In [2]:

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

In [3]:
```
sv(r"C:\Users\user\Downloads\C3_bot_detection_data - C3_bot_detection_data.csv")
```

Out[3]:

sername	Tweet	Retweet Count	Mention Count	Follower Count	Verified	Bot Label	Location	Created At	Hashtag
flong	Station activity person against natural majori...	85	1	2353	False	1	Adkinston	2020-05-11 15:29:50	Na
tephanie	Authority research natural life material staff...	55	5	9617	True	0	Sanderston	2022-11-26 05:18:10	both li
oberttran	Manage whose quickly especially foot none to g...	6	2	4363	True	0	Harrisonfurt	2022-08-08 03:16:54	phor ahea
pmason	Just cover eight opportunity strong policy which.	54	5	2242	True	1	Martinezberg	2021-08-14 22:27:05	ev quick nev
noah87	Animal sign six data good or.	26	3	8438	False	1	Camachoville	2020-04-13 21:24:21	foreig mentic
...	...	...	...	...	...	...	...	...	
uberg	Want but put card direction know miss former h...	64	0	9911	True	1	Lake Kimberlyburgh	2023-04-20 11:06:26	teac quality te educatic ar
camunoz	Provide whole maybe agree church respond most ...	18	5	9900	False	1	Greenbury	2022-10-18 03:57:35	add wa amor believ
ningham	Bring different everyone international capital...	43	3	6313	True	1	Deborahfort	2020-07-08 03:54:08	on adn artist fir
ompson	Than about single generation itself seek sell ...	45	1	6343	False	0	Stephenside	2022-03-22 12:13:44	st

sername	Tweet	Retweet Count	Mention Count	Follower Count	Verified	Bot Label	Location	Created At	Hashtag
daniel29	Here morning class various room human true bec...	91	4	4006	False	0	Novakberg	2022-12-03 06:11:07	hom

ns

In [5]:

```
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 [6]:

```
df['Bot Label'].value_counts()
```

Out[6]:

```
1    25018
0    24982
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 [11]:

```
g1={"1":{'1':0}}
df1=df1.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 [12]:
```
from sklearn.model_selection import train_test_split
x_train,x_test,y_train,y_test=train_test_split(x,y,test_size=45)
```

In [13]:
```
from sklearn.ensemble import RandomForestClassifier
rfc = RandomForestClassifier()
rfc.fit(x_train,y_train)
```

Out[13]:
```
RandomForestClassifier()
```

In [14]:
```
parameters = {'max_depth':[1,2,3,4,5],
    'min_samples_leaf':[5,10,15,20,25],
    'n_estimators':[10,20,30,40,50]}
```

In [15]:
```
from sklearn.model_selection import GridSearchCV

grid_search =  GridSearchCV(estimator=rfc,param_grid=parameters,cv=2,scoring='acc
grid_search.fit(x_train,y_train)
```

Out[15]:
```
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 [16]:
```
grid_search.best_score_
```

Out[16]:
```
0.5062556735477928
```

In [17]:
```
rfc_best = grid_search.best_estimator_
```

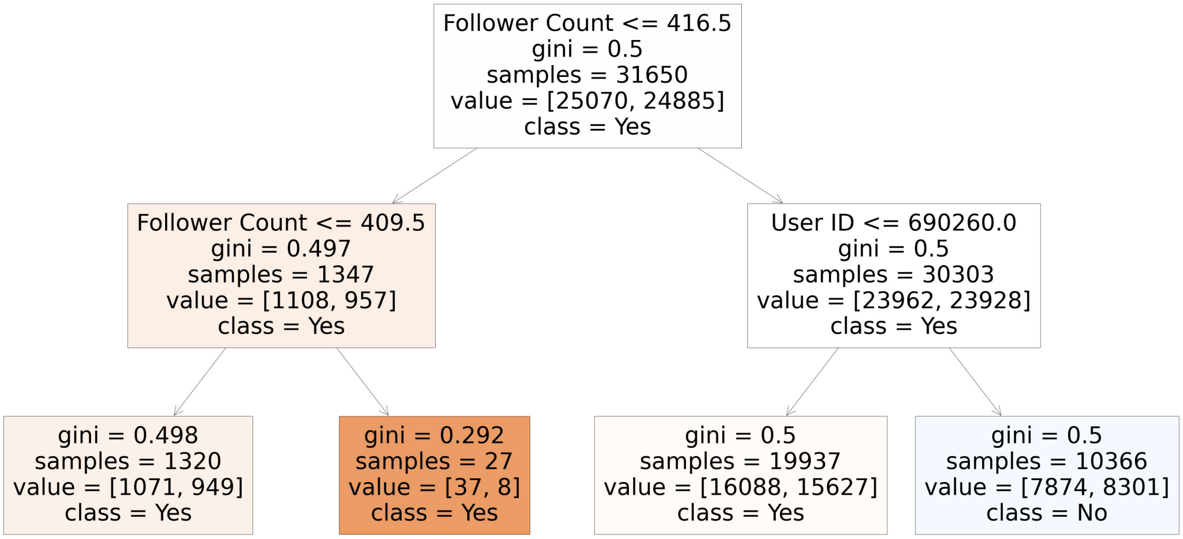
In [18]:

```
# drawing decision tree
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'
```

Out[18]:

```
[Text(2232.0, 1812.0, 'Follower Count <= 416.5\ngini = 0.5\nsamples = 31650\nva
lue = [25070, 24885]\nclass = Yes'),
 Text(1116.0, 1087.2, 'Follower Count <= 409.5\ngini = 0.497\nsamples = 1347\nv
alue = [1108, 957]\nclass = Yes'),
 Text(558.0, 362.39999999999986, 'gini = 0.498\nsamples = 1320\nvalue = [1071,
949]\nclass = Yes'),
 Text(1674.0, 362.39999999999986, 'gini = 0.292\nsamples = 27\nvalue = [37, 8]
\nclass = Yes'),
 Text(3348.0, 1087.2, 'User ID <= 690260.0\ngini = 0.5\nsamples = 30303\nvalue
= [23962, 23928]\nclass = Yes'),
 Text(2790.0, 362.39999999999986, 'gini = 0.5\nsamples = 19937\nvalue = [16088,
15627]\nclass = Yes'),
 Text(3906.0, 362.39999999999986, 'gini = 0.5\nsamples = 10366\nvalue = [7874,
8301]\nclass = No')]
```


Follower Count <= 416.5
gini = 0.5
samples = 31650
value = [25070, 24885]
class = Yes
Follower Count <= 409.5
gini = 0.497
samples = 1347
value = [1108, 957]
class = Yes
User ID <= 690260.0
gini = 0.5
samples = 30303
value = [23962, 23928]
class = Yes
gini = 0.498
samples = 1320
value = [1071, 949]
class = Yes
gini = 0.292
samples = 27
value = [37, 8]
class = Yes
gini = 0.5
samples = 19937
value = [16088, 15627]
class = Yes
gini = 0.5
samples = 10366
value = [7874, 8301]
class = No