In [1]:

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

```
In [2]: df=pd.read_csv(r"C:\Users\user\Downloads\C3_bot_detection_data.csv")
        df
```

Out[2]:

	User ID	Username	Tweet	Retweet Count	Mention Count	Follower Count	Verified	Bot Label	Lo
0	132131	flong	Station activity person against natural majori...	85	1	2353	False	1	Adl
1	289683	hinesstephanie	Authority research natural life material staff...	55	5	9617	True	0	Sanc
2	779715	roberttran	Manage whose quickly especially foot none to g...	6	2	4363	True	0	Harri
3	696168	pmason	Just cover eight opportunity strong policy which.	54	5	2242	True	1	Martin
4	704441	noah87	Animal sign six data good or.	26	3	8438	False	1	Camac
...	...	...	...	...	...	...	...	...	
49995	491196	uberg	Want but put card direction know miss former h...	64	0	9911	True	1	Kimberl
49996	739297	jessicamunoz	Provide whole maybe agree church respond most ...	18	5	9900	False	1	Gre
49997	674475	lynncunningham	Bring different everyone international capital...	43	3	6313	True	1	Debc
49998	167081	richardthompson	Than about single generation itself seek sell ...	45	1	6343	False	0	Steph
49999	311204	daniel29	Here morning class various room human true bec...	91	4	4006	False	0	Nov

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

Out[4]:
```
1    25018
0    24982
Name: Bot Label, dtype: int64
```

In [5]:
```
df1=df[['User ID','Retweet Count','Mention Count','Follower Count','Bot Label'
```

In [6]:
```
x=df1.drop('Bot Label',axis=1)
y=df['Bot Label']
```

In [7]:

```
g1={"1":{'0':1}}
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 [8]:

```
from sklearn.model_selection import train_test_split
x_train,x_test,y_train,y_test=train_test_split(x,y,test_size=0.30)
```

In [9]:

```
from sklearn.ensemble import RandomForestClassifier
rfc=RandomForestClassifier()
rfc.fit(x_train,y_train)
```

Out[9]:

```
RandomForestClassifier()
```

In [10]:

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

In [11]:

```
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[11]:

```
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 [12]:

```
grid_search.best_score_
```

Out[12]:

```
0.5052
```

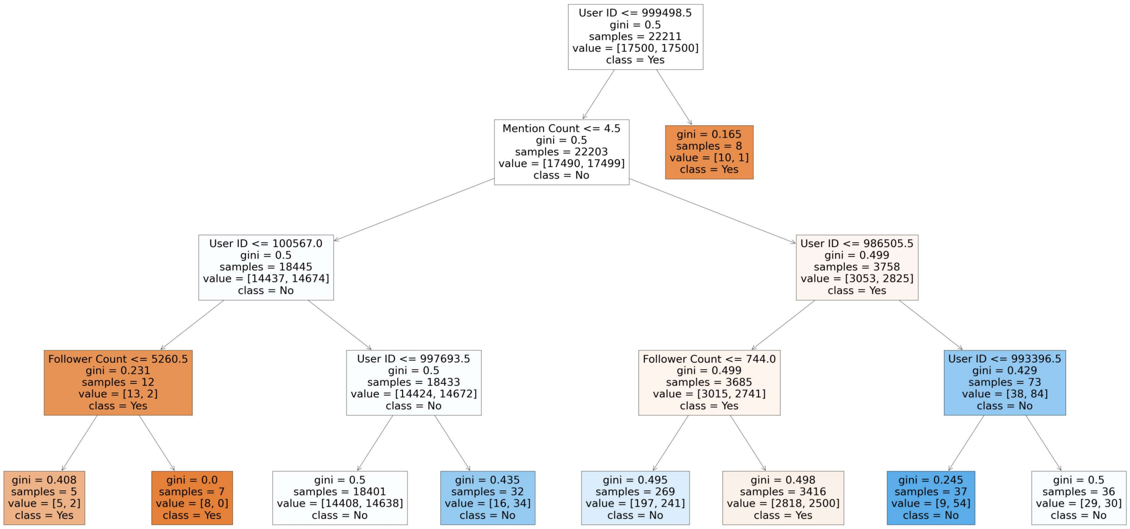
In [13]:

```
rfc_best=grid_search.best_estimator_
```

In [14]:
```
from sklearn.tree import plot_tree

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

Out[14]:
```
[Text(2511.0, 1956.96, 'User ID <= 999498.5\ngini = 0.5\nsamples = 22211\nval
ue = [17500, 17500]\nclass = Yes'),
 Text(2232.0, 1522.0800000000002, 'Mention Count <= 4.5\ngini = 0.5\nsamples
= 22203\nvalue = [17490, 17499]\nclass = No'),
 Text(1116.0, 1087.2, 'User ID <= 100567.0\ngini = 0.5\nsamples = 18445\nvalu
e = [14437, 14674]\nclass = No'),
 Text(558.0, 652.3200000000002, 'Follower Count <= 5260.5\ngini = 0.231\nsamp
les = 12\nvalue = [13, 2]\nclass = Yes'),
 Text(279.0, 217.44000000000005, 'gini = 0.408\nsamples = 5\nvalue = [5, 2]\n
class = Yes'),
 Text(837.0, 217.44000000000005, 'gini = 0.0\nsamples = 7\nvalue = [8, 0]\ncl
ass = Yes'),
 Text(1674.0, 652.3200000000002, 'User ID <= 997693.5\ngini = 0.5\nsamples =
18433\nvalue = [14424, 14672]\nclass = No'),
 Text(1395.0, 217.44000000000005, 'gini = 0.5\nsamples = 18401\nvalue = [1440
8, 14638]\nclass = No'),
 Text(1953.0, 217.44000000000005, 'gini = 0.435\nsamples = 32\nvalue = [16, 3
4]\nclass = No'),
 Text(3348.0, 1087.2, 'User ID <= 986505.5\ngini = 0.499\nsamples = 3758\nval
ue = [3053, 2825]\nclass = Yes'),
 Text(2790.0, 652.3200000000002, 'Follower Count <= 744.0\ngini = 0.499\nsamp
les = 3685\nvalue = [3015, 2741]\nclass = Yes'),
 Text(2511.0, 217.44000000000005, 'gini = 0.495\nsamples = 269\nvalue = [197,
241]\nclass = No'),
 Text(3069.0, 217.44000000000005, 'gini = 0.498\nsamples = 3416\nvalue = [281
8, 2500]\nclass = Yes'),
 Text(3906.0, 652.3200000000002, 'User ID <= 993396.5\ngini = 0.429\nsamples
= 73\nvalue = [38, 84]\nclass = No'),
 Text(3627.0, 217.44000000000005, 'gini = 0.245\nsamples = 37\nvalue = [9, 5
4]\nclass = No'),
 Text(4185.0, 217.44000000000005, 'gini = 0.5\nsamples = 36\nvalue = [29, 30]
\nclass = No'),
 Text(2790.0, 1522.0800000000002, 'gini = 0.165\nsamples = 8\nvalue = [10, 1]
\nclass = Yes')]
```


User ID <= 999498.5
gini = 0.5
samples = 22211
value = [17500, 17500]
class = Yes
Mention Count <= 4.5
gini = 0.5
samples = 22203
value = [17490, 17499]
class = No
gini = 0.165
samples = 8
value = [10, 1]
class = Yes
User ID <= 100567.0
gini = 0.5
samples = 18445
value = [14437, 14674]
class = No
User ID <= 986505.5
gini = 0.499
samples = 3758
value = [3053, 2825]
class = Yes
Follower Count <= 5260.5
gini = 0.231
samples = 12
value = [13, 2]
class = Yes
User ID <= 997693.5
gini = 0.5
samples = 18433
value = [14424, 14672]
class = No
Follower Count <= 744.0
gini = 0.499
samples = 3685
value = [3015, 2741]
class = Yes
User ID <= 993396.5
gini = 0.429
samples = 73
value = [38, 84]
class = No
gini = 0.408
samples = 5
value = [5, 2]
class = Yes
gini = 0.0
samples = 7
value = [8, 0]
class = Yes
gini = 0.5
samples = 18401
value = [14408, 14638]
class = No
gini = 0.435
samples = 32
value = [16, 34]
class = No
gini = 0.495
samples = 269
value = [197, 241]
class = No
gini = 0.498
samples = 3416
value = [2818, 2500]
class = Yes
gini = 0.245
samples = 37
value = [9, 54]
class = No
gini = 0.5
samples = 36
value = [29, 30]
class = No

In [ ]: