INTRUSION DETECTION ON CICIDS2017 DATASET

SANNI GANGWANI
T00668956

MACHINE LEARNING
COMP 4980

DECEMBER 2023

Notebook link:
https://colab.research.google.com/drive/17KTpxwVn5B7_h-yTqN2NeA2XQw7IM7tm?usp=sharing

Link to data set:
https://www.kaggle.com/datasets/cicdataset/cicids2017/code

Data set on the drive:
https://drive.google.com/drive/folders/1pMjhllcZAd3H72nDJKfxNa7oOjPkwbBa?usp=sharing

The CICIDS2017 dataset is a comprehensive resource focused on network traffic and intrusion detection. It uniquely combines normal and malicious network traffic, mirroring a typical organizational network. This inclusion of diverse cyber-attacks, such as Brute Force, Heartbleed, Botnets, DDoS, and Web Attacks, renders it exceptionally useful for research in network security, particularly in developing intrusion detection systems.

Data Composition

The dataset is comprised of scalar and categorical data. Scalar data encompasses quantitative metrics like byte counts, durations, and packet lengths. Conversely, categorical data includes qualitative aspects such as protocol types (TCP, UDP, etc.), service types (HTTP, FTP, etc.), and attack categories (normal, DDoS, Brute Force, etc.).

Volume and Handling

Encompassing roughly 300 MBs of data, the CICIDS2017 dataset presents a thorough perspective on network interactions and security threats. However, its size can pose challenges in analysis, particularly for systems with limited processing capabilities. A pragmatic approach might involve focusing on a subset of the data for a more manageable analysis, striking a balance between comprehensiveness and practicality.

Data Diversity and Variables

This dataset boasts a wide array of variables including IP addresses, port numbers, timestamps, packet and byte counts, flow duration, and various flag indicators. These variables, ranging in type and format, offer a rich landscape for analysis.

Format and Accessibility

Commonly available in CSV format, the CICIDS2017 dataset is compatible with numerous data analysis and machine learning tools, making it particularly accessible for Python programs using libraries like Pandas and Scikit-learn. However, its large size necessitates efficient data handling and preprocessing techniques.

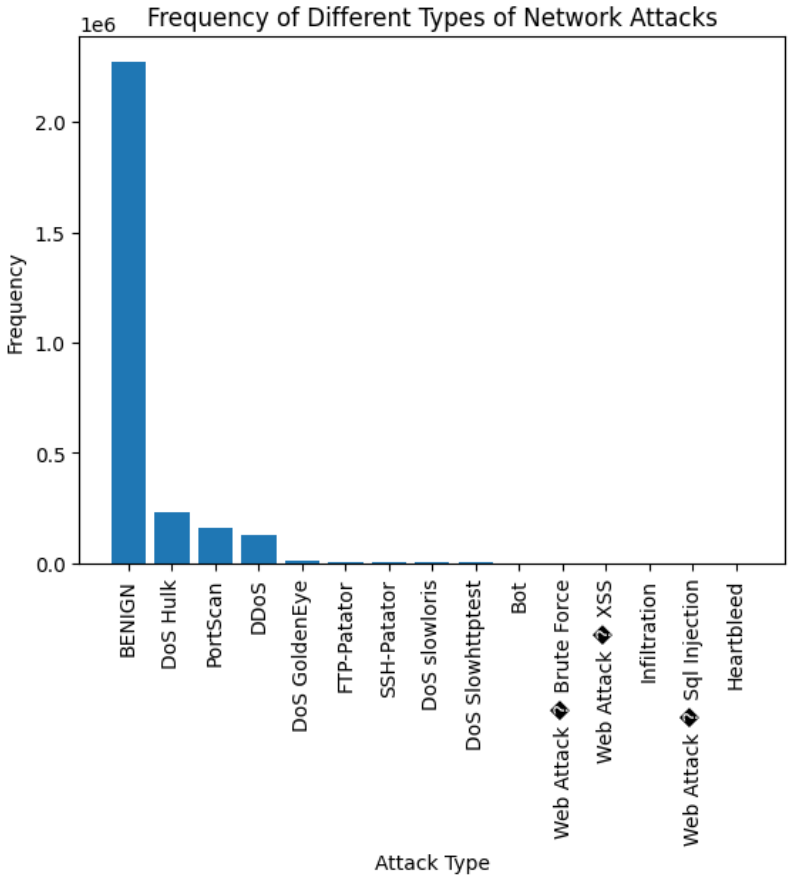
Practical Value

For practitioners and researchers in network security, the dataset is invaluable. It allows for the development and testing of network intrusion detection systems, provides insights into network security dynamics, and serves as an excellent educational resource in cybersecurity.

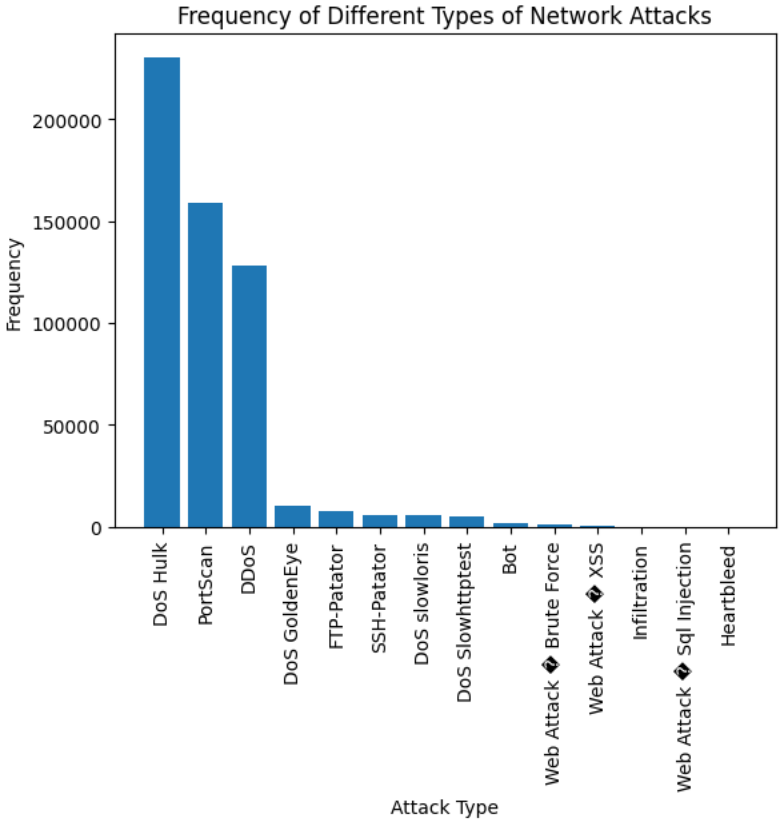
In essence, the CICIDS2017 dataset is a veritable goldmine for those venturing into the cybersecurity domain, particularly for students embarking on this journey. It reflects the complexity and unpredictability of internet traffic, offering hands-on experience with the kind of data cybersecurity professionals encounter daily. Far surpassing textbook knowledge, this dataset offers a practical, real-world learning experience, crucial for anyone aspiring to excel in the field of cybersecurity.

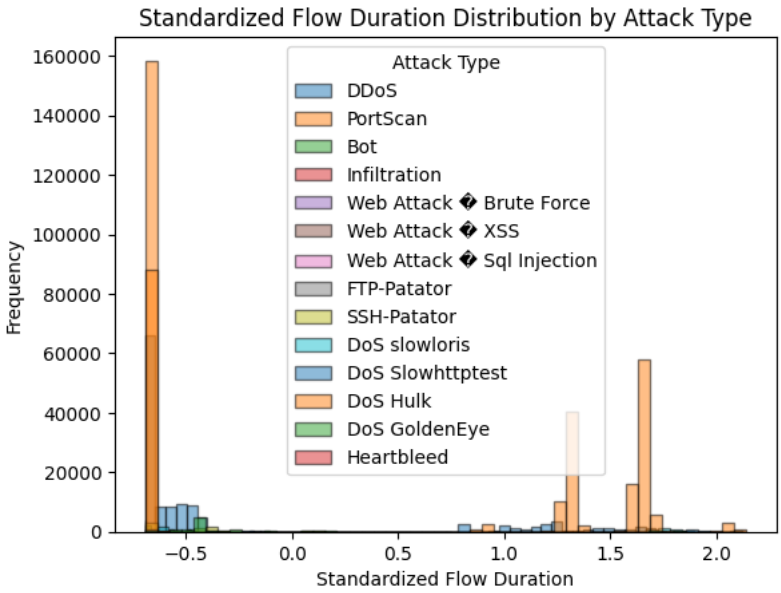
Data Analysis

The dataset being used has 2.8 millions rows and 79 columns. The datatypes are mostly integers and floats the last column(‘ Labels’) has object data type. This shall be one of our mainly focused columns. The dataset includes 15 unique **Label** values, indicating a wide range of traffic types, from benign to various forms of cyber attacks like 'DoS Hulk', 'PortScan', and 'DDoS'. The distribution of these labels is uneven, with 'BENIGN' traffic dominating, followed by various attack types. This imbalance is typical in cybersecurity datasets, reflecting the real-world scenario where normal traffic often exceeds malicious traffic.



The second graph helps us visualize what attacks are more frequent.



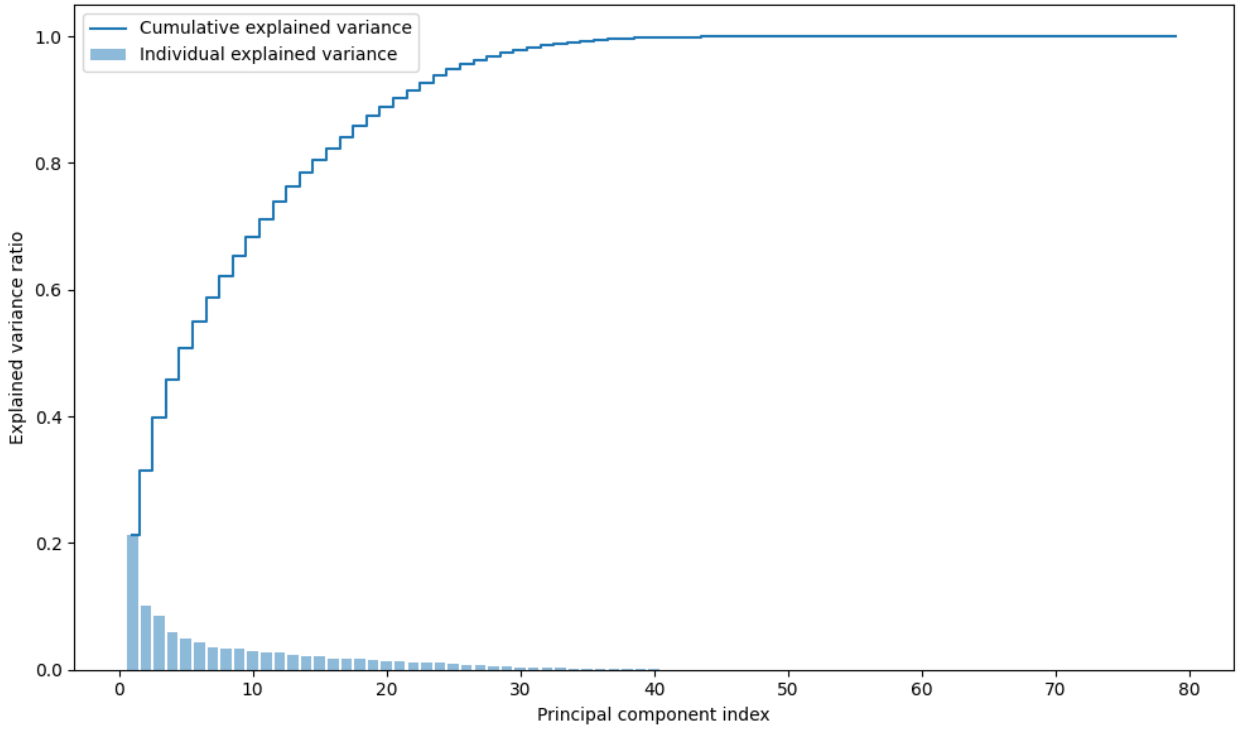


From the above graphs we see that normal traffic(‘BENIGN’) outweighs the malware attacks,'DoS Hulk' appears to be the most frequent attack type in the dataset, followed by 'PortScan' and 'DDoS'. The lower frequency of complex attacks like 'Infiltration', 'Heartbleed', or 'Web Attack SQL Injection' could indicate that they are less common or harder to execute. The

graph shows a clear deviation from the norm for certain types of attacks, particularly 'DoS Hulk' and 'PortScan', which seem to have longer flow durations. We will be using multiple Machine Learning Algorithms such as Random Forest Classifier, Random forest Reggresor, MLP to predict malicious activity i.e using flow duration to indicate malicious activity and predict the label column. Using the classifier algorithms.

DATA EXPLORATION

PCA



the first few principal components are crucial, with the first component alone explaining approximately 21.37% of the variance in your dataset. To explain 96% of the variance, you need to sum up the explained variance ratios until the cumulative variance exceeds this threshold, which is achieved by a specific number of components less than the total number of original features. This indicates that the dataset has intrinsic dimensionality that is lower than the number of original features, suggesting redundancy and potential correlation among them.

DESCISION TREES

machine learning hypothesis

Classify network activities into their respective categories based on features such as flow duration, protocol, and packet counts.

Experiment 1: Random Forest Classifier

- Preprocessing: Feature standardization and splitting the dataset.
- Model: RandomForestClassifier with 8 estimators and max depth of 4.
- Evaluation: Classification report, confusion matrix, and cross-validation.
- Results: High accuracy (96%) but poor performance in some classes (0 precision and recall).
- Refinement: Considered increasing estimators and depth for better classification of underrepresented classes.

Experiment 2: Random Forest Regressor

- Purpose: Predict the 'Flow Duration'.
- Model: RandomForestRegressor with 10 estimators and max depth of 5.
- Evaluation: Mean Squared Error and R-squared.
- Results: High R-squared (0.9998103) but large MSE (215199682269.3452).
- Refinement: Adjusted the number of estimators and depth, considering feature selection specific to flow duration.

Experiment 3: MLP Classifier

- Preprocessing: Standardization, handling inf/-inf, encoding categorical variables.
- Model: MLPClassifier with two hidden layers (10, 5 nodes) and 50 max iterations.
- Evaluation: Classification report and accuracy.
- Results: High accuracy (98.95%) but convergence issues and imbalanced classification.
- Refinement: Increased max iterations, considered different network architectures.

The project involved developing and testing machine learning models to analyze a dataset, with a focus on classification and regression tasks. The final methods chosen were a Random Forest Classifier, a Random Forest Regressor, and a Multi-layer Perceptron (MLP) classifier.

Random Forest Classifier

- **Preprocessing**: Standardization of features, handling of infinite and missing values.
- **Model**: RandomForestClassifier with 8 estimators and a maximum depth of 4.
- **Training & Prediction**: Trained on a split dataset and used to make predictions.
- **Evaluation**: Utilized classification report, confusion matrix, and cross-validation scores.

Random Forest Regressor

- **Target Variable**: 'Flow Duration', indicating potential malicious activity based on duration.
- **Model**: RandomForestRegressor with 10 estimators and a maximum depth of 5.
- **Evaluation**: Mean Squared Error (MSE) and R-squared metrics.

MLP Classifier

- **Preprocessing**: Similar to the Random Forest Classifier, with additional encoding of the target variable.
- **Model**: MLPClassifier with two hidden layers.
- **Evaluation**: Classification report and accuracy score.

Performance and Evaluation Metrics

Random Forest Classifier

- **Accuracy**: Approximately 96% across the dataset.
- **Precision & Recall**: Varied significantly across different classes, with some classes having 0 values in these metrics.
- **Cross-Validation Score**: Average of approximately 93.93%, with a standard deviation of 2.05%.

Random Forest Regressor

- **Mean Squared Error**: 215199682269.35
- **R-squared**: 0.9998103, indicating a good fit for the model.

MLP Classifier

- **Accuracy**: Approximately 98.95%.
- **Precision & Recall**: Similar to the Random Forest Classifier, varied across classes with some having 0 values.
- **Convergence Issues**: Indicated by warnings, suggesting a need for more iterations or tuning.

Interpretation of Results

- **Random Forest Classifier**: Demonstrated high accuracy but faced challenges with class imbalance, as indicated by low precision and recall in some classes.
- **Random Forest Regressor**: Showed an excellent fit for the continuous variable 'Flow Duration', although the high MSE suggests some errors in specific instances.
- **MLP Classifier**: Achieved high accuracy, but like the Random Forest Classifier, it struggled with class imbalance and convergence.