Random Forest

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

In [14]:
```
df1=pd.read_csv(r"C:\Users\user\Downloads\C3_bot_detection_data.csv")
df1
```

Out[14]:

	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	Adk
1	289683	hinesstephanie	Authority research natural life material staff...	55	5	9617	True	0	Sand
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	Camac
...	...	...	...	...	...	...	...	...	
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 [16]:
```
df=df1.head(10)
df
```

Out[16]:

	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	Adkinstc
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	Martinezbe
4	704441	noah87	Animal sign six data good or.	26	3	8438	False	1	Camachovil
5	570928	james00	See wonder travel this suffer less yard office...	41	4	3792	True	1	West Cheyenn
6	734182	leonard00	Door final sound my guess building rich.	54	0	10	True	0	South Donal
7	107312	lesterdaniel	Job phone price magazine worry stage check view.	64	0	1442	False	1	Smithhave
8	549888	kimberlymorris	Eye rest prove mission show floor.	25	2	836	False	0	Lak Brittanyvil
9	117640	schmittjill	Add letter year performance western what cultu...	67	3	6523	False	1	We Hannahboroug

In [17]:
```
df['Mention Count'].value_counts()
```

Out[17]:
```
0    2
2    2
3    2
5    2
1    1
4    1
Name: Mention Count, dtype: int64
```

In [18]:
```
x=df[['User ID', 'Retweet Count', 'Mention Count',
       'Follower Count', 'Verified', 'Bot Label']]
y=df['Mention Count']
```

In [19]:
```
g1={"Verified":{'True':1,'False':2}}
df=df.replace(g1)
print(df)
```

```
   User ID        Username                                              Tweet
\
0   132131           flong  Station activity person against natural majori...
1   289683  hinesstephanie  Authority research natural life material staff...
2   779715      roberttran  Manage whose quickly especially foot none to g...
3   696168           pmason  Just cover eight opportunity strong policy which.
4   704441           noah87                      Animal sign six data good or.
5   570928         james00  See wonder travel this suffer less yard office...
6   734182       leonard00          Door final sound my guess building rich.
7   107312    lesterdaniel   Job phone price magazine worry stage check view.
8   549888  kimberlymorris                Eye rest prove mission show floor.
9   117640     schmittjill  Add letter year performance western what cultu...

   Retweet Count  Mention Count  Follower Count  Verified  Bot Label  \
0             85              1            2353     False          1
1             55              5            9617      True          0
2              6              2            4363      True          0
3             54              5            2242      True          1
4             26              3            8438     False          1
5             41              4            3792      True          1
6             54              0              10      True          0
7             64              0            1442     False          1
8             25              2             836     False          0
9             67              3            6523     False          1

             Location           Created At                          Hashtags
0           Adkinston  2020-05-11 15:29:50                               NaN
1          Sanderston  2022-11-26 05:18:10                         both live
2        Harrisonfurt  2022-08-08 03:16:54                       phone ahead
3        Martinezberg  2021-08-14 22:27:05                 ever quickly new I
4        Camachoville  2020-04-13 21:24:21                    foreign mention
5       West Cheyenne  2023-05-07 22:24:47  anyone respond perhaps market run
6        South Donald  2021-01-21 03:02:53                          president
7          Smithhaven  2022-06-12 16:45:16              option husband admit
8  Lake Brittanyville  2021-12-19 19:00:16                               NaN
9  West Hannahborough  2022-07-30 05:39:16                  available thing
```

In [20]:
```
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 [21]:
```
from sklearn.ensemble import RandomForestClassifier

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

Out[21]:
```
RandomForestClassifier()
```

In [22]:
```
parameters = { 'max_depth':[1,2,3,4,5],
    'min_samples_leaf':[5,10,15,20,25],
               'n_estimators':[10,20,30,40,50]
}
```

In [23]:
```
from sklearn.model_selection import GridSearchCV

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

```
C:\ProgramData\Anaconda3\lib\site-packages\sklearn\model_selection\_split.py:
666: UserWarning: The least populated class in y has only 1 members, which is
less than n_splits=2.
  warnings.warn(("The least populated class in y has only %d"
```

Out[23]:
```
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 [24]:
```
rf_best=grid_search.best_estimator_
print(rf_best)
```

```
RandomForestClassifier(max_depth=1, min_samples_leaf=20, n_estimators=20)
```

In [25]:
```
from sklearn.tree import plot_tree

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

Out[25]:
```
[Text(2232.0, 1087.2, 'gini = 0.612\nsamples = 5\nvalue = [4, 0, 1, 1, 0, 1]
\nclass = Yes')]
```