Cybersecurity Threat Classification Using Machine Learning

1. Introduction

Cybersecurity threats are increasingly sophisticated, making automated detection essential. This project applies machine learning techniques to classify network intrusions using the CICIDS2017 dataset (or a similar dataset). The absence of predefined labels necessitated the use of pseudo-labeling via K-Means clustering.

2. Data Preprocessing

The dataset was preprocessed with the following steps:

- **Encoding Categorical Variables**: Features such as `protocol_type`, `service`, and `flag` were label-encoded.
- **Normalization**: StandardScaler was used to normalize numerical features.
- **Handling Missing Values**: No missing values were found.
- **Pseudo-Labeling**: K-Means clustering was applied to create labels (n_clusters=2).

3. Model Training and Selection

Three machine learning models were trained:

1. **Random Forest Classifier**: An ensemble learning method using 100 decision trees.
2. **Support Vector Machine (SVM)**: A linear kernel was used for classification.
3. **Neural Network**:
    - Architecture: 64 → 32 → 1 neurons with ReLU activations.
    - Optimizer: Adam, Loss: Binary Crossentropy.
    - Trained for 10 epochs.

4. Model Evaluation

Models were evaluated using accuracy, precision, recall, and F1-score.

Model	Accuracy	Precision	Recall	F1-score
Random Forest	**99.91%**	**1.00**	**1.00**	**1.00**
SVM	**99.87%**	**1.00**	**1.00**	**1.00**
Neural Network	**99.84%**	**0.99**	**1.00**	**1.00**

The confusion matrix was used to analyze prediction errors. Random Forest performed the best due to its ensemble learning approach, while SVM was computationally expensive. The neural network showed promise but required more training epochs for optimization.

5. Conclusion & Future Work

This study demonstrated the feasibility of using machine learning for cybersecurity threat classification. Future enhancements include:

- **Feature Engineering**: Identifying the most influential features.
- **Hyperparameter Tuning**: Optimizing model parameters for better accuracy.
- **Advanced Neural Architectures**: Exploring CNNs or RNNs for sequence-based network traffic analysis.

This report provides an overview of the methodology, findings, and future improvements for cybersecurity threat classification using ML.