**PAF-KIET**

**COCIS**

**PROJECT OF NP /ML**

Spring 2022

**Project Title**

**DETECTION SYSTEM FOR BOTNET**

**ATTACK**

Project Members Name

**GUFRAN BHATTI (64386)**

**SARIM AMIR (63686)**

**ABDUL BARI (10233)**

INTRODUCTION:

An intrusion detection system, is an intrusion detection system for detecting both network and computer intrusions and misuse by monitoring system activity and classifying it as either normal or anomalous. In order to positively identify attack traffic, the system must be taught to recognize normal system activity. The two phases of a majority of detection systems consist of the training phase (where a profile of normal behaviours is built) and testing phase (where current traffic is compared with the profile created in the training phase). Anomalies are detected in several ways, most often with artificial intelligence type techniques. Systems using artificial neural networks have been used to great effect. Another method is to define what normal usage of the system comprises using a strict mathematical model, and flag any deviation from this as an attack. Other techniques used to detect anomalies include data mining methods, grammar based methods, and Artificial Immune System.

Network-based intrusion detection systems often provide a second line of defence to detect anomalous traffic at the physical and network layers after it has passed through a firewall or other security appliance on the border of a network. Host-based anomalous intrusion detection systems are one of the last layers of defence and reside on computer end points. They allow for fine-tuned, granular protection of end points at the application level.

AIMS AND OBJECTIVE:

The purpose of an intrusion detection system (IDS) is to protect the confidentiality, integrity, and availability of a system. Intrusion detection systems (IDS) are designed to detect specific issues, and are categorized as signature-based (SIDS) or anomaly-based (AIDS). IDS can be software or hardware.

PROBLEM STATEMENT:

An IDS should be able to identify all abnormal patterns and traffic using monitoring, detecting and responding to unauthorized activities within the system. However, regarding its huge and unbalanced datasets, IDS encounters total data processing problem. Thus, different techniques have been presented which can handle this problem.

EXPERIMENTS:

We will be using four different types of algorithm to see which algorithm will be best for detecting an anomalous behaviour with the help of iot\_botnet dataset. The dataset contains 668522 rows and 46 columns.

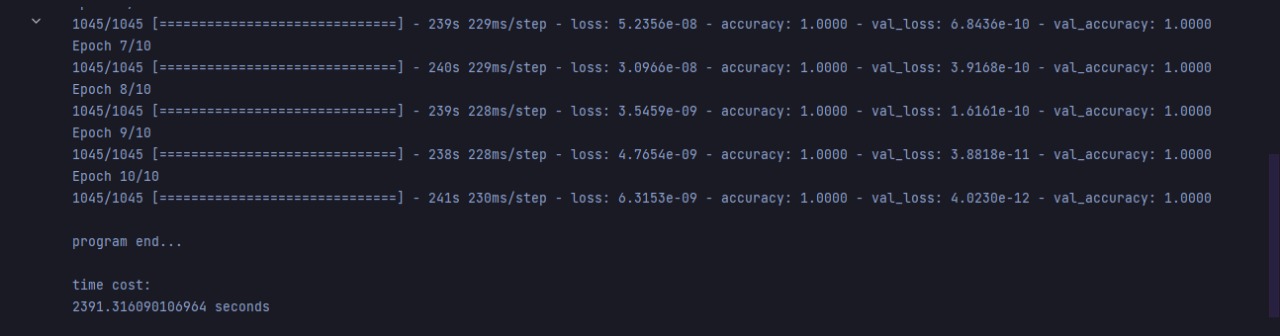
1. CNN (Convolution Neural Network)

CNN is mostly used in image recognition and processing that is specifically designed to process pixel data but it can also be used for any classification problem. Here we will see how our algorithm will perform on this dataset with different train/test ratio.

First we will read our dataset from the directory by using pandas library. Then we will be doing feature engineering on our dataset. Meaning that we will be dropping data from our dataset which doesn’t affect the result of target variable. Here our dropping feature are [‘stime’,’flgs’,’saddr’,’daddr’,’ltime’,’state’]. We will also replace the strings into integers in the following features [‘proto’,’category’,’subcategory’].

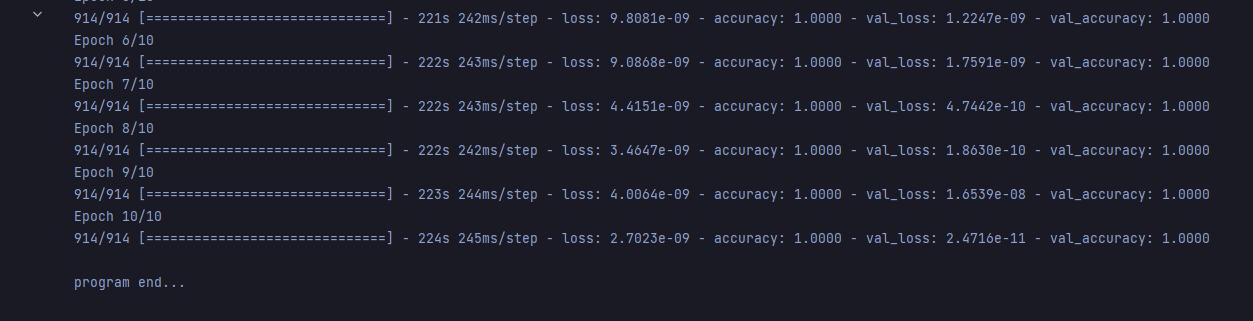
After that we will divide our dataset into x (dataset except target variable) and y (target variable). After that we will normalize our data (x and y) using the minmaxscalar() method. Now we will split our data into two different ratios 80/20 and 70/30 and see how our result is affected. We will run 10 epochs with batch size = 512.

**Results when ratio = 80/20 :**



**Results when ratio = 70/30:**

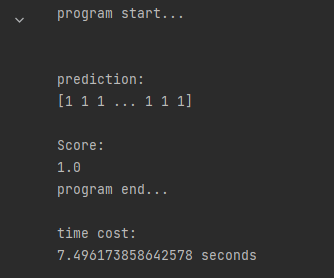


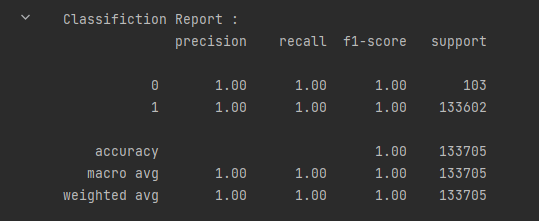


1. DT (Decision Tree):

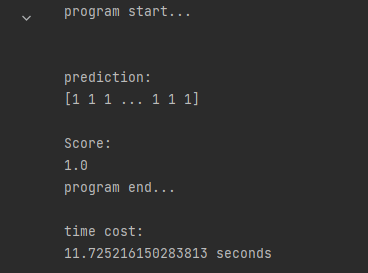
The decision tree Algorithm belongs to the family of supervised machine learning algorithms. It can be used for both a classification problem as well as for regression problem. Here the feature engineering process of the dataset will be the same as we did in CNN. After that we will train our model using DT algorithm and see the predictions and accuracy of this algorithm. Again, here we see the results first with 80/20 ratio then with 70/30 ratio.

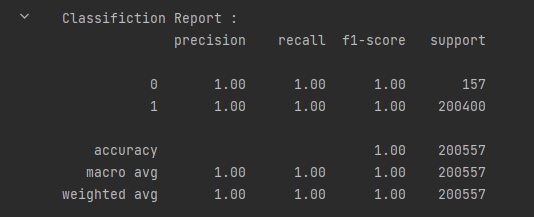
**Results when ratio = 80/20:**





**Results when ratio = 70/30:**

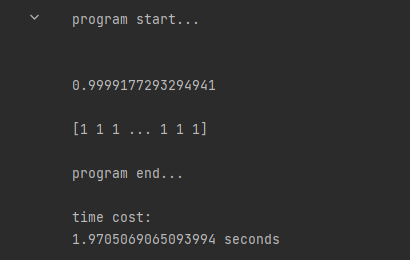


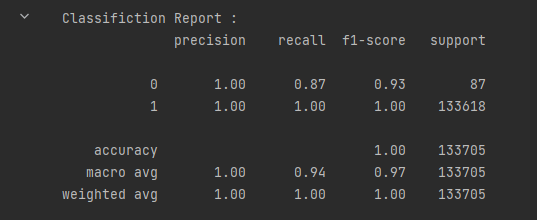


1. NB (Naïve Bayes):

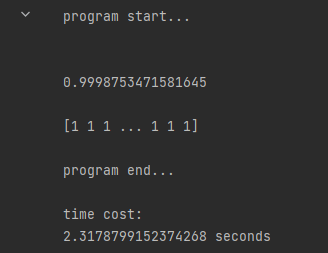
It is a classification technique based on Bayes' Theorem with an assumption of independence among predictors. In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature. Here the feature engineering process of the dataset will be the same as we did in DT. After that we will train our model using NB algorithm and see the predictions and accuracy of this algorithm. Again, here we see the results first with 80/20 ratio then with 70/30 ratio.

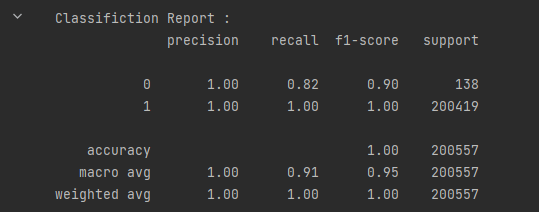
**Results when ratio = 80/20:**





**Results when ratio = 70/30:**

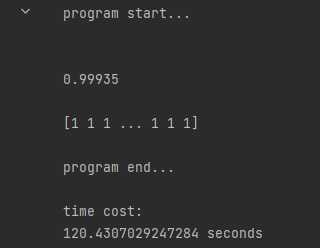


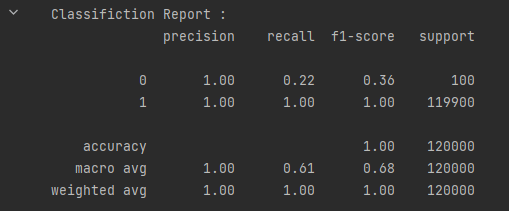


1. SVM (Support Vector Machine):

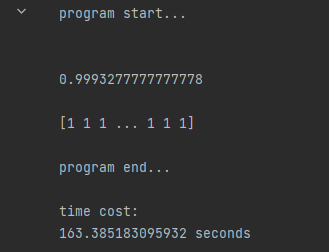
Support Vector Machine (SVM) is a supervised machine learning algorithm used for both classification and regression. Though we say regression problems as well its best suited for classification. The objective of SVM algorithm is to find a hyperplane in an N-dimensional space that distinctly classifies the data points. Here the feature engineering process of the dataset will be the same as we did in NB. There is one difference that we selected only the first 60000 rows of our dataset because it was taking a lot of time to train the model. After that we will train our model using SVM algorithm and see the predictions and accuracy of this algorithm. Again, here we see the results first with 80/20 ratio then with 70/30 ratio.

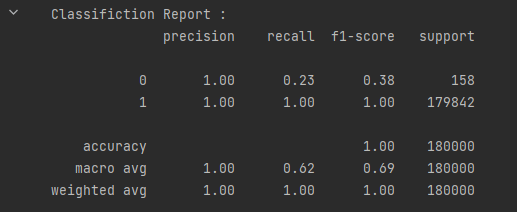
**Results when ratio = 80/20:**





**Results when ratio = 70/30:**





CONCLUSION:

If we observe the results of our experiments every algorithm had an accuracy of nearly 100%. It is because our dataset is labelled and all the algorithms is for supervised learning. This means our IDS is signature-based and it can only detect those anomalies which is present in the dataset. Our IDS cannot detect anomalies outside our dataset. For that we need to create an anomaly-based IDS which works on reinforcement learning. Also, if we observe we are getting more accuracy when the ratio is 80/20. The difference is very little but it is there.