Prediction of Malicious traffic in IoT network

Spoorthy Reddy Jarugu
Vellore Institute of Technology, India

Introduction

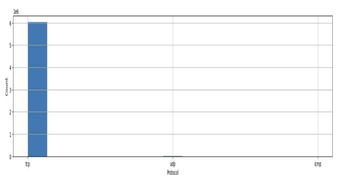
IoT devices are growing rapidly and due to there less computational power these devices are getting compromised easily. Task is to predict malicious traffic in the IoT network. Prediction is done based on the lables that are assigned is the dataset.

- In the data generated have 22 columns where time, IP address for originating point and responding point, protocol used for communication, duration, connection state, no of packets, label are recorded.
- With the combination of all the columns finally labels says the particular communication happened is Benign or Malicious.

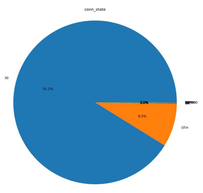
Data preprocessing refers to the steps and techniques used to prepare raw data for analysis. It involves a series of steps to transform the raw data into a clean, organized, and structured format that is suitable for analysis.

Data Preprocessing

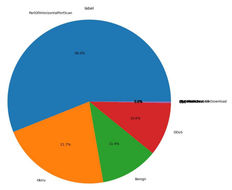
- Protocols can operate at different layers of the networking stack, from physical layer protocols that define electrical and physical aspects of communication, to higher-level protocols that deal with formatting and structuring of data, Here for our dataset we have TCP (Transmission Control Protocol), UDP(User Datagram Protocol), ICMP(Internet Control Message Protocol).
- Conn_state: Connection state of the network traffic is displayed in pi chart.
- Label is the main feature in our dataset which is used for prediction. this feature says weather the network traffic is malicious or benign. Below is the pi chart representation of label feature.



Protocol Visualization



Connection state pi chart representation



Pi chart representation of label

Label encoding

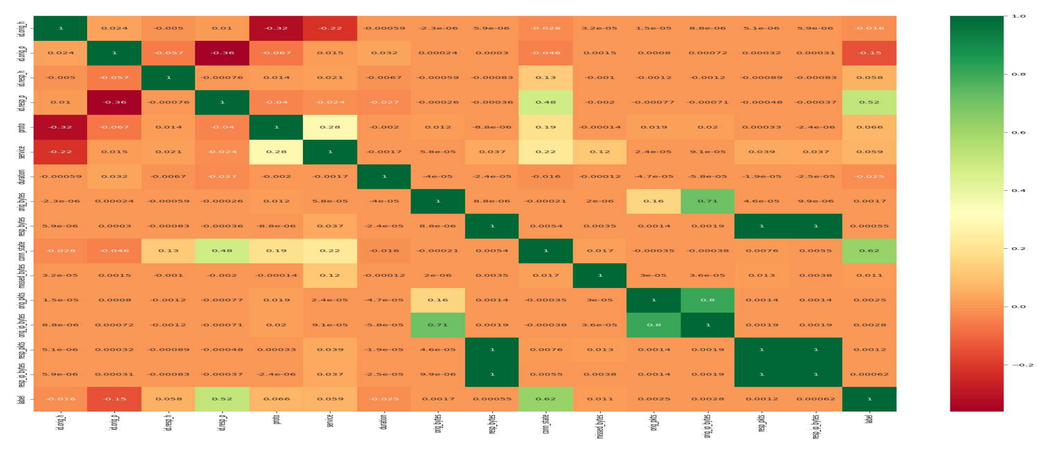
- Categorical data are those that are represented by labels or categories, such as gender, color, or type of product. Machine learning algorithms require numerical input, so numerical encoding is necessary to transform categorical data into numerical values.
- There are different types of encoding techniques that are available. We use label encoding to convert 'proto', 'service', 'conn_state', 'label' variables to numerical values.

Label Encoding

```
In [49]: pmap = {'tcp':0,'udp':1,'icmp':2}
         dd['proto']=dd['proto'].map(pmap)

In [50]: pmap = {'nil':0,'dns':1,'irc':2,'http':3,'dhcp':4,'ssl':5,'ssh':6}
         dd['service']=dd['service'].map(pmap)

In [51]: pmap = {'S0':0,'OTH':1,'SF':2,'REJ':3,'S3':4,'RSTR':5,'SH':6,'RSTO':7,'RSTOS0':8,'S1':9,'SHR':10,'S2':11,'RSTRH':12}
         dd['conn_state']=dd['conn_state'].map(pmap)

In [52]: pmap = {'PartOfAHorizontalPortScan':0,'Okiru':1,'Benign':2,'DDoS':3,'C&C':4,'C&C-HeartBeat':5,'Attack':6,'C&C-FileDownload':7,'C8
         dd['label']=dd['label'].map(pmap)
```

Correlation

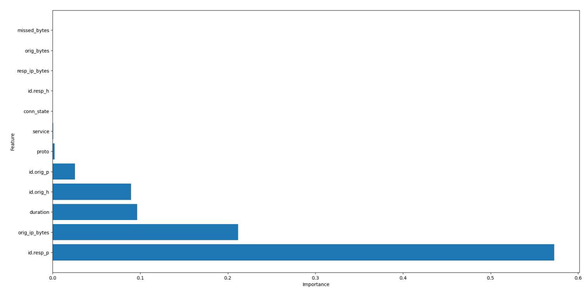
- Correlation is a statistical measure that describes the relationship between two variables. It is used to determine how strongly and in what way two variables are related to each other. Correlation coefficients range between -1 and +1.



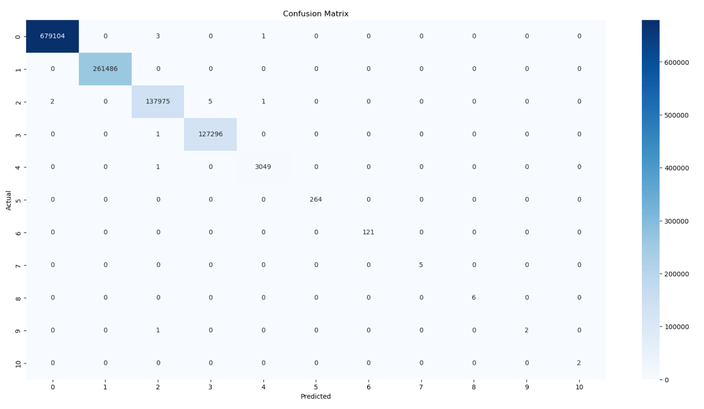
Evaluation

We used Decision Tree Classifier which is the supervised learning algorithm.

Evaluation Matric	Value
MSE	6.5325
RMSE	0.0080
MAE	2.0672
Accuracy	0.9999
Precision	0.9999
Recall	0.9696
F1-Score	0.9817



Important features used



Evaluation result

Conclusion

For this Flip00 task I have taken dataset from Kaggle. For which data preprocessing, data cleaning, splitted the data for train and test, finally applied Decision Tree Classifier algorithm.

Additionally used originator IP address (id.orig_h) and respondent IP address (id.resp_h) by converting them to integer form.

This prediction model helps to effectively predict the Labels of traffic.