In [1]:

```python
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
```

In [2]:

```python
from sklearn.linear_model import LogisticRegression
```

In [3]:

```python
df=pd.read_csv("C3_bot_detection_data.csv")
df
```

Out[3]:

	User ID	Username	Tweet	Retweet Count	Mention Count	Follower Count	Verified	Bot Label	Locatio
0	132131	flong	Station activity person against natural majori...	85	1	2353	False	1	Adkinsto
1	289683	hinesstephanie	Authority research natural life material staff...	55	5	9617	True	0	Sandersto
2	779715	roberttran	Manage whose quickly especially foot none to g...	6	2	4363	True	0	Harrisonfu
3	696168	pmason	Just cover eight opportunity strong policy which.	54	5	2242	True	1	Martinezber
4	704441	noah87	Animal sign six data good or.	26	3	8438	False	1	Camachovill
...	...	...	...	...	...	...	...	...	.
49995	491196	uberg	Want but put card direction know miss former h...	64	0	9911	True	1	Lak Kimberlyburg
49996	739297	jessicamunoz	Provide whole maybe agree church	18	5	9900	False	1	Greenbur

	User ID	Username	Tweet	Retweet Count	Mention Count	Follower Count	Verified	Bot Label	Locatio
			respond most ...						
49997	674475	lynncunningham	Bring different everyone international capital...	43	3	6313	True	1	Deborahfor
49998	167081	richardthompson	Than about single generation itself seek sell ...	45	1	6343	False	0	Stephensid
49999	311204	daniel29	Here morning class various room human true bec...	91	4	4006	False	0	Novakber

50000 rows × 11 columns

In [4]:

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 50000 entries, 0 to 49999
Data columns (total 11 columns):
 #   Column          Non-Null Count  Dtype 
---  ------          --------------  ----- 
 0   User ID         50000 non-null  int64 
 1   Username        50000 non-null  object
 2   Tweet           50000 non-null  object
 3   Retweet Count   50000 non-null  int64 
 4   Mention Count   50000 non-null  int64 
 5   Follower Count  50000 non-null  int64 
 6   Verified        50000 non-null  bool  
 7   Bot Label       50000 non-null  int64 
 8   Location        50000 non-null  object
 9   Created At      50000 non-null  object
 10  Hashtags        41659 non-null  object
dtypes: bool(1), int64(5), object(5)
memory usage: 3.9+ MB
```

In [5]:

```
df.columns
```

Out[5]:

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

In [6]:

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

In [7]:

```
feature_matrix.shape
```

Out[7]:

```
(50000, 5)
```

In [8]:

```
target_vector.shape
```

Out[8]:

```
(50000,)
```

In [9]:

```
from sklearn.preprocessing import StandardScaler
```

In [10]:

```
fs=StandardScaler().fit_transform(feature_matrix)
```

In [11]:

```
logr=LogisticRegression()
logr.fit(fs,target_vector)
```

Out[11]:

```
LogisticRegression()
```

In [12]:

```
observation=[[1,2,3,4,5]]
```

In [13]:

```
prediction=logr.predict(observation)
print(prediction)
```

```
[ True]
```

In [14]:

```
logr.classes_
```

Out[14]:

```
array([False,  True])
```

In [15]:

```
logr.predict_proba(observation)[0][0]
```

Out[15]:

```
0.4875957520146553
```

In [16]:

```
logr.predict_proba(observation)
```

Out[16]:

```
array([[0.48759575, 0.51240425]])
```

In [17]:

```
df['Verified'].value_counts()
```

Out[17]:

```
True     25004
False    24996
Name: Verified, dtype: int64
```

In [18]:

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

In [19]:

```
g1={ 'Verified':{'True':1,'False':2}}
df=df.replace(g1)
df
```

Out[19]:

	User ID	Username	Tweet	Retweet Count	Mention Count	Follower Count	Verified	Bot Label	Locatio
0	132131	flong	Station activity person against natural majori...	85	1	2353	False	1	Adkinsto
1	289683	hinesstephanie	Authority research natural life material staff...	55	5	9617	True	0	Sandersto
2	779715	roberttran	Manage whose quickly especially foot none to g...	6	2	4363	True	0	Harrisonfu
3	696168	pmason	Just cover eight opportunity strong policy which.	54	5	2242	True	1	Martinezber
4	704441	noah87	Animal sign six data good or.	26	3	8438	False	1	Camachovill
...	...	...	...	...	...	...	...	...	.
49995	491196	uberg	Want but put card direction know miss former h...	64	0	9911	True	1	Lak Kimberlyburg
49996	739297	jessicamunoz	Provide whole maybe agree church respond most ...	18	5	9900	False	1	Greenbur
49997	674475	lynncunningham	Bring different everyone international capital...	43	3	6313	True	1	Deborahfo

	User ID	Username	Tweet	Retweet Count	Mention Count	Follower Count	Verified	Bot Label	Locatio
49998	167081	richardthompson	Than about single generation itself seek sell ...	45	1	6343	False	0	Stephensid
49999	311204	daniel29	Here morning class various room human true bec...	91	4	4006	False	0	Novakber

50000 rows × 11 columns

In [20]:

```
from sklearn.model_selection import train_test_split
```

In [21]:

```
x_train,x_test,y_train,y_test=train_test_split(x,y,train_size=0.70)
```

In [22]:

```
from sklearn.ensemble import RandomForestClassifier
```

In [23]:

```
rfc=RandomForestClassifier()
rfc.fit(x_train,y_train)
```

Out[23]:

```
RandomForestClassifier()
```

In [24]:

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

In [25]:

```
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[25]:

```
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 [26]:

```
grid_search.best_score_
```

Out[26]:

```
0.5074285714285715
```

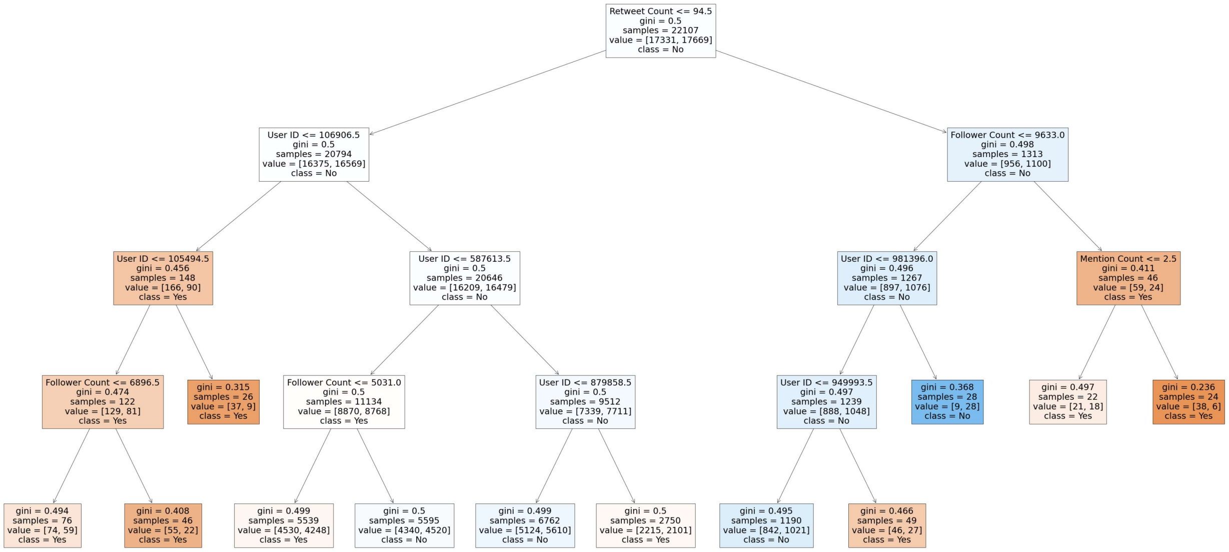
In [27]:

```
rfc_best=grid_search.best_estimator_
```

In [28]:
```
from sklearn.tree import plot_tree

plt.figure(figsize=(80,40))
plot_tree(rfc_best.estimators_[5],feature_names=x.columns,class_names=['Yes','No'],fill
```

Out[28]:
```
[Text(2391.4285714285716, 1956.96, 'Retweet Count <= 94.5\ngini = 0.5\nsamples = 22107\n
value = [17331, 17669]\nclass = No'),
 Text(1169.142857142857, 1522.0800000000002, 'User ID <= 106906.5\ngini = 0.5\nsamples =
20794\nvalue = [16375, 16569]\nclass = No'),
 Text(637.7142857142858, 1087.2, 'User ID <= 105494.5\ngini = 0.456\nsamples = 148\nvalu
e = [166, 90]\nclass = Yes'),
 Text(425.14285714285717, 652.3200000000002, 'Follower Count <= 6896.5\ngini = 0.474\nsa
mples = 122\nvalue = [129, 81]\nclass = Yes'),
 Text(212.57142857142858, 217.44000000000005, 'gini = 0.494\nsamples = 76\nvalue = [74,
59]\nclass = Yes'),
 Text(637.7142857142858, 217.44000000000005, 'gini = 0.408\nsamples = 46\nvalue = [55, 2
2]\nclass = Yes'),
 Text(850.2857142857143, 652.3200000000002, 'gini = 0.315\nsamples = 26\nvalue = [37, 9]
\nclass = Yes'),
 Text(1700.5714285714287, 1087.2, 'User ID <= 587613.5\ngini = 0.5\nsamples = 20646\nval
ue = [16209, 16479]\nclass = No'),
 Text(1275.4285714285716, 652.3200000000002, 'Follower Count <= 5031.0\ngini = 0.5\nsamp
les = 11134\nvalue = [8870, 8768]\nclass = Yes'),
 Text(1062.857142857143, 217.44000000000005, 'gini = 0.499\nsamples = 5539\nvalue = [453
0, 4248]\nclass = Yes'),
 Text(1488.0, 217.44000000000005, 'gini = 0.5\nsamples = 5595\nvalue = [4340, 4520]\ncla
ss = No'),
 Text(2125.714285714286, 652.3200000000002, 'User ID <= 879858.5\ngini = 0.5\nsamples =
9512\nvalue = [7339, 7711]\nclass = No'),
 Text(1913.1428571428573, 217.44000000000005, 'gini = 0.499\nsamples = 6762\nvalue = [51
24, 5610]\nclass = No'),
 Text(2338.285714285714, 217.44000000000005, 'gini = 0.5\nsamples = 2750\nvalue = [2215,
2101]\nclass = Yes'),
 Text(3613.714285714286, 1522.0800000000002, 'Follower Count <= 9633.0\ngini = 0.498\nsa
mples = 1313\nvalue = [956, 1100]\nclass = No'),
 Text(3188.571428571429, 1087.2, 'User ID <= 981396.0\ngini = 0.496\nsamples = 1267\nval
ue = [897, 1076]\nclass = No'),
 Text(2976.0, 652.3200000000002, 'User ID <= 949993.5\ngini = 0.497\nsamples = 1239\nval
ue = [888, 1048]\nclass = No'),
 Text(2763.4285714285716, 217.44000000000005, 'gini = 0.495\nsamples = 1190\nvalue = [84
2, 1021]\nclass = No'),
 Text(3188.571428571429, 217.44000000000005, 'gini = 0.466\nsamples = 49\nvalue = [46, 2
7]\nclass = Yes'),
 Text(3401.1428571428573, 652.3200000000002, 'gini = 0.368\nsamples = 28\nvalue = [9, 2
8]\nclass = No'),
 Text(4038.857142857143, 1087.2, 'Mention Count <= 2.5\ngini = 0.411\nsamples = 46\nvalu
e = [59, 24]\nclass = Yes'),
 Text(3826.2857142857147, 652.3200000000002, 'gini = 0.497\nsamples = 22\nvalue = [21, 1
8]\nclass = Yes'),
 Text(4251.428571428572, 652.3200000000002, 'gini = 0.236\nsamples = 24\nvalue = [38, 6]
\nclass = Yes')]
```


Retweet Count <= 94.5
gini = 0.5
samples = 22107
value = [17331, 17669]
class = No
User ID <= 106906.5
gini = 0.5
samples = 20794
value = [16375, 16569]
class = No
Follower Count <= 9633.0
gini = 0.498
samples = 1313
value = [956, 1100]
class = No
User ID <= 105494.5
gini = 0.456
samples = 148
value = [166, 90]
class = Yes
User ID <= 587613.5
gini = 0.5
samples = 20646
value = [16209, 16479]
class = No
User ID <= 981396.0
gini = 0.496
samples = 1267
value = [897, 1076]
class = No
Mention Count <= 2.5
gini = 0.411
samples = 46
value = [59, 24]
class = Yes
Follower Count <= 6896.5
gini = 0.474
samples = 122
value = [129, 81]
class = Yes
gini = 0.315
samples = 26
value = [37, 9]
class = Yes
Follower Count <= 5031.0
gini = 0.5
samples = 11134
value = [8870, 8768]
class = Yes
User ID <= 879858.5
gini = 0.5
samples = 9512
value = [7339, 7711]
class = No
User ID <= 949993.5
gini = 0.497
samples = 1239
value = [888, 1048]
class = No
gini = 0.368
samples = 28
value = [9, 28]
class = No
gini = 0.497
samples = 22
value = [21, 18]
class = Yes
gini = 0.236
samples = 24
value = [38, 6]
class = Yes
gini = 0.494
samples = 76
value = [74, 59]
class = Yes
gini = 0.408
samples = 46
value = [55, 22]
class = Yes
gini = 0.499
samples = 5539
value = [4530, 4248]
class = Yes
gini = 0.5
samples = 5595
value = [4340, 4520]
class = No
gini = 0.499
samples = 6762
value = [5124, 5610]
class = No
gini = 0.5
samples = 2750
value = [2215, 2101]
class = Yes
gini = 0.495
samples = 1190
value = [842, 1021]
class = No
gini = 0.466
samples = 49
value = [46, 27]
class = Yes