

CS-359 Machine Learning

Final Report

Introduction:

Intrusion detection/prevention systems (IDSs/IPSs) are essential perimeter security tools. They are considered the frontline defense tools against various cyberattacks. In this project, the CIC-IDS2017 dataset is used. The original dataset contains benign traffic and 14 attack traffic. The traffic is captured over five days. The dataset is available at <https://www.unb.ca/cic/datasets/ids-2017.html>

The dataset provides five large pcap files, one file per day. The pcap files include information about each exchanged packet. As a preprocessing stage, the pcap files are processed by the CICFlowMeter tool, which generates flows, each with 78 features. These features can be used as input to machine learning algorithms. Note that an internet traffic flow is identified by a unique combination of source and destination internet protocol numbers (IPs), source and destination port numbers, a timestamp, and transport protocol (TCP or UDP).

In this project, you will use the knowledge you acquired in this course to build classifiers that can efficiently distinguish denial of service (DoS) attacks from benign traffic. In the dataset, only Wednesday traffic includes DoS attacks and benign traffic. The Wednesday traffic includes:

- Benign
- DoS GoldenEye
- DoS Hulk
- DoS slowhttptest
- DoS slowloris
- Heartbleed

The last attack is not a denial of service, so the corresponding flows should be removed from the dataset. The DoS attacks should be grouped in a class labeled 'attack.'

Data Sources:

You will be provided with the dataset in the following formats:

- Several CSV files
- Several JSON files
- Several parquet files

The last column in the dataset is the class label.

Submission:

1. The completed architectural decision records document in PDF format
2. The Python notebook for the project in ipynb format
 - a. You should include comments and discussions at every stage, mainly when making critical decisions or evaluating model performances.
3. A snapshot of the first 20 rows of the loaded document in step 10 of the ADR document.
4. [Optional] A presentation file to present the content of the ADR document