Title: Network Intrusion Detection Using Machine Learning

Introduction

As cyber threats continue to evolve, the need for robust security measures has become paramount. Traditional rule-based intrusion detection systems often fail to detect new and complex attack patterns. Machine learning (ML) offers an efficient alternative by learning from historical data and identifying potential threats automatically. This project aims to build a Network Intrusion Detection System (NIDS) using ML techniques, specifically Random Forest and XGBoost, to classify network traffic as normal or malicious.

Methodology

Data Preprocessing

- **Dataset**: The model is trained on `Data.csv`, containing network traffic records.
- **Handling Missing Values**: Checked for missing values and applied necessary imputation.
- **Feature Encoding**: Categorical data was transformed using '`LabelEncoder`'.
- **Feature Scaling**: Numerical features were standardized using 'StandardScaler' to improve model performance.
- **Splitting**: The dataset was divided into training and testing sets using an 80-20 split.

Model Training

- **Random Forest Classifier**: A bagging ensemble model that builds multiple decision trees and aggregates their outputs for better accuracy and robustness.
- **XGBoost Classifier**: A gradient boosting algorithm known for its speed and predictive performance, optimizing weak learners iteratively.

Model Evaluation

- **Accuracy**: Measures the proportion of correctly classified instances.

- **Precision**: The ratio of correctly predicted positive observations to total predicted positives (important for reducing false positives).

- **Recall**: The ability of the model to detect all relevant instances (minimizing false negatives).

- **F1-Score**: The harmonic mean of precision and recall, providing a balanced measure.

- **Confusion Matrix**: A tabular representation of correct and incorrect predictions for further analysis.

Results

The models were evaluated on their effectiveness in detecting network intrusions. Key findings include:

- **Random Forest** achieved an accuracy of **98.74%**, with precision and recall values of **98.84%** and **98.80%**, respectively.

- **XGBoost** achieved an accuracy of **98.80%,** with precision and recall values of **99.39%** and **98.35%**, respectively.

The XGBoost model demonstrated superior performance due to its gradient boosting approach, which reduces errors progressively. However, Random Forest remains a strong alternative due to its robustness and ability to handle noisy data.

Conclusion

Machine learning models, especially XGBoost, demonstrate strong potential for network intrusion detection. Future improvements could include additional feature engineering, deep learning models, and real-time deployment for enhanced security monitoring.

Author: Riya Jaiswal Date: March 2025