NETWORK TRAFFIC ANALYSIS DASHBOARD USING MACHINE LEARNING ON UNSW-NB15 DATASET

1. INTRODUCTION

This project focuses on modern network intrusion detection by analyzing network flow data from the publicly available UNSW-NB15 dataset. The main goal was to explore, visualize, and apply machine learning methods to distinguish between normal and attack traffic, as well as to identify anomalies. A user-friendly interactive dashboard was created using Plotly Dash to effectively share the findings.

2. DATASET DESCRIPTION

- Dataset: UNSW-NB15 (training subset)
- Records: Approximately 175,000 flows
- Features: 45 columns including protocol, service, packet counts, durations, statistical rates, attack category (attack_cat), and a binary attack label (label).
- Source: UNSW Cyber Range Lab
- Target:
  - label: 0 = normal, 1 = attack
  - ttack_cat: multiple categories (DoS, Exploits, Fuzzers, etc.)

3. DATA CLEANING & PREPROCESSING

- Attack categories marked as '-' or missing were relabeled as Normal.
- All missing values were filled appropriately (or set to default Normal).
- Selected six numeric features for anomaly detection and supervised learning:
  - dur, sbytes, dbytes, rate, smean, dmean
- An anomaly_flag column was created using an Isolation Forest model to indicate potential anomalies.
- Standard splitting of features and targets was done for supervised classification.

4. EXPLORATORY DATA ANALYSIS (EDA)

EDA was conducted and integrated into the dashboard with these visualizations:

- Normal vs. attack flow counts
- Attack category distribution
- Top protocols by flow count
- Top services
- Histograms of numeric features

- Feature correlation heatmap

This helped identify patterns, such as the prevalence of TCP traffic, and highlighted the most common attack types.

5. ANOMALY DETECTION

An unsupervised Isolation Forest model was trained on the selected numeric features, with contamination set to 2%. It successfully flagged around 2% of records as potential anomalies. A separate bar chart displayed anomalies by attack category, and a scatter plot illustrated the difference in source versus destination bytes for normal and anomalous flows.

6. SUPERVISED MACHINE LEARNING

A Random Forest Classifier was trained to predict the label (normal vs. attack):

- Training/test split: 80/20
- Features used: the same six numeric features
- Model: RandomForestClassifier(random_state=42)
- Evaluation: confusion matrix shown on the dashboard
- Feature importances were also displayed, showing which numeric indicators were most predictive.

7. DASHBOARD FEATURES

The dashboard was developed using Dash and Dash Bootstrap Components. It includes:

- A clean layout displaying KPIs for total records, attacks, and anomalies
- Anomaly detection visualizations
- Supervised ML results (confusion matrix + feature importance)
- A top-10 anomalies data table
- Professional design with consistent colors, legends, and titles

8. RESULTS & FINDINGS

After analyzing and visualizing the dataset, several clear patterns emerged:

- Traffic Composition: Most flows were normal, reflecting a typical network environment with only a small fraction of labeled attacks. However, even a small number of attacks in absolute terms (thousands of records) remain a significant security concern.
- Attack Patterns: Among the attack categories, “Exploits” and “Fuzzers” were particularly frequent. This suggests that the network environment depicted in the dataset was vulnerable to software exploitation attempts and protocol fuzzing.

- Protocol Distribution: TCP overwhelmingly dominated as the primary protocol, which is realistic for modern network traffic. However, within TCP, some services like HTTP and FTP showed a high number of attack flows, indicating the need for closer inspection of common services rather than rare ports.
- Anomaly Detection: The Isolation Forest flagged about 2% of the flows as anomalous. Many of these flagged flows overlapped with labeled attack flows, validating the anomaly detection method. However, some of these anomalies were labeled as normal, suggesting the potential for uncovering hidden or unknown attacks—a crucial finding for security monitoring.
- Feature Importance: The Random Forest model determined that sbytes (source bytes), dbytes, and rate were the strongest predictors of attack traffic. This indicates that unusual traffic volumes and flow rates are key signs of suspicious behavior, aligning with traditional intrusion detection theories.
- Model Evaluation: The confusion matrix showed that the Random Forest did well at distinguishing normal from attack traffic, though some minority attack categories were misclassified. Further tuning or more balanced data might enhance this performance.
- Dashboard Usefulness: The dashboard offered an easy way to switch between EDA, anomaly results, and predictive modeling. It allowed users to quickly grasp network health through a combination of charts, KPIs, and an anomaly table. This makes it simple for a security analyst to explore which services or protocols might be under attack.

Overall, these results highlight that machine learning and interactive visualization can greatly improve monitoring, understanding, and responding to network security events, providing insights that go beyond static log files.

9. CONCLUSION

This project illustrates the entire process from raw network traffic logs to a clean, visual, and interactive dashboard, combining:

- data cleaning
- exploratory data analysis
- anomaly detection
- predictive classification
- professional visualization
- It demonstrates how modern machine learning techniques can be applied to cybersecurity data.
- Possible future improvements include:
- Integrating live data streams
- adding user-defined filters and time selectors
- more advanced classification models

10. REFERENCES

- UNSW-NB15 Dataset: https://research.unsw.edu.au/projects/unsw-nb15-dataset

- scikit-learn documentation
- Plotly Dash documentation