mk 02-09-2023

In [ ]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

In [77]:
```
from sklearn.linear_model import LogisticRegression
a=pd.read_csv(r"C:\USERS\user\Downloads\C3_bot_detection_data (1).csv")
a
```

Out[77]:

	User ID	Username	Tweet	Retweet Count	Mention Count	Follower Count	Verified	Bot Label	
0	132131	flong	Station activity person against natural majori...	85	1	2353	False	1	A
1	289683	hinesstephanie	Authority research natural life material staff...	55	5	9617	True	0	Sa
2	779715	roberttran	Manage whose quickly especially foot none to	6	2	4363	True	0	Ha

In [78]:
```
from sklearn.linear_model import LogisticRegression
```

In [79]:
```
a=a.head(10)
a
```

1	289683	hinesstephanie	Authority research natural life material staff...	55	5	9617	True	0	Sande
2	779715	roberttran	Manage whose quickly especially foot none to g...	6	2	4363	True	0	Harrisc
3	696168	pmason	Just cover eight opportunity strong policy which.	54	5	2242	True	1	Martinez
4	704441	noah87	Animal sign six data good or.	26	3	8438	False	1	Camach

In [81]:
```
a.columns
```

Out[81]:
```
Index(['User ID', 'Username', 'Tweet', 'Retweet Count', 'Mention Count',
       'Follower Count', 'Verified', 'Bot Label', 'Location', 'Created At',
       'Hashtags'],
      dtype='object')
```

In [83]:
```
b=a[['User ID','Retweet Count', 'Mention Count','Follower Count']]
b
```

Out[83]:

	User ID	Retweet Count	Mention Count	Follower Count
0	132131	85	1	2353
1	289683	55	5	9617
2	779715	6	2	4363
3	696168	54	5	2242
4	704441	26	3	8438
5	570928	41	4	3792
6	734182	54	0	10
7	107312	64	0	1442
8	549888	25	2	836
9	117640	67	3	6523

In [84]:
```
c=b.iloc[:,0:11]
d=b.iloc[:,-1]
```

In [85]:
```
c.shape
```

Out[85]:
```
(10, 4)
```

In [86]:
```
d.shape
```

Out[86]:
```
(10,)
```

In [ ]:
```

```

In [87]:
```
from sklearn.preprocessing import StandardScaler
fs=StandardScaler().fit_transform(c)
fs
```

Out[87]:
```
array([[-1.27874417,  1.67954984, -0.8588975 , -0.51903859],
       [-0.67927374,  0.32870546,  1.43149584,  1.82479847],
       [ 1.18525156, -1.87767368, -0.28629917,  0.12951765],
       [ 0.86736314,  0.28367732,  1.43149584, -0.55485438],
       [ 0.89884112, -0.97711076,  0.28629917,  1.44437668],
       [ 0.39083682, -0.30168858,  0.8588975 , -0.05472395],
       [ 1.01200281,  0.28367732, -1.43149584, -1.2750422 ],
       [-1.37317811,  0.73395878, -1.43149584, -0.81298621],
       [ 0.31078161, -1.02213891, -0.28629917, -1.00852108],
       [-1.33388105,  0.86904321,  0.28629917,  0.8264736 ]])
```

In [88]:
```
logr=LogisticRegression()
logr.fit(fs,d)
```

Out[88]:
```
LogisticRegression()
```

In [129]:
```
e=[[2,5,77,8]]
```

In [130]:
```
prediction=logr.predict(e)
prediction
```

Out[130]:
```
array([2242], dtype=int64)
```

In [131]:
```
logr.classes_
```

Out[131]:
```
array([  10,  836, 1442, 2242, 2353, 3792, 4363, 6523, 8438, 9617],
      dtype=int64)
```

In [133]:
```
logr.predict_proba(e)[0][0]
```

Out[133]:
```
8.966898605558299e-48
```

In [134]:
```
import re
from sklearn.datasets import load_digits
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

In [135]:
```
from sklearn.linear_model import LogisticRegression
from sklearn.model_selection import train_test_split
```

In [136]:
```
digits=load_digits()
digits
```

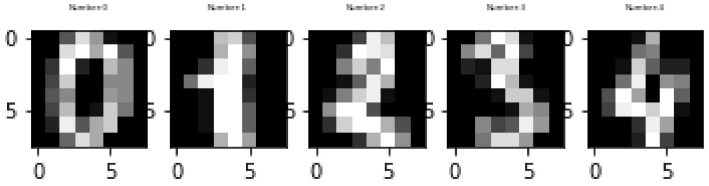
```
'pixel_0_6',
'pixel_0_7',
'pixel_1_0',
'pixel_1_1',
'pixel_1_2',
'pixel_1_3',
'pixel_1_4',
'pixel_1_5',
'pixel_1_6',
'pixel_1_7',
'pixel_2_0',
'pixel_2_1',
'pixel_2_2',
'pixel_2_3',
'pixel_2_4',
'pixel_2_5',
'pixel_2_6',
'pixel_2_7',
'pixel_3_0',
'pixel_3_1'
```

In [98]:
```
plt.figure(figsize=(20,4))
```

Out[98]:
```
<Figure size 1440x288 with 0 Axes>
```

```
<Figure size 1440x288 with 0 Axes>
```

In [99]:
```
for index,(image,label) in enumerate(zip(digits.data[0:5],digits.target[0:
    plt.subplot(1,5,index+1)
    plt.imshow(np.reshape(image,(8
                                 ,8)),cmap=plt.cm.gray)
    plt.title('Number:%i\n'%label,fontsize=4)
```



In [100]:
```
x_train,x_test,y_train,y_test=train_test_split(digits.data,digits.target,t
```

In [101]:
```
print(x_train.shape)
print(x_test.shape)
print(y_train.shape)
print(y_test.shape)
```

```
(1257, 64)
(540, 64)
(1257,)
(540,)
```

In [102]:
```
logre=LogisticRegression(max_iter=10000)
logre.fit(x_train,y_train)

```

Out[102]:
```
LogisticRegression(max_iter=10000)
```

In [103]:
```
print(logre.predict(x_test))
```

```
[7 4 5 4 5 6 7 6 4 6 7 5 7 7 7 9 4 7 6 5 8 0 7 1 6 2 7 4 7 7 6 7 3 1 2 9 7
 7 2 2 5 0 8 7 6 3 2 5 3 1 7 9 6 8 6 7 5 6 4 3 7 7 7 7 2 9 1 7 4 4 0 8 7 6
 0 6 1 3 5 5 2 9 4 0 2 3 3 4 9 4 8 1 4 2 8 8 0 9 1 6 1 4 9 0 7 9 1 3 3 0 7
 8 3 9 5 1 7 1 2 9 7 0 8 7 4 0 2 0 3 6 1 9 3 0 6 1 7 4 4 4 4 0 1 4 2 4 0 4
 4 6 7 3 6 8 6 8 8 6 1 0 9 9 6 8 2 9 5 2 5 0 7 3 9 6 1 5 0 9 6 9 1 7 2 4 0
 6 4 3 8 1 3 0 1 0 8 5 2 3 5 3 2 3 8 8 9 1 0 1 3 4 7 5 4 0 7 8 7 9 3 8 6 8
 7 3 5 2 7 1 4 4 5 3 0 6 4 5 9 3 5 1 1 8 9 8 7 3 0 8 4 1 9 4 8 4 3 7 6 7 9
 5 8 0 7 5 8 0 6 9 8 9 6 1 7 0 3 5 7 8 5 7 2 5 5 6 4 1 4 1 7 3 4 9 4 1 1 8
 9 6 5 9 2 8 3 5 9 6 1 4 2 8 0 7 7 8 3 3 8 8 9 9 0 8 8 4 4 8 6 5 2 6 6 7 0
 2 3 1 0 3 0 1 8 7 7 8 5 7 5 1 1 5 0 5 3 1 3 0 3 2 3 0 7 8 7 6 7 9 3 7 3 2
 1 4 6 7 7 8 8 7 5 8 4 5 0 4 9 3 2 9 8 1 7 4 6 5 4 0 4 8 0 5 0 1 5 4 6 9 8
 8 5 3 2 4 6 4 3 2 3 1 1 2 8 8 3 2 1 9 9 6 3 0 8 5 4 8 6 3 4 9 0 4 5 9 0 4
 5 6 2 8 1 2 5 6 2 8 9 2 7 0 5 3 4 8 3 8 1 1 0 1 4 5 1 4 1 9 0 1 6 0 3 3 5
 2 3 9 2 5 6 1 9 3 3 3 1 9 3 3 6 7 1 3 0 0 0 7 2 7 0 5 2 1 3 4 7 5 9 8 9 8
 2 3 1 2 6 6 8 8 9 2 4 3 4 7 1 3 7 1 4 0 3 4]
```

In [104]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

In [105]:
```
df=pd.read_csv(r"C:\USERS\user\Downloads\C3_bot_detection_data (1).csv")
```

In [140]:
```
df['Mention Count'].value_counts()
```

Out[140]:
```
4    8441
5    8436
2    8347
1    8294
3    8252
0    8230
Name: Mention Count, dtype: int64
```

In [142]:

```
g1={"Mention Count":{'Mention Count':1,'b':2}}
df=df.replace(g1)
print(df)
```

```
       User ID         Username  \
0       132131            flong
1       289683   hinesstephanie
2       779715       roberttran
3       696168            pmason
4       704441            noah87
...        ...               ...
49995   491196            uberg
49996   739297      jessicamunoz
49997   674475    lynncunningham
49998   167081  richardthompson
49999   311204         daniel29

                                                   Tweet  Retweet Count  \
0      Station activity person against natural majori...             85
1      Authority research natural life material staff...             55
2      Manage whose quickly especially foot none to g...              6
3      Just cover eight opportunity strong policy which.             54
4                          Animal sign six data good or.             26
...                                                  ...            ...
49995  Want but put card direction know miss former h...             64
49996  Provide whole maybe agree church respond most ...             18
49997  Bring different everyone international capital...             43
49998  Than about single generation itself seek sell ...             45
49999  Here morning class various room human true bec...             91

       Mention Count  Follower Count  Verified  Bot Label            Location
\
0                  1            2353     False          1           Adkinston
1                  5            9617      True          0          Sanderston
2                  2            4363      True          0        Harrisonfurt
3                  5            2242      True          1        Martinezberg
4                  3            8438     False          1        Camachoville
...              ...             ...       ...        ...                 ...
49995              0            9911      True          1  Lake Kimberlyburgh
49996              5            9900     False          1           Greenbury
49997              3            6313      True          1         Deborahfort
49998              1            6343     False          0         Stephenside
49999              4            4006     False          0           Novakberg

                Created At                         Hashtags
0      2020-05-11 15:29:50                              NaN
1      2022-11-26 05:18:10                        both live
2      2022-08-08 03:16:54                      phone ahead
3      2021-08-14 22:27:05                ever quickly new I
4      2020-04-13 21:24:21                   foreign mention
...                    ...                              ...
49995  2023-04-20 11:06:26  teach quality ten education any
49996  2022-10-18 03:57:35            add walk among believe
49997  2020-07-08 03:54:08          onto admit artist first
49998  2022-03-22 12:13:44                             star
49999  2022-12-03 06:11:07                             home

[50000 rows x 11 columns]
```

In [143]:
```
from sklearn.model_selection import train_test_split
x_train,x_test,y_train,y_test=train_test_split(x,y,train_size=0.70)
```

In [144]:
```
from sklearn.ensemble import RandomForestClassifier
```

In [145]:
```
rfc=RandomForestClassifier()
rfc.fit(x_train,y_train)
```

Out[145]:
```
RandomForestClassifier()
```

In [146]:
```
parameters={'max_depth':[1,2,3,4,5],
            'min_samples_leaf':[5,10,15,20,25],
            'n_estimators':[10,20,30,40,50]}
```

In [147]:
```
from sklearn.model_selection import GridSearchCV
```

In [148]:
```
grid_search=GridSearchCV(estimator=rfc,param_grid=parameters,cv=2,scoring=
grid_search.fit(x_train,y_train)
```

Out[148]:
```
GridSearchCV(cv=2, estimator=RandomForestClassifier(),
             param_grid={'max_depth': [1, 2, 3, 4, 5],
                         'min_samples_leaf': [5, 10, 15, 20, 25],
                         'n_estimators': [10, 20, 30, 40, 50]},
             scoring='accuracy')
```

In [149]:
```
grid_search.best_score_
```

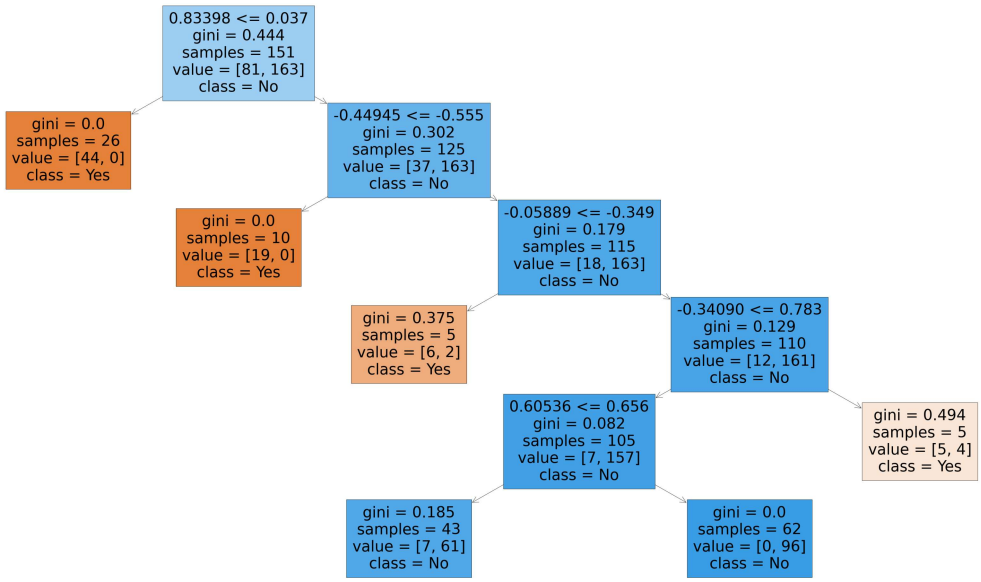
Out[149]:
```
0.9467213114754098
```

In [150]:
```
rfc_best=grid_search.best_estimator_
```

In [151]:
```
from sklearn.tree import plot_tree
```

In [152]:
```
plt.figure(figsize=(80,40))
plot_tree(rfc_best.estimators_[5],feature_names=x.columns,class_names=['Ye
```

Out[152]:
```
[Text(1275.4285714285713, 1993.2, '0.83398 <= 0.037\ngini = 0.444\nsamples =
151\nvalue = [81, 163]\nclass = No'),
 Text(637.7142857142857, 1630.8000000000002, 'gini = 0.0\nsamples = 26\nvalue
= [44, 0]\nclass = Yes'),
 Text(1913.1428571428569, 1630.8000000000002, '-0.44945 <= -0.555\ngini = 0.3
02\nsamples = 125\nvalue = [37, 163]\nclass = No'),
 Text(1275.4285714285713, 1268.4, 'gini = 0.0\nsamples = 10\nvalue = [19, 0]
\nclass = Yes'),
 Text(2550.8571428571427, 1268.4, '-0.05889 <= -0.349\ngini = 0.179\nsamples
= 115\nvalue = [18, 163]\nclass = No'),
 Text(1913.1428571428569, 906.0, 'gini = 0.375\nsamples = 5\nvalue = [6, 2]\n
class = Yes'),
 Text(3188.5714285714284, 906.0, '-0.34090 <= 0.783\ngini = 0.129\nsamples =
110\nvalue = [12, 161]\nclass = No'),
 Text(2550.8571428571427, 543.5999999999999, '0.60536 <= 0.656\ngini = 0.082
\nsamples = 105\nvalue = [7, 157]\nclass = No'),
 Text(1913.1428571428569, 181.19999999999982, 'gini = 0.185\nsamples = 43\nva
lue = [7, 61]\nclass = No'),
 Text(3188.5714285714284, 181.19999999999982, 'gini = 0.0\nsamples = 62\nvalu
e = [0, 96]\nclass = No'),
 Text(3826.2857142857138, 543.5999999999999, 'gini = 0.494\nsamples = 5\nvalu
e = [5, 4]\nclass = Yes')]
```



In [ ]:
```
1
```