In [1]:
```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
```

In [178]:
```
data=pd.read_csv(r"C:\Users\user\Desktop\vicky\C3_bot_detection_data.csv")
```

In [179]:
```
data.fillna(value=1)
```

Out[179]:

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

50000 rows × 11 columns

In [180]:
```
data.head()
```

Out[180]:

	User ID	Username	Tweet	Retweet Count	Mention Count	Follower Count	Verified	Bot Label	Location	Creat
0	132131	flong	Station activity person against natural majori...	85	1	2353	False	1	Adkinston	202 05- 15:29:
1	289683	hinesstephanie	Authority research natural life material staff...	55	5	9617	True	0	Sanderston	202 11- 05:18:
2	779715	roberttran	Manage whose quickly especially foot none to g...	6	2	4363	True	0	Harrisonfurt	202 08- 03:16:
3	696168	pmason	Just cover eight opportunity strong policy which.	54	5	2242	True	1	Martinezberg	202 08- 22:27:
4	704441	noah87	Animal sign six data good or.	26	3	8438	False	1	Camachoville	202 04- 21:24:

In [181]:
```
data.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 [186]:
```
data1=data[['User ID','Retweet Count','Mention Count','Mention Count','Bot Label']
```

In [187]:
```
data1['Bot Label'].value_counts()
```

Out[187]:
```
1    25018
0    24982
Name: Bot Label, dtype: int64
```

In [166]:
```
x=data1.drop('Bot Label',axis=1)
y=data1['Bot Label']
```

In [ ]:
```

```

In [188]:
```
g1={"Bot Label":{0:2,1:3}}
data1=data1.replace(g1)
print(data1)
```

```
       User ID  Retweet Count  Mention Count  Mention Count  Bot Label
0       132131             85              1              1          3
1       289683             55              5              5          2
2       779715              6              2              2          2
3       696168             54              5              5          3
4       704441             26              3              3          3
...        ...            ...            ...            ...        ...
49995   491196             64              0              0          3
49996   739297             18              5              5          3
49997   674475             43              3              3          3
49998   167081             45              1              1          2
49999   311204             91              4              4          2

[50000 rows x 5 columns]
```

In [189]:
```
from sklearn.model_selection import train_test_split
```

In [190]:
```
x_train,x_test,y_train,y_test=train_test_split(x,y,train_size=0.70)
```

In [191]:
```
from sklearn.ensemble import RandomForestClassifier
```

In [192]:
```
rfc=RandomForestClassifier()
rfc.fit(x_train,y_train)
```

Out[192]:
```
RandomForestClassifier()
```

In [193]:
```
parameters = {'max_depth':[1,2,3,4,5],
              'min_samples_leaf':[5,10,15,20,25],
              'n_estimators':[10,20,30,40,50]

}
```

In [194]:
```
from sklearn.model_selection import GridSearchCV
grid_search=GridSearchCV(estimator=rfc,param_grid=parameters,cv=2,scoring="accurac
grid_search.fit(x_train,y_train)
```

Out[194]:
```
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 [195]:
```
grid_search.best_score_
```

Out[195]:
```
0.7014387006348421
```

In [196]:
```
from sklearn.tree import plot_tree
```

In [197]:
```
rfc_best=grid_search.best_estimator_
```

In [198]:
```
plt.figure(figsize=(80,40))
plot_tree(rfc_best.estimators_[5],feature_names=x.columns,class_names=['Yes','No']
```

Out[198]:
```
[Text(2575.3846153846152, 1902.6000000000001, 'SibSp <= 0.5\ngini = 0.465\nsample
s = 390\nvalue = [229, 394]\nclass = No'),
 Text(1373.5384615384614, 1359.0, 'Parch <= 0.5\ngini = 0.403\nsamples = 273\nval
ue = [123, 317]\nclass = No'),
 Text(686.7692307692307, 815.4000000000001, 'Pclass <= 2.5\ngini = 0.358\nsamples
= 247\nvalue = [94, 308]\nclass = No'),
 Text(343.38461538461536, 271.79999999999995, 'gini = 0.45\nsamples = 88\nvalue =
[51, 98]\nclass = No'),
 Text(1030.1538461538462, 271.79999999999995, 'gini = 0.282\nsamples = 159\nvalue
= [43, 210]\nclass = No'),
 Text(2060.3076923076924, 815.4000000000001, 'PassengerId <= 164.0\ngini = 0.361
\nsamples = 26\nvalue = [29, 9]\nclass = Yes'),
 Text(1716.9230769230767, 271.79999999999995, 'gini = 0.408\nsamples = 5\nvalue =
[2, 5]\nclass = No'),
 Text(2403.6923076923076, 271.79999999999995, 'gini = 0.225\nsamples = 21\nvalue
= [27, 4]\nclass = Yes'),
 Text(3777.230769230769, 1359.0, 'SibSp <= 3.5\ngini = 0.487\nsamples = 117\nvalu
e = [106, 77]\nclass = Yes'),
 Text(3433.8461538461534, 815.4000000000001, 'SibSp <= 2.5\ngini = 0.484\nsamples
= 109\nvalue = [102, 71]\nclass = Yes'),
 Text(3090.461538461538, 271.79999999999995, 'gini = 0.488\nsamples = 97\nvalue =
[87, 64]\nclass = Yes'),
 Text(3777.230769230769, 271.79999999999995, 'gini = 0.434\nsamples = 12\nvalue =
[15, 7]\nclass = Yes'),
 Text(4120.615384615385, 815.4000000000001, 'gini = 0.48\nsamples = 8\nvalue =
[4, 6]\nclass = No')]
```

In [ ]:

In [ ]: