

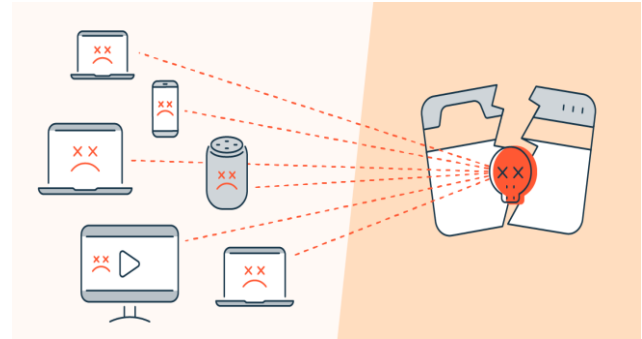
IDS 2018 Intrusion CSVs (CSE-CIC-IDS2018

Dataset!



INTRODUCTION

The **(CSE-CIC-IDS2018)** dataset is a large dataset created in 2018 by New Brunswick's University, **that aims to analyzing DDoS data**. Distributed denial of service (DDoS) attack defined as a **malicious attempt** to disrupt a targeted server's, service's, or network's normal traffic by flooding the target's Internet traffic or its surrounding infrastructure. This dataset based on the logs of the New Brunswick's University server, and it is separated into different files dependent on the date. Each file is unbalanced, and I will deal with this problem.



DATASET STRUCTURE

First, I got this dataset from Kaggle website. Generally, this dataset contains 80 columns, and 10,000 rows. I will choose the most important 10 columns from among them before I start executing.

QUESTIONS



Until now, there are two questions that I interested to answer them, which are:

Q1:

Which packets sent are malicious?


Q2:

Which Dst Port (Destination port) / protocol most of malicious packets came from?



Who benefits from exploring this question or building this model/system?

Anyone (organization or individuals) who would like to more accurately track **attacks** coming from **other networks**, hence **increase** protection of their information systems and assets against **DDoS attacks**.



TOOLS

Initially, I will be using the **Jupyter Notebook**.



WHAT I NEED TO DO?

I will follow the below methodology to achieve the goal:



1

Applying Exploratory data analysis to the dataset.

3

Features selection

2

Handling imbalanced data

4

Building the model
(Classification model)



I- Applying Exploratory data analysis to the dataset.

This might include:

- Description of the **data frame**, its **shape** and summary statistics.
- Handling **missing values** and categorical data.
- Maybe other **exploration** techniques.

2- Handling imbalanced data

This might include using some technique, such as **oversampling**, **bagging**, and **boosting** techniques.

3- Features selection

This step would include using **certain function** to choose only **most important features** as independent variables of the prediction process.

- .

4- Building the model

This step would include **training the model** to classify the data, and then **predict** the class of each server logs (on the testing data)

Finally, I will apply suitable **evaluation measurements** to prove the model robustness.