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

In [2]:
```
from sklearn.linear_model import LogisticRegression
```

In [3]:
```
df=pd.read_csv("detection.csv").dropna()

df
```

Out[3]:

	User ID	Username	Tweet	Retweet Count	Mention Count	Follower Count	Verified	Bot Label	L
1	289683	hinesstephanie	Authority research natural life material staff...	55	5	9617	True	0	Sa
2	779715	roberttran	Manage whose quickly especially foot none to g...	6	2	4363	True	0	Ha
3	696168	pmason	Just cover eight opportunity strong policy which.	54	5	2242	True	1	Mart

In [4]:
```
df.dropna(inplace=True)
```

In [5]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 41659 entries, 1 to 49999
Data columns (total 11 columns):
 #   Column          Non-Null Count  Dtype 
---  ------          --------------  ----- 
 0   User ID         41659 non-null  int64 
 1   Username        41659 non-null  object
 2   Tweet           41659 non-null  object
 3   Retweet Count   41659 non-null  int64 
 4   Mention Count   41659 non-null  int64 
 5   Follower Count  41659 non-null  int64 
 6   Verified        41659 non-null  bool  
 7   Bot Label       41659 non-null  int64 
 8   Location        41659 non-null  object
 9   Created At      41659 non-null  object
 10  Hashtags        41659 non-null  object
dtypes: bool(1), int64(5), object(5)
memory usage: 3.5+ MB
```

In [6]:
```
feature_matrix = df[['User ID','Retweet Count','Mention Count','Follower Count'
target_vector = df['Verified']
```

In [7]:
```
feature_matrix.shape
```

Out[7]:
```
(41659, 5)
```

In [8]:
```
target_vector.shape
```

Out[8]:
```
(41659,)
```

In [9]:
```
from sklearn.preprocessing import StandardScaler
```

In [10]:
```
fs = StandardScaler().fit_transform(feature_matrix)
```

In [11]:
```
logr = LogisticRegression()
logr.fit(fs,target_vector)
```

Out[11]:
```
LogisticRegression()
```

In [12]:
```
feature_matrix.shape
```

Out[12]:
```
(41659, 5)
```

In [13]:
```
target_vector.shape
```

Out[13]:
```
(41659,)
```

In [14]:
```
from sklearn.preprocessing import StandardScaler
```

In [15]:
```
fs = StandardScaler().fit_transform(feature_matrix)
```

In [16]:
```
logr = LogisticRegression()
logr.fit(fs,target_vector)
```

Out[16]:
```
LogisticRegression()
```

In [17]:
```
observation=df[['User ID','Retweet Count','Mention Count','Follower Count','Bot
```

In [18]:
```
prediction = logr.predict(observation)
prediction
```

Out[18]:
```
array([ True,  True,  True, ...,  True,  True,  True])
```

In [19]:
```
logr.classes_
```

Out[19]:
```
array([False,  True])
```

In [20]:
```
logr.predict_proba(observation)[0][1]
```

Out[20]:
```
1.0
```

In [21]:
```
df['Verified'].value_counts()
```

Out[21]:
```
True     20845
False    20814
Name: Verified, dtype: int64
```

In [26]:
```
x=df[['User ID','Retweet Count','Mention Count','Follower Count','Bot Label']]
y=df['Verified']
```

In [27]:
```
g1={'Verified':{'True':1, "False":2}}
df=df.replace(g1)
df
```

Out[27]:

	User ID	Username	Tweet	Retweet Count	Mention Count	Follower Count	Verified	Bot Label	L
1	289683	hinesstephanie	Authority research natural life material staff...	55	5	9617	True	0	Sa
2	779715	roberttran	Manage whose quickly especially foot none to g...	6	2	4363	True	0	Ha
3	696168	pmason	Just cover eight opportunity strong policy which.	54	5	2242	True	1	Mart

In [28]:
```
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 [29]:
```
from sklearn.ensemble import RandomForestClassifier
rfc = RandomForestClassifier()
rfc.fit(x_train,y_train)
```

Out[29]:
```
RandomForestClassifier()
```

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

In [31]:
```
from sklearn.model_selection import GridSearchCV
grid_search = GridSearchCV(estimator=rfc,param_grid= parameters,cv=2,scoring =
grid_search.fit(x_train,y_train)
```

Out[31]:
```
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')
```

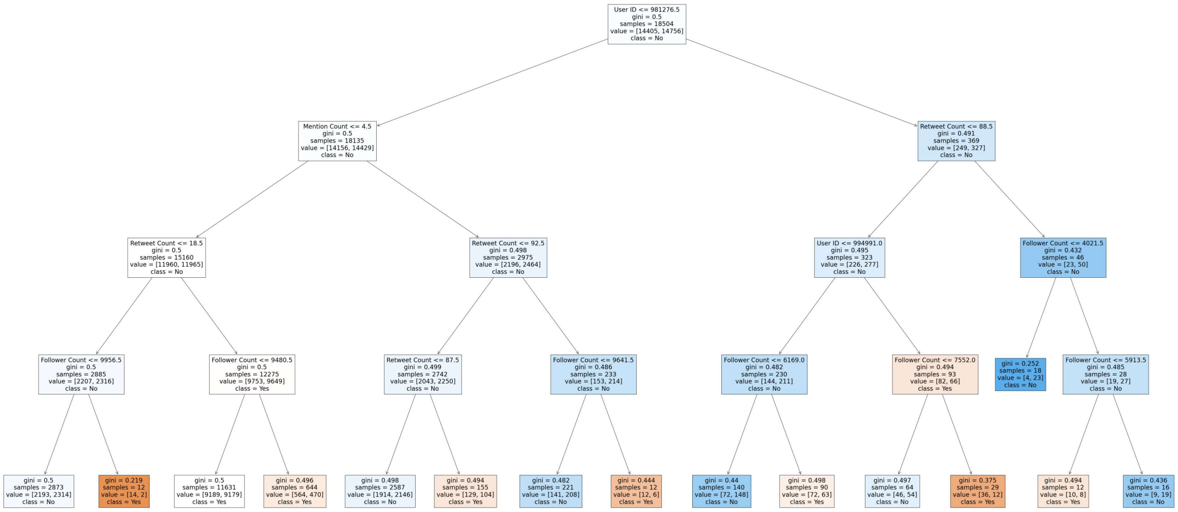
In [32]:
```
grid_search.best_score_
```

Out[32]:
```
0.508247229021664
```

In [33]:
```
rfc_best = grid_search.best_estimator_
```

In [34]:

```
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[34]: [Text(2431.285714285714, 1956.96, 'User ID <= 981276.5\ngini = 0.5\nsamples =
         18504\nvalue = [14405, 14756]\nclass = No'),
          Text(1275.4285714285713, 1522.0800000000002, 'Mention Count <= 4.5\ngini =
         0.5\nsamples = 18135\nvalue = [14156, 14429]\nclass = No'),
          Text(637.7142857142857, 1087.2, 'Retweet Count <= 18.5\ngini = 0.5\nsamples
         = 15160\nvalue = [11960, 11965]\nclass = No'),
          Text(318.85714285714283, 652.3200000000002, 'Follower Count <= 9956.5\ngini
         = 0.5\nsamples = 2885\nvalue = [2207, 2316]\nclass = No'),
          Text(159.42857142857142, 217.44000000000005, 'gini = 0.5\nsamples = 2873\nva
         lue = [2193, 2314]\nclass = No'),
          Text(478.2857142857142, 217.44000000000005, 'gini = 0.219\nsamples = 12\nval
         ue = [14, 2]\nclass = Yes'),
          Text(956.5714285714284, 652.3200000000002, 'Follower Count <= 9480.5\ngini =
         0.5\nsamples = 12275\nvalue = [9753, 9649]\nclass = Yes'),
          Text(797.1428571428571, 217.44000000000005, 'gini = 0.5\nsamples = 11631\nva
         lue = [9189, 9179]\nclass = Yes'),
          Text(1116.0, 217.44000000000005, 'gini = 0.496\nsamples = 644\nvalue = [564,
         470]\nclass = Yes'),
          Text(1913.1428571428569, 1087.2, 'Retweet Count <= 92.5\ngini = 0.498\nsampl
         es = 2975\nvalue = [2196, 2464]\nclass = No'),
          Text(1594.2857142857142, 652.3200000000002, 'Retweet Count <= 87.5\ngini =
         0.499\nsamples = 2742\nvalue = [2043, 2250]\nclass = No'),
          Text(1434.8571428571427, 217.44000000000005, 'gini = 0.498\nsamples = 2587\n
         value = [1914, 2146]\nclass = No'),
          Text(1753.7142857142856, 217.44000000000005, 'gini = 0.494\nsamples = 155\nv
         alue = [129, 104]\nclass = Yes'),
          Text(2232.0, 652.3200000000002, 'Follower Count <= 9641.5\ngini = 0.486\nsam
         ples = 233\nvalue = [153, 214]\nclass = No'),
          Text(2072.5714285714284, 217.44000000000005, 'gini = 0.482\nsamples = 221\nv
         alue = [141, 208]\nclass = No'),
          Text(2391.428571428571, 217.44000000000005, 'gini = 0.444\nsamples = 12\nval
         ue = [12, 6]\nclass = Yes'),
          Text(3587.142857142857, 1522.0800000000002, 'Retweet Count <= 88.5\ngini =
         0.491\nsamples = 369\nvalue = [249, 327]\nclass = No'),
          Text(3188.5714285714284, 1087.2, 'User ID <= 994991.0\ngini = 0.495\nsamples
         = 323\nvalue = [226, 277]\nclass = No'),
          Text(2869.7142857142853, 652.3200000000002, 'Follower Count <= 6169.0\ngini
         = 0.482\nsamples = 230\nvalue = [144, 211]\nclass = No'),
          Text(2710.285714285714, 217.44000000000005, 'gini = 0.44\nsamples = 140\nval
         ue = [72, 148]\nclass = No'),
          Text(3029.142857142857, 217.44000000000005, 'gini = 0.498\nsamples = 90\nval
         ue = [72, 63]\nclass = Yes'),
          Text(3507.428571428571, 652.3200000000002, 'Follower Count <= 7552.0\ngini =
         0.494\nsamples = 93\nvalue = [82, 66]\nclass = Yes'),
          Text(3347.9999999999995, 217.44000000000005, 'gini = 0.497\nsamples = 64\nva
         lue = [46, 54]\nclass = No'),
          Text(3666.8571428571427, 217.44000000000005, 'gini = 0.375\nsamples = 29\nva
         lue = [36, 12]\nclass = Yes'),
          Text(3985.7142857142853, 1087.2, 'Follower Count <= 4021.5\ngini = 0.432\nsa
         mples = 46\nvalue = [23, 50]\nclass = No'),
          Text(3826.2857142857138, 652.3200000000002, 'gini = 0.252\nsamples = 18\nval
         ue = [4, 23]\nclass = No'),
          Text(4145.142857142857, 652.3200000000002, 'Follower Count <= 5913.5\ngini =
         0.485\nsamples = 28\nvalue = [19, 27]\nclass = No'),
          Text(3985.7142857142853, 217.44000000000005, 'gini = 0.494\nsamples = 12\nva
         lue = [10, 8]\nclass = Yes'),
 Text(4304.571428571428, 217.44000000000005, 'gini = 0.436\nsamples = 16\nval
ue = [9, 19]\nclass = No')]
```


User ID <= 981276.5 gini = 0.5 samples = 18504 value = [14405, 14756] class = No
Mention Count <= 4.5 gini = 0.5 samples = 18135 value = [14156, 14429] class = No
Retweet Count <= 88.5 gini = 0.491 samples = 369 value = [249, 327] class = No
Retweet Count <= 18.5 gini = 0.5 samples = 15160 value = [11960, 11965] class = No
Retweet Count <= 92.5 gini = 0.498 samples = 2975 value = [2196, 2464] class = No
User ID <= 994991.0 gini = 0.495 samples = 323 value = [226, 277] class = No
Follower Count <= 4021.5 gini = 0.432 samples = 46 value = [23, 50] class = No
Follower Count <= 9956.5 gini = 0.5 samples = 2885 value = [2207, 2316] class = No
Follower Count <= 9480.5 gini = 0.5 samples = 12275 value = [9753, 9649] class = Yes
Retweet Count <= 87.5 gini = 0.499 samples = 2742 value = [2043, 2250] class = No
Follower Count <= 9641.5 gini = 0.486 samples = 233 value = [153, 214] class = No
Follower Count <= 6169.0 gini = 0.482 samples = 230 value = [144, 211] class = No
Follower Count <= 7552.0 gini = 0.494 samples = 93 value = [82, 66] class = Yes
gini = 0.252 samples = 18 value = [4, 23] class = No
Follower Count <= 5913.5 gini = 0.485 samples = 28 value = [19, 27] class = No
gini = 0.5 samples = 2873 value = [2193, 2314] class = No
gini = 0.219 samples = 12 value = [14, 2] class = Yes
gini = 0.5 samples = 11631 value = [9189, 9179] class = Yes
gini = 0.496 samples = 644 value = [564, 470] class = Yes
gini = 0.498 samples = 2587 value = [1914, 2146] class = No
gini = 0.494 samples = 155 value = [129, 104] class = Yes
gini = 0.482 samples = 221 value = [141, 208] class = No
gini = 0.444 samples = 12 value = [12, 6] class = Yes
gini = 0.44 samples = 140 value = [72, 148] class = No
gini = 0.498 samples = 90 value = [72, 63] class = Yes
gini = 0.497 samples = 64 value = [46, 54] class = No
gini = 0.375 samples = 29 value = [36, 12] class = Yes
gini = 0.494 samples = 12 value = [10, 8] class = Yes
gini = 0.436 samples = 16 value = [9, 19] class = No