Title: Enhancing Cybersecurity with Machine Learning-Based Intrusion Detection

IST.718.M001.SPRING25.Big Data Analytics

Team Members:

- Nithish Kumar Senthil Kumar
- Subhiksha Murugesan
- Rishi Manohar Manoharan
- Kunal Ahirrao

Objective

The objective of this project is to develop a machine learning model capable of detecting network intrusions using the **CSE-CIC-IDS2018** dataset. The dataset captures real-world network traffic, including both normal and malicious activities, making it a valuable resource for building an Intrusion Detection System (IDS). This project aims to analyze the dataset, extract meaningful features, and apply supervised learning techniques to classify network traffic as either normal or malicious.

Data Set Description

Overview/Description

The **CSE-CIC-IDS2018** dataset was created as a collaboration between the *Communications Security Establishment (CSE)* and the *Canadian Institute for Cybersecurity (CIC)*. It was designed to facilitate the evaluation and improvement of intrusion detection models. The dataset captures network traffic from a corporate environment under both normal conditions and various cyberattack scenarios.

The dataset includes seven types of attacks, including:

- Brute-force attacks (FTP, SSH, MySQL)
- Denial-of-Service (DoS) attacks (Slowloris, GoldenEye, Hulk)
- Distributed Denial-of-Service (DDoS) attacks (LOIC, HOIC)
- Botnet attacks (Zeus, Ares)
- Web application attacks (SQL injection, command injection, XSS)
- Heartbleed attack
- Network infiltration from inside

It contains raw network traffic logs (PCAP files), event logs from machines, and preprocessed feature-extracted CSV files.

Number of Rows and Columns

- The dataset consists of around **10 million** network flow records, split across **10** daily CSV files (close to a million records in each day/file).
- **81 features/columns** extracted from network traffic using CICFlowMeter-V3.

Sample Predictors

Some key predictors in the dataset include:

- **Flow Duration**: The length of a network connection.
- **Total Forward and Backward Packets**: Number of packets sent in both directions.
- **Packet Lengths**: Maximum, minimum, average, and standard deviation of packet lengths.
- **Flow Bytes Per Second**: Byte transfer rate.
- **Protocol**: The protocol used (e.g., TCP, UDP, ICMP).
- **Destination Port**: The port receiving the network traffic.
- **Label**: The classification of network traffic (Normal or Attack type).

Dataset Link

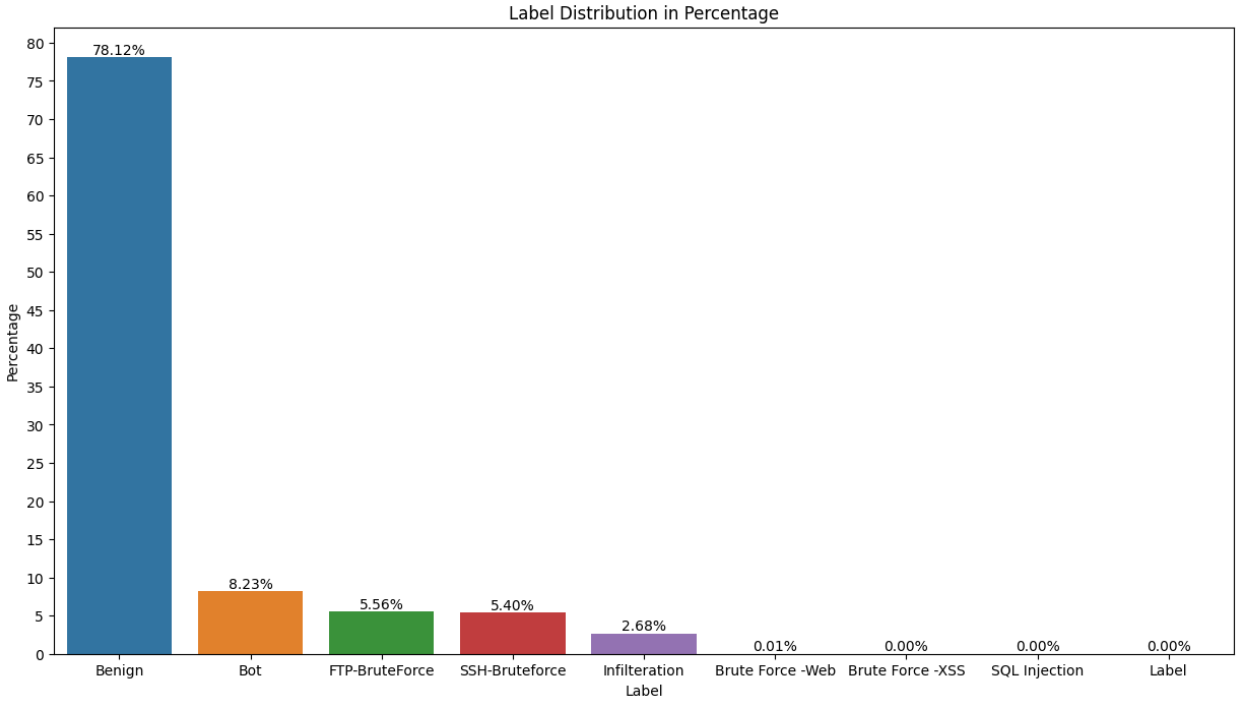
The dataset can be accessed via AWS Open Data: https://www.unb.ca/cic/datasets/ids-2018.html

Interesting or Surprising Aspects

- The dataset includes a wide range of realistic attack scenarios, making it one of the most comprehensive publicly available intrusion detection datasets.
- Some attack types (e.g., Heartbleed) involve exploits of specific vulnerabilities, making them harder to detect using traditional rule-based systems.
- The dataset is highly imbalanced, with benign traffic making up 78.12% of the total data, followed by Botnet (8.23%), FTP-BruteForce (5.56%), and other smaller attack categories.
- Handling class imbalance is a critical challenge and will require techniques such as resampling to improve model performance.

Preliminary Data Exploration

- The dataset is split into **daily CSV files**, requiring concatenation for analysis.
- **Memory optimization techniques** were applied to reduce system load while handling large data files.
- Data cleaning steps included:
  - Dropping unnecessary columns (e.g., timestamps).
  - Handling missing and infinite values.
  - Unifying attack labels based on research mappings.
- **Class distribution visualization** shows a strong imbalance, requiring careful preprocessing.




```python
def drop_infinate_null(df):
    print (df.shape)

    # replace infinity value as null value
    df = df.replace(["Infinity", "infinity"], np.inf)
    df = df.replace([np.inf, -np.inf], np.nan)

    # drop all null values
    df.dropna(inplace=True)

    print (df.shape)

    return df
```

```
## https://www.researchgate.net/figure/Attack-Types-in-CSE-CIC-IDS2018-dataset_tbl1_333894962

mapping= {'SSH-Bruteforce': 'Brute-force',
          'FTP-BruteForce': 'Brute-force',
          ################ Brute-force

          'Brute Force -XSS': 'Web attack',
          'Brute Force -Web': 'Web attack',
          'SQL Injection': 'Web attack',
          ################ Web attack

          'DoS attacks-Hulk': 'DoS attack',
          'DoS attacks-SlowHTTPTest': 'DoS attack',
          'DoS attacks-Slowloris': 'DoS attack',
          'DoS attacks-GoldenEye': 'DoS attack',
          ################ DoS attack

          'DDOS attack-HOIC': 'DDoS attack',
          'DDOS attack-LOIC-UDP': 'DDoS attack',
          'DDoS attacks-LOIC-HTTP': 'DDoS attack',
          ################ DDoS attack

          'Bot': 'Botnet',
          ################ Botnet

          'Infilteration': 'Infilteration',
          ################ Infilteration

          'Benign': 'Benign',
          'Label': 'Benign',
          ################ Infilteration
      }
```

```
1  %%time
2  df_d1 = pd.read_csv("/kaggle/input/ids-intrusion-csv/02-14-2018.csv", low_memory=False)
3  df_d2 = pd.read_csv("/kaggle/input/ids-intrusion-csv/02-15-2018.csv", low_memory=False)
4  df_d3 = pd.read_csv("/kaggle/input/ids-intrusion-csv/02-16-2018.csv", low_memory=False)
5  df_d4 = pd.read_csv("/kaggle/input/ids-intrusion-csv/02-20-2018.csv", low_memory=False)
6  df_d5 = pd.read_csv("/kaggle/input/ids-intrusion-csv/02-21-2018.csv", low_memory=False)
7  df_d6 = pd.read_csv("/kaggle/input/ids-intrusion-csv/02-22-2018.csv", low_memory=False)
8  df_d7 = pd.read_csv("/kaggle/input/ids-intrusion-csv/02-23-2018.csv", low_memory=False)
9  df_d8 = pd.read_csv("/kaggle/input/ids-intrusion-csv/02-28-2018.csv", low_memory=False)
10 df_d9 = pd.read_csv("/kaggle/input/ids-intrusion-csv/03-01-2018.csv", low_memory=False)
11 df_d10 = pd.read_csv("/kaggle/input/ids-intrusion-csv/03-02-2018.csv", low_memory=False)
```

```
CPU times: user 4min 30s, sys: 26.7 s, total: 4min 56s
Wall time: 5min 44s
```

Proposed Data Exploration

- **Feature Analysis**: Identify the most important network features for distinguishing normal and attack traffic.
- **Class Distribution**: Analyze the proportion of attack vs. normal traffic in the dataset.
- **Time-Series Analysis**: Study how different attack types evolve over time.
- **Correlation Analysis**: Check for feature dependencies and redundancies.
- **Protocol and Port Analysis**: Determine which protocols and ports are most commonly associated with attacks.

Proposed Predictions

- **Binary Classification**: Train a model to classify network traffic as either normal or attack.
- **Multi-Class Classification**: Develop a model to classify traffic into specific attack types (Brute-force, DoS, DDoS, Web attacks, etc.).
- **Deep Learning Approach**: Due to the large size and complexity of the dataset, deep learning models will be explored for anomaly detection and classification.
- **Anomaly Detection**: Implement unsupervised learning models to detect unusual patterns in network traffic.