**IBM PROJECT-MALWARE DETECTION**

GROUP 10

**Nithin Krishna**

**AIM**

Detection and Prevention of Advanced Persistent Threat (APT) activities in heterogeneous networks using SIEM and Deep Learning.

**FEATURES**

* Our observation also includes a multi class classified output with various types of attacks like DoS, Probe, U2R, etc…
* Hence, using these common features, we understand that when a new data point is provided and asked to classify under a type of attack with these respective columns’ data, we can find its type of attack.

1. Protocol
2. Service
3. Flag
4. Duration
5. level

**DATASET DESCRIPTION**

The dataset I’ve have considered is NSL-KDD

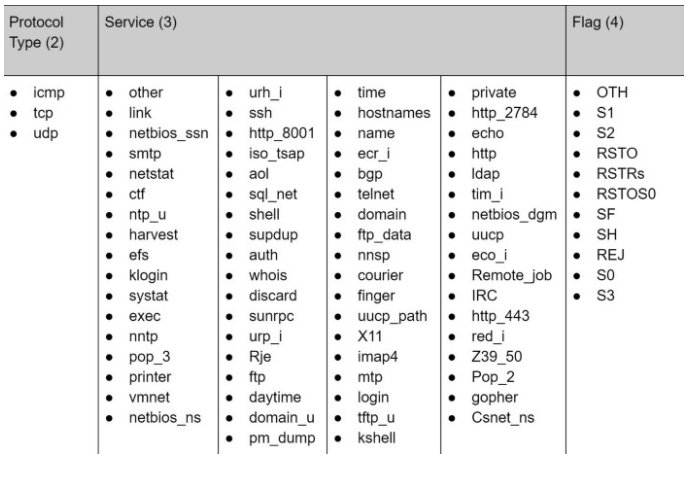
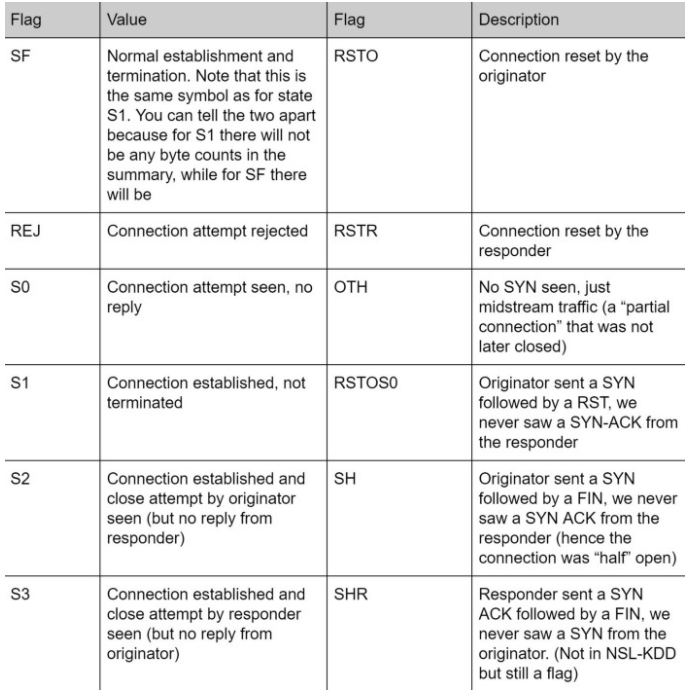
The feature types in this data set can be broken down into 4 types:

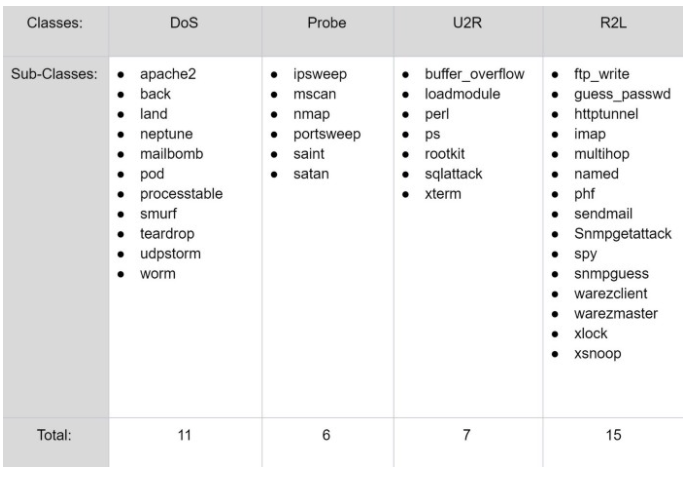
4 Categorical (Features: 2, 3, 4, 42)

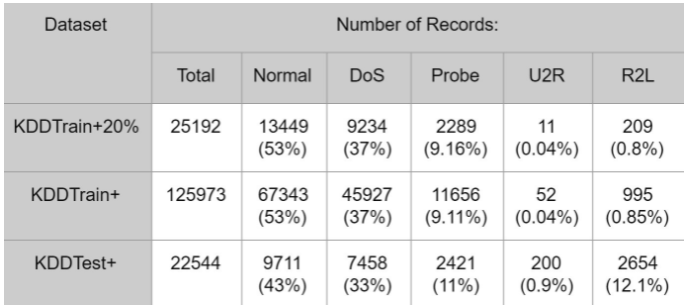
6 Binary (Features: 7, 12, 14, 20, 21, 22)

23 Discrete (Features: 8, 9, 15, 23–41, 43)

10 Continuous (Features: 1, 5, 6, 10, 11, 13, 16, 17, 18, 19)





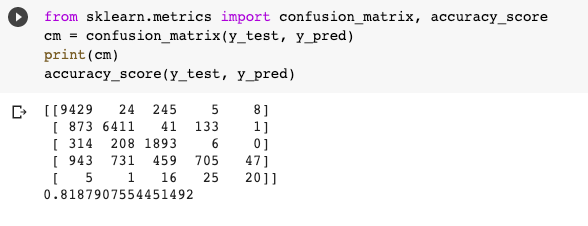


**ALGORITHM USED AND APPROACH**

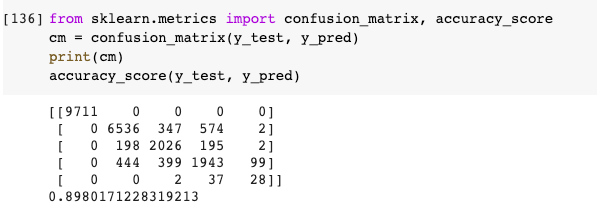
* We have used many algorithms on our dataset such as K Nearest Neighbour (KNN) , Decision Tree Classifier , XG Boost & Artificial Neural Network (ANN) on the dataset to obtain the confusion matrix and a classification report & an Apriori algorithm that searches for a series of frequent sets of items in the datasets.
* To begin, we have performed :
  1. Calculated variance — columns with 0 variance have been dropped
  2. Plotted a heat map to check for independent features that are highly correlated to each other and dropped them.
  3. Converted all the attacks into distinct categories(DOS,Probe,U2R,R2l)
  4. Calculated the impact of the independent features on Y and dropped the ones which do not constitute much of an effect.
  5. Applied one hot encoding on the categorical variables to convert strings to numbers.
  6. Split dataset into Test and Train datasets
  7. Scaling - Used StandardScaler() - to normalise all values
  8. Applied KNN ,Decision Tree Classifier, XG Boost , ANN & Apriori algorithms on the dataset and found the results.

**PERFORMANCE MATRIX**

1. Accuracy Score = 0.8187907554451492

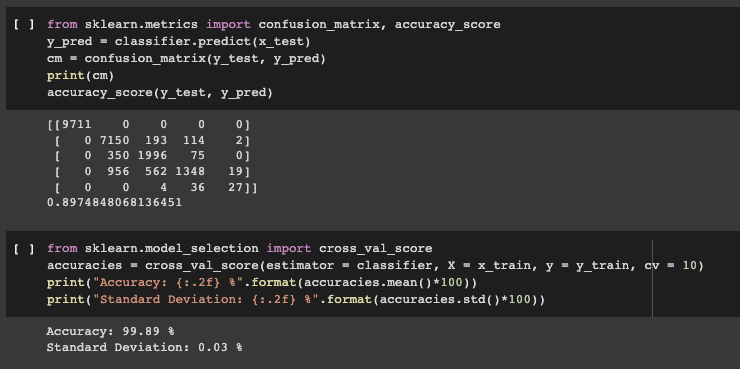
Algorithm Used : **KNN**

2. Accuracy Score = 0.8979727631637315

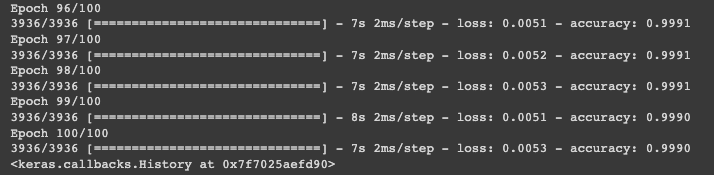
Algorithm Used : **Decision Tree Classifier**

3. Accuracy Score = 0.8979727631637315

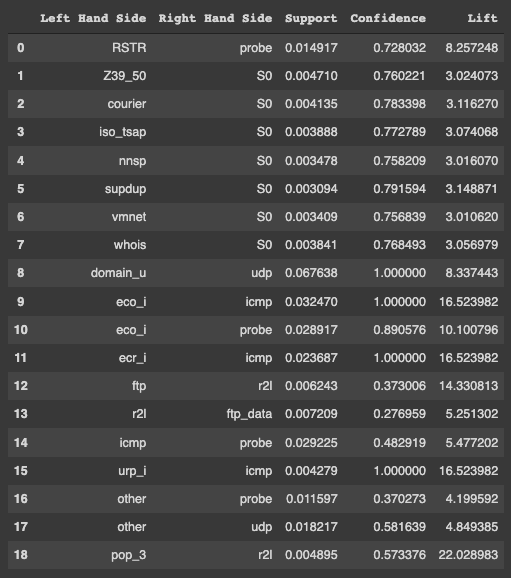
Cross Validation Score: 99.89 %

Algorithm Used : **XG Boost**

4. Algorithm Used : **ANN**

Accuracy Score = 0.990****

**Apriori** —

Over relational databases, Apriori is an algorithm for frequent item set mining and association rule learning. It works by recognising the most common individual items in the database and expanding them to bigger and larger item sets as long as those item sets exist in the database frequently enough. We can see the results we have obtained in the following picture below :

**REFERENCES:**

<https://towardsdatascience.com/a-deeper-dive-into-the-nsl-kdd-data-set-15c753364657>

**LINK TO DATASET:**

<https://www.unb.ca/cic/datasets/nsl.htm>

**GITHUB REPOSITORY :**

<https://github.com/nithin0905/malware_analysis>