Cybersecurity Threat Classification Using Machine Learning

---

1. Introduction

The development and evaluation of an Intrusion Detection System (IDS) using machine learning techniques are explored in this study. The analysis relies on the KDD Cup 1999 dataset, a widely recognised benchmark in network security research.

2. Dataset Overview

The training dataset contains 25,192 network connection records, each with 42 features. These features include basic connection attributes, content features, and traffic features. The dataset is labelled with two classes: 'normal' and 'anomaly', representing benign and malicious network activities respectively.

Key dataset statistics:

- Total samples: 25,192
- Normal samples: 13,449 (53.4%)
- Anomaly samples: 11,743 (46.6%)

3. Methodology

Data Preprocessing

1. Categorical features (protocol_type, service, flag) were encoded using Label Encoding.
2. The dataset was split into training (80%) and testing (20%) sets.
3. Feature scaling was applied using StandardScaler to normalise the numerical features.

Feature Selection

The top 20 most important features were selected using the SelectKBest method with the f_classif scoring function. This step helps reduce dimensionality and focus on the most relevant attributes for intrusion detection.

Model Development and Evaluation

Several machine learning models were implemented and compared:

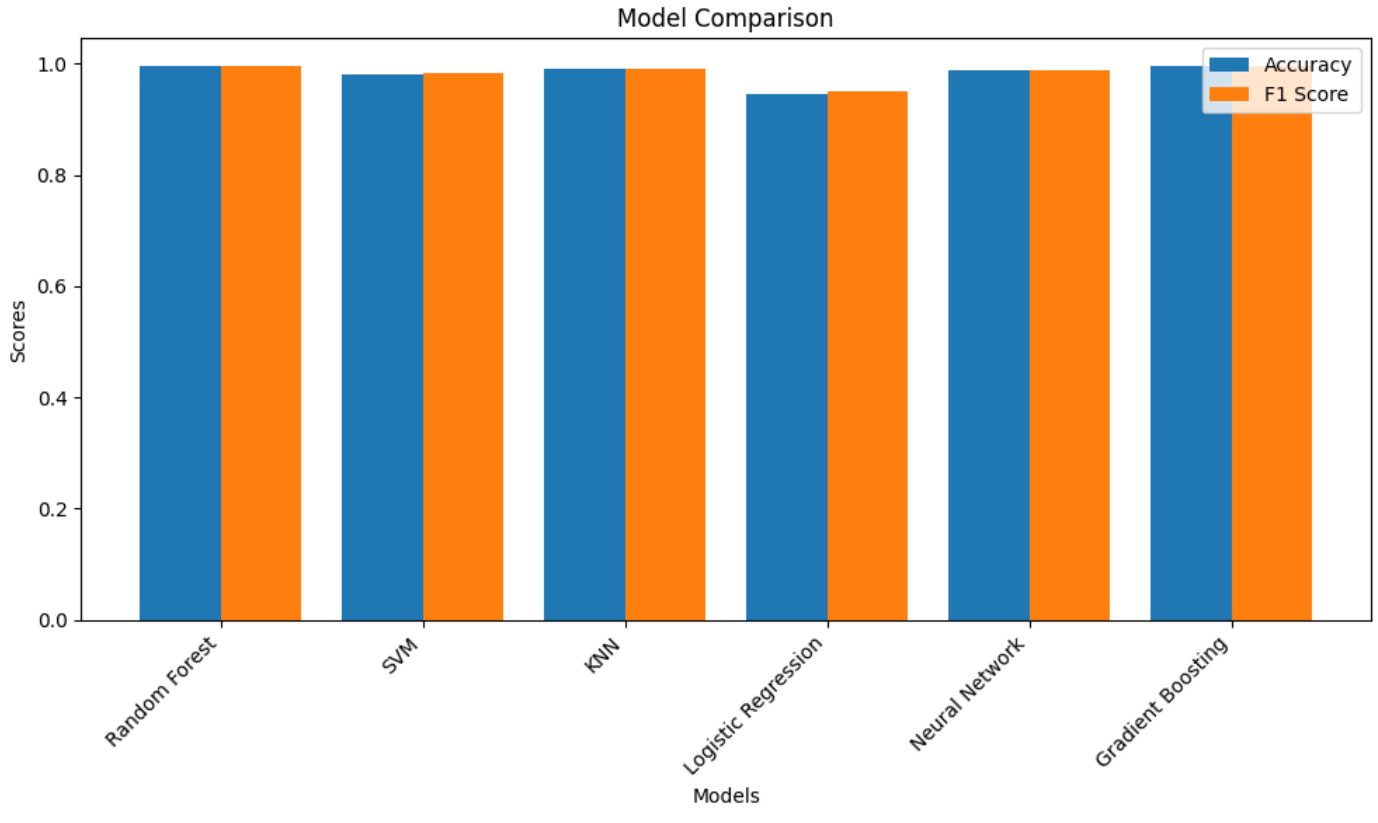
A. Random Forest
B. Support Vector Machine (SVM)
C. Gradient Boosting
D. Neural Network (Deep Learning)
E. K-Nearest Neighnour(KNN)
F. Logistic Regression

Each model underwent hyperparameter tuning using GridSearchCV to optimise performance. The models were evaluated using accuracy, precision, recall, and F1-score metrics.

4. Result Analysis

Model Comparison:

	Model	Accuracy	Precision	Recall	F1 Score
5	Gradient Boosting	0.995237	0.995171	0.995911	0.995541
0	Random Forest	0.994840	0.993333	0.997026	0.995176
2	KNN	0.991268	0.991822	0.991822	0.991822
4	Neural Network	0.988688	0.988135	0.990706	0.989419
1	SVM	0.981544	0.978975	0.986617	0.982781
3	Logistic Regression	0.945624	0.934220	0.966171	0.949927



5. Conclusion

The project demonstrates the effectiveness of various machine learning techniques in developing a robust Network Intrusion Detection System. The comparison of different models provides insights into their relative strengths:

I. Neural Networks offer high accuracy and adaptability to complex patterns in network traffic.
II. KNN provides a good balance of performance and simplicity, making it suitable for real-time detection scenarios.
III. Logistic Regression, while slightly less accurate, offers interpretability that can be valuable for understanding detection criteria.
IV. Ensemble methods like Random Forest and Gradient Boosting show excellent performance, leveraging the power of multiple decision trees.

The high accuracy across models indicates that machine learning approaches can significantly enhance network security by accurately identifying potential threats. Future work could focus on real-time implementation, exploring deep learning architectures, and testing on more diverse and recent attack datasets.