

```
In [2]: import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.linear_model import LogisticRegression
from sklearn.preprocessing import StandardScaler
import re
from sklearn.datasets import load_digits
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.ensemble import RandomForestClassifier
from sklearn.model_selection import GridSearchCV
from sklearn.tree import plot_tree
```

```
In [3]: df=pd.read_csv("C3_bot_detection_data - C3_bot_detection_data.csv")
df
```

Out[3]:

| | 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 | Harrisonfuri |
| 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 | Deborahfori |
| 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 [4]: df1=df.fillna(value=0)
df1
```

Out[4]:

| | 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 | Harrisonfuri |
| 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 | Deborahfori |
| 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 [5]: df1.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              50000 non-null  object
dtypes: bool(1), int64(5), object(5)
memory usage: 3.9+ MB
```

In [6]: df1.columns

Out[6]: Index(['User ID', 'Username', 'Tweet', 'Retweet Count', 'Mention Count', 'Follower Count', 'Verified', 'Bot Label', 'Location', 'Created At', 'Hashtags'], dtype='object')

In [7]: df2=df1[['User ID', 'Retweet Count', 'Mention Count', 'Follower Count', 'Bot Label']]
df2

Out[7]:

| | User ID | Retweet Count | Mention Count | Follower Count | Bot Label |
|-------|---------|---------------|---------------|----------------|-----------|
| 0 | 132131 | 85 | 1 | 2353 | 1 |
| 1 | 289683 | 55 | 5 | 9617 | 0 |
| 2 | 779715 | 6 | 2 | 4363 | 0 |
| 3 | 696168 | 54 | 5 | 2242 | 1 |
| 4 | 704441 | 26 | 3 | 8438 | 1 |
| ... | ... | ... | ... | ... | ... |
| 49995 | 491196 | 64 | 0 | 9911 | 1 |
| 49996 | 739297 | 18 | 5 | 9900 | 1 |
| 49997 | 674475 | 43 | 3 | 6313 | 1 |
| 49998 | 167081 | 45 | 1 | 6343 | 0 |
| 49999 | 311204 | 91 | 4 | 4006 | 0 |

50000 rows × 5 columns

```
In [9]: df2['Bot Label'].value_counts()
```

```
Out[9]: 1    25018
        0    24982
        Name: Bot Label, dtype: int64
```

```
In [10]: x=df2.drop('Bot Label',axis=1)
         y=df2['Bot Label']
```

```
In [12]: x_train,x_test,y_train,y_test = train_test_split(x,y,test_size=0.70)
```

```
In [13]: rfc=RandomForestClassifier()
         rfc.fit(x_train,y_train)
```

```
Out[13]: RandomForestClassifier()
```

```
In [14]: parameters = {'max_depth':[1,2,3,4,5],
                        'min_samples_leaf':[5,10,15,20,25],
                        'n_estimators':[10,20,30,40,50]}
```

```
In [15]: grid_search = GridSearchCV(estimator=rfc,param_grid=parameters,cv=2,scoring='accuracy')
         grid_search.fit(x_train,y_train)
```

```
Out[15]: 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 [16]: grid_search.best_score_
```

```
Out[16]: 0.5100666666666667
```

```
In [21]: rfc_best =grid_search.best_estimator_
```



```
In [22]: plt.figure(figsize=(80,50))
plot_tree(rfc_best.estimators_[5],feature_names=x.columns,class_names=['Yes','No'])
```

```
Out[22]: [Text(2697.0, 2446.2, 'User ID <= 192194.0\ngini = 0.5\nsamples = 9449\nvalue =
[7347, 7653]\n\nclass = No'),
Text(1488.0, 1902.6, 'User ID <= 126452.0\ngini = 0.492\nsamples = 950\nvalue
= [664, 857]\n\nclass = No'),
Text(744.0, 1359.0, 'Retweet Count <= 60.5\ngini = 0.5\nsamples = 252\nvalue =
[208, 204]\n\nclass = Yes'),
Text(372.0, 815.3999999999999, 'Retweet Count <= 7.5\ngini = 0.493\nsamples =
157\nvalue = [113, 143]\n\nclass = No'),
Text(186.0, 271.79999999999997, 'gini = 0.444\nsamples = 24\nvalue = [30, 15]\n
class = Yes'),
Text(558.0, 271.79999999999997, 'gini = 0.477\nsamples = 133\nvalue = [83, 128]
\n\nclass = No'),
Text(1116.0, 815.3999999999999, 'Follower Count <= 5697.5\ngini = 0.476\nsampl
es = 95\nvalue = [95, 61]\n\nclass = Yes'),
Text(930.0, 271.79999999999997, 'gini = 0.5\nsamples = 52\nvalue = [43, 44]\ncl
ass = No'),
Text(1302.0, 271.79999999999997, 'gini = 0.371\nsamples = 43\nvalue = [52, 17]
\n\nclass = Yes'),
Text(2232.0, 1359.0, 'Mention Count <= 4.5\ngini = 0.484\nsamples = 698\nvalue
[456, 653]\n\nclass = Yes')]
```

```
In [ ]:
```