BUAN 6341 APPLIED MACHINE LEARNING

GROUP PROJECT PROPOSAL

GROUP NO: 4

GROUP MEMBERS: Tara Canugovi, Shriya Reddy Kolan, Vidhya Bharati Kulkarni, Mehul Singh Rajaputhra, Riddhima Reddy Ramasahayam

Target Problem:

Cybersecurity intrusion detection is a critical challenge for organizations of all sizes. Intrusion detection systems (IDS) are essential for identifying malicious activities and protecting sensitive data. This project aims to develop a predictive model to accurately detect network intrusions and determine the key features indicative of cyberattacks.

Motivation:

- Understanding the patterns and characteristics of network intrusions can significantly enhance cybersecurity defenses.
- A data-driven approach can enable proactive identification of threats, minimize potential damage, and improve the effectiveness of security measures.

Data Source:

The dataset used in this project is the **Cybersecurity Intrusion Detection Dataset**, available on Kaggle (https://www.kaggle.com/datasets/dnkumars/cybersecurity-intrusion-detection-dataset).

Dataset Description:

The dataset consists of 9500+ network traffic *records and 10 features* with various features including:

- Network connection features (e.g., duration, protocol type, service, flag)
- Content features (e.g., number of failed logins, number of file creations)
- Traffic features (e.g., source bytes, destination bytes)
- Host-based features (e.g., number of connections to the same host)
- Time-based traffic features (e.g., number of connections in a specific time window)
- Target Variable: Intrusion type (e.g., normal, attack categories)

Project Plan:

1. **Data Preprocessing & Exploration:**
    - Clean and preprocess the dataset (handle missing values, encode categorical variables).
    - Perform Exploratory Data Analysis (EDA) to identify trends, correlations, and outliers in network traffic patterns.
2. **Feature Engineering:**
    - Identify the most significant predictors of network intrusions using correlation analysis and feature importance techniques.
    - Engineer new features that may enhance intrusion detection, such as aggregated traffic statistics or anomaly scores.
3. **Model Development:**
    - Train classification models (Logistic Regression, Random Forest, XGBoost, Neural Networks) to predict network intrusions.
    - Use Neural Networks (DNN, LSTM, CNN) for pattern recognition.
    - If attack labels are missing, anomaly detection can be used: Autoencoders, Isolation Forest, and One-Class SVM can be used to identify deviations from normal network behavior.
    - If certain thresholds are met (e.g., failed_logins > 10 & ip_reputation_score > 0.8), an alert is triggered.
4. **Evaluation & Insights:**
    - Compare model performance using accuracy, precision, recall, and F1-score.
    - Identify actionable insights for security analysts to improve intrusion detection and response strategies, such as key indicators of specific attack types.

Expected Outcome:

- A machine learning model that can accurately detect network intrusions.
- Actionable insights on the key features and patterns that indicate malicious network activity, helping security teams to better defend against cyberattacks.