RANDOM FOREST - 3

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"C:\Users\BHOOMISH\Downloads\C3_bot_detection_data.csv")
df
```

Out[2]:

	User ID	Username	Tweet	Retweet Count	Mention Count	Follower Count	Verified	Bot Label	Location	Created At	Hashtags
0	132131	flong	Station activity person against natural majori...	85	1	2353	False	1	Adkinston	2020-05-11 15:29:50	NaN
1	289683	hinesstephanie	Authority research natural life material staff...	55	5	9617	True	0	Sanderston	2022-11-26 05:18:10	both live
2	779715	roberttran	Manage whose quickly especially foot none to g...	6	2	4363	True	0	Harrisonfurt	2022-08-08 03:16:54	phone ahead
3	696168	pmason	Just cover eight opportunity strong policy which.	54	5	2242	True	1	Martinezberg	2021-08-14 22:27:05	ever quickly new l
4	704441	noah87	Animal sign six data good or.	26	3	8438	False	1	Camachoville	2020-04-13 21:24:21	foreign mention
...	...	...	...	...	...	...	...	...	...	...	...
49995	491196	uberg	Want but put card direction know miss former h...	64	0	9911	True	1	Lake Kimberlyburgh	2023-04-20 11:06:26	teach quality ten education any
49996	739297	jessicamunoz	Provide whole maybe agree church respond most ...	18	5	9900	False	1	Greenbury	2022-10-18 03:57:35	add walk among believe
49997	674475	lynncunningham	Bring different everyone international capital...	43	3	6313	True	1	Deborahfort	2020-07-08 03:54:08	onto admit artist first
49998	167081	richardthompson	Than about single generation itself seek sell ...	45	1	6343	False	0	Stephenside	2022-03-22 12:13:44	star
49999	311204	daniel29	Here morning class various room human true bec...	91	4	4006	False	0	Novakberg	2022-12-03 06:11:07	home

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

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]}
```

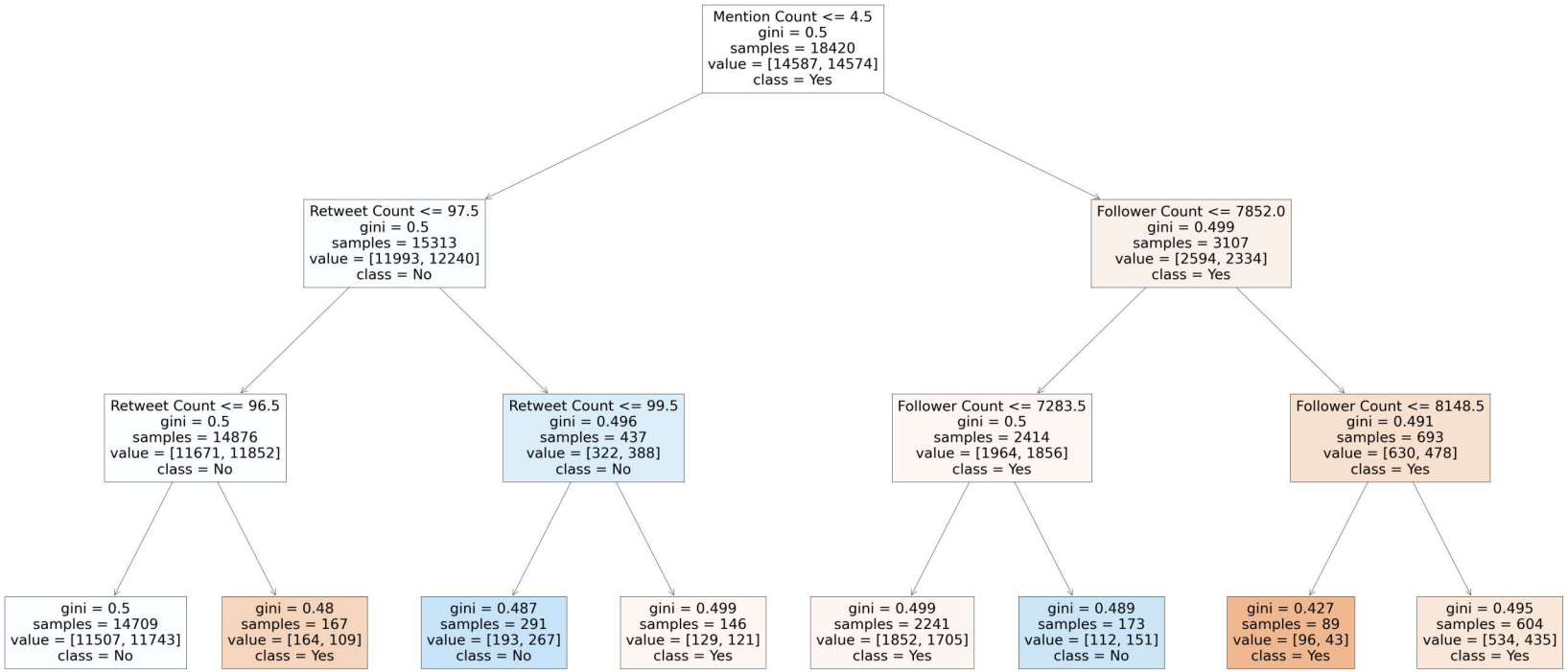
In [*]:
```
from sklearn.model_selection import GridSearchCV
grid_search=GridSearchCV(estimator=rfc,param_grid=parameters,cv=2,scoring="accuracy")
grid_search.fit(x_train,y_train)
```

In [*]:
```
grid_search.best_score_
```

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

In [20]:
```
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'],filled=True)
```

Out[20]:
```
[Text(2232.0, 1902.6000000000001, 'Mention Count <= 4.5\ngini = 0.5\nsamples = 18420\nvalue = [14587, 14574]\nclass = Yes'),
 Text(1116.0, 1359.0, 'Retweet Count <= 97.5\ngini = 0.5\nsamples = 15313\nvalue = [11993, 12240]\nclass = No'),
 Text(558.0, 815.4000000000001, 'Retweet Count <= 96.5\ngini = 0.5\nsamples = 14876\nvalue = [11671, 11852]\nclass = No'),
 Text(279.0, 271.79999999999995, 'gini = 0.5\nsamples = 14709\nvalue = [11507, 11743]\nclass = No'),
 Text(837.0, 271.79999999999995, 'gini = 0.48\nsamples = 167\nvalue = [164, 109]\nclass = Yes'),
 Text(1674.0, 815.4000000000001, 'Retweet Count <= 99.5\ngini = 0.496\nsamples = 437\nvalue = [322, 388]\nclass = No'),
 Text(1395.0, 271.79999999999995, 'gini = 0.487\nsamples = 291\nvalue = [193, 267]\nclass = No'),
 Text(1953.0, 271.79999999999995, 'gini = 0.499\nsamples = 146\nvalue = [129, 121]\nclass = Yes'),
 Text(3348.0, 1359.0, 'Follower Count <= 7852.0\ngini = 0.499\nsamples = 3107\nvalue = [2594, 2334]\nclass = Yes'),
 Text(2790.0, 815.4000000000001, 'Follower Count <= 7283.5\ngini = 0.5\nsamples = 2414\nvalue = [1964, 1856]\nclass = Yes'),
 Text(2511.0, 271.79999999999995, 'gini = 0.499\nsamples = 2241\nvalue = [1852, 1705]\nclass = Yes'),
 Text(3069.0, 271.79999999999995, 'gini = 0.489\nsamples = 173\nvalue = [112, 151]\nclass = No'),
 Text(3906.0, 815.4000000000001, 'Follower Count <= 8148.5\ngini = 0.491\nsamples = 693\nvalue = [630, 478]\nclass = Yes'),
 Text(3627.0, 271.79999999999995, 'gini = 0.427\nsamples = 89\nvalue = [96, 43]\nclass = Yes'),
 Text(4185.0, 271.79999999999995, 'gini = 0.495\nsamples = 604\nvalue = [534, 435]\nclass = Yes')]
```


Mention Count <= 4.5
gini = 0.5
samples = 18420
value = [14587, 14574]
class = Yes
Retweet Count <= 97.5
gini = 0.5
samples = 15313
value = [11993, 12240]
class = No
Follower Count <= 7852.0
gini = 0.499
samples = 3107
value = [2594, 2334]
class = Yes
Retweet Count <= 96.5
gini = 0.5
samples = 14876
value = [11671, 11852]
class = No
Retweet Count <= 99.5
gini = 0.496
samples = 437
value = [322, 388]
class = No
Follower Count <= 7283.5
gini = 0.5
samples = 2414
value = [1964, 1856]
class = Yes
Follower Count <= 8148.5
gini = 0.491
samples = 693
value = [630, 478]
class = Yes
gini = 0.5
samples = 14709
value = [11507, 11743]
class = No
gini = 0.48
samples = 167
value = [164, 109]
class = Yes
gini = 0.487
samples = 291
value = [193, 267]
class = No
gini = 0.499
samples = 146
value = [129, 121]
class = Yes
gini = 0.499
samples = 2241
value = [1852, 1705]
class = Yes
gini = 0.489
samples = 173
value = [112, 151]
class = No
gini = 0.427
samples = 89
value = [96, 43]
class = Yes
gini = 0.495
samples = 604
value = [534, 435]
class = Yes

In [ ]: