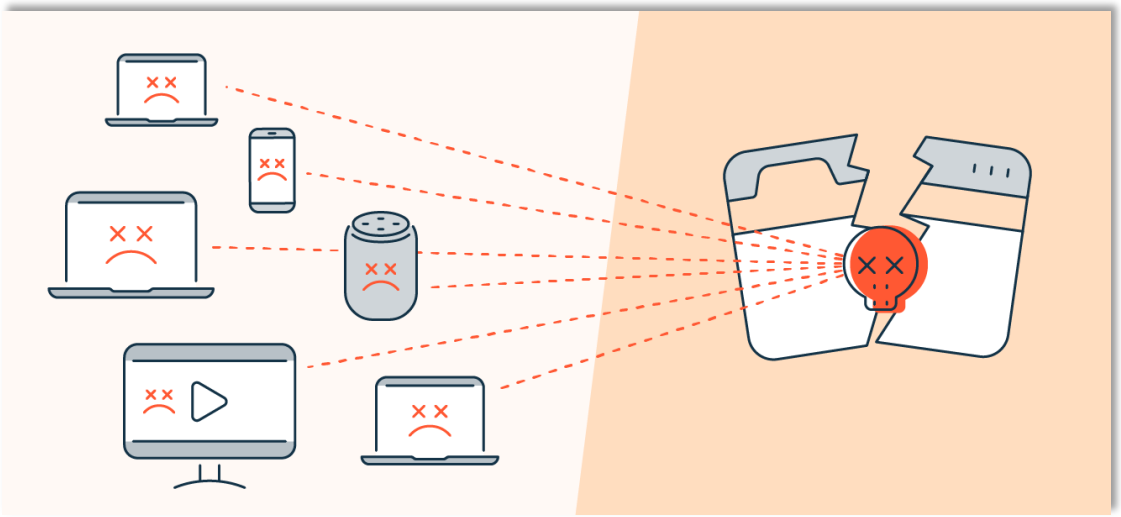
Intrusion Detection



Abstract

The primary project goal is to use a classification model to analyzing DDoS data. So, this project aims to detect DDoS attacks of various formats. I worked with a large data provided from the New Brunswick's University, this data was separated into different files dependent on the date. I choose one file of them to work on it. And then I dealt with a problem of unbalanced data as well as applied some data cleaning to prepare the data to the classification model step as it is the last step in this project, also 3 different classification models were implemented.

Design

This project deal with a one of data science problems (classification problem). I followed some organized steps to deal with this problem so, started with Importing the needed libraries and loading the dataset, then exploring the data an, data cleaning and some preprocessing before applying the classification models

Data

First, I got this dataset from Kaggle website in a .csv format. Generally, this dataset contains 80 columns, and 1048575 rows. 2 of which are categorical, one of them is the target column ‘Label’ and it has a binary class "Benign, Bot" for DDoS. All The rest is a numerical feature such as Destination Port, Protocol, Total Forward Packets, and I was select the most important 20 features from among them before I start executed.

Algorithms

I used three different models, The first one is the Logistic Regression, the second one is Random Forest, and the last one is k-nearest Neighbors classifiers. Each of them tested on the accuracy, precision, recall, and F1 score. The following table will summarize the results of the three applied models:

Model	Accuracy	Precision	Recall	F1 score
Logistic Regression	97%	100%	94%	97%
Random Forest	100%	100%	100%	100%
k-nearest Neighbors	100%	100%	100%	100%

Tools

I was used some of python libraries such as:

- NumPy and Pandas to manipulate data
- Matplotlib and Seaborn for plotting
- Scikit-learn to apply and evaluate the models

Communication

In addition to the PowerPoint slides and visuals presented, I will present my work to my classmates, and I will embed it on the GitHub.

In conclusion, adapted classification models results satisfies the main goal of this project, in terms of high ability to learn the data pattern thus the precise prediction process.

In future, further steps to manipulate data will be taken, beside trying to use another classifier as ANN (artificial neural network)