In [32]:

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
sns.set_style("darkgrid")
pd.set_option('display.max_columns',50)
pd.set_option('display.max_rows',50000000)

from sklearn.naive_bayes import GaussianNB
from sklearn import metrics

from sklearn.model_selection import train_test_split
```

In [3]:

```
data = pd.read_csv("kddcup99_csv.csv")
```

In [5]:

```
data.shape
```

Out[5]:

```
(494020, 42)
```

In [55]:

```
data.head(10)
```

Out[55]:

	duration	src_bytes	dst_bytes	land	wrong_fragment	urgent	hot	num_failed_logins	logge
0	0	181	5450	0	0	0	0	0	
1	0	239	486	0	0	0	0	0	
2	0	235	1337	0	0	0	0	0	
3	0	219	1337	0	0	0	0	0	
4	0	217	2032	0	0	0	0	0	
5	0	217	2032	0	0	0	0	0	
6	0	212	1940	0	0	0	0	0	
7	0	159	4087	0	0	0	0	0	
8	0	210	151	0	0	0	0	0	
9	0	212	786	0	0	0	1	0	

In [30]:

```
label = []
for i in range (494020):
    if(data['label'][i]=='normal'):
        label.append(0)
    else:
        label.append(1)

del data['label'] # deleting label column

data['label']=label # adding label list as column

print(data.info())
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 494020 entries, 0 to 494019
Data columns (total 42 columns):
 #   Column                       Non-Null Count   Dtype  
---  ------                       --------------   -----  
 0   duration                     494020 non-null  int64  
 1   protocol_type                494020 non-null  object 
 2   service                      494020 non-null  object 
 3   flag                         494020 non-null  object 
 4   src_bytes                    494020 non-null  int64  
 5   dst_bytes                    494020 non-null  int64  
 6   land                         494020 non-null  int64  
 7   wrong_fragment               494020 non-null  int64  
 8   urgent                       494020 non-null  int64  
 9   hot                          494020 non-null  int64  
 10  num_failed_logins            494020 non-null  int64  
 11  logged_in                    494020 non-null  int64  
 12  lnum_compromised             494020 non-null  int64  
 13  lroot_shell                  494020 non-null  int64  
 14  lsu_attempted                494020 non-null  int64  
 15  lnum_root                    494020 non-null  int64  
 16  lnum_file_creations          494020 non-null  int64  
 17  lnum_shells                  494020 non-null  int64  
 18  lnum_access_files            494020 non-null  int64  
 19  lnum_outbound_cmds           494020 non-null  int64  
 20  is_host_login                494020 non-null  int64  
 21  is_guest_login               494020 non-null  int64  
 22  count                        494020 non-null  int64  
 23  srv_count                    494020 non-null  int64  
 24  serror_rate                  494020 non-null  float64
 25  srv_serror_rate              494020 non-null  float64
 26  rerror_rate                  494020 non-null  float64
 27  srv_rerror_rate              494020 non-null  float64
 28  same_srv_rate                494020 non-null  float64
 29  diff_srv_rate                494020 non-null  float64
 30  srv_diff_host_rate           494020 non-null  float64
 31  dst_host_count               494020 non-null  int64  
 32  dst_host_srv_count           494020 non-null  int64  
 33  dst_host_same_srv_rate       494020 non-null  float64
 34  dst_host_diff_srv_rate       494020 non-null  float64
 35  dst_host_same_src_port_rate  494020 non-null  float64
 36  dst_host_srv_diff_host_rate  494020 non-null  float64
 37  dst_host_serror_rate         494020 non-null  float64
 38  dst_host_srv_serror_rate     494020 non-null  float64
 39  dst_host_rerror_rate         494020 non-null  float64
 40  dst_host_srv_rerror_rate     494020 non-null  float64
```

```
 41  label                        494020 non-null  int64
dtypes: float64(15), int64(24), object(3)
memory usage: 158.3+ MB
None
```

In [31]:

```
1  print(data.columns)
```

```
Index(['duration', 'protocol_type', 'service', 'flag', 'src_bytes',
       'dst_bytes', 'land', 'wrong_fragment', 'urgent', 'hot',
       'num_failed_logins', 'logged_in', 'lnum_compromised', 'lroot_shell',
       'lsu_attempted', 'lnum_root', 'lnum_file_creations', 'lnum_shells',
       'lnum_access_files', 'lnum_outbound_cmds', 'is_host_login',
       'is_guest_login', 'count', 'srv_count', 'serror_rate',
       'srv_serror_rate', 'rerror_rate', 'srv_rerror_rate', 'same_srv_rate',
       'diff_srv_rate', 'srv_diff_host_rate', 'dst_host_count',
       'dst_host_srv_count', 'dst_host_same_srv_rate',
       'dst_host_diff_srv_rate', 'dst_host_same_src_port_rate',
       'dst_host_srv_diff_host_rate', 'dst_host_serror_rate',
       'dst_host_srv_serror_rate', 'dst_host_rerror_rate',
       'dst_host_srv_rerror_rate', 'label'],
      dtype='object')
```

In [48]:

```
1  Y = data.label
2
3  del data['label']
4
5  X= data
6  X_train, X_test, Y_train, Y_test= train_test_split(X, Y, test_size=0.4, random_state=1)
7
8  print(X_train.shape)
9  print(X_test.shape)
10 print(Y_train.shape)
11 print(Y_test.shape)
```

```
(296412, 38)
(197608, 38)
(296412,)
(197608,)
```

In [49]:

```
1  gnb = GaussianNB()
2  y1_pred = gnb.fit(X_train, Y_train).predict(X_test)
3  print("Accuracy:",metrics.accuracy_score(Y_test, y1_pred))
```

```
Accuracy: 0.9789077365288855
```

In [43]:

```
1  del data['protocol_type']
2  del data['service']
3  del data['flag']
```

In [45]:

```
1 print(data.info())
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 494020 entries, 0 to 494019
Data columns (total 39 columns):
 #   Column                       Non-Null Count   Dtype  
---  ------                       --------------   -----  
 0   duration                     494020 non-null  int64  
 1   src_bytes                    494020 non-null  int64  
 2   dst_bytes                    494020 non-null  int64  
 3   land                         494020 non-null  int64  
 4   wrong_fragment               494020 non-null  int64  
 5   urgent                       494020 non-null  int64  
 6   hot                          494020 non-null  int64  
 7   num_failed_logins            494020 non-null  int64  
 8   logged_in                    494020 non-null  int64  
 9   lnum_compromised             494020 non-null  int64  
 10  lroot_shell                  494020 non-null  int64  
 11  lsu_attempted                494020 non-null  int64  
 12  lnum_root                    494020 non-null  int64  
 13  lnum_file_creations          494020 non-null  int64  
 14  lnum_shells                  494020 non-null  int64  
 15  lnum_access_files            494020 non-null  int64  
 16  lnum_outbound_cmds           494020 non-null  int64  
 17  is_host_login                494020 non-null  int64  
 18  is_guest_login               494020 non-null  int64  
 19  count                        494020 non-null  int64  
 20  srv_count                    494020 non-null  int64  
 21  serror_rate                  494020 non-null  float64
 22  srv_serror_rate              494020 non-null  float64
 23  rerror_rate                  494020 non-null  float64
 24  srv_rerror_rate              494020 non-null  float64
 25  same_srv_rate                494020 non-null  float64
 26  diff_srv_rate                494020 non-null  float64
 27  srv_diff_host_rate           494020 non-null  float64
 28  dst_host_count               494020 non-null  int64  
 29  dst_host_srv_count           494020 non-null  int64  
 30  dst_host_same_srv_rate       494020 non-null  float64
 31  dst_host_diff_srv_rate       494020 non-null  float64
 32  dst_host_same_src_port_rate  494020 non-null  float64
 33  dst_host_srv_diff_host_rate  494020 non-null  float64
 34  dst_host_serror_rate         494020 non-null  float64
 35  dst_host_srv_serror_rate     494020 non-null  float64
 36  dst_host_rerror_rate         494020 non-null  float64
 37  dst_host_srv_rerror_rate     494020 non-null  float64
 38  label                        494020 non-null  int64  
dtypes: float64(15), int64(24)
memory usage: 147.0 MB
None
```

In [50]:

```
1 from sklearn.preprocessing import StandardScaler
```

In [53]:

```
1 x = StandardScaler().fit_transform(data)
```

In [54]:

```
1 data.head(10)
```

Out[54]:

	duration	src_bytes	dst_bytes	land	wrong_fragment	urgent	hot	num_failed_logins	logge
0	0	181	5450	0	0	0	0	0	
1	0	239	486	0	0	0	0	0	
2	0	235	1337	0	0	0	0	0	
3	0	219	1337	0	0	0	0	0	
4	0	217	2032	0	0	0	0	0	
5	0	217	2032	0	0	0	0	0	
6	0	212	1940	0	0	0	0	0	
7	0	159	4087	0	0	0	0	0	
8	0	210	151	0	0	0	0	0	
9	0	212	786	0	0	0	1	0	

In [56]:

```
1 from sklearn.decomposition import PCA
2
```

In [57]:

```
1 pca = PCA(n_components=5)
```

In [58]:

```
1 principalComponents = pca.fit_transform(x)
```

In [60]:

```
1 principalDf = pd.DataFrame(data = principalComponents
2              , columns = ['principal component 1', 'principal component 2', 'principal
```

In [66]:

```
1
```

	principal component 1	principal component 2	principal component 3	principal component 4	principal component 5
10	0.088392	1.289797	3.939393	-1.424584	-0.522773
11	0.102584	1.252110	3.896632	-1.405588	-0.491536
12	0.054376	1.216496	3.815990	-1.378223	-0.479459
13	-0.089294	1.279506	3.940688	-1.423432	-0.521793
14	-0.008495	1.281276	3.977258	-1.438595	-0.520729
15	-0.044831	1.249219	3.910163	-1.412952	-0.509908
16	-0.051893	1.232859	3.881073	-1.404304	-0.501710
17	-0.102108	1.166052	3.758386	-1.357802	-0.464341
18	-0.111284	1.317311	4.192792	-1.524404	-0.630445
19	-0.156242	1.118001	3.659098	-1.322474	-0.444572

In [69]:

```
Y2 = label

X2= principalDf

X2_train, X2_test, Y2_train, Y2_test= train_test_split(X2, Y2, test_size=0.4, random_st

gnb = GaussianNB()
y2_pred = gnb.fit(X2_train, Y2_train).predict(X2_test)
print("Accuracy:",metrics.accuracy_score(Y2_test, y2_pred))
```

```
Accuracy: 0.9472136755596939
```

In [ ]:

```
1
```