MID -REPORT

Problem Statement: Network Anomaly Detection System

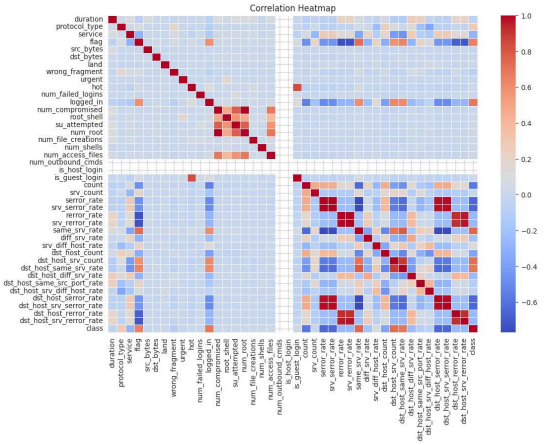
As cyber threats continue to evolve, traditional rule-based security systems struggle to detect novel attacks. A **Network Anomaly Detection System (NADS)** focuses on identifying abnormal network behavior by analyzing traffic patterns and detecting deviations from expected activity. **NADS** employs **machine learning** techniques to distinguish between normal and anomalous network traffic, aiding in the detection of potential cyber threats such as malware infections and unauthorized access attempts.

In this project, we develop an **ML-based mode**l for anomaly detection by training it on network traffic data. Our approach involves feature extraction, model selection, and performance evaluation to differentiate between normal and suspicious activity. This model demonstrates how machine learning can enhance cybersecurity by recognizing unusual patterns, forming the foundation for a more advanced anomaly detection system.

Dataset Overview

This dataset comprises **25,192 instances** and **42 attributes**, designed for **network intrusion detection**. The target variable, **`class`**, labels traffic as **normal** or **anomalous**. It includes:

- **3 Categorical Features** (e.g., `protocol_type`, `service`, `flag`).
- **38 Numerical Features**
- **1 Target label**(`class`)



Key Observations:

- There's a statistically significant difference in duration between the two classes.
- Several numerical features exhibit strong positive or negative correlations (absolute value > 0.7). However, there are many feature pairs that do not have any correlations(NaN).
- The 'protocol_type' column shows 'tcp' as the most frequent protocol. The 'service' column has many unique values, with 'http' being the most common.

This dataset is well-suited for classification and anomaly detection but requires feature selection and preprocessing for optimal model performance.

ML Techniques Used:

1. Random Forest (RF)

- Random Forest is an ensemble learning method that constructs multiple decision trees and combines their outputs to improve prediction accuracy.

- Can handle large and complex datasets with high-dimensional features.

2. Bayesian Methods

- Bayesian methods use probability theory to make predictions based on prior knowledge and observed data.
- Useful for anomaly detection by comparing the likelihood of an event occurring under normal vs. malicious conditions.

3. K-Nearest Neighbors (KNN)

- KNN is an instance-based learning algorithm that classifies data points based on the majority class of their nearest neighbors.
- Simple and effective for detecting anomalies based on similarity to known attack patterns.

4. Support Vector Machine (SVM)

- SVM is a supervised learning algorithm that finds the optimal hyperplane to separate normal and malicious network traffic.
- Provides a robust classification boundary for distinguishing between normal and attack traffic.

5. K-Means Clustering

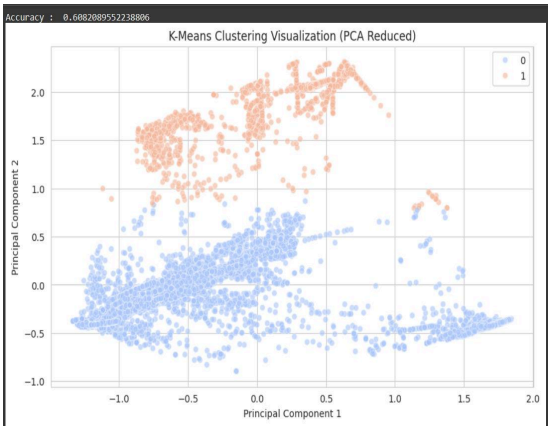
- K-Means is an unsupervised learning algorithm that groups network traffic into clusters based on similarities.

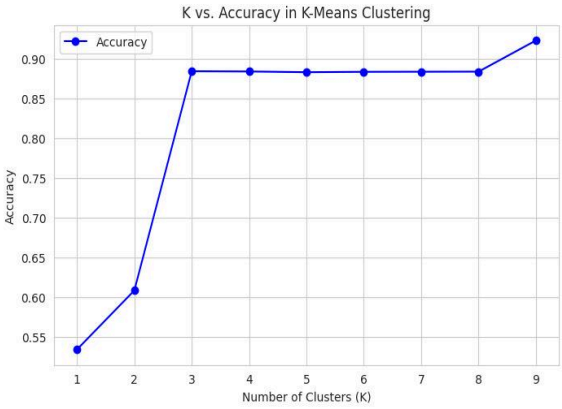
6. Logistic Regression

- Logistic regression is a classification algorithm that predicts whether network traffic is normal (0) or an intrusion (1) based on extracted features.
- It helps understand which network features contribute most to detecting intrusions by analyzing their impact on the probability of an attack.

Implemented Technique: K-Means Clustering

Accuracy achieved - 60.8%, when value of k is 2.
In this dataset, the "class" column represents a binary classification problem (normal vs. attack). Using k=2 aligns with this natural division, allowing K-Means to effectively separate normal traffic from anomalies.While increasing kkk may refine the clusters, it can lead to over-segmentation, making interpretation more complex. k=2 provides a balance, capturing the core distinction in the data while maintaining interpretability.





Link to the data set:https://www.kaggle.com/datasets/sampadab17/network-intrusion-detection

Colab link :
https://colab.research.google.com/drive/1sqtFuSb91WTceVbdQMIgPcVM3io1VHrl

Group Members:

Vadlamudi Jyothsna(B23CS1076)
Nishu Verma(B23CS1045)
Maurya Reshma Rajendra(B23CS1034)
Pradeepika Nori(B23CS1047)
Bhagya Shree(B23CS1007)
Nagma Saj(B23EE1043)