Anomaly Detection in Network Traffic Logs Using Machine Learning

Name: Mridul Roy

Roll: 40018, Reg: H-145

PMICS, Department of Computer Science and Engineering

University of Dhaka

# Chapter 1: Introduction

In today's digital world, cyber-attacks are becoming increasingly common and dangerous.   
Detecting abnormal (anomalous) behavior in network traffic is essential to maintaining a secure digital infrastructure.   
This project presents a machine learning-based solution to detect anomalies in network traffic using a supervised Random Forest classifier trained on synthetically labeled data derived from Isolation Forest.

# Chapter 2: Dataset Description

We used a sample CSV dataset containing network flow data. It included features such as packet size, flow duration, inter-arrival time, and byte/packet rates.  
The dataset had no initial labels, so we first applied unsupervised anomaly detection (Isolation Forest) to generate labels ('normal' and 'anomaly'). These labels were then used for supervised learning.  
Total records: 20056

# Chapter 3: Algorithms Used

1. Isolation Forest (Unsupervised): Used to generate anomaly labels.  
2. Random Forest (Supervised): Trained to predict anomalies based on features.  
3. Evaluation Metrics: Accuracy, F1-score, Confusion Matrix.

# Chapter 4: Implementation and Output

After training the Random Forest model, we used it to predict on the same dataset. The outputs were:  
  
- True Positives (TP): 158 (Correct anomaly detections)  
  
- True Negatives (TN): 18209 (Correct normal detections)  
  
- False Positives (FP): 845 (Normal wrongly flagged as anomaly)  
  
- False Negatives (FN): 844 (Missed anomalies)

# Chapter 5: Mathematical Explanation

Accuracy = (TP + TN) / (TP + TN + FP + FN) = (158 + 18209) / 20056 ≈ 91.58%  
  
Precision = TP / (TP + FP) = 158 / 1003 ≈ 0.1575  
  
Recall = TP / (TP + FN) = 158 / 1002 ≈ 0.1577  
  
F1 Score = 2 \* (Precision \* Recall) / (Precision + Recall) ≈ 0.1576

# Chapter 6: Graphs and Interpretation

The confusion matrix showed model performance in classifying anomalies and normal traffic.  
  
- The top-left cell shows True Negatives (normal correctly predicted)  
  
- The bottom-right shows True Positives (anomalies correctly predicted)  
  
- Other cells reflect misclassifications (FP, FN)  
  
Graph annotations clearly labeled each cell and provided summary metrics like accuracy and F1 Score.

# Chapter 7: Applications and Benefits

This project can be used in real-time network intrusion detection systems (NIDS), firewalls, and endpoint security tools.  
It’s especially useful in:  
  
- Detecting DDoS attacks  
  
- Spotting unauthorized access  
  
- Identifying abnormal user behavior  
  
Future improvement can include using deep learning, real-time streaming data, and adaptive models for better precision.