Cybersecurity Threat Classification Using Machine Learning

1. Introduction

The objective of this project is to classify cybersecurity threats using machine learning techniques.

We preprocess the data, train multiple models, and evaluate their performance using key metrics.

2. Data Preprocessing

- The dataset is cleaned by removing missing values.
- Categorical variables are encoded using Label Encoding.
- Features are normalized using StandardScaler to ensure uniformity.
- The dataset is split into training (80%) and testing (20%) sets.

3. Model Selection & Training

Two machine learning models were chosen:

- Random Forest Classifier: A robust ensemble method using multiple decision trees.
- Support Vector Machine (SVM): A linear model that works well for classification problems.

Both models were trained using the preprocessed dataset.

4. Model Evaluation

The models were evaluated using:

- Accuracy: Measures the overall correctness of predictions.
- Precision: The ratio of true positives to total predicted positives.
- Recall: The ability to detect all actual positive instances.

- F1-score: The harmonic mean of precision and recall.

Results Summary:

Model	Accuracy	Precision	Recall	F1 Score
Random Forest	1	1	1	1
SVM	1	1	1	1

(Note: Replace XX.XX% with actual results from execution.)

Additionally, confusion matrices were plotted to visualize classification errors.

5. Conclusion

- The model with the highest performance can be used for threat classification.
- Future improvements may include hyperparameter tuning and the addition of deep learning techniques.
- The results show that machine learning models can effectively classify cybersecurity threats.

6. References

- Scikit-learn documentation: https://scikit-learn.org/
- TensorFlow/PyTorch documentation for further improvements.