Anomaly Detection in Network Traffic Using Isolation Forest

Project Description:

This project implements anomaly detection in network traffic using the Isolation Forest algorithm. The KDD Cup 1999 dataset is used, and unsupervised learning is applied to detect potential network intrusions.

Objectives:

1. Detect anomalous patterns in network traffic data.
2. Use unsupervised learning for security threat identification.
3. Evaluate model performance using confusion matrix and classification metrics.

Methodology:

1. Load and preprocess the dataset (`corrected.gz`).
2. One-hot encode categorical features (protocol_type, service, flag).
3. Scale features using StandardScaler.
4. Apply Isolation Forest for anomaly detection.
5. Evaluate using confusion matrix and classification report.

Tools and Libraries Used:

Python, Pandas, NumPy, Scikit-learn, Matplotlib, Seaborn

Results:

The model successfully detected anomalies with reasonable accuracy. The confusion matrix and classification report highlighted the model's capability to differentiate between normal and attack traffic.

Conclusion:

Isolation Forests offer a robust way to detect anomalies in network traffic without prior labeled data. Further improvements could include trying other models like Autoencoders and comparing performance.

Resource Link:

Dataset: https://www.kaggle.com/datasets/galaxyh/kdd-cup-1999-data