In [1]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.linear_model import LogisticRegression
from sklearn.preprocessing import StandardScaler
```

In [2]:
```
from sklearn.linear_model import LogisticRegression
```

In [3]:
```
df=pd.read_csv(r"E:\154\C3_bot_detection_data - C3_bot_detection_data.csv").dropna()
df
```

Out[3]:

	User ID	Username	Tweet	Retweet Count	Mention Count	Follower Count	Verified	Bot Label	Location	Created At	Hashtags
1	289683	hinesstephanie	Authority research natural life material staff...	55	5	9617	True	0	Sanderston	2022-11-26 05:18:10	both live
2	779715	roberttran	Manage whose quickly especially foot none to g...	6	2	4363	True	0	Harrisonfurt	2022-08-08 03:16:54	phone ahead
3	696168	pmason	Just cover eight opportunity strong policy which.	54	5	2242	True	1	Martinezberg	2021-08-14 22:27:05	eve quickly new
4	704441	noah87	Animal sign six data good or.	26	3	8438	False	1	Camachoville	2020-04-13 21:24:21	foreign mentior
5	570928	james00	See wonder travel this suffer less yard office...	41	4	3792	True	1	West Cheyenne	2023-05-07 22:24:47	anyone respond perhaps marke run
...	...	...	...	...	...	...	...	...	...	...	..
49995	491196	uberg	Want but put card direction know miss former h...	64	0	9911	True	1	Lake Kimberlyburgh	2023-04-20 11:06:26	teach quality ter education any
49996	739297	jessicamunoz	Provide whole maybe agree church respond most ...	18	5	9900	False	1	Greenbury	2022-10-18 03:57:35	add walk among believe
49997	674475	lynncunningham	Bring different everyone international capital...	43	3	6313	True	1	Deborahfort	2020-07-08 03:54:08	onto admi artist firs
49998	167081	richardthompson	Than about single generation itself seek sell ...	45	1	6343	False	0	Stephenside	2022-03-22 12:13:44	sta
49999	311204	daniel29	Here morning class various room human true bec...	91	4	4006	False	0	Novakberg	2022-12-03 06:11:07	home

41659 rows × 11 columns

In [4]:
```
df.head()
```

Out[4]:

	User ID	Username	Tweet	Retweet Count	Mention Count	Follower Count	Verified	Bot Label	Location	Created At	Hashtags
1	289683	hinesstephanie	Authority research natural life material staff...	55	5	9617	True	0	Sanderston	2022-11-26 05:18:10	both live
2	779715	roberttran	Manage whose quickly especially foot none to g...	6	2	4363	True	0	Harrisonfurt	2022-08-08 03:16:54	phone ahead
3	696168	pmason	Just cover eight opportunity strong policy which.	54	5	2242	True	1	Martinezberg	2021-08-14 22:27:05	ever quickly new l
4	704441	noah87	Animal sign six data good or.	26	3	8438	False	1	Camachoville	2020-04-13 21:24:21	foreign mention
5	570928	james00	See wonder travel this suffer less yard office...	41	4	3792	True	1	West Cheyenne	2023-05-07 22:24:47	anyone respond perhaps market run

In [5]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 41659 entries, 1 to 49999
Data columns (total 11 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   User ID         41659 non-null  int64
 1   Username        41659 non-null  object
 2   Tweet           41659 non-null  object
 3   Retweet Count   41659 non-null  int64
 4   Mention Count   41659 non-null  int64
 5   Follower Count  41659 non-null  int64
 6   Verified        41659 non-null  bool
 7   Bot Label       41659 non-null  int64
 8   Location        41659 non-null  object
 9   Created At      41659 non-null  object
 10  Hashtags        41659 non-null  object
dtypes: bool(1), int64(5), object(5)
memory usage: 3.5+ MB
```

In [6]:
```
df.describe()
```

Out[6]:

	User ID	Retweet Count	Mention Count	Follower Count	Bot Label
count	41659.000000	41659.000000	41659.000000	41659.000000	41659.000000
mean	548640.613097	49.950911	2.515207	4990.867928	0.500204
std	259990.806985	29.195286	1.709249	2880.947193	0.500006
min	100025.000000	0.000000	0.000000	0.000000	0.000000
25%	321829.500000	25.000000	1.000000	2493.500000	0.000000
50%	548396.000000	50.000000	3.000000	4997.000000	1.000000
75%	772751.500000	75.000000	4.000000	7475.500000	1.000000
max	999995.000000	100.000000	5.000000	10000.000000	1.000000

In [7]:
```
df.columns
```

Out[7]:
```
Index(['User ID', 'Username', 'Tweet', 'Retweet Count', 'Mention Count',
       'Follower Count', 'Verified', 'Bot Label', 'Location', 'Created At',
       'Hashtags'],
      dtype='object')
```

In [8]:
```
feature_matrix = df[['User ID','Retweet Count','Mention Count','Follower Count','Bot Label']]
target_vector = df[["Verified"]]
```

In [9]:
```
fs=StandardScaler().fit_transform(feature_matrix)
logr=LogisticRegression()
logr.fit(fs,target_vector)
```

```
C:\ProgramData\Anaconda3\lib\site-packages\sklearn\utils\validation.py:63: DataConversionWarning:
A column-vector y was passed when a 1d array was expected. Please change the shape of y to (n_sam
ples, ), for example using ravel().
  return f(*args, **kwargs)
```

Out[9]:
```
LogisticRegression()
```

In [10]:
```
observation=[[1,2,3,4,5]]
```

In [11]:
```
prediction=logr.predict(observation)
print(prediction)
```

```
[False]
```

In [12]:
```
logr.classes_
```

Out[12]:
```
array([False,  True])
```

In [13]:
```
logr.predict_proba(observation)[0][0]
```

Out[13]:
```
0.504915130281248
```

In [14]:
```
logr.predict_proba(observation)[0][1]
```

Out[14]:
```
0.49508486971875193
```

Random Forest

In [15]:
```
df['Verified'].value_counts()
```

Out[15]:
```
True     20845
False    20814
Name: Verified, dtype: int64
```

In [16]:
```
x=df[['User ID','Retweet Count','Mention Count','Follower Count','Bot Label']]
y=df['Verified']
```

In [17]:
```
g1={'Verified':{'True':1, "False":2}}
df=df.replace(g1)
df
```

Out[17]:

	User ID	Username	Tweet	Retweet Count	Mention Count	Follower Count	Verified	Bot Label	Location	Created At	Hashtags
1	289683	hinesstephanie	Authority research natural life material staff...	55	5	9617	True	0	Sanderston	2022-11-26 05:18:10	both live
2	779715	roberttran	Manage whose quickly especially foot none to g...	6	2	4363	True	0	Harrisonfurt	2022-08-08 03:16:54	phone ahead
3	696168	pmason	Just cover eight opportunity strong policy which.	54	5	2242	True	1	Martinezberg	2021-08-14 22:27:05	eve quickly new
4	704441	noah87	Animal sign six data good or.	26	3	8438	False	1	Camachoville	2020-04-13 21:24:21	foreign mentior
5	570928	james00	See wonder travel this suffer less yard office...	41	4	3792	True	1	West Cheyenne	2023-05-07 22:24:47	anyone respond perhaps marke run
...	...	...	...	...	...	...	...	...	...	...	..
49995	491196	uberg	Want but put card direction know miss former h...	64	0	9911	True	1	Lake Kimberlyburgh	2023-04-20 11:06:26	teach quality ter educatior any
49996	739297	jessicamunoz	Provide whole maybe agree church respond most ...	18	5	9900	False	1	Greenbury	2022-10-18 03:57:35	add walk among believe
49997	674475	lynncunningham	Bring different everyone international capital...	43	3	6313	True	1	Deborahfort	2020-07-08 03:54:08	onto admi artist firs
49998	167081	richardthompson	Than about single generation itself seek sell ...	45	1	6343	False	0	Stephenside	2022-03-22 12:13:44	sta
49999	311204	daniel29	Here morning class various room human true bec...	91	4	4006	False	0	Novakberg	2022-12-03 06:11:07	home

41659 rows × 11 columns

In [18]:
```
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 [19]:
```
from sklearn.ensemble import RandomForestClassifier
rfc = RandomForestClassifier()
rfc.fit(x_train,y_train)
```

Out[19]:
```
RandomForestClassifier()
```

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

In [21]:
```
from sklearn.model_selection import GridSearchCV
grid_search = GridSearchCV(estimator=rfc,param_grid=parameters,cv=2,scoring="accuracy")
grid_search.fit(x_train,y_train)
```

Out[21]:
```
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')
```

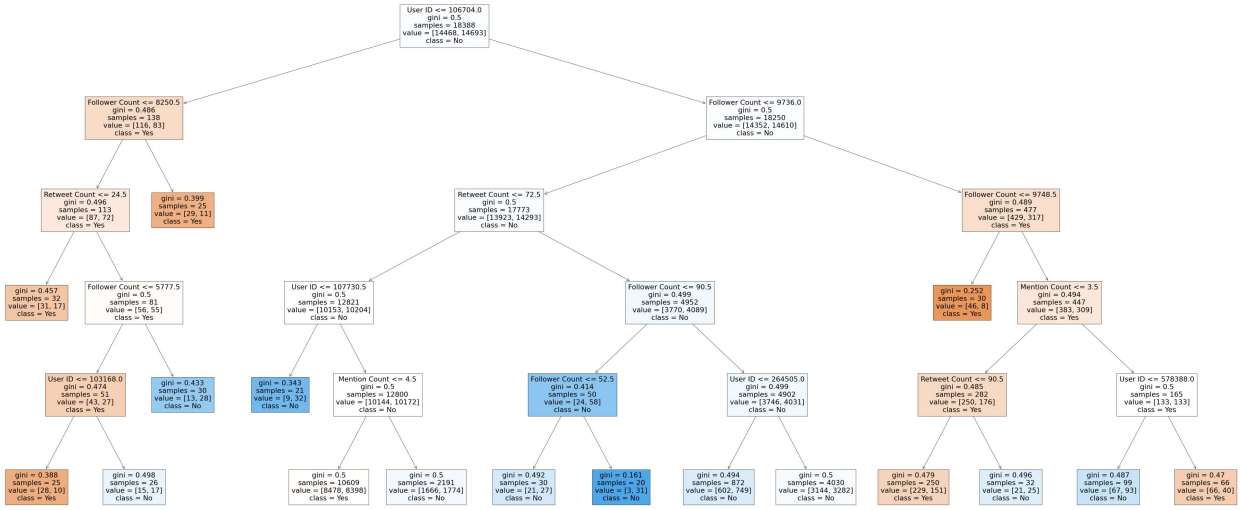
In [22]:
```
grid_search.best_score_
```

Out[22]:
```
0.5035833975646568
```

In [23]:
```
rfc_best = grid_search.best_estimator_
```

In [24]:
```
from sklearn.tree import plot_tree
plt.figure(figsize=(89,40))
plot_tree(rfc_best.estimators_[5], feature_names=x.columns, class_names=['Yes', 'No'], filled=True
```

Out[24]:
```
[Text(1790.6971153846155, 1993.2, 'User ID <= 106704.0\ngini = 0.5\nsamples = 18388\nvalue = [144
68, 14693]\nclass = No'),
 Text(573.0230769230769, 1630.8000000000002, 'Follower Count <= 8250.5\ngini = 0.486\nsamples = 1
38\nvalue = [116, 83]\nclass = Yes'),
 Text(382.0153846153846, 1268.4, 'Retweet Count <= 24.5\ngini = 0.496\nsamples = 113\nvalue = [8
7, 72]\nclass = Yes'),
 Text(191.0076923076923, 906.0, 'gini = 0.457\nsamples = 32\nvalue = [31, 17]\nclass = Yes'),
 Text(573.0230769230769, 906.0, 'Follower Count <= 5777.5\ngini = 0.5\nsamples = 81\nvalue = [56,
55]\nclass = Yes'),
 Text(382.0153846153846, 543.5999999999999, 'User ID <= 103168.0\ngini = 0.474\nsamples = 51\nval
ue = [43, 27]\nclass = Yes'),
 Text(191.0076923076923, 181.19999999999982, 'gini = 0.388\nsamples = 25\nvalue = [28, 10]\nclass
= Yes'),
 Text(573.0230769230769, 181.19999999999982, 'gini = 0.498\nsamples = 26\nvalue = [15, 17]\nclass
= No'),
 Text(764.0307692307692, 543.5999999999999, 'gini = 0.433\nsamples = 30\nvalue = [13, 28]\nclass
= No'),
 Text(764.0307692307692, 1268.4, 'gini = 0.399\nsamples = 25\nvalue = [29, 11]\nclass = Yes'),
 Text(3008.371153846154, 1630.8000000000002, 'Follower Count <= 9736.0\ngini = 0.5\nsamples = 182
50\nvalue = [14352, 14610]\nclass = No'),
 Text(2005.5807692307692, 1268.4, 'Retweet Count <= 72.5\ngini = 0.5\nsamples = 17773\nvalue = [1
3923, 14293]\nclass = No'),
 Text(1337.0538461538463, 906.0, 'User ID <= 107730.5\ngini = 0.5\nsamples = 12821\nvalue = [1015
3, 10204]\nclass = No'),
 Text(1146.0461538461539, 543.5999999999999, 'gini = 0.343\nsamples = 21\nvalue = [9, 32]\nclass
= No'),
 Text(1528.0615384615385, 543.5999999999999, 'Mention Count <= 4.5\ngini = 0.5\nsamples = 12800\n
value = [10144, 10172]\nclass = No'),
 Text(1337.0538461538463, 181.19999999999982, 'gini = 0.5\nsamples = 10609\nvalue = [8478, 8398]
\nclass = Yes'),
 Text(1719.0692307692307, 181.19999999999982, 'gini = 0.5\nsamples = 2191\nvalue = [1666, 1774]\n
class = No'),
 Text(2674.1076923076926, 906.0, 'Follower Count <= 90.5\ngini = 0.499\nsamples = 4952\nvalue =
[3770, 4089]\nclass = No'),
 Text(2292.0923076923077, 543.5999999999999, 'Follower Count <= 52.5\ngini = 0.414\nsamples = 50
\nvalue = [24, 58]\nclass = No'),
 Text(2101.0846153846155, 181.19999999999982, 'gini = 0.492\nsamples = 30\nvalue = [21, 27]\nclas
s = No'),
 Text(2483.1, 181.19999999999982, 'gini = 0.161\nsamples = 20\nvalue = [3, 31]\nclass = No'),
 Text(3056.123076923077, 543.5999999999999, 'User ID <= 264505.0\ngini = 0.499\nsamples = 4902\nv
alue = [3746, 4031]\nclass = No'),
 Text(2865.1153846153848, 181.19999999999982, 'gini = 0.494\nsamples = 872\nvalue = [602, 749]\nc
lass = No'),
 Text(3247.130769230769, 181.19999999999982, 'gini = 0.5\nsamples = 4030\nvalue = [3144, 3282]\nc
lass = No'),
 Text(4011.1615384615384, 1268.4, 'Follower Count <= 9748.5\ngini = 0.489\nsamples = 477\nvalue =
[429, 317]\nclass = Yes'),
 Text(3820.153846153846, 906.0, 'gini = 0.252\nsamples = 30\nvalue = [46, 8]\nclass = Yes'),
 Text(4202.169230769231, 906.0, 'Mention Count <= 3.5\ngini = 0.494\nsamples = 447\nvalue = [383,
309]\nclass = Yes'),
 Text(3820.153846153846, 543.5999999999999, 'Retweet Count <= 90.5\ngini = 0.485\nsamples = 282\n
value = [250, 176]\nclass = Yes'),
 Text(3629.146153846154, 181.19999999999982, 'gini = 0.479\nsamples = 250\nvalue = [229, 151]\ncl
ass = Yes'),
 Text(4011.1615384615384, 181.19999999999982, 'gini = 0.496\nsamples = 32\nvalue = [21, 25]\nclas
s = No'),
 Text(4584.184615384615, 543.5999999999999, 'User ID <= 578388.0\ngini = 0.5\nsamples = 165\nvalu
e = [133, 133]\nclass = Yes'),
 Text(4393.176923076923, 181.19999999999982, 'gini = 0.487\nsamples = 99\nvalue = [67, 93]\nclass
= No'),
 Text(4775.192307692308, 181.19999999999982, 'gini = 0.47\nsamples = 66\nvalue = [66, 40]\nclass
= Yes')]
```


User ID <= 106704.0
gini = 0.5
samples = 18388
value = [14468, 14693]
class = No
Follower Count <= 8250.5
gini = 0.486
samples = 138
value = [116, 83]
class = Yes
Follower Count <= 9736.0
gini = 0.5
samples = 18250
value = [14352, 14610]
class = No
Retweet Count <= 24.5
gini = 0.496
samples = 113
value = [87, 72]
class = Yes
gini = 0.399
samples = 25
value = [29, 11]
class = Yes
Retweet Count <= 72.5
gini = 0.5
samples = 17773
value = [13923, 14293]
class = No
Follower Count <= 9748.5
gini = 0.489
samples = 477
value = [429, 317]
class = Yes
gini = 0.457
samples = 32
value = [31, 17]
class = Yes
Follower Count <= 5777.5
gini = 0.5
samples = 81
value = [56, 55]
class = Yes
User ID <= 107730.5
gini = 0.5
samples = 12821
value = [10153, 10204]
class = No
Follower Count <= 90.5
gini = 0.499
samples = 4952
value = [3770, 4089]
class = No
gini = 0.252
samples = 30
value = [46, 8]
class = Yes
Mention Count <= 3.5
gini = 0.494
samples = 447
value = [383, 309]
class = Yes
User ID <= 103168.0
gini = 0.474
samples = 51
value = [43, 27]
class = Yes
gini = 0.433
samples = 30
value = [13, 28]
class = No
gini = 0.343
samples = 21
value = [9, 32]
class = No
Mention Count <= 4.5
gini = 0.5
samples = 12800
value = [10144, 10172]
class = No
Follower Count <= 52.5
gini = 0.414
samples = 50
value = [24, 58]
class = No
User ID <= 264505.0
gini = 0.499
samples = 4902
value = [3746, 4031]
class = No
Retweet Count <= 90.5
gini = 0.485
samples = 282
value = [250, 176]
class = Yes
User ID <= 578388.0
gini = 0.5
samples = 165
value = [133, 133]
class = Yes
gini = 0.388
samples = 25
value = [28, 10]
class = Yes
gini = 0.498
samples = 26
value = [15, 17]
class = No
gini = 0.5
samples = 10609
value = [8478, 8398]
class = Yes
gini = 0.5
samples = 2191
value = [1666, 1774]
class = No
gini = 0.492
samples = 30
value = [21, 27]
class = No
gini = 0.161
samples = 20
value = [3, 31]
class = No
gini = 0.494
samples = 872
value = [602, 749]
class = No
gini = 0.5
samples = 4030
value = [3144, 3282]
class = No
gini = 0.479
samples = 250
value = [229, 151]
class = Yes
gini = 0.496
samples = 32
value = [21, 25]
class = No
gini = 0.487
samples = 99
value = [67, 93]
class = No
gini = 0.47
samples = 66
value = [66, 40]
class = Yes

In [ ]: