CSE4020 LAB1

# M SHARATH SRIVATSAN 19BCE1688

In [3]:

```
import os
import pandas as pd
df=pd.read_csv("/content/drive/MyDrive/KDD_Train.csv")
```

In [5]:

```
from google.colab import drive
drive.mount('/content/drive')
```

```
Drive already mounted at /content/drive; to attempt to forcibly remount, call drive.mount
("/content/drive", force_remount=True).
```

KDDCUP DATASET

In [21]:

```
df.shape
df
```

Out[21]:

| | duration | protocol_type | service | flag | src_bytes | dst_bytes | land | wrong_fragment | urgent | hot | ... | dst_host_srv_cou |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 0 | tcp | ftp_data | SF | 491 | 0 | 0 | 0 | 0 | 0 | ... | 2 |
| **1** | 0 | udp | other | SF | 146 | 0 | 0 | 0 | 0 | 0 | ... | |
| **2** | 0 | tcp | private | S0 | 0 | 0 | 0 | 0 | 0 | 0 | ... | 2 |
| **3** | 0 | tcp | http | SF | 232 | 8153 | 0 | 0 | 0 | 0 | ... | 25 |
| **4** | 0 | tcp | http | SF | 199 | 420 | 0 | 0 | 0 | 0 | ... | 25 |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | |
| **125968** | 0 | tcp | private | S0 | 0 | 0 | 0 | 0 | 0 | 0 | ... | 2 |
| **125969** | 8 | udp | private | SF | 105 | 145 | 0 | 0 | 0 | 0 | ... | 24 |
| **125970** | 0 | tcp | smtp | SF | 2231 | 384 | 0 | 0 | 0 | 0 | ... | 3 |
| **125971** | 0 | tcp | klogin | S0 | 0 | 0 | 0 | 0 | 0 | 0 | ... | |
| **125972** | 0 | tcp | ftp_data | SF | 151 | 0 | 0 | 0 | 0 | 0 | ... | 7 |

125973 rows × 42 columns

In [20]:

```
#finding count of 'normal' and 'anomaly' using a for loop
countn=0
counta=0
for item in df['class']:
  if item=='normal':
    countn+=1
  else:
    counta+=1
print(countn)
print(counta)
```

```
0
```

125973

In [7]:

```
#finding count of 'normal' and 'anomaly' using value_counts
df['class'].value_counts()
```

Out[7]:

```
normal     67343
anomaly    58630
Name: class, dtype: int64
```

In [9]:

```
df['class']=df['class'].replace('normal',0).replace('anomaly',1)
#or- df['class'].replace(('normal','anomaly'),(0,1),inplace=True)
```

In [25]:

```
#splitting the data set into two based on class
halfn=df[df['class']==0]
halfa=df[df['class']==1]
halfn
```

Out[25]:

| | duration | protocol_type | service | flag | src_bytes | dst_bytes | land | wrong_fragment | urgent | hot | ... | dst_host_srv_cou |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 0 | tcp | ftp_data | SF | 491 | 0 | 0 | 0 | 0 | 0 | ... | 2 |
| **1** | 0 | udp | other | SF | 146 | 0 | 0 | 0 | 0 | 0 | ... | |
| **3** | 0 | tcp | http | SF | 232 | 8153 | 0 | 0 | 0 | 0 | ... | 25 |
| **4** | 0 | tcp | http | SF | 199 | 420 | 0 | 0 | 0 | 0 | ... | 25 |
| **12** | 0 | tcp | http | SF | 287 | 2251 | 0 | 0 | 0 | 0 | ... | 21 |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | |
| **125965** | 0 | tcp | smtp | SF | 2233 | 365 | 0 | 0 | 0 | 0 | ... | |
| **125967** | 0 | tcp | http | SF | 359 | 375 | 0 | 0 | 0 | 0 | ... | 25 |
| **125969** | 8 | udp | private | SF | 105 | 145 | 0 | 0 | 0 | 0 | ... | 24 |
| **125970** | 0 | tcp | smtp | SF | 2231 | 384 | 0 | 0 | 0 | 0 | ... | 3 |
| **125972** | 0 | tcp | ftp_data | SF | 151 | 0 | 0 | 0 | 0 | 0 | ... | 7 |

67343 rows × 42 columns

In [26]:

```
halfa
```

Out[26]:

| | duration | protocol_type | service | flag | src_bytes | dst_bytes | land | wrong_fragment | urgent | hot | ... | dst_host_srv_c |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **2** | 0 | tcp | private | S0 | 0 | 0 | 0 | 0 | 0 | 0 | ... | |
| **5** | 0 | tcp | private | REJ | 0 | 0 | 0 | 0 | 0 | 0 | ... | |
| **6** | 0 | tcp | private | S0 | 0 | 0 | 0 | 0 | 0 | 0 | ... | |
| **7** | 0 | tcp | private | S0 | 0 | 0 | 0 | 0 | 0 | 0 | ... | |
| **8** | 0 | tcp | remote_job | S0 | 0 | 0 | 0 | 0 | 0 | 0 | ... | |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | |
| **125958** | 0 | tcp | private | S0 | 0 | 0 | 0 | 0 | 0 | 0 | ... | |
| **125964** | 0 | tcp | private | S0 | 0 | 0 | 0 | 0 | 0 | 0 | ... | |
| **125966** | 0 | tcp | private | S0 | 0 | 0 | 0 | 0 | 0 | 0 | ... | |

| | duration | protocol_type | service | flag | src_bytes | dst_bytes | land | wrong_fragment | urgent | hot | ... | dst_host_srv_c |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **125968** | 0 | tcp | ftp_data | SF | 0 | 0 | 0 | 0 | 0 | 0 | ... | |
| **125971** | 0 | tcp | klogin | S0 | 0 | 0 | 0 | 0 | 0 | 0 | ... | |

58630 rows × 42 columns

In [10]:

```
#splitting the dataset into two halves, half1 and half2
total_count=0
for item in df['class']:
  total_count+=1
half1=df.head(int(total_count/2))
half2=df.tail(int(total_count/2))
```

In [12]:

```
# a represents number of 'normal' in half1 and b is number of 'anomaly' in half1
a,b=half1['class'].value_counts()
```

In [13]:

```
# a represents number of 'normal' in half2 and b is number of 'anomaly' in half2
c,d=half2['class'].value_counts()
```

In [32]:

```
#finding accuracy by comapring the class values in each half
acc = half1['class'] == half2['class'].reset_index(drop=True)
acc.mean()
```

Out[32]:

```
0.5024132346870733
```

In [15]:

```
#comparing the number of 'normal' and 'anomaly' in half1 and half2
cmp_n=100-int(c-a)/int(a+c)
print(accuracy_n)
cmp_a=100-int(b-d)/int(b+d)
print(accuracy_a)
```

```
99.99333264036352
99.99234167391563
```

## CICIDS DATASET

In [4]:

```
df1=pd.read_csv("/content/drive/MyDrive/CICIDS.csv")
df1
```

Out[4]:

| | Dst Port | Protocol | Timestamp | Flow Duration | Tot Fwd Pkts | Tot Bwd Pkts | TotLen Fwd Pkts | TotLen Bwd Pkts | Fwd Pkt Len Max | Fwd Pkt Len Min | ... | Fwd Seg Size Min | Active Mean | Active Std | Active Max |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 0 | 0 | 14/02/2018 08:31:01 | 112641719 | 3 | 0 | 0 | 0 | 0 | 0 | ... | 0 | 0.0 | 0.0 | 0 |
| **1** | 0 | 0 | 14/02/2018 08:33:50 | 112641466 | 3 | 0 | 0 | 0 | 0 | 0 | ... | 0 | 0.0 | 0.0 | 0 |
| **2** | 0 | 0 | 14/02/2018 08:36:39 | 112638623 | 3 | 0 | 0 | 0 | 0 | 0 | ... | 0 | 0.0 | 0.0 | 0 |
| **3** | 22 | 6 | 14/02/2018 08:40:13 | 6453966 | 15 | 10 | 1239 | 2273 | 744 | 0 | ... | 32 | 0.0 | 0.0 | 0 |

| | Dst Port | Protocol | Timestamp | Flow Duration | Tot Fwd Pkts | Tot Bwd Pkts | TotLen Fwd Pkts | TotLen Bwd Pkts | Fwd Pkt Len Max | Fwd Pkt Len Min | ... | Fwd Seg Size Min | Active Mean | Active Std | Active Max |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **4** | 22 | 6 | 14/02/2018 08:40:23 | 8804066 | 14 | 11 | 1143 | 2209 | 744 | 0 | ... | 32 | 0.0 | 0.0 | 0 |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| **1048570** | 80 | 6 | 14/02/2018 10:53:23 | 10156986 | 5 | 5 | 1089 | 1923 | 587 | 0 | ... | 20 | 0.0 | 0.0 | 0 |
| **1048571** | 80 | 6 | 14/02/2018 10:53:33 | 117 | 2 | 0 | 0 | 0 | 0 | 0 | ... | 20 | 0.0 | 0.0 | 0 |
| **1048572** | 80 | 6 | 14/02/2018 10:53:28 | 5095331 | 3 | 1 | 0 | 0 | 0 | 0 | ... | 20 | 0.0 | 0.0 | 0 |
| **1048573** | 80 | 6 | 14/02/2018 10:53:28 | 5235511 | 3 | 1 | 0 | 0 | 0 | 0 | ... | 20 | 0.0 | 0.0 | 0 |
| **1048574** | 443 | 6 | 14/02/2018 10:53:28 | 5807256 | 6 | 4 | 327 | 145 | 245 | 0 | ... | 20 | 291569.0 | 0.0 | 291569 |

1048575 rows × 80 columns

In [5]:

```
df1.shape
```

Out[5]:

```
(1048575, 80)
```

In [6]:

```
#finding count of the different Labels in the dataset
df1['Label'].value_counts()
```

Out[6]:

```
Benign             667626
FTP-BruteForce     193360
SSH-Bruteforce     187589
Name: Label, dtype: int64
```

In [10]:

```
#splitting the dataset into two halves, half3 and half4
total_count1=0
for item in df1['Label']:
  total_count1+=1
half3=df1.head(int(total_count1/2))
half4=df1.tail(int(total_count1/2))
```

In [12]:

```
#finding accuracy by comapring the class values in each half
acc1= half3['Label'] == half4['Label'].reset_index(drop=True)
acc1.mean()
```

Out[12]:

```
0.27339605979167897
```