**False Positive Reduction in Intrusion Detection Systems**

Intrusion Detection Systems (IDS) are essential to securing modern network infrastructures. However, high false positive rates challenge their effectiveness, burdening security teams and masking real threats. This project addresses false positive reduction by employing six machine learning classifiers: Support Vector Machine (SVM), Naïve Bayes, Logistic Regression, Decision Tree, Random Forest, and Artificial Neural Networks (ANN), trained on the CICIDS2017 dataset. A comprehensive preprocessing pipeline was used, including normalization, feature selection, and class balancing with SMOTE. Ensemble-based models, particularly Random Forest and ANN, demonstrated the highest accuracy and significantly reduced false positives. Future work will focus on enhancing real-time capabilities and resilience to evolving cyber threats.

**Dataset**

The CICIDS2017 dataset offers comprehensive labeled traffic, including Denial of Service (DoS), brute force, and web intrusions. Its diversity and real-world characteristics make it ideal for supervised learning.

**Preprocessing**

* **Concatenation**: Combined multiple session captures.
* **Cleaning**: Removed missing values, duplicates, and low variance features.
* **Normalization**: Scaled numerical features between [0,1] range.
* **Missing Values**: Median imputation for robustness.
* **Dimensionality Reduction**: Applied Principal Component Analysis (PCA).
* **Balancing**: Synthetic Minority Over-sampling Technique (SMOTE) for class imbalance.

**Model Training**

* **Algorithms Used**: SVM, Naïve Bayes, Logistic Regression, Decision Tree, Random Forest, ANN.
* **Hyperparameter Tuning**: Grid Search with Cross Validation.
* **Evaluation Metrics**: Accuracy, Precision, Recall, F1-Score, Training Time.

| **Model** | **Accuracy** | **Precision** | **Recall** | **F1-Score** | **Training Time (s)** |
| --- | --- | --- | --- | --- | --- |
| SVM | 0.957 | 0.962 | 0.957 | 0.957 | 1787 |
| Naïve Bayes | 0.716 | 0.970 | 0.716 | 0.814 | 3.64 |
| Logistic Regression | 0.958 | 0.960 | 0.958 | 0.957 | 428.22 |
| Decision Tree | 0.995 | 0.993 | 0.995 | 0.994 | 50.80 |
| Random Forest | 0.998 | 0.998 | 0.998 | 0.997 | 909.37 |
| ANN | 0.997 | 0.997 | 0.997 | 0.997 | 10971.13 |

**Key Findings**

* **Random Forest and ANN** models outperformed other classifiers with near-perfect scores.
* **Naïve Bayes** achieved the fastest training but with lower accuracy.
* **Decision Tree** models were accurate but prone to overfitting.
* **Ensemble-based models** are most effective for reducing false positives without sacrificing detection capability.

**Conclusion**

Machine learning techniques, especially ensemble and neural network models, offer promising pathways for reducing false positives in IDS. While traditional classifiers like SVM and Logistic Regression perform reasonably well, advanced models provide better accuracy and resilience.

**Future Work:**

* Develop lightweight models for real-time deployment.
* Enhance model robustness against adversarial attacks.
* Integrate multi-modal data sources (logs, user behavior, threat intelligence).