INTRUSION DETECTION SYSTEM USING DEEP LEARNING

Project Review 2 for
CSE3501- Information Security Analysis and Audit (G2)

By:
YUVRAJ KUMAR: 18BIT0276
(CNN Implementation)

Complete Team Members:

AMAN AGARWAL: 18BIT0256
SHIVANSH SRIVASTAVA: 18BIT0324
YUVRAJ KUMAR: 18BIT0276

Submitted to

PROF. SUMAIYA THASEEN I.
School of Information Technology and Engineering,
Vellore Institute of Technology, Vellore-632014

FALL 2020-21

1. Design and Description of the System

Network Intrusion Detection Systems (NIDSs) are essential tools for the network system administrators to detect various security breaches inside an organization's network. An NIDS monitors and analyzes the network traffic entering into or exiting from the network devices of an organization and raises alarms if an intrusion is observed. Based on the methods of intrusion detection, NIDSs are categorized into two classes:

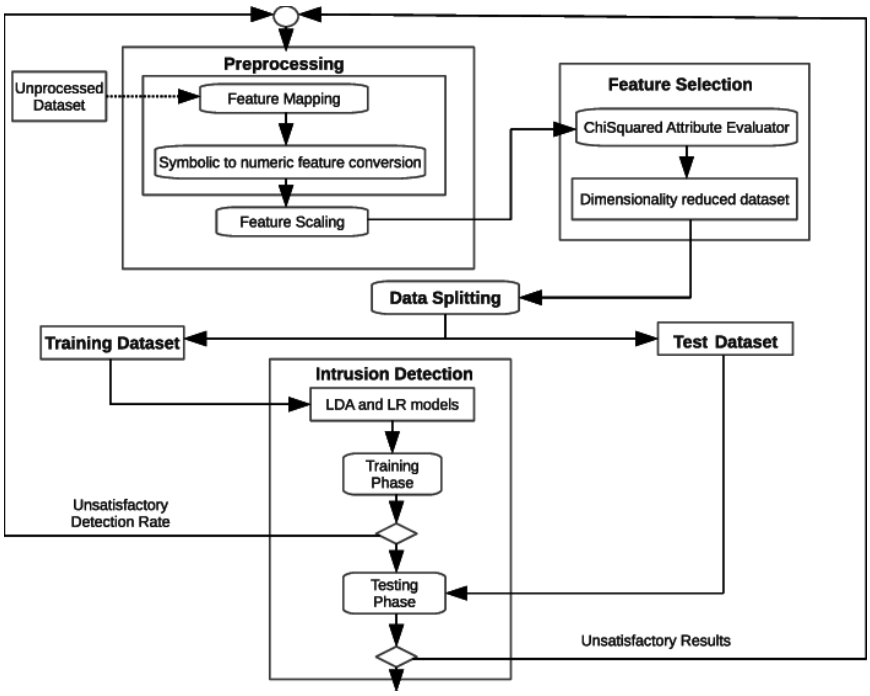
i) signature (misuse) based NIDS (SNIDS),
ii) anomaly detection-based NIDS (ADNIDS).

There are primarily two challenges that arise while developing an efficient and flexible NIDS for unknown future attacks. First, proper feature selections from the network traffic dataset for anomaly detection is difficult. The features selected for one class of attack may not work well for other categories of attacks due to continuously changing and evolving attack scenarios. Second, unavailability of labeled traffic dataset from real networks for developing an NIDS. Immense efforts are required to produce such a labeled dataset from the raw network traffic traces collected over a period or in real-time. Additionally, to preserve the confidentiality of the internal organizational network structure as well as the privacy of various users, network administrators are reluctant towards reporting any intrusion that might have occurred in their networks.

It is envisioned that the deep learning based approaches can help to overcome the challenges of developing an efficient NIDS . We can collect unlabeled network traffic data from different network sources and a good feature representation from these datasets using deep learning techniques can be obtained. These features can, then, be applied for supervised classification to a small, but labeled traffic dataset consisting of normal as well as anomalous traffic records. The traffic data for labeled dataset can be collected in a confined, isolated and private network environment. With this motivation, we use self-taught learning, a deep learning technique based on sparse autoencoder and soft-max regression, to develop an NIDS. We verify the usability of the self-taught learning based NIDS by applying on NSL-KDD intrusion dataset, an improved version of the benchmark dataset for various NIDS evaluations - KDD Cup 99. We provide a comparison of our current work with other techniques as well.

In the most widely used approach, the training data is used for both training and testing either using n-fold cross-validation or splitting the training data into training, cross-validation, and test sets. NIDSs based on this approach achieved very high accuracy and less false-alarm rates. The second approach uses the training and test data separately for the training and testing. Since the training and test data were collected in different environments, the accuracy obtained using the second approach is not as high as in the first approach. Therefore, we emphasize on the results of the second approach in our work for accurate evaluation of NIDS. However, we present the results of the first approach as well for completeness. We describe our NIDS implementation before discussing the results.

Architecture

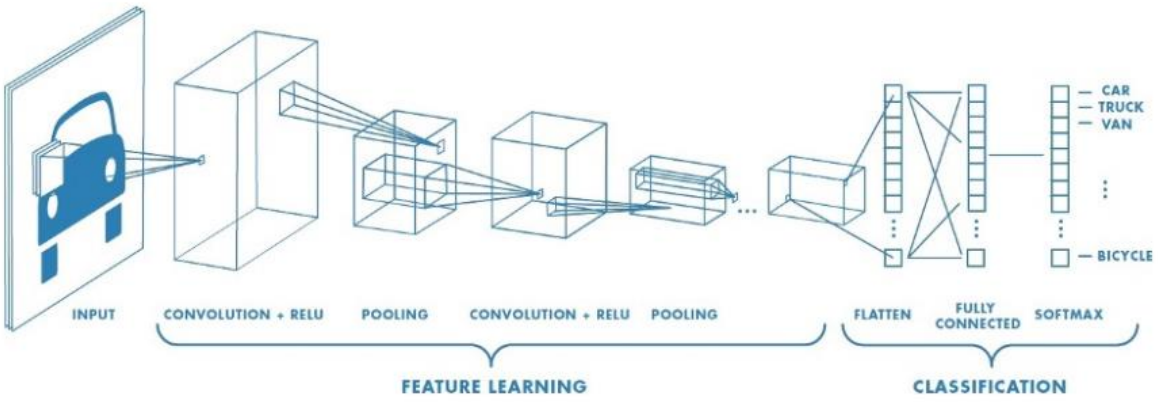


A Network Intrusion Detection System (NIDS) helps system administrators to detect network security breaches in their organizations. However, many challenges arise while developing a flexible and efficient NIDS for unforeseen and unpredictable attacks. We propose a deep learning based approach for developing such an efficient and flexible NIDS. We present the performance of our approach and compare it with a few previous work. Compared metrics include accuracy, precision, recall, and f-measure values.

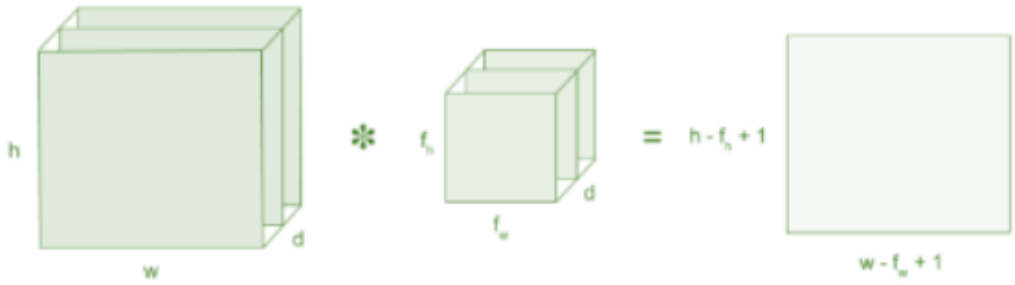
Convolutional Neural Network (CNN) — Deep Learning

In neural networks, Convolutional neural network (ConvNets or CNNs) is one of the main categories to do images recognition, images classifications. Objects detections, recognition faces etc., are some of the areas where CNNs are widely used.

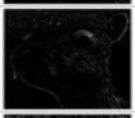
Technically, deep learning CNN models to train and test, each input image will pass it through a series of convolution layers with filters (Kernals), Pooling, fully connected layers (FC) and apply Softmax function to classify an object with probabilistic values between 0 and 1. The below figure is a complete flow of CNN to process an input image and classifies the objects based on values.



- An image matrix (volume) of dimension **(h x w x d)**
- A filter **(f_h x f_w x d)**
- Outputs a volume dimension **(h - f_h + 1) x (w - f_w + 1) x 1**



Convolution of an image with different filters can perform operations such as edge detection, blur and sharpen by applying filters. The below example shows various convolution image after applying different types of filters (Kernels).

Operation	Filter	Convolved Image
Identity	$\begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}$	
Edge detection	$\begin{bmatrix} 1 & 0 & -1 \\ 0 & 0 & 0 \\ -1 & 0 & 1 \end{bmatrix}$	
	$\begin{bmatrix} 0 & 1 & 0 \\ 1 & -4 & 1 \\ 0 & 1 & 0 \end{bmatrix}$	
	$\begin{bmatrix} -1 & -1 & -1 \\ -1 & 8 & -1 \\ -1 & -1 & -1 \end{bmatrix}$	
Sharpen	$\begin{bmatrix} 0 & -1 & 0 \\ -1 & 5 & -1 \\ 0 & -1 & 0 \end{bmatrix}$	
Box blur (normalized)	$\frac{1}{9}\begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}$	
Gaussian blur (approximation)	$\frac{1}{16}\begin{bmatrix} 1 & 2 & 1 \\ 2 & 4 & 2 \\ 1 & 2 & 1 \end{bmatrix}$	

2. PRE-PROCESSING (Normalization, Sampling – Oversampling, Undersampling (technique used for sampling - atleast 70%balanced)

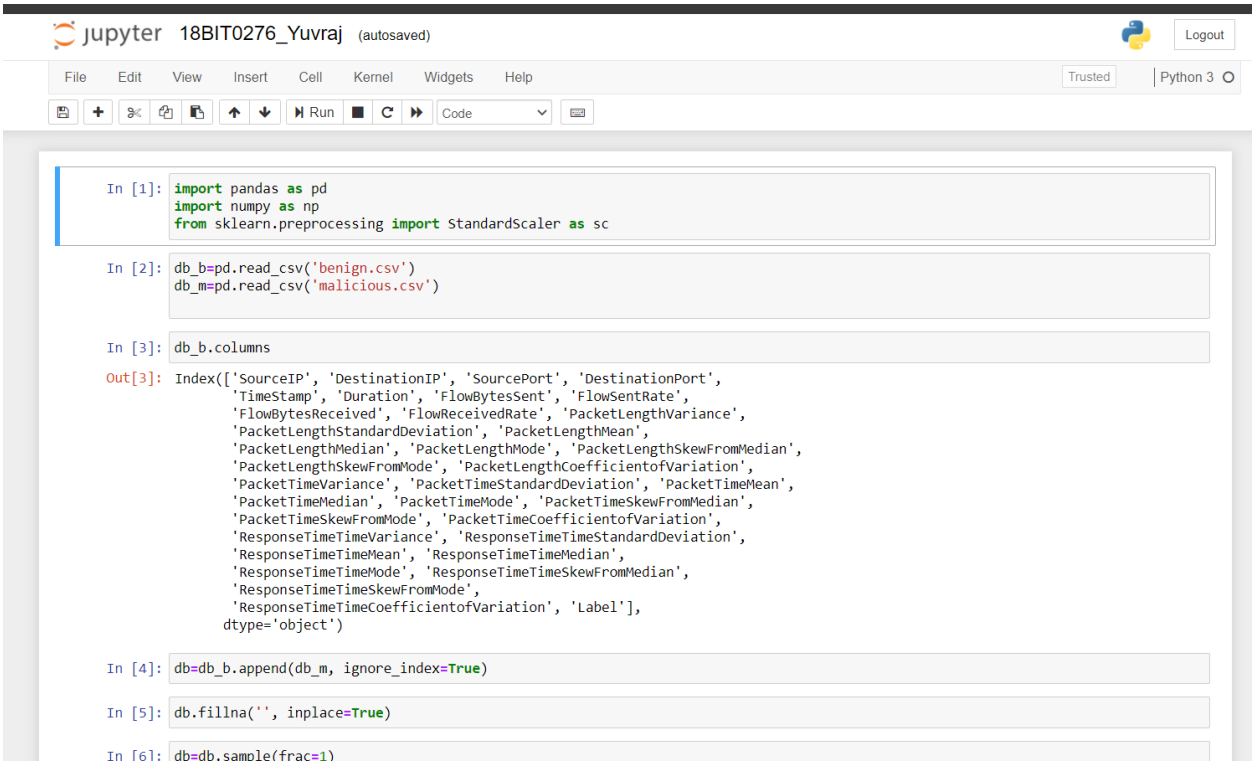
***Sampling is done before model creation. ->Pg.26*

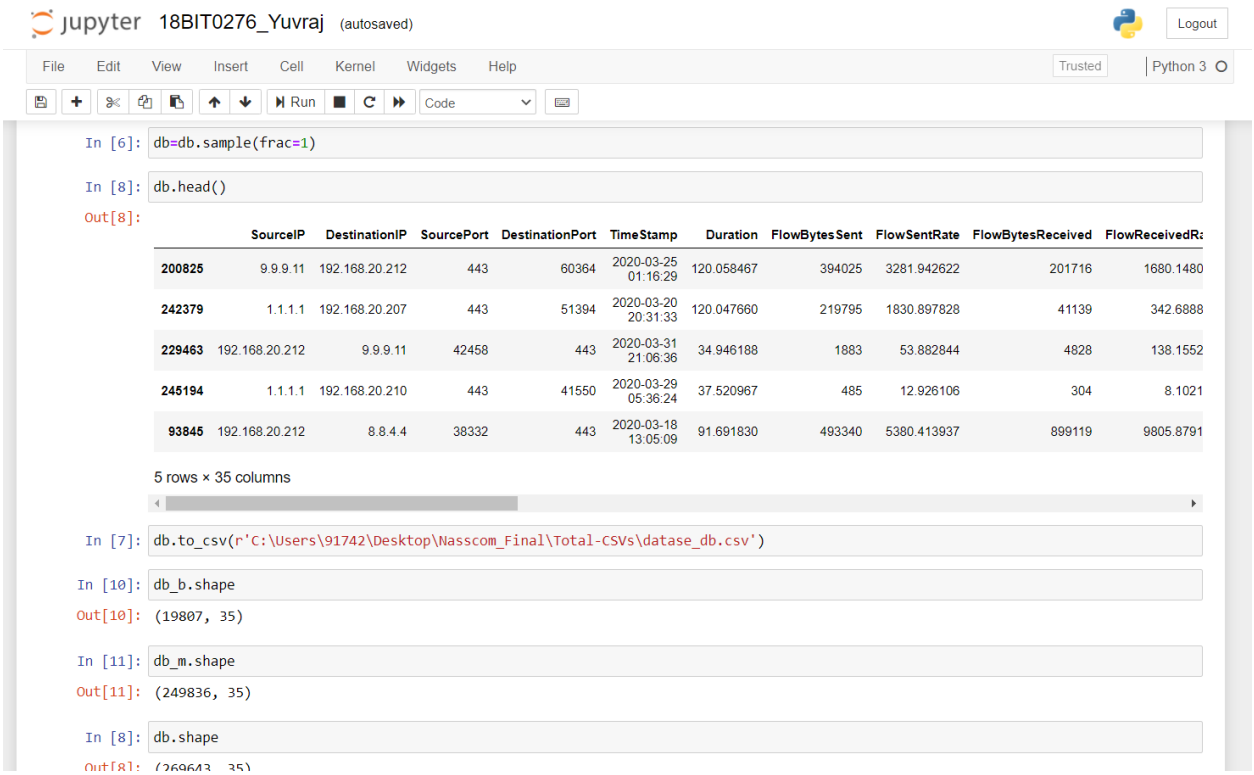
There were 4 datasets in a Total CSVs ZIP file provided to us by the Faculty.

Link to the Dataset :

https://www.unb.ca/cic/datasets/dohbrw-2020.html

One having **l2-benign** label and other containing **l2-malicious** were assigned to me. So, first step of processing was two combine two datasets into one. The combined dataset will have all Benign entries on top and all Malicious entries at the bottom or vice versa. This could create a problem will splitting into training and testing data. To avoid such problem I randomly shuffled the dataset so that both the labels get mixed.





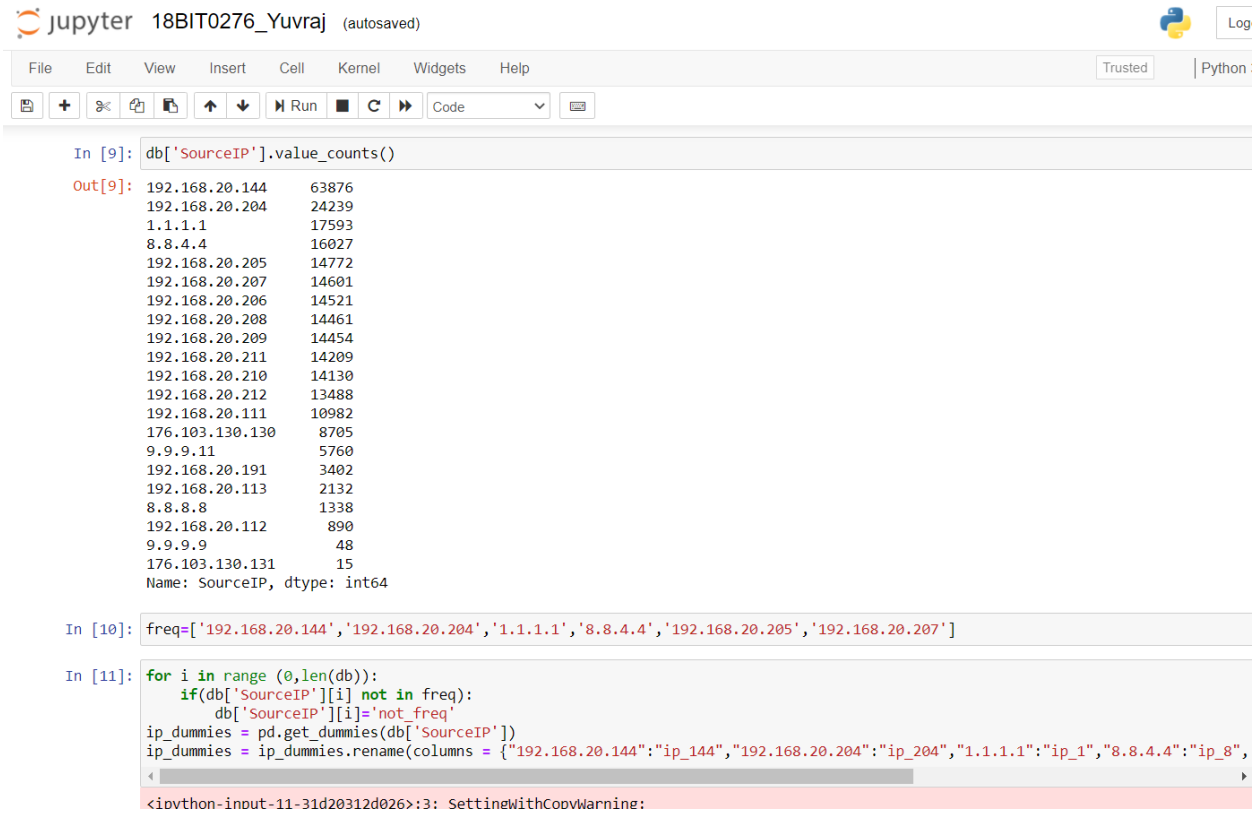
Now, the dataset basically had 2 kind of features, namely categorical and numeric. Categorical features comprised **SourceIP, DestinationIP, SourcePort, DestinationPort etc.** and numerical features included **Duration, FlowSentRate, FlowReceived, FlowBytesRate etc.** I have used different techniques to handle both kind of data.

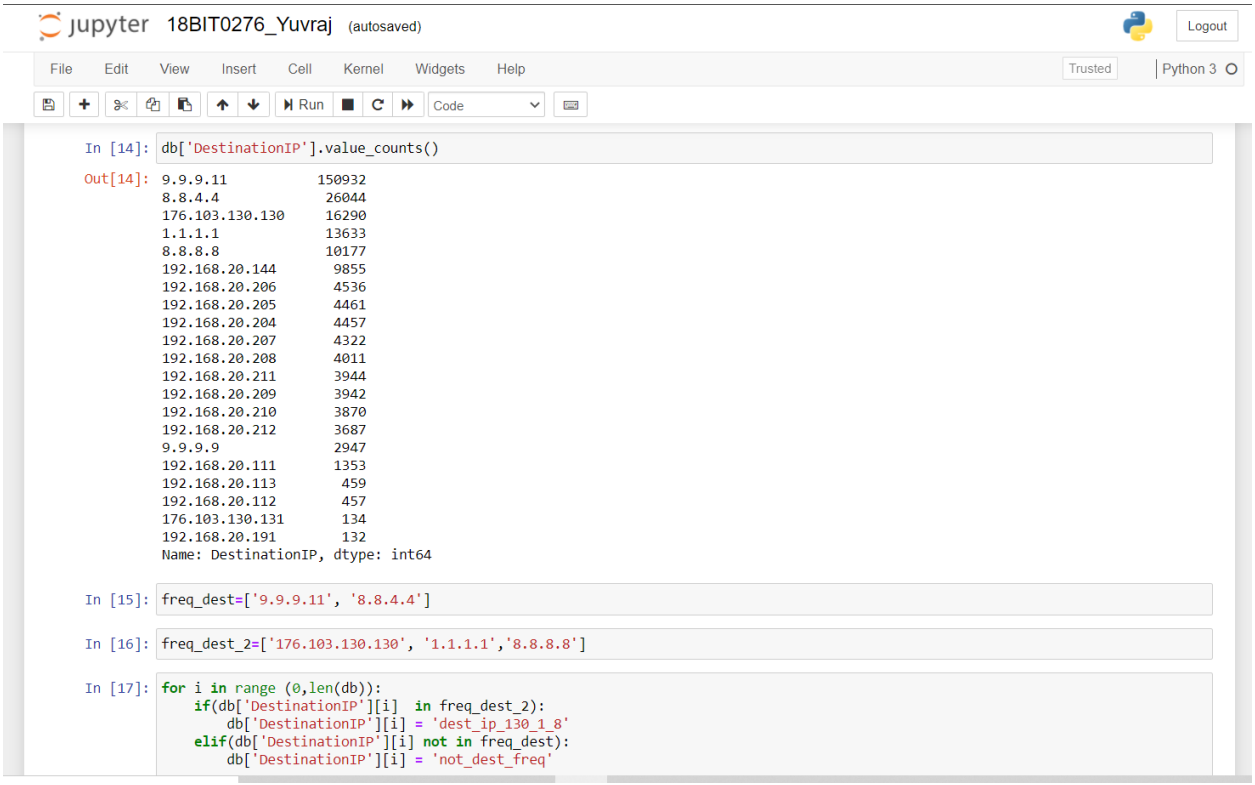
HANDLING CATEGORICAL DATA

Since total number of distinct values in each of the column are large, therefore encoding all them would have create a great sparsity in the dataset. So, in order to overcome the problem of sparsity, I analyzed each feature(column) of the dataset and choose a suitable threshold for the same. If the value is not in the threshold or given range, then I categorized it as other's. Different threshold criteria was chosen for each of the feature based on it's analysis. Then I one-hot encoded all of them so that I can feed them into my neural network. Below is the demonstration for the same.

This was done on the following column:

- **SOURCE IP COLUMN**
- **DESTINATION IP COLUMN**





SIMILARLY, AS ABOVE, THE SOURCE PORT AND DESTINATION PORT FEATURES HAVE BEEN PRE-PROCESSED AND CONCATENATING PRE-PROCESSED DATA INTO A NEW DATAFRAME

In [18]:
```
destination_dummy=pd.get_dummies(db['DestinationIP'])
```

In [19]:
```
destination_dummy = destination_dummy.rename(columns = {"9.9.9.11":"ip_11","8.8.4.4":"ip_4"})
```

In [20]:
```
db_create=pd.concat([db_create,destination_dummy], axis = 1)
```

In [21]:
```
db['SourcePort'].value_counts()
```

Out[21]:
```
443      49486
58759       73
41618       60
44886       60
40378       53
           ...
63314        1
64686        1
61443        1
60121        1
59999        1
Name: SourcePort, Length: 15446, dtype: int64
```

In [22]:
```
for i in range (0,len(db)):
    if(db['SourcePort'][i] != 443):
        db['SourcePort'][i] = 'Wrong_port'
    else:
        db['SourcePort'][i] = 'port_is_443'
```

```
<ipython-input-22-522c55c71581>:3: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  db['SourcePort'][i] = 'Wrong_port'
C:\Users\91742\anaconda3\lib\site-packages\pandas\core\indexing.py:671: SettingWithCopyWarning:
  db['SourcePort'][i] = 'port_is_443'
```

In [23]:
```
for i in range (0,len(db)):
    if(db['DestinationPort'][i] != 443):
        db['DestinationPort'][i] = 'Wrong_port'
    else:
        db['DestinationPort'][i] = 'port_443'
```

```
<ipython-input-23-ec0a6933b03d>:5: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  db['DestinationPort'][i] = 'port_443'
<ipython-input-23-ec0a6933b03d>:3: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  db['DestinationPort'][i] = 'Wrong_port'
```

In [24]:
```
source_port_dummy = pd.get_dummies(db['SourcePort'])
```

In [25]:
```
destination_port_dummy = pd.get_dummies(db['DestinationPort'])
```

In [26]:
```
destination_port_dummy = destination_port_dummy.rename(columns = {"port_443":"destination_port_443"})
```

In [27]:
```
db_create = pd.concat([db_create,source_port_dummy],axis=1)
db_create = pd.concat([db_create,destination_port_dummy],axis=1)
```

HANDLING NUMERIC FEATURES / NORMALIZATION

Some values were really large and some values were small. To avoid biasness and create a balanced dataset, I used Standard Scaler to normalize the numeric data.

I WILL BE TAKING CARE OF ALL THE NUMERIC FEATURE BELOW LIKE, FlowSentRate, ByteRate, Duration etc. using different Methodologies like Median, Standard Scaling, Sc_Fit_Transform etc.

- **DURATION COLUMN**
- **FLOW BYTES SENT**
- **FLOW SENT RATE**
- **FLOW BYTES RECEIVED**
- **PACKET LENGTH (*USING MEDIAN METHOD*)**
- **PACKET LENGTH CoV**
- **PACKET TIME VARIANCE**
- **PACKET LENGTH MEAN**
- **OTHER SPECIAL NUMERIC FEATURES**

In [28]:
```
from sklearn.preprocessing import StandardScaler
```

In [29]:
```
sc=StandardScaler()
```

In [30]:
```
drn_db = db[["Duration"]]
drn_db = sc.fit_transform(drn_db)
drn_db = pd.DataFrame(drn_db)
```

In [31]:
```
drn_db = drn_db.rename(columns = {0:"Duration"})
```

In [32]:
```
db_create = pd.concat([db_create,drn_db],axis=1)
```

In [33]:
```
db['FlowBytesSent'].value_counts()
```

Out[33]:
```
1807      44384
1875      15189
1806       7627
1085       7474
1739       7471
          ...
211692        1
199402        1
201449        1
217825        1
421554        1
Name: FlowBytesSent, Length: 72805, dtype: int64
```

In [34]:
```
fs = db[['FlowBytesSent']]
```

In [35]:
```
fs.mean()
```

Out[35]:
```
FlowBytesSent    65757.889977
```

jupyter 18BIT0276_Yuvraj (autosaved)

Logout

File Edit View Insert Cell Kernel Widgets Help

Trusted | Python 3

Run Code

0	1807

In [47]:
```
for i in range (0,len(db)):
    if(db['FlowBytesSent'][i] == 1807):
        db['FlowBytesSent'][i] = 'fs_1807'
    elif(db['FlowBytesSent'][i] >= 1810):
        df['FlowBytesSent'][i] = 'fsless'
    else:
        db['FlowBytesSent'][i] = 'fs_more'
```

```
---------------------------------------------------------------------------
NameError                                 Traceback (most recent call last)
<ipython-input-47-67f2d0160206> in <module>
      3         db['FlowBytesSent'][i] = 'fs_1807'
      4     elif(db['FlowBytesSent'][i] >= 1810):
----> 5         df['FlowBytesSent'][i] = 'fsless'
      6     else:
      7         db['FlowBytesSent'][i] = 'fs_more'

NameError: name 'df' is not defined
```

In [38]:
```
for i in range (0,len(db)):
    if(db['FlowBytesSent'][i] == 1807):
        db['FlowBytesSent'][i] = 'fs_1807'
    elif(db['FlowBytesSent'][i] >= 1810):
        db['FlowBytesSent'][i] = 'fsless'
    else:
        db['FlowBytesSent'][i] = 'fs_more'
```

```
<ipython-input-38-4a8b68076bc7>:5: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
```

jupyter 18BIT0276_Yuvraj (autosaved)

File Edit View Insert Cell Kernel Widgets Help Trusted

Run Code

```
rsus-a-copy
  db['FlowBytesSent'][i] = 'fs_1807'
```

In [39]:
```
fs_dummy = pd.get_dummies(db['FlowBytesSent'])
```

In [40]:
```
destination_port_dummy.head()
```

Out[40]:

	Wrong_port	destination_port_443
251170	0	1
250984	0	1
90316	0	1
6366	0	1
135318	0	1

In [41]:
```
db_create = pd.concat([db_create,fs_dummy],axis=1)
```

In [42]:
```
fs_r = db[["FlowSentRate"]]
```

In [43]:
```
fs_r = sc.fit_transform(fs_r)
```

In [44]:
```
fs_r = pd.DataFrame(fs_r)
```

In [45]:
```
fs_r = fs_r.rename(columns = {0:"FlowSentRate"})
```

In [46]:
```
db_create = pd.concat([db_create,fs_r],axis=1)
```

jupyter 18BIT0276_Yuvraj (autosaved)

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3

Run Code

```
See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  db["FlowBytesReceived"][115] = 0
```

In [48]:
```
frec = db[['FlowBytesReceived']]
```

In [49]:
```
frec.mean()
```

Out[49]:
```
FlowBytesReceived    67362.488946
dtype: float64
```

In [50]:
```
frec.median()
```

Out[50]:
```
FlowBytesReceived    4896.0
dtype: float64
```

In [51]:
```
for i in range (0,len(db)):
    if(isinstance(db['FlowBytesReceived'][i], str)):
        db['FlowBytesReceived'][i] = 0

    if(db['FlowBytesReceived'][i] == 98):
        db['FlowBytesReceived'][i] = 'fr_98'
    elif(db['FlowBytesReceived'][i] > 4900):
        db['FlowBytesReceived'][i] = 'fr_more'
    else:
        db['FlowBytesReceived'][i] = 'fr_less'
```

```
<ipython-input-51-b979bb7d5b2d>:8: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
```

jupyter 18BIT0276_Yuvraj (autosaved)

Logout

File Edit View Insert Cell Kernel Widgets Help

Trusted | Python 3

Run Code

In [52]:
```
frec_dummies = pd.get_dummies(db['FlowBytesReceived'])
```

In [64]:
```
db_new = pd.concat([db_new,frec_dummies],axis=1)
```

```
---------------------------------------------------------------------------
NameError                                 Traceback (most recent call last)
<ipython-input-64-34d07a978212> in <module>
----> 1 db_new = pd.concat([db_new,frec_dummies],axis=1)

NameError: name 'db_new' is not defined
```

In [53]:
```
bd_create = pd.concat([db_create,frec_dummies],axis=1)
```

In [54]:
```
frc = db[["FlowReceivedRate"]]
frc = sc.fit_transform(frc)
frc = pd.DataFrame(frc)
```

In [55]:
```
frc = frc.rename(columns = {0:"FlowRecvRate"})
```

In [56]:
```
db_create = pd.concat([db_create,frc],axis=1)
```

In [57]:
```
plv = db[["PacketLengthVariance"]]
plv = sc.fit_transform(plv)
plv = pd.DataFrame(plv)
```

In [58]:
```
plv = plv.rename(columns = {0:"Pcket_lenVar"})
```

In [59]:
```
db_create = pd.concat([db_create,plv],axis=1)
```

In [60]:
```
plsd = db[["PacketLengthStandardDeviation"]]
```

jupyter 18BIT0276_Yuvraj (autosaved)

Logout

File Edit View Insert Cell Kernel Widgets Help

Trusted | Python 3

Run Code

In [62]:
```
db_create = pd.concat([db_create,plsd],axis=1)
```

In [63]:
```
for i in range (0,len(db)):
    if(isinstance(db['PacketLengthMean'][i], str)):
        df['PacketLengthMean'][i] = 0
```

In [64]:
```
plm = db[['PacketLengthMean']]
plm = sc.fit_transform(plm)
plm = pd.DataFrame(plm)
```

In [65]:
```
plm = plm.rename(columns = {0:"Length_mean"})
```

In [66]:
```
db_create = pd.concat([db_create,plm],axis=1)
```

In [67]:
```
db_temp_half = db_create
```

In [68]:
```
for i in range (0,len(db)):
    if(isinstance(db['PacketLengthMedian'][i], str)):
        db['PacketLengthMedian'][i] = 0
```

In [81]:
```
pl_med = df[['PacketLengthMedian']]
```

```
---------------------------------------------------------------------------
NameError                                 Traceback (most recent call last)
<ipython-input-81-c6e68f44c154> in <module>
----> 1 pl_med = df[['PacketLengthMedian']]
```

jupyter 18BIT0276_Yuvraj (autosaved)

File Edit View Insert Cell Kernel Widgets Help Trusted

Run Code

In [69]:
```
pl_med = db[['PacketLengthMedian']]
```

In [70]:
```
pl_med = sc.fit_transform(pl_med)
pl_med = pd.DataFrame(pl_med)
```

In [71]:
```
pl_med = pl_med.rename(columns = {0:"L_median"})
```

In [72]:
```
db_create = pd.concat([db_create,pl_med],axis=1)
```

In [73]:
```
for i in range (0,len(db)):
    if(isinstance(db['PacketLengthMode'][i], str)):
        db['PacketLengthMode'][i] = 0
```

In [74]:
```
pl_mode = db[['PacketLengthMode']]
pl_mode = sc.fit_transform(pl_mode)
pl_mode = pd.DataFrame(pl_mode)
```

In [75]:
```
pl_mode = pl_mode.rename(columns = {0:"L_mode"})
```

In [76]:
```
db_create = pd.concat([db_create,pl_mode],axis=1)
```

In [77]:
```
for i in range (0,len(db)):
    if(isinstance(db['PacketLengthSkewFromMedian'][i], str)):
        db['PacketLengthSkewFromMedian'][i] = 0
```

In [78]:
```
s_med = db[['PacketLengthSkewFromMedian']]
s_med = sc.fit_transform(s_med)
s_med = pd.DataFrame(s_med)
```

jupyter 18BIT0276_Yuvraj (autosaved)

File Edit View Insert Cell Kernel Widgets Help Trusted

Run Code

In [79]:
```
s_med = s_med.rename(columns = {0:"s_median"})
```

In [80]:
```
db_create = pd.concat([db_create,s_med],axis=1)
```

In [81]:
```
for i in range (0,len(db)):
    if(isinstance(db['PacketLengthSkewFromMode'][i], str)):
        db['PacketLengthSkewFromMode'][i] = 0
```

In [82]:
```
s_mode = db[['PacketLengthSkewFromMode']]
s_mode = sc.fit_transform(s_mode)
s_mode = pd.DataFrame(s_mode)
```

In [83]:
```
s_mode = s_mode.rename(columns = {0:"s_mode"})
```

In [84]:
```
db_create = pd.concat([db_create,s_mode],axis=1)
```

In [85]:
```
for i in range (0,len(db)):
    if(isinstance(db['PacketLengthCoefficientofVariation'][i], str)):
        db['PacketLengthCoefficientofVariation'][i] = 0
```

In [86]:
```
plcv = db[['PacketLengthCoefficientofVariation']]
plcv = sc.fit_transform(plcv)
plcv = pd.DataFrame(plcv)
```

In [87]:
```
plcv = plcv.rename(columns = {0:"PLCV"})
```

In [88]:
```
db_create = pd.concat([db_create,plcv],axis=1)
```

jupyter 18BIT0276_Yuvraj (autosaved)

File Edit View Insert Cell Kernel Widgets Help Trusted

Run Code

```
        db['PacketTimeVariance'][i] = 0
```

In [90]:
```
ptv = db[['PacketTimeVariance']]
ptv = sc.fit_transform(ptv)
ptv = pd.DataFrame(ptv)
```

In [91]:
```
ptv = ptv.rename(columns = {0:"PTV"})
```

In [92]:
```
db_create = pd.concat([db_create,ptv],axis=1)
```

In [93]:
```
for i in range (0,len(db)):
    if(isinstance(db['PacketTimeStandardDeviation'][i], str)):
        db['PacketTimeStandardDeviatione'][i] = 0
```

In [94]:
```
pt_std = db[['PacketTimeStandardDeviation']]
pt_std = sc.fit_transform(pt_std)
pt_std = pd.DataFrame(pt_std)
```

In [95]:
```
pt_std = pt_std.rename(columns = {0:"PT_std"})
db_create = pd.concat([db_create,pt_std],axis=1)
```

In [96]:
```
for i in range (0,len(db)):
    if(isinstance(db['PacketTimeMean'][i], str)):
        db['PacketTimeMean'][i] = 0

pt_mean = db[['PacketTimeMean']]
pt_mean = sc.fit_transform(pt_mean)
pt_mean = pd.DataFrame(pt_mean)
```

In [97]:
```
pt_mean = pt_mean.rename(columns = {0:"PT_mean"})
db_create = pd.concat([db_create,pt_mean],axis=1)
```

jupyter 18BIT0276_Yuvraj (autosaved) Logout

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3

Run Code

In [98]:
```
for i in range (0,len(db)):
    if(isinstance(db['PacketTimeMedian'][i], str)):
        db['PacketTimeMedian'][i] = 0

pt_med = db[['PacketTimeMedian']]
pt_med = sc.fit_transform(pt_med)
pt_med = pd.DataFrame(pt_med)
```

In [99]:
```
pt_med = pt_med.rename(columns = {0:"PT_median"})
db_create = pd.concat([db_create,pt_med],axis=1)
```

In [100]:
```
for i in range (0,len(db)):
    if(isinstance(db['PacketTimeMode'][i], str)):
        db['PacketTimeMode'][i] = 0

packt_time_mode = db[['PacketTimeMode']]
packt_time_mode = sc.fit_transform(packt_time_mode)
packt_time_mode = pd.DataFrame(packt_time_mode)
```

In [101]:
```
packt_time_mode = packt_time_mode.rename(columns = {0:"Packt_Time_mode"})
db_create = pd.concat([db_create,packt_time_mode],axis=1)
```

In [102]:
```
for i in range (0,len(db)):
    if(isinstance(db['PacketTimeSkewFromMedian'][i], str)):
        db['PacketTimeSkewFromMedian'][i] = 0

s_med_time = db[['PacketTimeSkewFromMedian']]
s_med_time = sc.fit_transform(s_med_time)
s_med_time = pd.DataFrame(s_med_time)
```

In [103]:
```
s_med_time = s_med_time.rename(columns = {0:"s_med_time"})
dfb_create = pd.concat([db_create,s_med_time],axis=1)
```

jupyter 18BIT0276_Yuvraj (autosaved)

File Edit View Insert Cell Kernel Widgets Help

Trusted | Pytho

Run Code

In [104]:
```
for i in range (0,len(db)):
    if(isinstance(db['PacketTimeSkewFromMode'][i], str)):
        db['PacketTimeSkewFromMode'][i] = 0

s_mode_time = db[['PacketTimeSkewFromMode']]
s_mode_time = sc.fit_transform(s_mode_time)
s_mode_time = pd.DataFrame(s_mode_time)
```

In [105]:
```
s_mode_time = s_mode_time.rename(columns = {0:"s_mode_time"})
```

In [106]:
```
db_create = pd.concat([db_create,s_mode_time],axis=1)
```

In [107]:
```
for i in range (0,len(db)):
    if(isinstance(db['PacketTimeCoefficientofVariation'][i], str)):
        db['PacketTimeCoefficientofVariation'][i] = 0

pkt_cov_time = db[['PacketTimeCoefficientofVariation']]
pkt_cov_time = sc.fit_transform(pkt_cov_time)
pkt_cov_time = pd.DataFrame(pkt_cov_time)
```

In [108]:
```
pjt_cov_time = pkt_cov_time.rename(columns = {0:"pkt_cov_time"})
db_create = pd.concat([db_create,pkt_cov_time],axis=1)
```

In [109]:
```
for i in range (0,len(db)):
    if(isinstance(db['ResponseTimeTimeVariance'][i], str)):
        db['ResponseTimeTimeVariance'][i] = 0

r_time_var = db[['ResponseTimeTimeVariance']]
r_time_var = sc.fit_transform(r_time_var)
r_time_var = pd.DataFrame(r_time_var)
```

jupyter 18BIT0276_Yuvraj (autosaved)

File Edit View Insert Cell Kernel Widgets Help

Trusted | Python 3

Run Code

In [110]:
```
r_time_var = r_time_var.rename(columns = {0:"r_time_var"})
db_create = pd.concat([db_create,r_time_var],axis=1)
```

In [111]:
```
for i in range (0,len(db)):
    if(isinstance(db['ResponseTimeTimeStandardDeviation'][i], str)):
        db['ResponseTimeTimeStandardDeviation'][i] = 0

r_time_std = db[['ResponseTimeTimeStandardDeviation']]
r_time_std = sc.fit_transform(r_time_std)
r_time_std = pd.DataFrame(r_time_std)
```

In [112]:
```
r_time_std = r_time_std.rename(columns = {0:"r_time_std"})
```

In [113]:
```
db_create = pd.concat([db_create,r_time_std],axis=1)
```

In [114]:
```
for i in range (0,len(db)):
    if(isinstance(db['ResponseTimeTimeMean'][i], str)):
        db['ResponseTimeTimeMean'][i] = 0
```

In [115]:
```
r_time_mean = db[['ResponseTimeTimeMean']]
r_time_mean = sc.fit_transform(r_time_mean)
r_time_mean = pd.DataFrame(r_time_mean)
```

In [116]:
```
r_time_mean = r_time_mean.rename(columns = {0:"r_time_mean"})
```

In [117]:
```
db_create = pd.concat([db_create,r_time_mean],axis=1)
```

In [118]:
```
for i in range (0,len(db)):
    if(isinstance(db['ResponseTimeTimeMedian'][i], str)):
        db['ResponseTimeTimeMedian'][i] = 0

r_time_med = db[['ResponseTimeTimeMedian']]
```

```
r_time_med = sc.fit_transform(r_time_med)
r_time_med = pd.DataFrame(r_time_med)
```

```
<ipython-input-118-4b739f030769>:3: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-vi
rsus-a-copy
  db['ResponseTimeTimeMedian'][i] = 0
```

In [119]:
```
r_time_med = r_time_med.rename(columns = {0:"r_time_med"})
db_create = pd.concat([db_create,r_time_med],axis=1)
```

In [120]:
```
for i in range (0,len(db)):
    if(isinstance(db['ResponseTimeTimeMode'][i], str)):
        db['ResponseTimeTimeMode'][i] = 0

r_time_mode = db[['ResponseTimeTimeMode']]
r_time_mode = sc.fit_transform(r_time_mode)
r_time_mode = pd.DataFrame(r_time_mode)
```

In [121]:
```
r_time_mode = r_time_mode.rename(columns = {0:"r_time_mode"})
db_create = pd.concat([db_create,r_time_mode],axis=1)
```

In [122]:
```
for i in range (0,len(db)):
    if(isinstance(db['ResponseTimeTimeSkewFromMedian'][i], str)):
        db['ResponseTimeTimeSkewFromMedian'][i] = 0

r_s_med_time = db[['ResponseTimeTimeSkewFromMedian']]
r_s_med_time = sc.fit_transform(r_s_med_time)
r_s_med_time = pd.DataFrame(r_s_med_time)
```

```
<ipython-input-122-1abd97a78dfc>:3: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame
```

jupyter 18BIT0276_Yuvraj (autosaved)

File Edit View Insert Cell Kernel Widgets Help

Run Code

In [123]:
```
r_s_med_time = r_s_med_time.rename(columns = {0:"r_s_med_time"})
db_create = pd.concat([db_create,r_s_med_time],axis=1)
```

In [124]:
```
for i in range (0,len(db)):
    if(isinstance(db['ResponseTimeTimeSkewFromMode'][i], str)):
        db['ResponseTimeTimeSkewFromMode'][i] = 0
```

In [125]:
```
r_s_mode_time = db[['ResponseTimeTimeSkewFromMode']]
r_s_mode_time = sc.fit_transform(r_s_mode_time)
r_s_mode_time = pd.DataFrame(r_s_mode_time)
```

In [126]:
```
r_s_mode_time = r_s_mode_time.rename(columns = {0:"r_s_mode_time"})
db_create = pd.concat([db_create,r_s_mode_time],axis=1)
```

In [127]:
```
for i in range (0,len(db)):
    if(isinstance(db['ResponseTimeTimeCoefficientofVariation'][i], str)):
        db['ResponseTimeTimeCoefficientofVariation'][i] = 0
```

In [128]:
```
r_c_time = db[['ResponseTimeTimeCoefficientofVariation']]
r_c_time = sc.fit_transform(r_c_time)
r_c_time = pd.DataFrame(r_c_time)
```

In [129]:
```
r_c_time = r_c_time.rename(columns = {0:"r_c_time"})
db_create = pd.concat([db_create,r_c_time],axis=1)
```

In [142]:
```
labels = pd.get_dummies(db['Label'])
```

In [143]:
```
db_create = pd.concat([db_create,labels],axis=1)
```

jupyter 18BIT0276_Yuvraj (autosaved)

Logout

File Edit View Insert Cell Kernel Widgets Help

Trusted | Python 3

Run | Code

In [146]:
```
db_create.drop('Malicious' , axis=1)
```

Out[146]:

	ip_1	ip_144	ip_204	ip_205	ip_207	ip_8	not_freq	ip_4	ip_11	dest_ip_130_1_8	...	0	r_time_var	r_time_std	r_time_mean	r_time_med	r_
0	0	0	0	0	0	0	1	0	0	0	...	0.754077	-0.178941	-0.244670	-0.239715	-0.221598	
1	0	0	0	0	0	0	1	0	0	0	...	-0.814219	-0.132047	0.277059	-0.183655	-0.225091	
2	0	0	0	0	0	0	1	0	0	0	...	-0.817715	-0.178940	-0.244316	-0.238336	-0.221617	
3	0	0	0	0	0	0	1	0	0	0	...	-0.316040	-0.178946	-0.250146	-0.238243	-0.221523	
4	0	0	0	0	0	0	1	0	0	0	...	-0.770578	-0.178937	-0.242955	-0.240378	-0.221754	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
269638	1	0	0	0	0	0	0	0	0	0	...	-0.997728	-0.178940	-0.244337	-0.240169	-0.221636	
269639	1	0	0	0	0	0	0	0	0	0	...	0.947832	-0.178940	-0.244243	-0.240186	-0.221621	
269640	1	0	0	0	0	0	0	0	0	0	...	-0.858274	-0.178946	-0.250175	-0.243940	-0.226451	
269641	1	0	0	0	0	0	0	0	0	0	...	0.898332	-0.178940	-0.244267	-0.240184	-0.221634	
269642	1	0	0	0	0	0	0	0	0	0	...	-0.588891	-0.178785	-0.219341	-0.231216	-0.218044	

269643 rows × 45 columns

In [147]:
```
db_create['Benign'] = db_create['Benign'].replace(['1'],'Benign')
db_create['Benign'] = db_create['Benign'].replace (['0'],'Malicious')
```

In [148]:
```
db_create.head()
```

Out[148]:

	ip_1	ip_144	ip_204	ip_205	ip_207	ip_8	not_freq	ip_4	ip_11	dest_ip_130_1_8	...	r_time_var	r_time_std	r_time_mean	r_time_med	r_time_mode	r_s_
0	0	0	0	0	0	0	1	0	0	0	...	-0.178941	-0.244670	-0.239715	-0.221598	-0.159623	
1	0	0	0	0	0	0	1	0	0	0	...	-0.132047	0.277059	-0.183655	-0.225091	-0.159603	

jupyter 18BIT0276_Yuvraj (autosaved)

Logout

File Edit View Insert Cell Kernel Widgets Help

Trusted | Python 3

Run | Code

In [150]:
```
db_create.drop('Malicious' , inplace=True , axis=1)
```

In [151]:
```
db_create.head()
```

Out[151]:

	ip_1	ip_144	ip_204	ip_205	ip_207	ip_8	not_freq	ip_4	ip_11	dest_ip_130_1_8	...	0	r_time_var	r_time_std	r_time_mean	r_time_med	r_time_
0	0	0	0	0	0	0	1	0	0	0	...	0.754077	-0.178941	-0.244670	-0.239715	-0.221598	-0.1
1	0	0	0	0	0	0	1	0	0	0	...	-0.814219	-0.132047	0.277059	-0.183655	-0.225091	-0.1
2	0	0	0	0	0	0	1	0	0	0	...	-0.817715	-0.178940	-0.244316	-0.238336	-0.221617	-0.1
3	0	0	0	0	0	0	1	0	0	0	...	-0.316040	-0.178946	-0.250146	-0.238243	-0.221523	-0.1
4	0	0	0	0	0	0	1	0	0	0	...	-0.770578	-0.178937	-0.242955	-0.240378	-0.221754	-0.1

5 rows × 45 columns

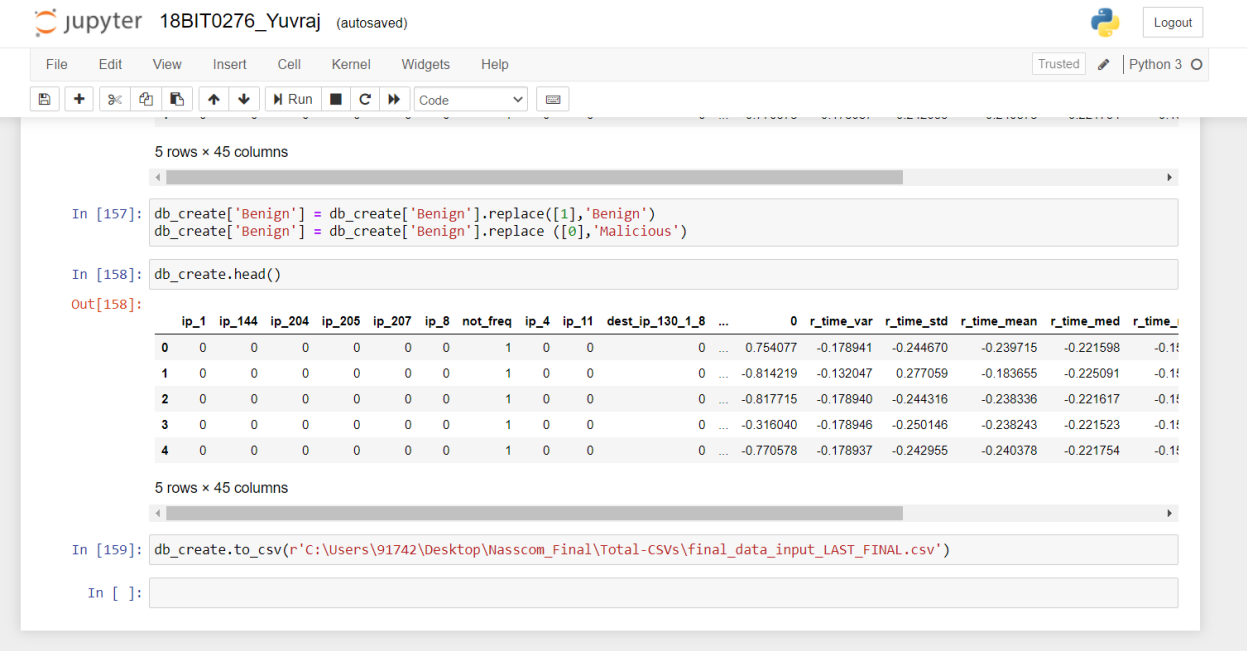
In [152]:
```
db_create.to_csv(r'C:\Users\91742\Desktop\Nasscom_Final\Total-CSVs\final_data_input_FINAL_Semifinal.csv')
```

In [153]:
```
db_create['Benign'] = db_create['Benign'].replace(['1'],'Benign')
db_create['Benign'] = db_create['Benign'].replace (['0'],'Malicious')
```

In [154]:
```
db_create.head()
```

Out[154]:

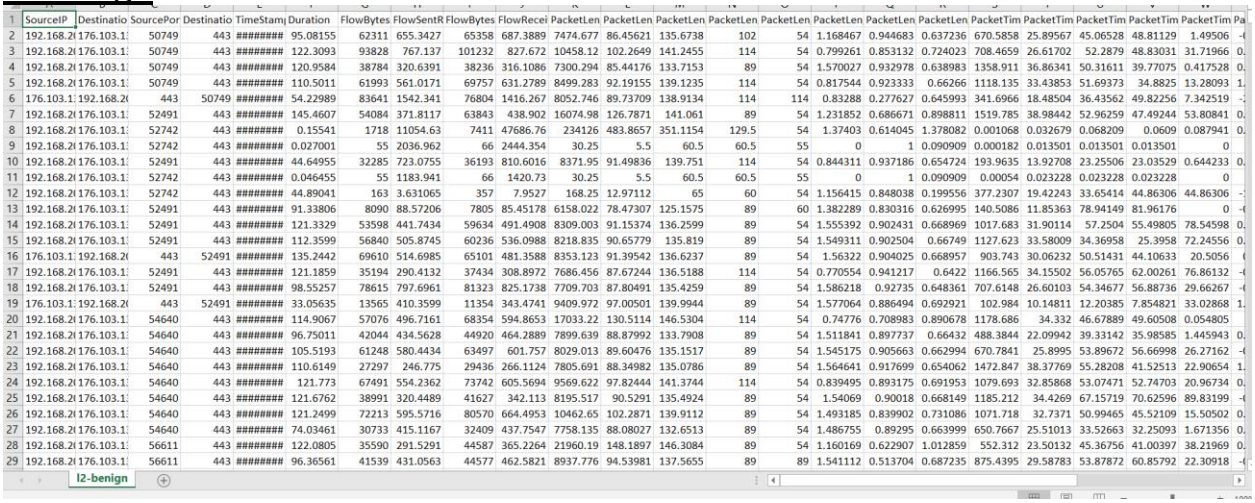
	ip_1	ip_144	ip_204	ip_205	ip_207	ip_8	not_freq	ip_4	ip_11	dest_ip_130_1_8	...	0	r_time_var	r_time_std	r_time_mean	r_time_med	r_time_
0	0	0	0	0	0	0	1	0	0	0	...	0.754077	-0.178941	-0.244670	-0.239715	-0.221598	-0.1
1	0	0	0	0	0	0	1	0	0	0	...	-0.814219	-0.132047	0.277059	-0.183655	-0.225091	-0.1
2	0	0	0	0	0	0	1	0	0	0	...	-0.817715	-0.178940	-0.244316	-0.238336	-0.221617	-0.1
3	0	0	0	0	0	0	1	0	0	0	...	-0.316040	-0.178946	-0.250146	-0.238243	-0.221523	-0.1



As I had 269643 records in DoH and only 19807 records in benign, I tried Under Sampling and went Ahead with 19807 Records and upon Random Sampling, Each Dataset Picked 19807 records to form 39614 Record Data Together.

DATASETS:

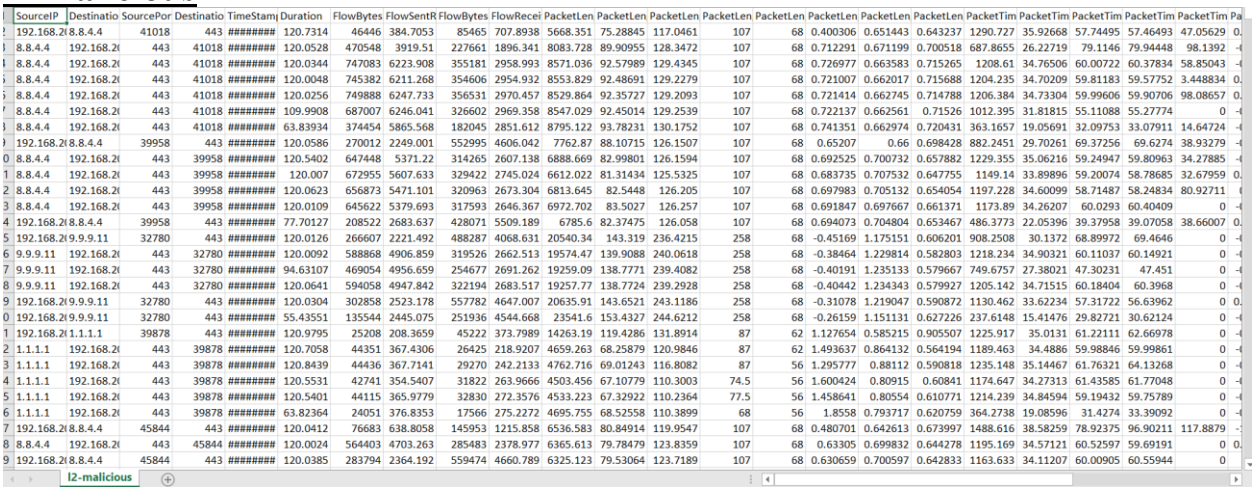
l2-benign



SourceIP	Destinatio	SourcePor	Destinatio	TimeStam	Duration	FlowBytes	FlowSentR	FlowBytes	FlowRecei	PacketLen	PacketLen	PacketLen	PacketLen	PacketLen	PacketLen	PacketLen	PacketLen	PacketTim	PacketTim	PacketTim	PacketTim	PacketTim	Pa
192.168.2(	176.103.1	50749	443	########	95.08155	62311	655.3427	65358	687.3889	7474.677	86.45621	135.6738	102	54	1.168467	0.944683	0.637236	670.5858	25.89567	45.06528	48.81129	1.49506	-(
192.168.2(	176.103.1	50749	443	########	122.3093	93828	767.137	101232	827.672	10458.12	102.2649	141.2455	114	54	0.799261	0.853132	0.724023	708.4659	26.61702	52.2879	48.83031	31.71966	0.
192.168.2(	176.103.1	50749	443	########	120.9584	38784	320.6391	38236	316.1086	7300.294	85.44176	133.7153	89	54	1.570027	0.932978	0.638983	1358.911	36.86341	50.31611	39.77075	0.417528	0.
192.168.2(	176.103.1	50749	443	########	110.5011	61993	561.0171	69757	631.2789	8499.283	92.19155	139.1235	114	54	0.817544	0.923333	0.66266	1118.135	33.43853	51.69373	34.8825	13.28093	1.
176.103.1	192.168.2(	443	50749	########	54.22989	83641	1542.341	76804	1416.267	8052.746	89.73709	138.9134	114	114	0.83288	0.277627	0.645993	341.6966	18.48504	36.43562	49.82256	7.342519	-.
192.168.2(	176.103.1	52491	443	########	145.4607	54084	371.8117	63843	438.902	16074.98	126.7871	141.061	89	54	1.231852	0.686671	0.898811	1519.785	38.98442	52.96259	47.49244	53.80841	0.
192.168.2(	176.103.1	52742	443	########	0.15541	1718	11054.63	7411	47686.76	234126	483.8657	351.1154	129.5	54	1.37403	0.614045	1.378082	0.001068	0.032679	0.068209	0.0609	0.087941	0.
192.168.2(	176.103.1	52742	443	########	0.027001	55	2036.962	66	2444.354	30.25	5.5	60.5	60.5	55	0	1	0.090909	0.000182	0.013501	0.013501	0.013501	0	
192.168.2(	176.103.1	52491	443	########	44.64955	32285	723.0755	36193	810.6016	8371.95	91.49836	139.751	114	54	0.844311	0.937186	0.654724	193.9635	13.92708	23.25506	23.03529	0.644233	0.
192.168.2(	176.103.1	52742	443	########	0.046455	55	1183.941	66	1420.73	30.25	5.5	60.5	60.5	55	0	1	0.090909	0.00054	0.023228	0.023228	0.023228	0	
192.168.2(	176.103.1	52742	443	########	44.89041	163	3.631065	357	7.9527	168.25	12.97112	65	60	54	1.156415	0.848038	0.199556	377.2307	19.42243	33.65414	44.86306	44.86306	-:
192.168.2(	176.103.1	52491	443	########	91.33806	8090	88.57206	7805	85.45178	6158.022	78.47307	125.1575	89	60	1.382289	0.830316	0.626995	140.5086	11.85363	78.94149	81.96176	0	-(
192.168.2(	176.103.1	52491	443	########	121.3329	53598	441.7434	59634	491.4908	8309.003	91.15374	136.2599	89	54	1.555392	0.902431	0.668969	1017.683	31.90114	57.2504	55.49805	78.54598	0.
192.168.2(	176.103.1	52491	443	########	112.3599	56840	505.8745	60236	536.0988	8218.835	90.65779	135.819	89	54	1.549311	0.902504	0.66749	1127.623	33.58009	34.36958	25.3958	72.24556	0.
176.103.1	192.168.2(	443	52491	########	135.2442	69610	514.6985	65101	481.3588	8353.123	91.39542	136.6237	89	54	1.56322	0.904025	0.668957	903.743	30.06232	50.51431	44.10633	20.5056	(
192.168.2(	176.103.1	52491	443	########	121.1859	35194	290.4132	37434	308.8972	7686.456	87.67244	136.5188	114	54	0.770554	0.941217	0.6422	1166.565	34.15502	56.05765	62.00261	76.86132	-(
192.168.2(	176.103.1	52491	443	########	98.55257	78615	797.6961	81323	825.1738	7709.703	87.80491	135.4259	89	54	1.586218	0.92735	0.648361	707.6148	26.60103	54.34677	56.88736	29.66267	-(
176.103.1	192.168.2(	443	52491	########	33.05635	13565	410.3599	11354	343.4741	9409.972	97.00501	139.9944	89	54	1.577064	0.886494	0.692921	102.984	10.14811	12.20385	7.854821	33.02868	1.
192.168.2(	176.103.1	54640	443	########	114.9067	57076	496.7161	68354	594.8653	17033.22	130.5114	146.5304	114	54	0.74776	0.708983	0.890678	1178.686	34.332	46.67889	49.60508	0.054805	
192.168.2(	176.103.1	54640	443	########	96.75011	42044	434.5628	44920	464.2889	7899.639	88.87992	133.7908	89	54	1.511841	0.897737	0.66432	488.3844	22.09942	39.33142	35.98585	1.445943	0.
192.168.2(	176.103.1	54640	443	########	105.5193	61248	580.4434	63497	601.757	8029.013	89.60476	135.1517	89	54	1.545175	0.905663	0.662994	670.7841	25.8995	53.89672	56.66998	26.27162	-(
192.168.2(	176.103.1	54640	443	########	110.6149	27297	246.775	29436	266.1124	7805.691	88.34982	135.0786	89	54	1.564641	0.917699	0.654062	1472.847	38.37769	55.28208	41.52513	22.90654	1.
192.168.2(	176.103.1	54640	443	########	121.773	67491	554.2362	73742	605.5694	9569.622	97.82444	141.3744	114	54	0.839495	0.893175	0.691953	1079.693	32.85868	53.07471	52.74703	20.96734	0.
192.168.2(	176.103.1	54640	443	########	121.6762	38991	320.4489	41627	342.113	8195.517	90.5291	135.4924	89	54	1.54069	0.90018	0.668149	1185.212	34.4269	67.15719	70.62596	89.83199	-(
192.168.2(	176.103.1	54640	443	########	121.2499	72213	595.5716	80570	664.4953	10462.65	102.2871	139.9112	89	54	1.493185	0.839902	0.731086	1071.718	32.7371	50.99465	45.52109	15.50502	0.
192.168.2(	176.103.1	54640	443	########	74.03461	30733	415.1167	32409	437.7547	7758.135	88.08027	132.6513	89	54	1.486755	0.89295	0.663999	650.7667	25.51013	33.52663	32.25093	1.671356	0.
192.168.2(	176.103.1	56611	443	########	122.0805	35590	291.5291	44587	365.2264	21960.19	148.1897	146.3084	89	54	1.160169	0.622907	1.012859	552.312	23.50132	45.36756	41.00397	38.21969	0.
192.168.2(	176.103.1	56611	443	########	96.36561	41539	431.0563	44577	462.5821	8937.776	94.53981	137.5655	89	89	1.541112	0.513704	0.687235	875.4395	29.58783	53.87872	60.85792	22.30918	-(

l2-benign

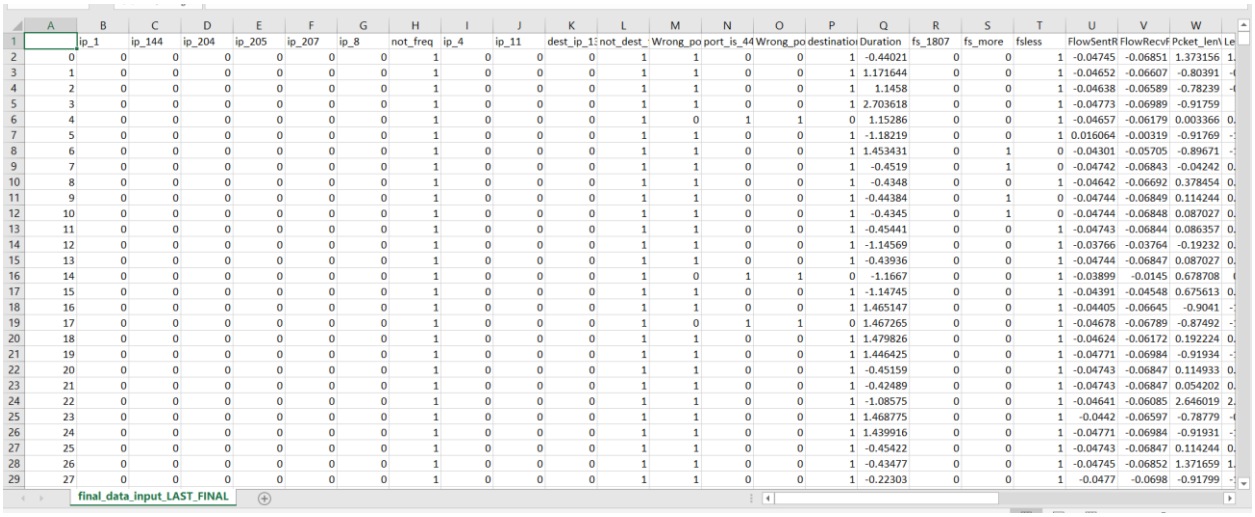
l2-malicious



SourceIP	Destinatio	SourcePor	Destinatio	TimeStam	Duration	FlowBytes	FlowSentR	FlowBytes	FlowRecei	PacketLen	PacketLen	PacketLen	PacketLen	PacketLen	PacketLen	PacketLen	PacketLen	PacketTim	PacketTim	PacketTim	PacketTim	PacketTim	Pa
192.168.2(	8.8.4.4	41018	443	########	120.7314	46446	384.7053	85465	707.8938	5668.351	75.28845	117.0461	107	68	0.400306	0.651443	0.643237	1290.727	35.92668	57.74495	57.46493	47.05629	0.
8.8.4.4	192.168.2(	443	41018	########	120.0528	470548	3919.51	227661	1896.341	8083.728	89.90955	128.3472	107	68	0.712291	0.671199	0.700518	687.8655	26.22719	79.1146	79.94448	98.1392	-(
8.8.4.4	192.168.2(	443	41018	########	120.0344	747083	6223.908	355181	2958.993	8571.036	92.57989	129.4345	107	68	0.726977	0.663583	0.715265	1208.61	34.76506	60.00722	60.37834	58.85043	-(
8.8.4.4	192.168.2(	443	41018	########	120.0048	745382	6211.268	354606	2954.932	8553.829	92.48691	129.2279	107	68	0.721007	0.662017	0.715688	1204.235	34.70209	59.81183	59.57752	3.448834	0.
8.8.4.4	192.168.2(	443	41018	########	120.0256	749888	6247.733	356531	2970.457	8529.864	92.35727	129.2093	107	68	0.721414	0.662745	0.714788	1206.384	34.73304	59.99606	59.90706	98.08657	0.
8.8.4.4	192.168.2(	443	41018	########	109.9908	687007	6246.041	326602	2969.358	8547.029	92.45014	129.2539	107	68	0.722137	0.662561	0.71526	1012.395	31.81815	55.11088	55.27774	0	-(
8.8.4.4	192.168.2(	443	41018	########	63.83934	374454	5865.568	182045	2851.612	8795.122	93.78231	130.1752	107	68	0.741351	0.662974	0.720431	363.1657	19.05691	32.09753	33.07911	14.64724	-(
192.168.2(	8.8.4.4	39958	443	########	120.0586	270012	2249.001	552995	4606.042	7762.87	88.10715	126.1507	107	68	0.65207	0.66	0.698428	882.2451	29.70261	69.37256	69.6274	38.93279	-(
8.8.4.4	192.168.2(	443	39958	########	120.5402	647448	5371.22	314265	2607.138	6888.669	82.99801	126.1594	107	68	0.692525	0.700732	0.657882	1229.355	35.06216	59.24947	59.80963	34.27885	-(
8.8.4.4	192.168.2(	443	39958	########	120.007	672955	5607.633	329422	2745.024	6612.022	81.31434	125.5325	107	68	0.683735	0.707532	0.647755	1149.14	33.89896	59.20074	58.78685	32.67959	0.
8.8.4.4	192.168.2(	443	39958	########	120.0623	656873	5471.101	320963	2673.304	6813.645	82.5448	126.205	107	68	0.697983	0.705132	0.654054	1197.228	34.60099	58.71487	58.24834	80.92711	(
8.8.4.4	192.168.2(	443	39958	########	120.0109	645622	5379.693	317593	2646.367	6972.702	83.5027	126.257	107	68	0.691847	0.697667	0.661371	1173.89	34.26207	60.0293	60.40409	0	-(
192.168.2(	8.8.4.4	39958	443	########	77.70127	208522	2683.637	428071	5509.189	6785.6	82.37475	126.058	107	68	0.694073	0.704804	0.653467	486.3773	22.05396	39.37958	39.07058	38.66007	0.
192.168.2(	9.9.9.11	32780	443	########	120.0126	266607	2221.492	488287	4068.631	20540.34	143.319	236.4215	258	68	-0.45169	1.175151	0.606201	908.2508	30.1372	68.89972	69.4646	0	-(
9.9.9.11	192.168.2(	443	32780	########	120.0092	588868	4906.859	319526	2662.513	19574.47	139.9088	240.0618	258	68	-0.38464	1.229814	0.582803	1218.234	34.90321	60.11037	60.14921	0	-(
9.9.9.11	192.168.2(	443	32780	########	94.63107	469054	4956.659	254677	2691.262	19259.09	138.7771	239.4082	258	68	-0.40191	1.235133	0.579667	749.6757	27.38021	47.30231	47.451	0	-(
9.9.9.11	192.168.2(	443	32780	########	120.0641	594058	4947.842	322194	2683.517	19257.77	138.7724	239.2928	258	68	-0.40442	1.234343	0.579927	1205.142	34.71515	60.18404	60.3968	0	-(
192.168.2(	9.9.9.11	32780	443	########	120.0304	302858	2523.178	557782	4647.007	20635.91	143.6521	243.1186	258	68	-0.31078	1.219047	0.590872	1130.462	33.62234	57.31722	56.63962	0	0.
192.168.2(	9.9.9.11	32780	443	########	55.43551	135544	2445.075	251936	4544.668	23541.6	153.4327	244.6212	258	68	-0.26159	1.151131	0.627226	237.6148	15.41476	29.82721	30.62124	0	-(
192.168.2(	1.1.1.1	39878	443	########	120.9795	25208	208.3659	45222	373.7989	14263.19	119.4286	131.8914	87	62	1.127654	0.585215	0.905507	1225.917	35.0131	61.22111	62.66978	0	-(
1.1.1.1	192.168.2(	443	39878	########	120.7058	44351	367.4306	26425	218.9207	4659.263	68.25879	120.9846	87	62	1.493637	0.864132	0.564194	1189.463	34.4886	59.98846	59.99861	0	-(
1.1.1.1	192.168.2(	443	39878	########	120.8439	44436	367.7141	29270	242.2133	4762.716	69.01243	116.8082	87	56	1.295777	0.88112	0.590818	1235.148	35.14467	61.76321	64.13268	0	-(
1.1.1.1	192.168.2(	443	39878	########	120.5531	42741	354.5407	31822	263.9666	4503.456	67.10779	110.3003	74.5	56	1.600424	0.80915	0.60841	1174.647	34.27313	61.43585	61.77048	0	-(
1.1.1.1	192.168.2(	443	39878	########	120.5401	44115	365.9779	32830	272.3576	4533.223	67.32922	110.2364	77.5	56	1.458641	0.80554	0.610771	1214.239	34.84594	59.19432	59.75789	0	-(
1.1.1.1	192.168.2(	443	39878	########	63.82364	24051	376.8353	17566	275.2272	4695.755	68.52558	110.3899	68	56	1.8558	0.793717	0.620759	364.2738	19.08596	31.4274	33.39092	0	-(
192.168.2(	8.8.4.4	45844	443	########	120.0412	76683	638.8058	145953	1215.858	6536.583	80.84914	119.9547	107	68	0.480701	0.642613	0.673997	1488.616	38.58259	78.92375	96.90211	117.8879	-:
8.8.4.4	192.168.2(	443	45844	########	120.0024	564403	4703.263	285483	2378.977	6365.613	79.78479	123.8359	107	68	0.63305	0.699832	0.644278	1195.169	34.57121	60.52597	59.69191	0	0.
192.168.2(	8.8.4.4	45844	443	########	120.0385	283794	2364.192	559474	4660.789	6325.123	79.53064	123.7189	107	68	0.630659	0.700597	0.642833	1163.633	34.11207	60.00905	60.55944	0	

l2-malicious

Combined Final dataset

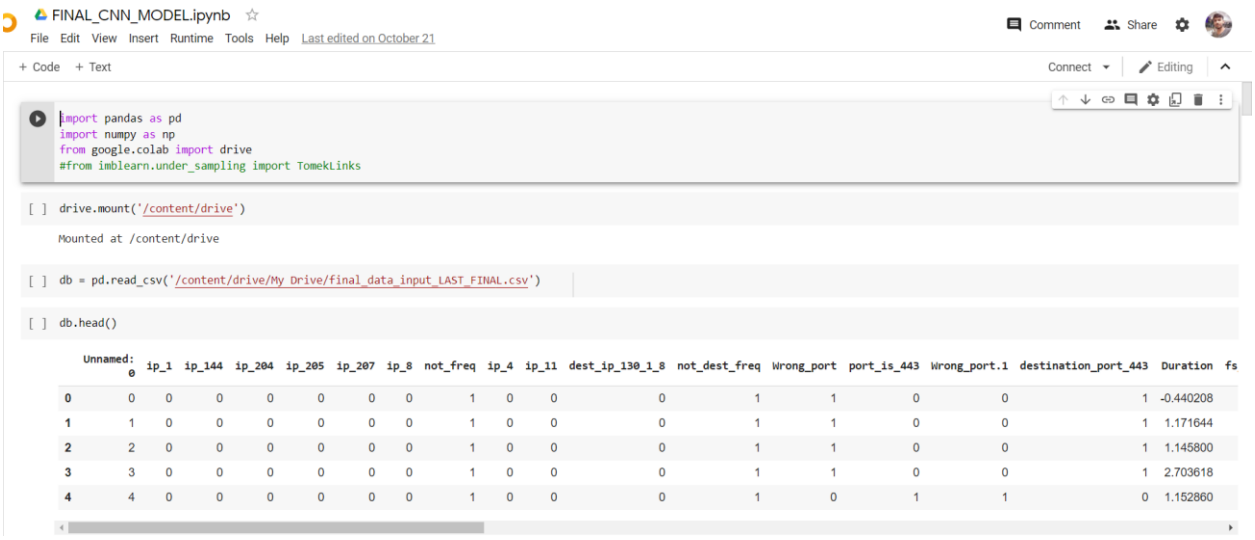


	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W
1		ip_1	ip_144	ip_204	ip_205	ip_207	ip_8	not_freq	ip_4	ip_11	dest_ip_1	not_dest_	Wrong_po	port_is_44	Wrong_po	destinatio	Duration	fs_1807	fs_more	fsless	FlowSentR	FlowRecvF	Pcket_len\
2	0	0	0	0	0	0	0	1	0	0	0	1	1	0	0	1	-0.44021	0	0	1	-0.04745	-0.06851	1.373156
3	1	0	0	0	0	0	0	1	0	0	0	1	1	0	0	1	1.171644	0	0	1	-0.04652	-0.06607	-0.80391
4	2	0	0	0	0	0	0	1	0	0	0	1	1	0	0	1	1.1458	0	0	1	-0.04638	-0.06589	-0.78239
5	3	0	0	0	0	0	0	1	0	0	0	1	1	0	0	1	2.703618	0	0	1	-0.04773	-0.06989	-0.91759
6	4	0	0	0	0	0	0	1	0	0	0	1	0	1	1	0	1.15286	0	0	1	-0.04657	-0.06179	0.003366
7	5	0	0	0	0	0	0	1	0	0	0	1	1	0	0	1	-1.18219	0	0	1	0.016064	-0.00319	-0.91769
8	6	0	0	0	0	0	0	1	0	0	0	1	1	0	0	1	1.453431	0	1	0	-0.04301	-0.05705	-0.89671
9	7	0	0	0	0	0	0	1	0	0	0	1	1	0	0	1	-0.4519	0	1	0	-0.04742	-0.06843	-0.04242
10	8	0	0	0	0	0	0	1	0	0	0	1	1	0	0	1	-0.4348	0	0	1	-0.04642	-0.06692	0.378454
11	9	0	0	0	0	0	0	1	0	0	0	1	1	0	0	1	-0.44384	0	1	0	-0.04744	-0.06849	0.114244
12	10	0	0	0	0	0	0	1	0	0	0	1	1	0	0	1	-0.4345	0	1	0	-0.04744	-0.06848	0.087027
13	11	0	0	0	0	0	0	1	0	0	0	1	1	0	0	1	-0.45441	0	0	1	-0.04743	-0.06844	0.086357
14	12	0	0	0	0	0	0	1	0	0	0	1	1	0	0	1	-1.14569	0	0	1	-0.03766	-0.03764	-0.19232
15	13	0	0	0	0	0	0	1	0	0	0	1	1	0	0	1	-0.43936	0	0	1	-0.04744	-0.06847	0.087027
16	14	0	0	0	0	0	0	1	0	0	0	1	0	1	1	0	-1.1667	0	0	1	-0.03899	-0.0145	0.678708
17	15	0	0	0	0	0	0	1	0	0	0	1	1	0	0	1	-1.14745	0	0	1	-0.04391	-0.04548	0.675613
18	16	0	0	0	0	0	0	1	0	0	0	1	1	0	0	1	1.465147	0	0	1	-0.04405	-0.06645	-0.9041
19	17	0	0	0	0	0	0	1	0	0	0	1	0	1	1	0	1.467265	0	0	1	-0.04678	-0.06789	-0.87492
20	18	0	0	0	0	0	0	1	0	0	0	1	1	0	0	1	1.479826	0	0	1	-0.04624	-0.06172	0.192224
21	19	0	0	0	0	0	0	1	0	0	0	1	1	0	0	1	1.446425	0	0	1	-0.04771	-0.06984	-0.91934
22	20	0	0	0	0	0	0	1	0	0	0	1	1	0	0	1	-0.45159	0	0	1	-0.04743	-0.06847	0.114933
23	21	0	0	0	0	0	0	1	0	0	0	1	1	0	0	1	-0.42489	0	0	1	-0.04743	-0.06847	0.054202
24	22	0	0	0	0	0	0	1	0	0	0	1	1	0	0	1	-1.08575	0	0	1	-0.04641	-0.06085	2.646019
25	23	0	0	0	0	0	0	1	0	0	0	1	1	0	0	1	1.468775	0	0	1	-0.0442	-0.06597	-0.78779
26	24	0	0	0	0	0	0	1	0	0	0	1	1	0	0	1	1.439916	0	0	1	-0.04771	-0.06984	-0.91931
27	25	0	0	0	0	0	0	1	0	0	0	1	1	0	0	1	-0.45422	0	0	1	-0.04743	-0.06847	0.114244
28	26	0	0	0	0	0	0	1	0	0	0	1	1	0	0	1	-0.43477	0	0	1	-0.04745	-0.06852	1.371659
29	27	0	0	0	0	0	0	1	0	0	0	1	1	0	0	1	-0.22303	0	0	1	-0.0477	-0.0698	-0.91799

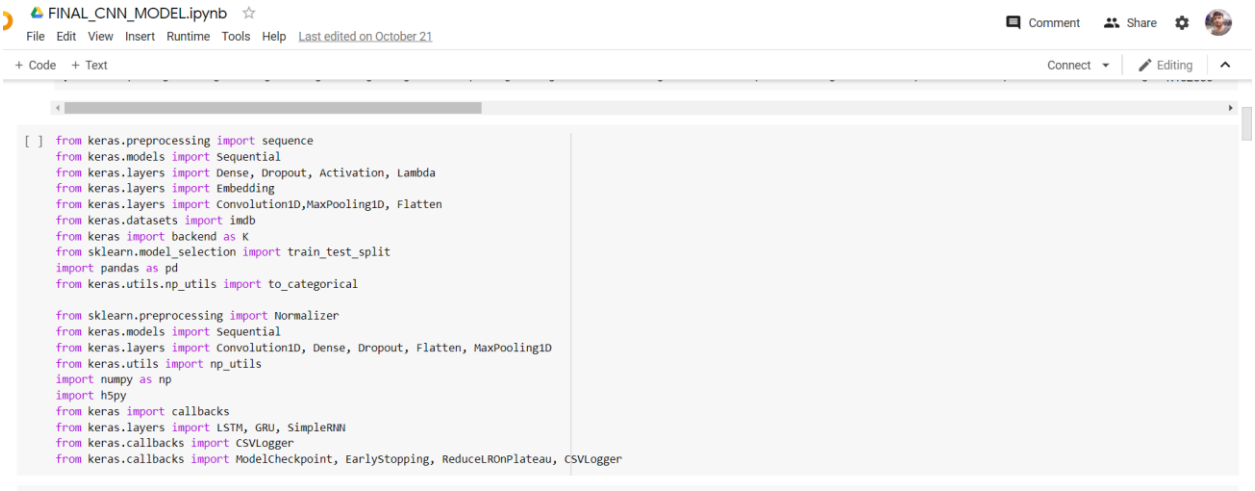
final_data_input_LAST_FINAL

3. TRAINING AND CROSS VALIDATION

IMPORTING NECESSARY LIBRARIES AND PACKAGES ON GOOGLE COLAB



	Unnamed: 0	ip_1	ip_144	ip_204	ip_205	ip_207	ip_8	not_freq	ip_4	ip_11	dest_ip_130_1_8	not_dest_freq	Wrong_port	port_is_443	Wrong_port.1	destination_port_443	Duration	fs
0	0	0	0	0	0	0	0	1	0	0	0	1	1	0	0	1	-0.440208	
1	1	0	0	0	0	0	0	1	0	0	0	1	1	0	0	1	1.171644	
2	2	0	0	0	0	0	0	1	0	0	0	1	1	0	0	1	1.145800	
3	3	0	0	0	0	0	0	1	0	0	0	1	1	0	0	1	2.703618	
4	4	0	0	0	0	0	0	1	0	0	0	1	0	1	1	0	1.152860	



FINAL_CNN_MODEL.ipynb ☆

File Edit View Insert Runtime Tools Help Last edited on October 21

Comment Share

+ Code + Text

Connect Editing

```
[ ] from keras.layers import LSTM, GRU, SimpleRNN
    from keras.callbacks import CSVLogger
    from keras.callbacks import ModelCheckpoint, EarlyStopping, ReduceLROnPlateau, CSVLogger
```

```
[ ] db = db.iloc[:,1:46]
```

```
[ ] db = db.iloc[:,0:45]
```

```
[ ] db
```

	ip_1	ip_144	ip_204	ip_205	ip_207	ip_8	not_freq	ip_4	ip_11	dest_ip_130_1_8	not_dest_freq	Wrong_port	port_is_443	Wrong_port.1	destination_port_443	Duration	fs_1807
0	0	0	0	0	0	0	1	0	0	0	1	1	0	0	1	-0.440208	0
1	0	0	0	0	0	0	1	0	0	0	1	1	0	0	1	1.171644	0
2	0	0	0	0	0	0	1	0	0	0	1	1	0	0	1	1.145800	0
3	0	0	0	0	0	0	1	0	0	0	1	1	0	0	1	2.703618	0
4	0	0	0	0	0	0	1	0	0	0	1	0	1	1	0	1.152860	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
269638	1	0	0	0	0	0	0	0	0	0	1	0	1	1	0	-1.180485	0
269639	1	0	0	0	0	0	0	0	0	0	1	0	1	1	0	-0.434267	0
269640	1	0	0	0	0	0	0	0	0	0	1	0	1	1	0	1.466744	0
269641	1	0	0	0	0	0	0	0	0	0	1	0	1	1	0	-0.446174	0
269642	1	0	0	0	0	0	0	0	0	0	1	0	1	1	0	-0.853504	0

269643 rows × 45 columns

FINAL_CNN_MODEL.ipynb ☆

File Edit View Insert Runtime Tools Help Last edited on October 21

Comment Share

+ Code + Text

Connect Editing

```
[ ] x1 = db.iloc[:,0:44]
```

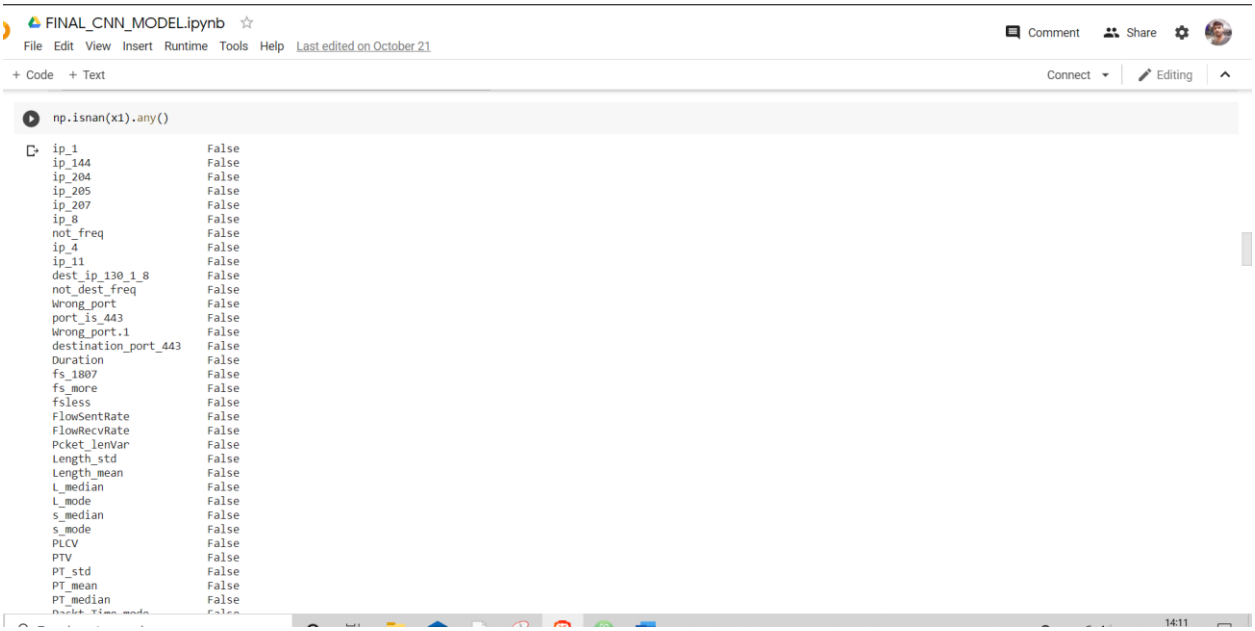
```
[ ] y1 = db.iloc[:,44]
```

```
[ ] x1
```

	ip_1	ip_144	ip_204	ip_205	ip_207	ip_8	not_freq	ip_4	ip_11	dest_ip_130_1_8	not_dest_freq	Wrong_port	port_is_443	Wrong_port.1	destination_port_443	Duration	fs_1807
0	0	0	0	0	0	0	1	0	0	0	1	1	0	0	1	-0.440208	0
1	0	0	0	0	0	0	1	0	0	0	1	1	0	0	1	1.171644	0
2	0	0	0	0	0	0	1	0	0	0	1	1	0	0	1	1.145800	0
3	0	0	0	0	0	0	1	0	0	0	1	1	0	0	1	2.703618	0
4	0	0	0	0	0	0	1	0	0	0	1	0	1	1	0	1.152860	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
269638	1	0	0	0	0	0	0	0	0	0	1	0	1	1	0	-1.180485	0
269639	1	0	0	0	0	0	0	0	0	0	1	0	1	1	0	-0.434267	0
269640	1	0	0	0	0	0	0	0	0	0	1	0	1	1	0	1.466744	0
269641	1	0	0	0	0	0	0	0	0	0	1	0	1	1	0	-0.446174	0
269642	1	0	0	0	0	0	0	0	0	0	1	0	1	1	0	-0.853504	0

269643 rows × 44 columns

Now I checked any invalid python data type or any number is present in my data which I will be using or if array contains at least one non-numeric value.



Sampling Of the Dataset

Dataset given is highly unbalanced. So, to prevent overfitting, sampling was done. Also since this was network data, therefore creating dummy data might not be feasible in real time. Therefore I preferred undersampling instead of over-sampling. To overcome this, I used technique of random sampling

As I had 249836 records in malicious and only 19807 records in benign, I tried Under Sampling and went Ahead with 19807 Records and upon Random

Sampling, Each Dataset Picked 19807 records to form 39614 Record Data Together.

FINAL_CNN_MODEL.ipynb

File Edit View Insert Runtime Tools Help Last edited on October 21

Comment Share

+ Code + Text

Connect Editing

```
[ ] df_c_0 = db[db['Benign'] == 'Malicious']
    df_c_1 = db[db['Benign'] == 'Benign']

[ ] df_c_0_under = df_c_0.sample(c_class_1)

[ ] df_under = pd.concat([df_c_0_under, df_c_1], axis=0)

[ ] print(df_under.Benign.value_counts())

    Benign       19807
    Malicious    19807
    Name: Benign, dtype: int64

[ ] df_Totoal = df_under

[ ] df_Totoal.head()
```

	ip_1	ip_144	ip_204	ip_205	ip_207	ip_8	not_freq	ip_4	ip_11	dest_ip_130_1_8	not_dest_freq	Wrong_port	port_is_443	Wrong_port.1	destination_port_443	Duration	fs_1807
253816	0	1	0	0	0	0	0	0	1	0	0	1	0	0	1	1.443449	0
131414	0	1	0	0	0	0	0	1	0	0	0	1	0	0	1	-1.135324	0
144159	0	0	0	0	0	0	1	0	0	1	0	1	0	0	1	-0.440212	0
167495	0	0	0	1	0	0	0	0	1	0	0	1	0	0	1	-1.150709	0
253272	0	0	0	0	0	0	1	0	0	1	0	1	0	0	1	1.140970	0

TRAINING AND TESTING ACCURACY

```
[ ]  x_train,x_test,y_train,y_test = train_test_split(x,y,train_size=0.8,random_state=42)
```

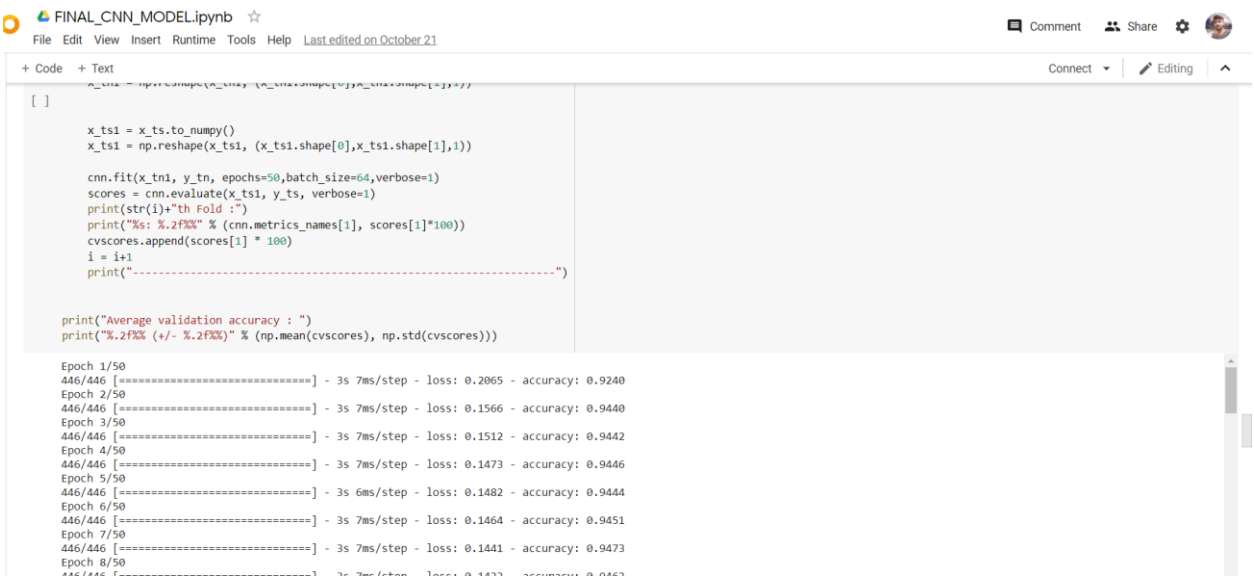
```
[ ]  from sklearn.model_selection import StratifiedKFold
     kfold = StratifiedKFold(n_splits=10, shuffle=True, random_state=42)
```

MODEL CREATION

I also added two layers in my CNN model which are:

Convolution1D and MaxPooling1D

And the optimizer used here is “adam”




```
[ ] Epoch 36/50
    446/446 [==============================] - 3s 7ms/step - loss: 0.1385 - accuracy: 0.9480
    Epoch 37/50
    446/446 [==============================] - 3s 7ms/step - loss: 0.1382 - accuracy: 0.9483
    Epoch 38/50
    446/446 [==============================] - 3s 7ms/step - loss: 0.1378 - accuracy: 0.9480
    Epoch 39/50
    446/446 [==============================] - 3s 7ms/step - loss: 0.1375 - accuracy: 0.9484
    Epoch 40/50
    446/446 [==============================] - 3s 7ms/step - loss: 0.1376 - accuracy: 0.9487
    Epoch 41/50
    446/446 [==============================] - 3s 7ms/step - loss: 0.1371 - accuracy: 0.9475
    Epoch 42/50
    446/446 [==============================] - 3s 7ms/step - loss: 0.1371 - accuracy: 0.9486
    Epoch 43/50
    446/446 [==============================] - 3s 7ms/step - loss: 0.1384 - accuracy: 0.9483
    Epoch 44/50
    446/446 [==============================] - 3s 7ms/step - loss: 0.1390 - accuracy: 0.9484
    Epoch 45/50
    446/446 [==============================] - 3s 7ms/step - loss: 0.1373 - accuracy: 0.9485
    Epoch 46/50
    446/446 [==============================] - 3s 7ms/step - loss: 0.1363 - accuracy: 0.9486
    Epoch 47/50
    446/446 [==============================] - 3s 7ms/step - loss: 0.1378 - accuracy: 0.9481
    Epoch 48/50
    446/446 [==============================] - 3s 7ms/step - loss: 0.1376 - accuracy: 0.9488
    Epoch 49/50
    446/446 [==============================] - 3s 7ms/step - loss: 0.1372 - accuracy: 0.9486
    Epoch 50/50
    446/446 [==============================] - 3s 7ms/step - loss: 0.1377 - accuracy: 0.9487
    100/100 [==============================] - 0s 2ms/step - loss: 0.1527 - accuracy: 0.9382
    9th Fold :
    accuracy: 93.82%
    ------------------------------------------------------------------
    Average validation accuracy :
    94.63% (+/- 0.47%)
```

Validation Accuracy is **: 94.63% (+/-0.47%):-**

Result of 10-fold cross validation with 60 epochs each:

94.63% (+/-0.47%)

Training accuracy: 94.83449459075928

Test accuracy: 94.44654583930969

HYPER PARAMETER TUNING

Parameters for the model like number of hidden units, number of stacked units, choice of optimizer, loss function chosen, value of Dropouts, number of epochs, folds variations and choice of metrics has been tuned to get the final model with better accuracy.

```
[ ] x_tr = x_train.to_numpy()
    x_tr = np.reshape(x_tr, (x_tr.shape[0], x_tr.shape[1],1))

[ ] x_ts = x_test.to_numpy()
    x_ts = np.reshape(x_ts, (x_ts.shape[0], x_ts.shape[1],1))

[ ] y_tr = pd.get_dummies(y_train)
    y_ts = pd.get_dummies(y_test)

[ ] _, train_acc = cnn.evaluate(x_tr, y_tr, verbose=1)
    _, test_acc = cnn.evaluate(x_ts, y_ts, verbose=1)

    991/991 [==============================] - 1s 2ms/step - loss: 0.1343 - accuracy: 0.9483
    248/248 [==============================] - 0s 2ms/step - loss: 0.1407 - accuracy: 0.9445

[ ] print("Training accuracy: "+str(train_acc*100))
    print("Test accuracy: "+str(test_acc*100))

    Training accuracy: 94.83449459075928
    Test accuracy: 94.44654583930969

[ ] y_probs = cnn.predict(x_ts, verbose=1).ravel()
    y_classes = cnn.predict_classes(x_ts, verbose=1)

    248/248 [==============================] - 0s 2ms/step
    WARNING:tensorflow:From <ipython-input-41-ed2a250cf2a8>:2: Sequential.predict_classes (from tensorflow.python.keras.engine.sequential) is deprecated and will be removed after 2021-01-0
    Instructions for updating:
    Please use instead:* `np.argmax(model.predict(x), axis=-1)`,   if your model does multi-class classification   (e.g. if it uses a `softmax` last-layer activation).* `(model.predict(x)
    248/248 [==============================] - 0s 2ms/step
```

PERFORMANCE METRICS (Precision, Recall, F1-Score, Support)

```
[ ] from sklearn.metrics import classification_report
    print(classification_report(y_ts,y_pred))

                  precision    recall  f1-score   support

               0       0.92      0.97      0.95      3926
               1       0.97      0.92      0.94      3997

       micro avg       0.94      0.94      0.94      7923
       macro avg       0.95      0.94      0.94      7923
    weighted avg       0.95      0.94      0.94      7923
     samples avg       0.94      0.94      0.94      7923

[ ] from sklearn.preprocessing import LabelEncoder

[ ] le = LabelEncoder()

[ ] y_test_le = le.fit_transform(y_test)

[ ] from sklearn.metrics import roc_curve
    fpr_keras, tpr_keras, thresholds_keras = roc_curve(y_test_le, y_classes)

[ ] from sklearn.metrics import auc
    auc_keras = auc(fpr_keras, tpr_keras)

[ ] from matplotlib import pyplot
```

NOTE : THE F-SCORE VALUE NEAR 1 (~0.95) PROVES THAT THE MODEL IS BEST TRAINED, NEITHER UNDER-FITTED OR OVER-FITTED.

AUC:

0.9446646242960366

```
print(auc_keras)

0.9446646242960366
```

ROC CURVE:

TOTAL COMPUTATION TIME :

ON LOCAL MACHINE (JUPYTER NOTEBOOK) : ~ 9-10 Hours (TOTAL TIME)

ON GOOGLE COLAB: <1 Hours

Overview of results:

(10 folds and 50 epochs)

```
------------------------------------
Average validation accuracy :
94.63% (+/- 0.47%)
```

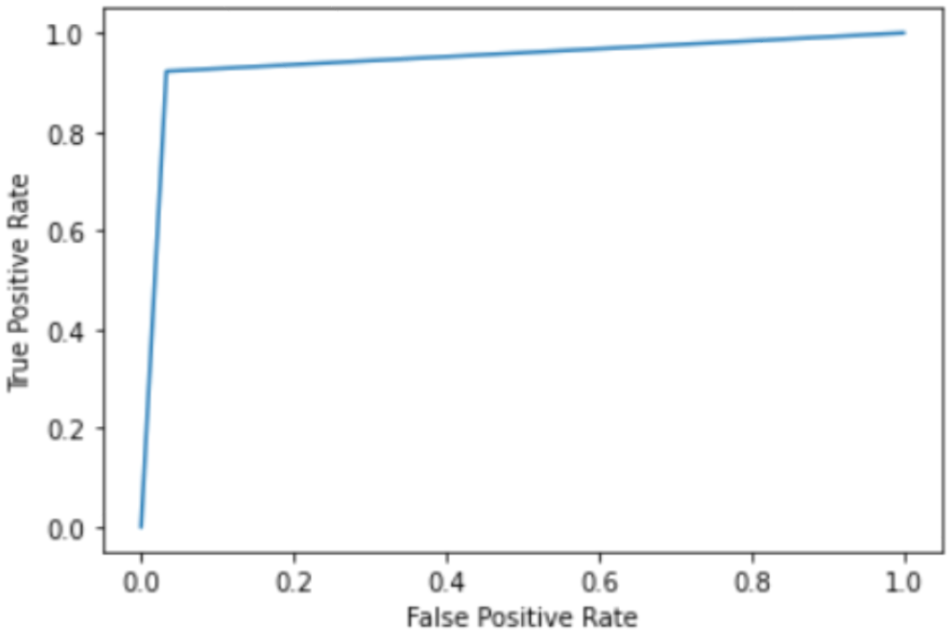
```
Training accuracy: 94.83449459075928
Test accuracy: 94.44654583930969
```

	precision	recall	f1-score	support
0	0.92	0.97	0.95	3926
1	0.97	0.92	0.94	3997
micro avg	0.94	0.94	0.94	7923
macro avg	0.95	0.94	0.94	7923
weighted avg	0.95	0.94	0.94	7923
samples avg	0.94	0.94	0.94	7923

AUC:

```
0.9446646242960366
```

```
pyplot.plot(fpr_keras, tpr_keras)
pyplot.xlabel('False Positive Rate')
pyplot.ylabel('True Positive Rate')
pyplot.show()
```



Yuvraj Kumar
(18BIT0276)

Student's Signature