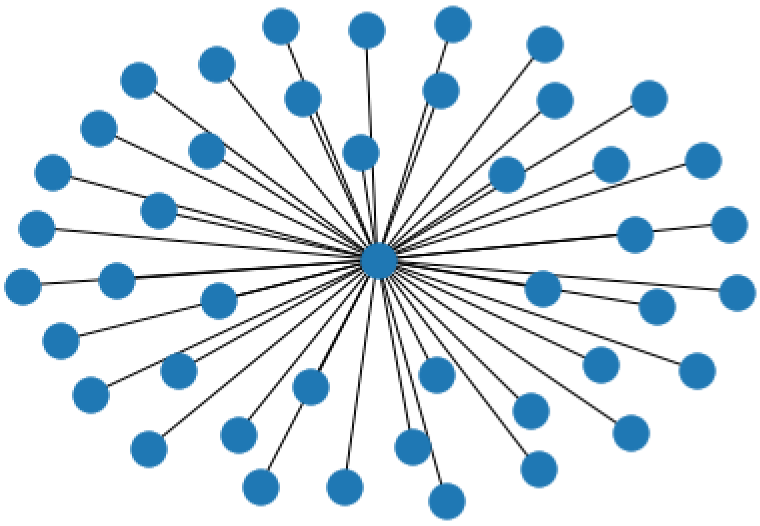
```
0 upgraded, 4 newly installed, 0 to remove and 49 not upgraded.
Need to get 91.4 MB of archives.
After this operation, 309 MB of additional disk space will be used.
Get:1 http://archive.ubuntu.com/ubuntu bionic-updates/universe amd64 chromium-codecs-ffmpeg-ext
Get:2 http://archive.ubuntu.com/ubuntu bionic-updates/universe amd64 chromium-browser amd64 104
Get:3 http://archive.ubuntu.com/ubuntu bionic-updates/universe amd64 chromium-browser-l10n all
Get:4 http://archive.ubuntu.com/ubuntu bionic-updates/universe amd64 chromium-chromedriver amd6
Fetched 91.4 MB in 1s (61.3 MB/s)
Selecting previously unselected package chromium-codecs-ffmpeg-extra.
(Reading database ... 155685 files and directories currently installed.)
Preparing to unpack .../chromium-codecs-ffmpeg-extra_104.0.5112.101-0ubuntu0.18.04.1_amd64.deb
Unpacking chromium-codecs-ffmpeg-extra (104.0.5112.101-0ubuntu0.18.04.1) ...
Selecting previously unselected package chromium-browser.
Preparing to unpack .../chromium-browser_104.0.5112.101-0ubuntu0.18.04.1_amd64.deb ...
Unpacking chromium-browser (104.0.5112.101-0ubuntu0.18.04.1) ...
Selecting previously unselected package chromium-browser-l10n.
Preparing to unpack .../chromium-browser-l10n_104.0.5112.101-0ubuntu0.18.04.1_all.deb ...
Unpacking chromium-browser-l10n (104.0.5112.101-0ubuntu0.18.04.1) ...
Selecting previously unselected package chromium-chromedriver.
Preparing to unpack .../chromium-chromedriver_104.0.5112.101-0ubuntu0.18.04.1_amd64.deb ...
Unpacking chromium-chromedriver (104.0.5112.101-0ubuntu0.18.04.1) ...
Setting up chromium-codecs-ffmpeg-extra (104.0.5112.101-0ubuntu0.18.04.1) ...
Setting up chromium-browser (104.0.5112.101-0ubuntu0.18.04.1) ...
update-alternatives: using /usr/bin/chromium-browser to provide /usr/bin/x-www-browser (x-www-b
update-alternatives: using /usr/bin/chromium-browser to provide /usr/bin/gnome-www-browser (gno
Setting up chromium-chromedriver (104.0.5112.101-0ubuntu0.18.04.1) ...
Setting up chromium-browser-l10n (104.0.5112.101-0ubuntu0.18.04.1) ...
Processing triggers for man-db (2.8.3-2ubuntu0.1) ...
Processing triggers for hicolor-icon-theme (0.17-2) ...
Processing triggers for mime-support (3.60ubuntu1) ...
Processing triggers for libc-bin (2.27-3ubuntu1.5) ...
cp: cannot stat '/usr/local/bin/chromedriver.exe': No such file or directory
```

```
import sys
sys.path.insert(0,'/usr/local/bin/chromedriver.exe')
options = webdriver.ChromeOptions()
options.add_argument('--headless')
options.add_argument('--no-sandbox')
options.add_argument('--disable-dev-shm-usage')
browser= webdriver.Chrome('chromedriver',options=options)
browser.get("https://www.website.com")
import requests
import textwrap
list_urls = ['http://www.ezeephones.com/','http://www.ezeephones.com/about-us'] #here i take phishing s
pagelinks = []
for url in list_urls:
    browser.get(url)
    soup = BeautifulSoup(browser.page_source,"html.parser")
    for line in soup.find_all('a'):
        href = line.get('href')
        pagelinks.append([url, href])
df = pd.DataFrame(pagelinks, columns=["from", "to"])
```

```
df.head()
```

	from	to
0	http://www.ezeephones.com/	/
1	http://www.ezeephones.com/	javascript:void(0);
2	http://www.ezeephones.com/	/news
3	http://www.ezeephones.com/	/Entertainment
4	http://www.ezeephones.com/	/Money

```
GA = nx.from_pandas_edgelist(df, source="from", target="to")
nx.draw(GA, with_labels=False)
```



```
cv = CountVectorizer()
feature = cv.fit_transform(data.text_sent)
feature[:5].toarray()
```

```
     array([[0, 0, 0, ..., 0, 0, 0],
            [0, 0, 0, ..., 0, 0, 0],
            [0, 0, 0, ..., 0, 0, 0],
            [0, 0, 0, ..., 0, 0, 0],
            [0, 0, 0, ..., 0, 0, 0]])
```

```
from sklearn.datasets import make_classification
X,Y= make_classification(n_samples=100000, n_features=20, n_informative=17, n_redundant=3, random_state
```

```
Xtrain, Xtest, Ytrain, Ytest = train_test_split(X,Y,test_size=0.3, random_state=1)
```

```
Xtrain.shape
```

```
     (70000, 20)
```

```
Ytrain.shape
```

```
     (70000,)
```

```
Xtest.shape
```

```
     (30000, 20)
```

```
Ytest.shape
```

```
     (30000,)
```

```
#importing packages
from sklearn.metrics import accuracy_score
# Creating holders to store the model performance results
ML_Model = []
acc_train = []
acc_test = []

#function to call for storing the results
def storeResults(model, a,b):
  ML_Model.append(model)
  acc_train.append(round(a,2))
  acc_test.append(round(b,2))
```

LOGISTIC REGRESSION

```
lr = LogisticRegression()
lr.fit(Xtrain,Ytrain)
```

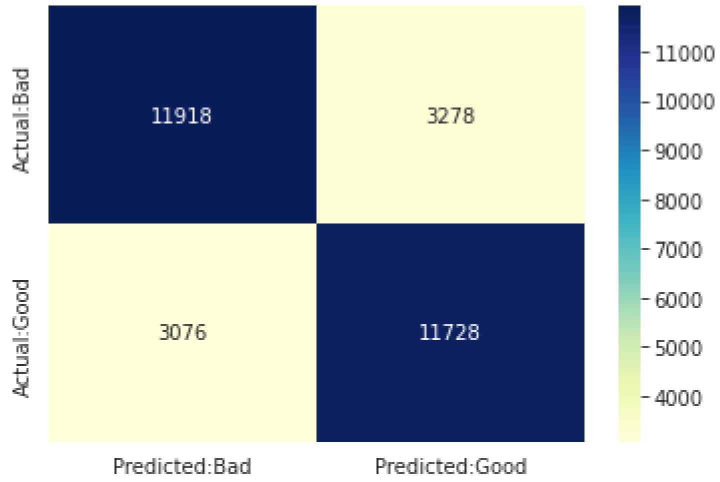
```
LogisticRegression()
```

```
lr.score(Xtest,Ytest)
```

```
0.7882
```

```
Scores_ml = {}
Scores_ml['Logistic Regression'] = np.round(lr.score(Xtest,Ytest),2)
print('Training Accuracy :',lr.score(Xtrain,Ytrain))
print('Testing Accuracy :',lr.score(Xtest,Ytest))
con_mat = pd.DataFrame(confusion_matrix(lr.predict(Xtest),Ytest),
            columns = ['Predicted:Bad', 'Predicted:Good'],
            index = ['Actual:Bad', 'Actual:Good'])
print('\nCLASSIFICATION REPORT\n')
print(classification_report(lr.predict(Xtest),Ytest,target_names =['Bad','Good']))
print('\nCONFUSION MATRIX')
plt.figure(figsize= (6,4))
sns.heatmap(con_mat, annot = True,fmt='d',cmap="YlGnBu")
```

```
Training Accuracy : 0.7892428571428571
Testing Accuracy : 0.7882

CLASSIFICATION REPORT

              precision    recall  f1-score   support

         Bad       0.79      0.78      0.79     15196
        Good       0.78      0.79      0.79     14804

    accuracy                           0.79     30000
   macro avg       0.79      0.79      0.79     30000
weighted avg       0.79      0.79      0.79     30000


CONFUSION MATRIX
<matplotlib.axes._subplots.AxesSubplot at 0x7f93cf8b7a10>
```



```
acc_train_lr = lr.score(Xtrain,Ytrain)
acc_test_lr = lr.score(Xtest,Ytest)
storeResults('Logistic Regression', acc_train_lr, acc_test_lr)
```

MULTINOMIAL NB

```
from sklearn.preprocessing import MinMaxScaler #fixed import
scaler = MinMaxScaler()
Xtrain = scaler.fit_transform(Xtrain)
Xtest = scaler.transform(Xtest)
mnb = MultinomialNB()
mnb.fit(Xtrain,Ytrain)
```

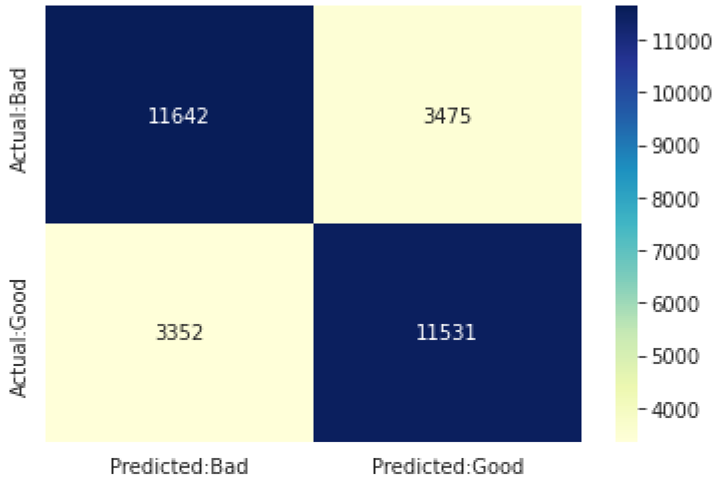
```
MultinomialNB()
```

```
mnb.score(Xtest,Ytest)
```

```
0.7724333333333333
```

```
Scores_ml['MultinomialNB'] = np.round(mnb.score(Xtest,Ytest),2)
print('Training Accuracy :',mnb.score(Xtrain,Ytrain))
print('Testing Accuracy :',mnb.score(Xtest,Ytest))
con_mat = pd.DataFrame(confusion_matrix(mnb.predict(Xtest),Ytest),
            columns = ['Predicted:Bad', 'Predicted:Good'],
            index = ['Actual:Bad', 'Actual:Good'])
print('\nCLASSIFICATION REPORT\n')
print(classification_report(mnb.predict(Xtest),Ytest,target_names =['Bad','Good']))
print('\nCONFUSION MATRIX')
plt.figure(figsize= (6,4))
sns.heatmap(con_mat, annot = True,fmt='d',cmap="YlGnBu")
```

```
Training Accuracy : 0.7709857142857143
Testing Accuracy : 0.7724333333333333

CLASSIFICATION REPORT

              precision    recall  f1-score   support

         Bad       0.78      0.77      0.77     15117
        Good       0.77      0.77      0.77     14883

    accuracy                           0.77     30000
   macro avg       0.77      0.77      0.77     30000
weighted avg       0.77      0.77      0.77     30000


CONFUSION MATRIX
<matplotlib.axes._subplots.AxesSubplot at 0x7f93cf880650>
```



```
acc_train_mnb = mnb.score(Xtrain,Ytrain)
acc_test_mnb = mnb.score(Xtest,Ytest)
storeResults('multinomial NB', acc_train_mnb, acc_test_mnb)
```

DECISION TREE CLASSIFIER

```
from sklearn.tree import DecisionTreeClassifier
tree = DecisionTreeClassifier(max_depth = 5)
tree.fit(Xtrain, Ytrain)
```

```
DecisionTreeClassifier(max_depth=5)
```

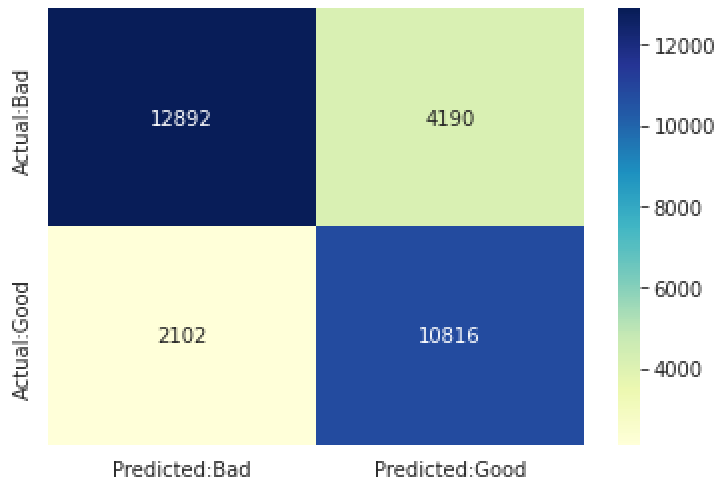
```
tree.score(Xtest,Ytest)
```

```
0.7902666666666667
```

```
Scores_ml['Decision Tree Classifier'] = np.round(tree.score(Xtest,Ytest),2)
```

```
print('Training Accuracy :',tree.score(Xtrain,Ytrain))
print('Testing Accuracy :',tree.score(Xtest,Ytest))
con_mat = pd.DataFrame(confusion_matrix(tree.predict(Xtest), Ytest),
            columns = ['Predicted:Bad', 'Predicted:Good'],
            index = ['Actual:Bad', 'Actual:Good'])


print('\nCLASSIFICATION REPORT\n')
print(classification_report(tree.predict(Xtest), Ytest,
                            target_names =['Bad','Good']))

print('\nCONFUSION MATRIX')
plt.figure(figsize= (6,4))
sns.heatmap(con_mat, annot = True,fmt='d',cmap="YlGnBu")
```

```
Training Accuracy : 0.7941857142857143
Testing Accuracy : 0.7902666666666667

CLASSIFICATION REPORT

              precision    recall  f1-score   support

         Bad       0.86      0.75      0.80     17082
        Good       0.72      0.84      0.77     12918

    accuracy                           0.79     30000
   macro avg       0.79      0.80      0.79     30000
weighted avg       0.80      0.79      0.79     30000


CONFUSION MATRIX
<matplotlib.axes._subplots.AxesSubplot at 0x7f93cf338510>
```



```
acc_train_dtc = tree.score(Xtrain,Ytrain)
acc_test_dtc = tree.score(Xtest,Ytest)
storeResults('Decision Tree Classifier', acc_train_dtc, acc_test_dtc)
```

RANDOM FOREST CLASSIFIER

```
from sklearn.ensemble import RandomForestClassifier
forest = RandomForestClassifier(max_depth=5)
forest.fit(Xtrain, Ytrain)
```

```
RandomForestClassifier(max_depth=5)
```

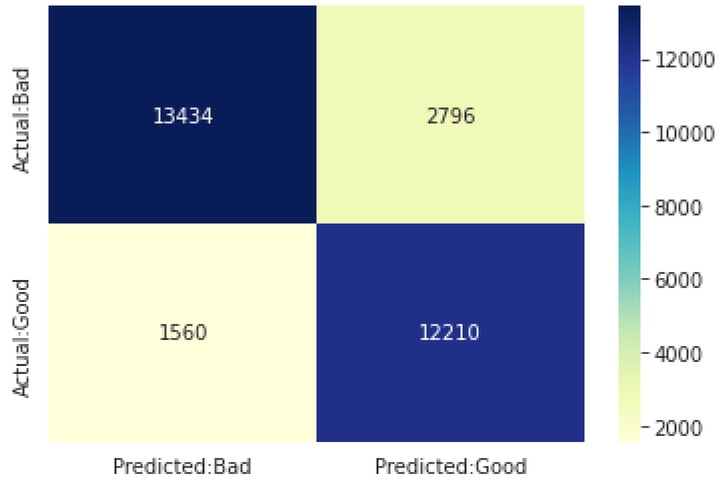
```
forest.score(Xtest,Ytest)
```

```
0.8548
```

```
Scores_ml['Random Forest Classifier'] = np.round(forest.score(Xtest,Ytest),2)
```

```
print('Training Accuracy :',forest.score(Xtrain,Ytrain))
print('Testing Accuracy :',forest.score(Xtest,Ytest))
con_mat = pd.DataFrame(confusion_matrix(forest.predict(Xtest), Ytest),
            columns = ['Predicted:Bad', 'Predicted:Good'],
            index = ['Actual:Bad', 'Actual:Good'])

print('\nCLASSIFICATION REPORT\n')
print(classification_report(forest.predict(Xtest), Ytest,target_names =['Bad','Good']))

print('\nCONFUSION MATRIX')
plt.figure(figsize= (6,4))
sns.heatmap(con_mat, annot = True,fmt='d',cmap="YlGnBu")
```

```
Training Accuracy : 0.8590857142857142
Testing Accuracy : 0.8548

CLASSIFICATION REPORT

              precision    recall  f1-score   support

         Bad       0.90      0.83      0.86     16230
        Good       0.81      0.89      0.85     13770

    accuracy                           0.85     30000
   macro avg       0.85      0.86      0.85     30000
weighted avg       0.86      0.85      0.86     30000


CONFUSION MATRIX
<matplotlib.axes._subplots.AxesSubplot at 0x7f93cf23dd50>
```



```
acc_train_rf = forest.score(Xtrain,Ytrain)
acc_test_rf = forest.score(Xtest,Ytest)
storeResults('Random Forest', acc_train_rf, acc_test_rf)
```

SUPPORT VECTOR CLASSIFICATION

```
!wget https://developer.nvidia.com/compute/cuda/9.0/Prod/local_installers/cuda-repo-ubuntu1704-9-0-loca
!ls  # Check if required cuda 9.0 amd64-deb file is downloaded
!dpkg -i cuda-repo-ubuntu1704-9-0-local_9.0.176-1_amd64-deb
!ls /var/cuda-repo-9-0-local | grep .pub
!apt-key add /var/cuda-repo-9-0-local/7fa2af80.pub
!apt-get update
!sudo apt-get install cuda-9.0
```

```
Unpacking cuda-visual-tools-9-0 (9.0.176-1) ...
Selecting previously unselected package cuda-toolkit-9-0.
Preparing to unpack .../28-cuda-toolkit-9-0_9.0.176-1_amd64.deb ...
Unpacking cuda-toolkit-9-0 (9.0.176-1) ...
Selecting previously unselected package cuda-libraries-9-0.
Preparing to unpack .../29-cuda-libraries-9-0_9.0.176-1_amd64.deb ...
Unpacking cuda-libraries-9-0 (9.0.176-1) ...
Selecting previously unselected package cuda-runtime-9-0.
Preparing to unpack .../30-cuda-runtime-9-0_9.0.176-1_amd64.deb ...
Unpacking cuda-runtime-9-0 (9.0.176-1) ...
```

```
Selecting previously unselected package cuda-demo-suite-9-0.
Preparing to unpack .../31-cuda-demo-suite-9-0_9.0.176-1_amd64.deb ...
Unpacking cuda-demo-suite-9-0 (9.0.176-1) ...
Selecting previously unselected package cuda-9-0.
Preparing to unpack .../32-cuda-9-0_9.0.176-1_amd64.deb ...
Unpacking cuda-9-0 (9.0.176-1) ...
Setting up cuda-license-9-0 (9.0.176-1) ...
*** LICENSE AGREEMENT ***
By using this software you agree to fully comply with the terms and
conditions of the EULA (End User License Agreement). The EULA is located
at /usr/local/cuda-9.0/doc/EULA.txt. The EULA can also be found at
http://docs.nvidia.com/cuda/eula/index.html. If you do not agree to the
terms and conditions of the EULA, do not use the software.

Setting up cuda-cusparse-9-0 (9.0.176-1) ...
Setting up cuda-cudart-9-0 (9.0.176-1) ...
Setting up cuda-nvrtc-9-0 (9.0.176-1) ...
Setting up cuda-cusparse-dev-9-0 (9.0.176-1) ...
Setting up cuda-cufft-9-0 (9.0.176-1) ...
Setting up cuda-cusolver-9-0 (9.0.176-1) ...
Setting up cuda-nvml-dev-9-0 (9.0.176-1) ...
Setting up cuda-npp-9-0 (9.0.176-1) ...
Setting up cuda-cusolver-dev-9-0 (9.0.176-1) ...
Setting up cuda-misc-headers-9-0 (9.0.176-1) ...
Setting up cuda-cublas-9-0 (9.0.176-1) ...
Setting up cuda-nvrtc-dev-9-0 (9.0.176-1) ...
Setting up cuda-driver-dev-9-0 (9.0.176-1) ...
Setting up cuda-curand-9-0 (9.0.176-1) ...
Setting up cuda-nvgraph-9-0 (9.0.176-1) ...
Setting up cuda-core-9-0 (9.0.176-1) ...
Setting up cuda-libraries-9-0 (9.0.176-1) ...
Setting up cuda-runtime-9-0 (9.0.176-1) ...
Setting up cuda-cudart-dev-9-0 (9.0.176-1) ...
Setting up cuda-cufft-dev-9-0 (9.0.176-1) ...
Setting up cuda-npp-dev-9-0 (9.0.176-1) ...
Setting up cuda-curand-dev-9-0 (9.0.176-1) ...
Setting up cuda-cublas-dev-9-0 (9.0.176-1) ...
Setting up cuda-nvgraph-dev-9-0 (9.0.176-1) ...
Setting up cuda-command-line-tools-9-0 (9.0.176-1) ...
Setting up cuda-demo-suite-9-0 (9.0.176-1) ...
Setting up cuda-visual-tools-9-0 (9.0.176-1) ...
Setting up cuda-samples-9-0 (9.0.176-1) ...
Setting up cuda-libraries-dev-9-0 (9.0.176-1) ...
Setting up cuda-documentation-9-0 (9.0.176-1) ...
Setting up cuda-toolkit-9-0 (9.0.176-1) ...
Setting up cuda-9-0 (9.0.176-1) ...
Processing triggers for libc-bin (2.27-3ubuntu1.5) ...
```

```
!nvcc --version
```

```
nvcc: NVIDIA (R) Cuda compiler driver
Copyright (c) 2005-2017 NVIDIA Corporation
Built on Fri_Sep__1_21:08:03_CDT_2017
Cuda compilation tools, release 9.0, V9.0.176
```

```
!pip install thundersvm
```

```
Looking in indexes: https://pypi.org/simple, https://us-python.pkg.dev/colab-wheels/public/simple
Collecting thundersvm
  Downloading thundersvm-0.3.12-py3-none-any.whl (507 kB)
     |████████████████████████████████| 507 kB 25.5 MB/s
Requirement already satisfied: scikit-learn in /usr/local/lib/python3.7/dist-packages (from thunde
Requirement already satisfied: numpy in /usr/local/lib/python3.7/dist-packages (from thundersvm)
Requirement already satisfied: scipy in /usr/local/lib/python3.7/dist-packages (from thundersvm)
Requirement already satisfied: threadpoolctl>=2.0.0 in /usr/local/lib/python3.7/dist-packages (fro
Requirement already satisfied: joblib>=0.11 in /usr/local/lib/python3.7/dist-packages (from scikit
Installing collected packages: thundersvm
Successfully installed thundersvm-0.3.12
```

```
!pip install thundersvm-cpu
```

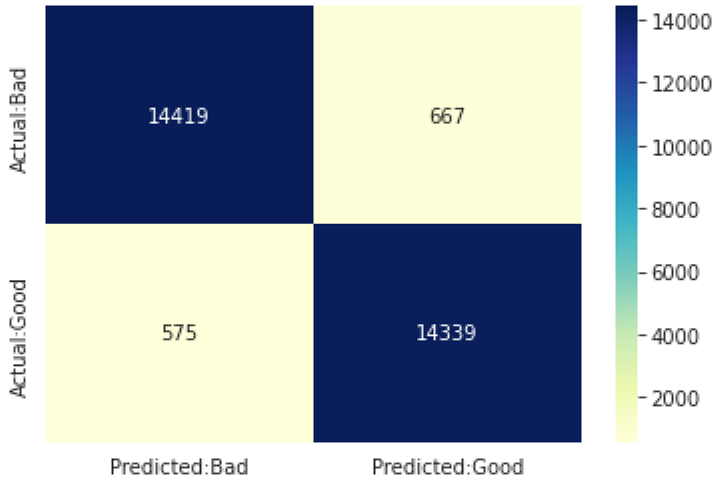
```
Looking in indexes: https://pypi.org/simple, https://us-python.pkg.dev/colab-wheels/public/simple
Collecting thundersvm-cpu
  Downloading thundersvm_cpu-0.3.3-py3-none-any.whl (227 kB)
     |████████████████████████████████| 227 kB 26.9 MB/s
Requirement already satisfied: scikit-learn in /usr/local/lib/python3.7/dist-packages (from thunde
Requirement already satisfied: numpy in /usr/local/lib/python3.7/dist-packages (from thundersvm-c
Requirement already satisfied: scipy in /usr/local/lib/python3.7/dist-packages (from thundersvm-c
Requirement already satisfied: joblib>=0.11 in /usr/local/lib/python3.7/dist-packages (from sciki
Requirement already satisfied: threadpoolctl>=2.0.0 in /usr/local/lib/python3.7/dist-packages (fr
Installing collected packages: thundersvm-cpu
Successfully installed thundersvm-cpu-0.3.3
```

```
from thundersvm import SVC
svcs = SVC(C=100, kernel='rbf')
svcs.fit(Xtrain, Ytrain)
svcs.score(Xtest, Ytest)
```

```
0.9586
```

```
Scores_ml['Support Vector Classification'] = np.round(svcs.score(Xtest,Ytest),2)
print('Training Accuracy :',svcs.score(Xtrain,Ytrain))
print('Testing Accuracy :',svcs.score(Xtest,Ytest))
con_mat = pd.DataFrame(confusion_matrix(svcs.predict(Xtest), Ytest),
            columns = ['Predicted:Bad', 'Predicted:Good'],
            index = ['Actual:Bad', 'Actual:Good'])


print('\nCLASSIFICATION REPORT\n')
print(classification_report(svcs.predict(Xtest), Ytest,
                            target_names =['Bad','Good']))

print('\nCONFUSION MATRIX')
plt.figure(figsize= (6,4))
sns.heatmap(con_mat, annot = True,fmt='d',cmap="YlGnBu")
```

```
Training Accuracy : 0.9565428571428571
Testing Accuracy : 0.9586

CLASSIFICATION REPORT

              precision    recall  f1-score   support

         Bad       0.96      0.96      0.96     15086
        Good       0.96      0.96      0.96     14914

    accuracy                           0.96     30000
   macro avg       0.96      0.96      0.96     30000
weighted avg       0.96      0.96      0.96     30000


CONFUSION MATRIX
<matplotlib.axes._subplots.AxesSubplot at 0x7f93cf23dc90>
```



```
acc_train_svcs = svcs.score(Xtrain,Ytrain)
acc_test_svcs = svcs.score(Xtest,Ytest)
storeResults('Support Vector Classification', acc_train_svcs, acc_test_svcs)
```

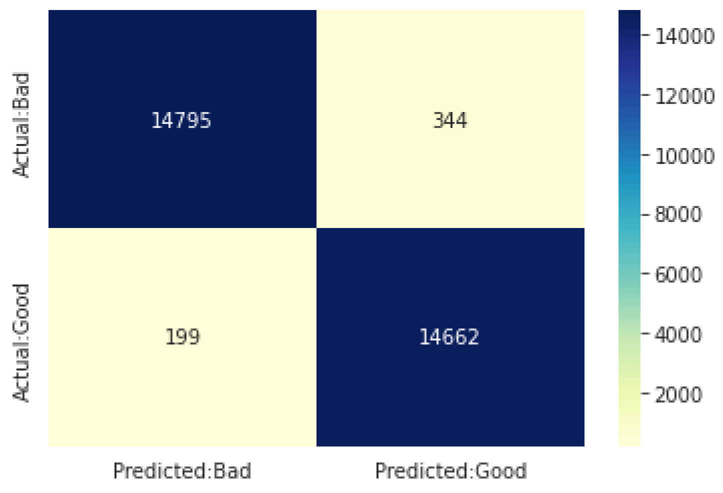
K-Nearest Neighbor

```
from sklearn.neighbors import KNeighborsClassifier
knn = KNeighborsClassifier(n_neighbors = 5)
knn.fit(Xtrain,Ytrain)
knn.score(Xtest,Ytest)
```

```
    0.9819
```

```
Scores_ml['K-Nearest Neighbor'] = np.round(knn.score(Xtest,Ytest),2)
print('Training Accuracy :',knn.score(Xtrain,Ytrain))
print('Testing Accuracy :',knn.score(Xtest,Ytest))
con_mat = pd.DataFrame(confusion_matrix(knn.predict(Xtest), Ytest),
            columns = ['Predicted:Bad', 'Predicted:Good'],
            index = ['Actual:Bad', 'Actual:Good'])


print('\nCLASSIFICATION REPORT\n')
print(classification_report(knn.predict(Xtest), Ytest,target_names =['Bad','Good']))

print('\nCONFUSION MATRIX')
plt.figure(figsize= (6,4))
sns.heatmap(con_mat, annot = True,fmt='d',cmap="YlGnBu")
```

```
    Training Accuracy : 0.9866285714285714
    Testing Accuracy : 0.9819

    CLASSIFICATION REPORT

                  precision    recall  f1-score   support

             Bad       0.99      0.98      0.98     15139
            Good       0.98      0.99      0.98     14861

        accuracy                           0.98     30000
       macro avg       0.98      0.98      0.98     30000
    weighted avg       0.98      0.98      0.98     30000


    CONFUSION MATRIX
    <matplotlib.axes._subplots.AxesSubplot at 0x7f93cf5a5450>
```



```
acc_train_knm = knn.score(Xtrain,Ytrain)
acc_test_knm = knn.score(Xtest,Ytest)
storeResults('K-Nearest Algorithm', acc_train_knm, acc_test_knm)
```

```
results = pd.DataFrame({ 'ML Model': ML_Model,
    'Train Accuracy': acc_train,
    'Test Accuracy': acc_test})
results
```

	ML Model	Train Accuracy	Test Accuracy
0	Logistic Regression	0.79	0.79
1	multinomial NB	0.77	0.77
2	Decision Tree Classifier	0.79	0.79
3	Random Forest	0.86	0.85
4	Support Vector Classification	0.96	0.96
5	K-Nearest Algorithm	0.99	0.98

```
#Sorting the datafram on accuracy
results.sort_values(by=['Test Accuracy', 'Train Accuracy'], ascending=False)
```

	ML Model	Train Accuracy	Test Accuracy
5	K-Nearest Algorithm	0.99	0.98
4	Support Vector Classification	0.96	0.96
3	Random Forest	0.86	0.85
0	Logistic Regression	0.79	0.79
2	Decision Tree Classifier	0.79	0.79
1	multinomial NB	0.77	0.77

```
pipeline_ls = make_pipeline(CountVectorizer(tokenizer = RegexpTokenizer(r'[A-Za-z]+').tokenize,stop_wor
```

```
Xtrain, Xtest, Ytrain, Ytest = train_test_split(data.URL,data.Label)
```

```
pipeline_ls.fit(Xtrain,Ytrain)
```

```
Pipeline(steps=[('countvectorizer',
                 CountVectorizer(stop_words='english',
                                 tokenizer=<bound method RegexpTokenizer.tokenize of
RegexpTokenizer(pattern='[A-Za-z]+', gaps=False, discard_empty=True, flags=
<RegexFlag.UNICODE|DOTALL|MULTILINE: 56>)>)),
                ('kneighborsclassifier', KNeighborsClassifier())])
```

```
pickle.dump(pipeline_ls,open('phishing.pkl','wb'))
```

```
predict_bad = ['steamglfts.hut2.ru/']
predict_good = ['www.auburnmedia.com/wordpress/']
loaded_model = pickle.load(open('phishing.pkl', 'rb'))
result = loaded_model.predict(predict_bad)
result2 = loaded_model.predict(predict_good)
print(result)
print("-"*10)
print(result2)
```

```
['bad']
----------
['good']
```