جامعة الأمير سلطان
PRINCE SULTAN UNIVERSITY

Lab 1: Botnet Detection Using Supervised Learning

Objective:

We have studied in the past weeks supervised learning, which can be used for cybersecurity classification tasks, including intrusion detection, malware classification, and botnet detection. This lab applies these concepts to a real-world problem of network security.

This lab focuses on detecting botnet traffic in network flows using supervised machine learning. We will utilize the CTU-13 dataset, a well-known benchmark for botnet detection, and apply Logistic Regression, Decision Tree, and Support Vector Machine (SVM) classifiers. The goal is to analyze network traffic patterns, extract meaningful features, and train classifiers to distinguish between normal and botnet activities.

Tasks:

- Preprocess and clean network flow data to ensure high-quality input for modeling.
- Extract and engineer relevant features from raw NetFlow data, including traffic volume, packet size, flow duration, and byte transmission patterns.
- Train and compare multiple machine learning models, including Logistic Regression, Decision Tree, and Support Vector Machine (SVM), to determine which algorithm performs best in botnet classification.
- Evaluate model performance using classification reports, focusing on precision, recall, and F1-score rather than just accuracy.
- Analyze false positives and false negatives through confusion matrices to understand model weaknesses and trade-offs in botnet detection.
- Investigate feature importance to identify which network flow attributes contribute the most to detecting botnets.
- Explore hyperparameter tuning strategies to improve model performance by optimizing key parameters such as decision tree depth, SVM kernel type, and logistic regression regularization strength.

Step by step procedure:

First: read the webpage of CTU-13 dataset:

https://www.stratosphereips.org/datasets-ctu13

1. **Data Preprocessing**:
    - Load and clean the **CTU-13 NetFlow** dataset.
    - Handle missing values and normalize features to improve model performance.
2. **Data Analysis:**
    - Perform exploratory data analysis (EDA) on the CTU-13 NetFlow dataset.
    - Use Pandas API functions such as groupby, value_counts, describe, and correlation matrix to understand the dataset.
    - Visualize key traffic statistics using matplotlib and seaborn, including distribution plots, boxplots, and bar charts. Make sure your figures are well organized with labels in X and Y-axis with clear ticks and save them in pdf format.
    - Identify patterns in botnet and normal traffic to guide feature extraction.

3. **Feature Engineering**:
    - Extract primary features like **Duration, SrcBytes, DstBytes, and Packets**.
    - Derive **new features** to capture network traffic patterns, including:
        - **BytesPerPacket**: Helps detect irregular packet sizes.
        - **FlowRate & PacketRate**: Identifies traffic anomalies, useful for DDoS detection.
        - **ByteRatio & PacketToByteRatio**: Measures the ratio of sent/received data to detect unusual communication.
        - **Log Transformations**: Applies to **Duration** and **Packets** to reduce skewness.
        - **Packet Size Variance**: Detects abnormal variations in traffic flow.
4. **Model Training & Evaluation**:
    - Split dataset into train and test sets.
    - Train **Logistic Regression, Decision Tree, and SVM** models using only train set.
    - Evaluate models on test set with **classification reports (Precision, Recall, F1-score, FPR)** instead of just accuracy.
5. **Feature Importance Analysis**:
    - Analyze which features contribute most to botnet detection.

      Hint: Decision Tree **feature importance ranking** provide insights into key traffic characteristics.
6. **Writing-up:** Write a short Latex-based report (https://www.overleaf.com) of your findings with the following sections:
    - Introduction.
    - Problem statement.
    - Data Analysis.
    - Feature Engineering.
    - Results and Discussion
    - Conclusion