UNSW-NB15 Preprocessing & Train/Val/Test Split

Week 4.1 - Lab

Day 1 – Security Datasets + Feature Engineering Basics

Lead: *Muhammad Tayyab*
Athar Imran
Buildables Cybersecurity Fellow

29th of *September, 2025 - Wednesday*

Dataset Acquisition:

Dataset Name: UNSW-NB15 Intrusion Detection Dataset
Source: UNSW Canberra Cyber
Year: 2015
Size: ~2.5M records, 49 features (42 traffic features + 2 labels + 5 metadata)
Attack Types: 9 categories (Fuzzers, Analysis, Backdoors, DoS, Exploits, Generic, Reconnaissance, Shellcode, Worms) + Benign
Format: CSV (already feature-engineered using Argus & Bro [Zeek] + custom tools)

License and Usage

- The dataset is **publicly available for research purposes**.
- Redistribution of the raw dataset is restricted; one must download from the official UNSW site.
- Cite the creators when publishing:
  Moustafa, N., & Slay, J. (2015). UNSW-NB15: A comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set). Military Communications and Information Systems Conference (MilCIS).

Python Notebook Code:

1. Import both datasets (train & test) to Google Drive.
2. Go to Google CoLab.
3. Write the code that will:
    a. **Import** Pandas, Numpy, sklearn.model_selection (train_test_split), sklearn.preprocessing (LabelEncoder) and Drive from google.colab
    b. **Load** both datasets into the notebook
    c. **Clean** the dataset by removing the index column and replacing empty values with NaNs.
    d. **Encode Labels** that indicate the number of benign vs. attack samples. Giving us two targets:
        i. Binary classification: attack vs benign.
        ii. Multiclass classification: attack types.
    e. **Select Features** to ensure a clean numerical feature matrix.
    f. **Split Data** into 70/15/15.
    g. **Save** splits.

(Code is provided along with the report.)

The code will output train, val, and test datasets split into feature and label.