**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(r"C:\Users\user\Desktop\C3\_bot\_detection\_data.csv")

df

Out[3]:

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

50000 rows × 11 columns

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]:

df**.**columns

Out[6]:

Index(['User ID', 'Username', 'Tweet', 'Retweet Count', 'Mention Count',

'Follower Count', 'Verified', 'Bot Label', 'Location', 'Created At',

'Hashtags'],

dtype='object')

In [14]:

feature\_matrix **=** df[['User ID', 'Retweet Count', 'Mention Count',

'Follower Count', 'Verified', 'Bot Label']]

target\_vector **=** df['Verified']

In [15]:

feature\_matrix**.**shape

Out[15]:

(41659, 6)

In [16]:

target\_vector**.**shape

Out[16]:

(41659,)

In [17]:

**from** sklearn.preprocessing **import** StandardScaler

In [18]:

fs **=** StandardScaler()**.**fit\_transform(feature\_matrix)

In [19]:

logr**=** LogisticRegression()

logr**.**fit(fs,target\_vector)

Out[19]:

LogisticRegression()

In [23]:

observation**=**df[['User ID', 'Retweet Count', 'Mention Count',

'Follower Count', 'Verified', 'Bot Label']]

In [24]:

prediction **=** logr**.**predict(observation)

prediction

Out[24]:

array([ True, True, True, ..., True, True, True])

In [25]:

logr**.**classes\_

Out[25]:

array([False, True])

In [26]:

logr**.**predict\_proba(observation)[0][1]

Out[26]:

1.0

In [27]:

df['Verified']**.**value\_counts()

Out[27]:

True 20845

False 20814

Name: Verified, dtype: int64

In [28]:

x**=**df[['User ID', 'Retweet Count', 'Mention Count',

'Follower Count', 'Verified', 'Bot Label']]

y**=**df['Verified']

In [29]:

g1**=**{'Verified':{'True':1,"False":2}}

df**=**df**.**replace(g1)

df

Out[29]:

|  | **User ID** | **Username** | **Tweet** | **Retweet Count** | **Mention Count** | **Follower Count** | **Verified** | **Bot Label** | **Location** | **Created At** | **Hashtags** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **1** | 289683 | hinesstephanie | Authority research natural life material staff... | 55 | 5 | 9617 | True | 0 | Sanderston | 26-11-2022 05:18 | both live |
| **2** | 779715 | roberttran | Manage whose quickly especially foot none to g... | 6 | 2 | 4363 | True | 0 | Harrisonfurt | 08-08-2022 03:16 | phone ahead |
| **3** | 696168 | pmason | Just cover eight opportunity strong policy which. | 54 | 5 | 2242 | True | 1 | Martinezberg | 14-08-2021 22:27 | ever quickly new I |
| **4** | 704441 | noah87 | Animal sign six data good or. | 26 | 3 | 8438 | False | 1 | Camachoville | 13-04-2020 21:24 | foreign mention |
| **5** | 570928 | james00 | See wonder travel this suffer less yard office... | 41 | 4 | 3792 | True | 1 | West Cheyenne | 07-05-2023 22:24 | anyone respond perhaps market run |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| **49995** | 491196 | uberg | Want but put card direction know miss former h... | 64 | 0 | 9911 | True | 1 | Lake Kimberlyburgh | 20-04-2023 11:06 | teach quality ten education any |
| **49996** | 739297 | jessicamunoz | Provide whole maybe agree church respond most ... | 18 | 5 | 9900 | False | 1 | Greenbury | 18-10-2022 03:57 | add walk among believe |
| **49997** | 674475 | lynncunningham | Bring different everyone international capital... | 43 | 3 | 6313 | True | 1 | Deborahfort | 08-07-2020 03:54 | onto admit artist first |
| **49998** | 167081 | richardthompson | Than about single generation itself seek sell ... | 45 | 1 | 6343 | False | 0 | Stephenside | 22-03-2022 12:13 | star |
| **49999** | 311204 | daniel29 | Here morning class various room human true bec... | 91 | 4 | 4006 | False | 0 | Novakberg | 03-12-2022 06:11 | home |

41659 rows × 11 columns

In [30]:

**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 [31]:

**from** sklearn.ensemble **import** RandomForestClassifier

rfc **=** RandomForestClassifier()

rfc**.**fit(x\_train,y\_train)

Out[31]:

RandomForestClassifier()

In [32]:

parameters **=**{'max\_depth':[1, 2, 3, 4,5],'min\_samples\_leaf':[5,10,15,20,25],

'n\_estimators':[10,20,30,40,50]}

In [33]:

**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)

Out[33]:

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 [34]:

grid\_search**.**best\_score\_

Out[34]:

1.0

In [35]:

rfc\_best **=**grid\_search**.**best\_estimator\_

In [36]:

**from** sklearn.tree **import** plot\_tree

plt**.**figure(figsize**=**(89,40))

plot\_tree(rfc\_best**.**estimators\_[5], feature\_names**=**x**.**columns, class\_names**=**['Yes','No'], filled**=True**)

Out[36]:

[Text(2483.1, 1630.8000000000002, 'Verified <= 0.5\ngini = 0.5\nsamples = 18480\nvalue = [14562, 14599]\nclass = No'),

Text(1241.55, 543.5999999999999, 'gini = 0.0\nsamples = 9251\nvalue = [14562, 0]\nclass = Yes'),

Text(3724.6499999999996, 543.5999999999999, 'gini = 0.0\nsamples = 9229\nvalue = [0, 14599]\nclass = No')]

In [ ]:

