In [1]:

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

In [2]:

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

Out[2]:

	User ID	Username	Tweet	Retweet Count	Mention Count	Follower Count	Verified	Bot Label	Loc
0	132131	flong	Station activity person against natural majori...	85	1	2353	False	1	Adki
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	Harris
3	696168	pmason	Just cover eight opportunity strong policy which.	54	5	2242	True	1	Martine
4	704441	noah87	Animal sign six data good or.	26	3	8438	False	1	Camach
...	...	...	...	...	...	...	...	...	
49995	491196	uberg	Want but put card direction know miss former h...	64	0	9911	True	1	Kimberly
49996	739297	jessicamunoz	Provide whole maybe agree church respond most ...	18	5	9900	False	1	Gree
49997	674475	lynncunningham	Bring different everyone international capital...	43	3	6313	True	1	Debor
49998	167081	richardthompson	Than about single generation itself seek sell ...	45	1	6343	False	0	Stephe
49999	311204	daniel29	Here morning class various room human true bec...	91	4	4006	False	0	Nova

50000 rows × 11 columns

```
In [3]: 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 [4]: df=df.dropna()
```

```
In [5]: df.isnull().sum()
```

```
Out[5]: User ID           0
        Username          0
        Tweet             0
        Retweet Count     0
        Mention Count     0
        Follower Count    0
        Verified          0
        Bot Label         0
        Location          0
        Created At        0
        Hashtags          0
        dtype: int64
```

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["Bot Label"].value_counts()
```

Out[7]:
```
1    20838
0    20821
Name: Bot Label, dtype: int64
```

In [8]:
```
df1=df[['User ID','Retweet Count','Mention Count','Follower Count','Bot Label']
```

In [9]:
```
x=df1.drop('Bot Label',axis=1)
y=df1['Bot Label']
```

In [10]:
```
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 [11]:
```
from sklearn.ensemble import RandomForestClassifier
rfc=RandomForestClassifier()
rfc.fit(x_train,y_train)
```

Out[11]:
```
RandomForestClassifier()
```

In [12]:
```
parameters={'max_depth':[1,2,3,4,5],
            'min_samples_leaf':[5,10,15,20,25],
            'n_estimators':[10,20,30,40,50]}
```

In [13]:
```
from sklearn.model_selection import GridSearchCV
grid_search=GridSearchCV(estimator=rfc,param_grid=parameters,cv=2,scoring="accu
grid_search.fit(x_train,y_train)
```

Out[13]:
```
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 [14]:

```
grid_search.best_score_
```

Out[14]:

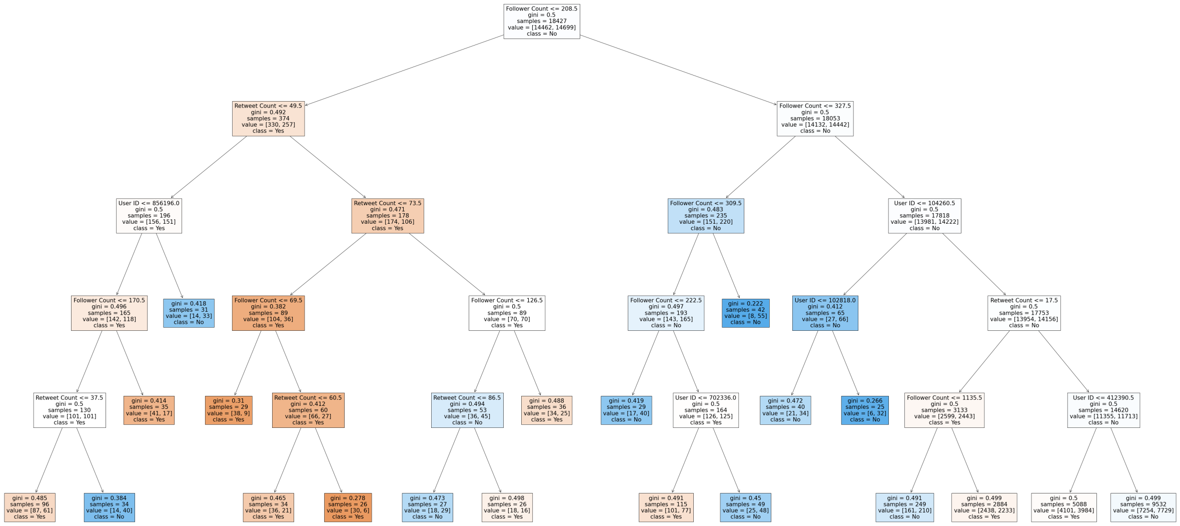
```
0.5105106975846294
```

In [15]:

```
rfc_best=grid_search.best_estimator_
```

In [16]:

```
from sklearn.tree import plot_tree
plt.figure(figsize=(80,40))
plot_tree(rfc_best.estimators_[5],feature_names=x.columns,class_names=['Yes','N
Out[16]: [Text(2064.6000000000004, 1993.2, 'Follower Count <= 208.5\ngini = 0.5\nsampl
         es = 18427\nvalue = [14462, 14699]\nclass = No'),
          Text(1041.6000000000001, 1630.8000000000002, 'Retweet Count <= 49.5\ngini =
         0.492\nsamples = 374\nvalue = [330, 257]\nclass = Yes'),
          Text(595.2, 1268.4, 'User ID <= 856196.0\ngini = 0.5\nsamples = 196\nvalue =
         [156, 151]\nclass = Yes'),
          Text(446.40000000000003, 906.0, 'Follower Count <= 170.5\ngini = 0.496\nsamp
         les = 165\nvalue = [142, 118]\nclass = Yes'),
          Text(297.6, 543.5999999999999, 'Retweet Count <= 37.5\ngini = 0.5\nsamples =
         130\nvalue = [101, 101]\nclass = Yes'),
          Text(148.8, 181.19999999999982, 'gini = 0.485\nsamples = 96\nvalue = [87, 6
         1]\nclass = Yes'),
          Text(446.40000000000003, 181.19999999999982, 'gini = 0.384\nsamples = 34\nva
         lue = [14, 40]\nclass = No'),
          Text(595.2, 543.5999999999999, 'gini = 0.414\nsamples = 35\nvalue = [41, 17]
         \nclass = Yes'),
          Text(744.0, 906.0, 'gini = 0.418\nsamples = 31\nvalue = [14, 33]\nclass = N
         o'),
          Text(1488.0, 1268.4, 'Retweet Count <= 73.5\ngini = 0.471\nsamples = 178\nva
         lue = [174, 106]\nclass = Yes'),
          Text(1041.6000000000001, 906.0, 'Follower Count <= 69.5\ngini = 0.382\nsampl
         es = 89\nvalue = [104, 36]\nclass = Yes'),
          Text(892.8000000000001, 543.5999999999999, 'gini = 0.31\nsamples = 29\nvalue
         = [38, 9]\nclass = Yes'),
          Text(1190.4, 543.5999999999999, 'Retweet Count <= 60.5\ngini = 0.412\nsample
         s = 60\nvalue = [66, 27]\nclass = Yes'),
          Text(1041.6000000000001, 181.19999999999982, 'gini = 0.465\nsamples = 34\nva
         lue = [36, 21]\nclass = Yes'),
          Text(1339.2, 181.19999999999982, 'gini = 0.278\nsamples = 26\nvalue = [30,
         6]\nclass = Yes'),
          Text(1934.4, 906.0, 'Follower Count <= 126.5\ngini = 0.5\nsamples = 89\nvalu
         e = [70, 70]\nclass = Yes'),
          Text(1785.6000000000001, 543.5999999999999, 'Retweet Count <= 86.5\ngini =
         0.494\nsamples = 53\nvalue = [36, 45]\nclass = No'),
          Text(1636.8000000000002, 181.19999999999982, 'gini = 0.473\nsamples = 27\nva
         lue = [18, 29]\nclass = No'),
          Text(1934.4, 181.19999999999982, 'gini = 0.498\nsamples = 26\nvalue = [18, 1
         6]\nclass = Yes'),
          Text(2083.2000000000003, 543.5999999999999, 'gini = 0.488\nsamples = 36\nval
         ue = [34, 25]\nclass = Yes'),
          Text(3087.6000000000004, 1630.8000000000002, 'Follower Count <= 327.5\ngini
         = 0.5\nsamples = 18053\nvalue = [14132, 14442]\nclass = No'),
          Text(2678.4, 1268.4, 'Follower Count <= 309.5\ngini = 0.483\nsamples = 235\n
         value = [151, 220]\nclass = No'),
          Text(2529.6000000000004, 906.0, 'Follower Count <= 222.5\ngini = 0.497\nsamp
         les = 193\nvalue = [143, 165]\nclass = No'),
          Text(2380.8, 543.5999999999999, 'gini = 0.419\nsamples = 29\nvalue = [17, 4
         0]\nclass = No'),
          Text(2678.4, 543.5999999999999, 'User ID <= 702336.0\ngini = 0.5\nsamples =
         164\nvalue = [126, 125]\nclass = Yes'),
          Text(2529.6000000000004, 181.19999999999982, 'gini = 0.491\nsamples = 115\nv
         alue = [101, 77]\nclass = Yes'),
          Text(2827.2000000000003, 181.19999999999982, 'gini = 0.45\nsamples = 49\nval
         ue = [25, 48]\nclass = No'),
          Text(2827.2000000000003, 906.0, 'gini = 0.222\nsamples = 42\nvalue = [8, 55]
         \nclass = No'),
          Text(3496.8, 1268.4, 'User ID <= 104260.5\ngini = 0.5\nsamples = 17818\nvalu
         e = [13981, 14222]\nclass = No'),
          Text(3124.8, 906.0, 'User ID <= 102818.0\ngini = 0.412\nsamples = 65\nvalue
         = [27, 66]\nclass = No'),
          Text(2976.0, 543.5999999999999, 'gini = 0.472\nsamples = 40\nvalue = [21, 3
         4]\nclass = No'),
          Text(3273.6000000000004, 543.5999999999999, 'gini = 0.266\nsamples = 25\nval
         ue = [6, 32]\nclass = No'),
          Text(3868.8, 906.0, 'Retweet Count <= 17.5\ngini = 0.5\nsamples = 17753\nval
         ue = [13954, 14156]\nclass = No'),
          Text(3571.2000000000003, 543.5999999999999, 'Follower Count <= 1135.5\ngini
         = 0.5\nsamples = 3133\nvalue = [2599, 2443]\nclass = Yes'),
          Text(3422.4, 181.19999999999982, 'gini = 0.491\nsamples = 249\nvalue = [161,
         210]\nclass = No'),
          Text(3720.0000000000005, 181.19999999999982, 'gini = 0.499\nsamples = 2884\n
         value = [2438, 2233]\nclass = Yes'),
          Text(4166.400000000001, 543.5999999999999, 'User ID <= 412390.5\ngini = 0.5
         \nsamples = 14620\nvalue = [11355, 11713]\nclass = No'),
          Text(4017.6000000000004, 181.19999999999982, 'gini = 0.5\nsamples = 5088\nva
         lue = [4101, 3984]\nclass = Yes'),
          Text(4315.200000000001, 181.19999999999982, 'gini = 0.499\nsamples = 9532\nv
         alue = [7254, 7729]\nclass = No')]
```


Follower Count <= 208.5 gini = 0.5 samples = 18427 value = [14462, 14699] class = No
Retweet Count <= 49.5 gini = 0.492 samples = 374 value = [330, 257] class = Yes
Follower Count <= 327.5 gini = 0.5 samples = 18053 value = [14132, 14442] class = No
User ID <= 856196.0 gini = 0.5 samples = 196 value = [156, 151] class = Yes
Retweet Count <= 73.5 gini = 0.471 samples = 178 value = [174, 106] class = Yes
Follower Count <= 309.5 gini = 0.483 samples = 235 value = [151, 220] class = No
User ID <= 104260.5 gini = 0.5 samples = 17818 value = [13981, 14222] class = No
Follower Count <= 170.5 gini = 0.496 samples = 165 value = [142, 118] class = Yes
gini = 0.418 samples = 31 value = [14, 33] class = No
Follower Count <= 69.5 gini = 0.382 samples = 89 value = [104, 36] class = Yes
Follower Count <= 126.5 gini = 0.5 samples = 89 value = [70, 70] class = Yes
Follower Count <= 222.5 gini = 0.497 samples = 193 value = [143, 165] class = No
gini = 0.222 samples = 42 value = [8, 55] class = No
User ID <= 102818.0 gini = 0.412 samples = 65 value = [27, 66] class = No
Retweet Count <= 17.5 gini = 0.5 samples = 17753 value = [13954, 14156] class = No
Retweet Count <= 37.5 gini = 0.5 samples = 130 value = [101, 101] class = Yes
gini = 0.414 samples = 35 value = [41, 17] class = Yes
gini = 0.31 samples = 29 value = [38, 9] class = Yes
Retweet Count <= 60.5 gini = 0.412 samples = 60 value = [66, 27] class = Yes
Retweet Count <= 86.5 gini = 0.494 samples = 53 value = [36, 45] class = No
gini = 0.488 samples = 36 value = [34, 25] class = Yes
gini = 0.419 samples = 29 value = [17, 40] class = No
User ID <= 702336.0 gini = 0.5 samples = 164 value = [126, 125] class = Yes
gini = 0.472 samples = 40 value = [21, 34] class = No
gini = 0.266 samples = 25 value = [6, 32] class = No
Follower Count <= 1135.5 gini = 0.5 samples = 3133 value = [2599, 2443] class = Yes
User ID <= 412390.5 gini = 0.5 samples = 14620 value = [11355, 11713] class = No
gini = 0.485 samples = 96 value = [87, 61] class = Yes
gini = 0.384 samples = 34 value = [14, 40] class = No
gini = 0.465 samples = 34 value = [36, 21] class = Yes
gini = 0.278 samples = 26 value = [30, 6] class = Yes
gini = 0.473 samples = 27 value = [18, 29] class = No
gini = 0.498 samples = 26 value = [18, 16] class = Yes
gini = 0.491 samples = 115 value = [101, 77] class = Yes
gini = 0.45 samples = 49 value = [25, 48] class = No
gini = 0.491 samples = 249 value = [161, 210] class = No
gini = 0.499 samples = 2884 value = [2438, 2233] class = Yes
gini = 0.5 samples = 5088 value = [4101, 3984] class = Yes
gini = 0.499 samples = 9532 value = [7254, 7729] class = No

```
In [ ]:
```