Intrusion Detection Datasets – EDA and Preprocessing Report

Lukas W. Blochmann

December 11, 2025

1 Introduction

This report presents the complete exploratory data analysis (EDA) and preprocessing pipeline for the intrusion detection datasets used by this project:

- CICIDS2017 [1]
- NSL-KDD [2]

Both datasets contain labeled network traffic but differ significantly in structure, size, feature space, and attack taxonomy. A unified preprocessing strategy ensures comparability in subsequent machine learning experiments.

2 CICIDS2017

2.1 Dataset Overview

The CICIDS2017 dataset contains realistic network traffic covering benign traffic and seven attack categories. The merged dataset contains:

- Rows: 2,830,743
- Columns: 79

This dataset is characterized by high dimensionality and severe class imbalance.

2.2 Exploratory Data Analysis

EDA revealed several structural issues:

- 1358 Missing Values in `Flow Bytes/s`
- 308,381 duplicate rows
- Infinite values in `Flow Bytes/s` and `Flow Packets/s`
- Extremely skewed distributions
- Strong class imbalance (BENIGN dominating)

2.3 Data Cleaning

Actions executed on the original data included:

- Drop all duplicate rows
- Convert infinity values to NaN
- Impute all NaN with feature's median value
- Drop non variance features

Actions executed to generate new information on existing data:

- A binary indicator for attack/no-attack was constructed:

$$\text{anomaly\_bool} = \begin{cases} 0 & \text{BENIGN} \\ 1 & \text{otherwise} \end{cases}$$

- Mapped detailed attack labels to more general categories

```
attack_types = {
    'BENIGN': 'BENIGN',
    'DDoS': 'DDoS',
    'DoS␣Hulk': 'DoS',
    'DoS␣GoldenEye': 'DoS',
    'DoS␣slowloris': 'DoS',
    'DoS␣Slowhttptest': 'DoS',
    'PortScan': 'PortScan',
    'FTP–Patator': 'BruteForce',
    'SSH–Patator': 'BruteForce',
    'Bot': 'Bot',
    'Web␣Attack␣?␣Brute␣Force': 'WebAttack',
    'Web␣Attack␣?␣XSS': 'WebAttack',
    'Web␣Attack␣?␣Sql␣Injection': 'WebAttack',
    'Infiltration': 'Infiltration',
    'Heartbleed': 'Heartbleed'
}
```

- Target column `Label` was label encoded to new column `Attack Number`

```
0: BENIGN
1: Bot
2: BruteForce
3: DDoS
4: DoS
5: Heartbleed
6: Infiltration
7: PortScan
8: WebAttack
```

2.4 Preprocessing Pipeline

The following steps were applied:

2.4.1 Standardization

All numerical features were scaled using:

$$z = \frac{x - \mu}{\sigma}.$$

Scaled data was saved to `SCA_processed.csv`.

2.4.2 Incremental PCA

Dimensionality reduction was performed using Incremental PCA:

$$k = \left\lfloor \frac{n_{\text{features}}}{2} \right\rfloor,$$

retaining approximately 99.23% of the variance.
The reduced dataset is stored as `PCA_processed.csv`. Both scaler and PCA models were exported as `.pkl` files to ensure reproducibility.

2.4.3 Balancing

Due to extreme imbalance, a balanced dataset of 15,000 samples was created through controlled undersampling:

$$\text{Attack Type} = \begin{cases} 0 & \text{BENIGN} \\ 1 & \text{attack} \end{cases}$$

Balanced outputs:

- `SCA_balanced.csv`
- `PCA_balanced.csv`

applied on the normally scaled dataset and the PCA transformed dataset.

2.5 Suggestions for Model Training

Binary Classification (Benign vs. Attack)

The most discriminative features are global flow statistics, packet-size distributions, inter-arrival-time metrics and aggregate TCP-flag counts. Key features include: *Flow Duration*, *Flow Bytes/s*, *Flow Packets/s*, *Total Fwd Packets*, *Fwd/Bwd Packet Length Mean/Std*, *Flow IAT Mean/Std*, and *SYN/ACK/PSH Flag Counts*. They capture strong deviations in traffic volume, timing and connection behavior typical for DoS, DDoS, brute-force and botnet attacks.

Multi-Class Classification

In addition to the binary features, fine-grained directional, timing and flag-based metrics become essential for distinguishing attack categories. Important features include: *Fwd/Bwd IAT Max*, *Active/Idle Mean/Std*, and individual *SYN/FIN/RST Flag Counts*. These features separate, for example, high-rate DDoS traffic from low-rate SlowHTTP attacks or scan patterns from infiltration flows.

3 NSL-KDD

3.1 Dataset Overview

The NSL-KDD dataset contains realistic network traffic covering benign traffic and four attack categories. The merged dataset contains:

- Rows: 185,559
- Columns: 44

3.2 Exploratory Data Analysis

EDA revealed the following data insights:

- All multi-class datasets share identical columns.
- Binary datasets differ only by the presence of a simple attack indicator.
- Missing values exist only in `attack_category` ($\approx 4\%$).
- 33,912 duplicate entries
- No infinite values
- Skewed features
- Medium class imbalance regarding the target column `class`

3.3 Data Cleaning

Actions executed on the original data included:

- Drop all duplicate rows
- Drop rows with missing values
- Drop non variance features
- Label encoded categorical features

```
Applied Label-Encoding on 'protocol_type' (3 unique classes)
Applied Label-Encoding on 'service' (70 unique classes)
Applied Label-Encoding on 'flag' (11 unique classes)
Applied Label-Encoding on 'attack_label' (23 unique classes)
```

Actions executed to generate new information on existing data:

- A binary indicator for attack/no-attack was constructed:

$$\text{anomaly\_bool} = \begin{cases} 0 & \text{normal} \\ 1 & \text{otherwise} \end{cases}$$

- Target column `attack_category` was label encoded to the new column `attack_number`:

```
0: dos
1: normal
2: probe
3: r2l
4: u2r
```

3.4 Preprocessing Pipeline

The following steps were applied:

3.4.1 Standardization

All numerical features were scaled using:

$$z = \frac{x - \mu}{\sigma}.$$

Scaled data was saved to $\texttt{SCA}_processed.csv$.

3.4.2 Balancing

A balanced dataset of 15,000 samples was created through controlled undersampling:

$$\texttt{anomaly\_bool} = \begin{cases} 0 & \texttt{normal} \\ 1 & \texttt{attack} \end{cases}$$

Balanced outputs:

- `SCA_balanced.csv`

applied on the scaled dataset.

3.5 Suggestions for Model Training

Binary Classification (Normal vs. Attack)

Binary classification is dominated by basic connection attributes and short-term traffic statistics. The most relevant features are: *duration*, *protocol_type*, *service*, *src_bytes*, *dst_bytes*, and the 2-second window features *count*, *srv_count*, *serror_rate*, *srv_serror_rate*, *same_srv_rate*. These efficiently detect DoS bursts, scanning behavior and abnormal host interaction patterns.

Multi-Class Classification

When distinguishing DoS, Probe, R2L, and U2R, content-level and host-based features become critical. Key features include: *logged_in*, *num_failed_logins*, *hot*, *num_root*, *root_shell*, *dst_host_count*, *dst_host_srv_count*, and *dst_host_serror_rate* They reveal login misuse (R2L), privilege escalation attempts (U2R), scanning diversity (Probe) and high-frequency connection patterns (DoS).

4 Conclusion

Both datasets underwent a complete preprocessing pipeline, resulting in:

1. Cleaned and inspected datasets
2. Standardized numerical feature spaces
3. Dimensionality reduction for CICIDS2017
4. Balanced datasets for fair binary classification
5. Reproducible preprocessing models (scaler, PCA)

These prepared datasets are now suitable for training machine learning models for intrusion detection systems.

References

[1] Canadian Institute for Cybersecurity. CICIDS2017 Dataset. https://www.unb.ca/cic/datasets/ids-2017.html

[2] University of New Brunswick. NSL-KDD Dataset. https://github.com/Jehuty4949/NSL_KDD

Appendix

Used Flags

- SCA_* = scaled
- PCA_* = pca transformed

Figure 1: Flowchart visualizing key actions executed on the dataset **CICIDS2017**

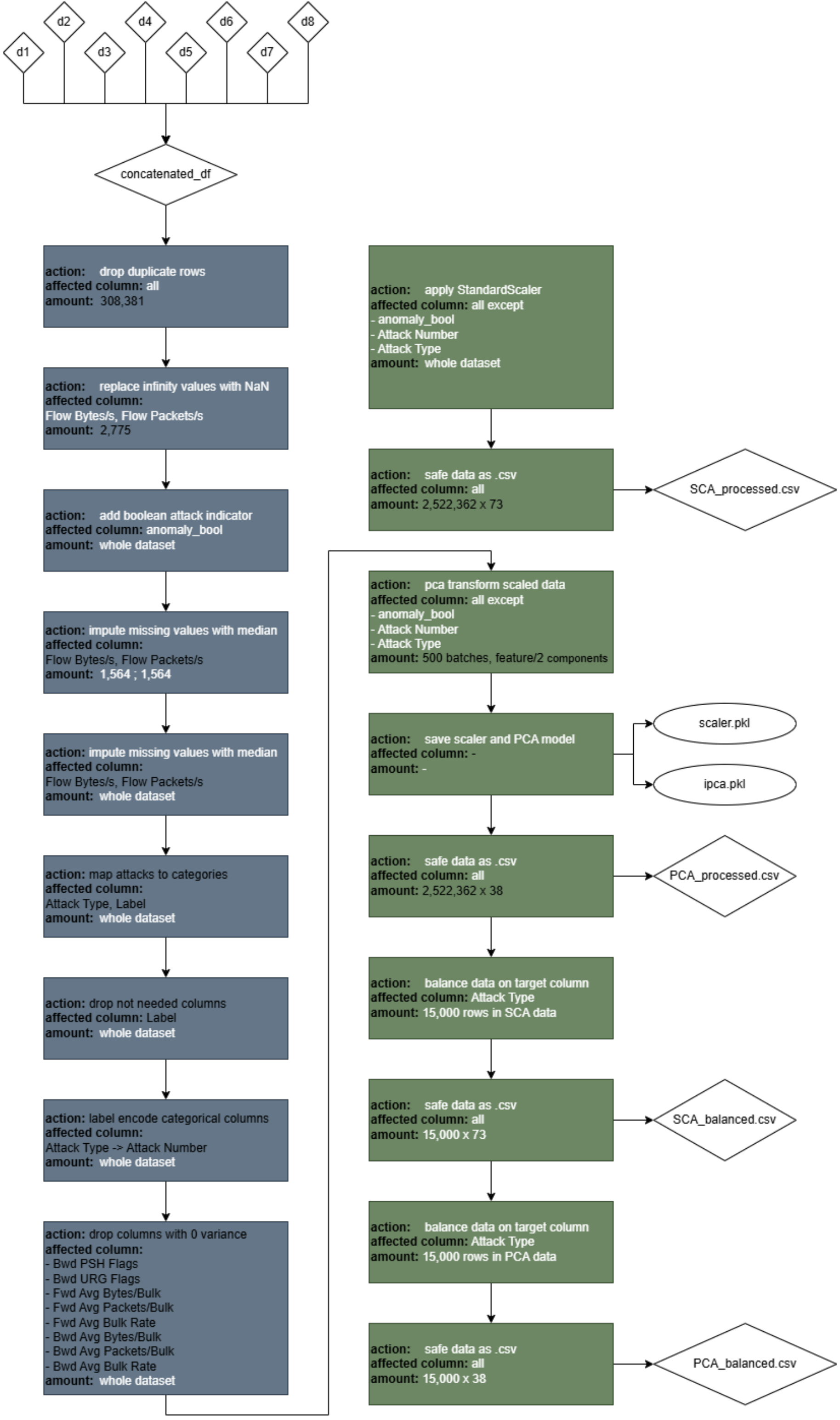
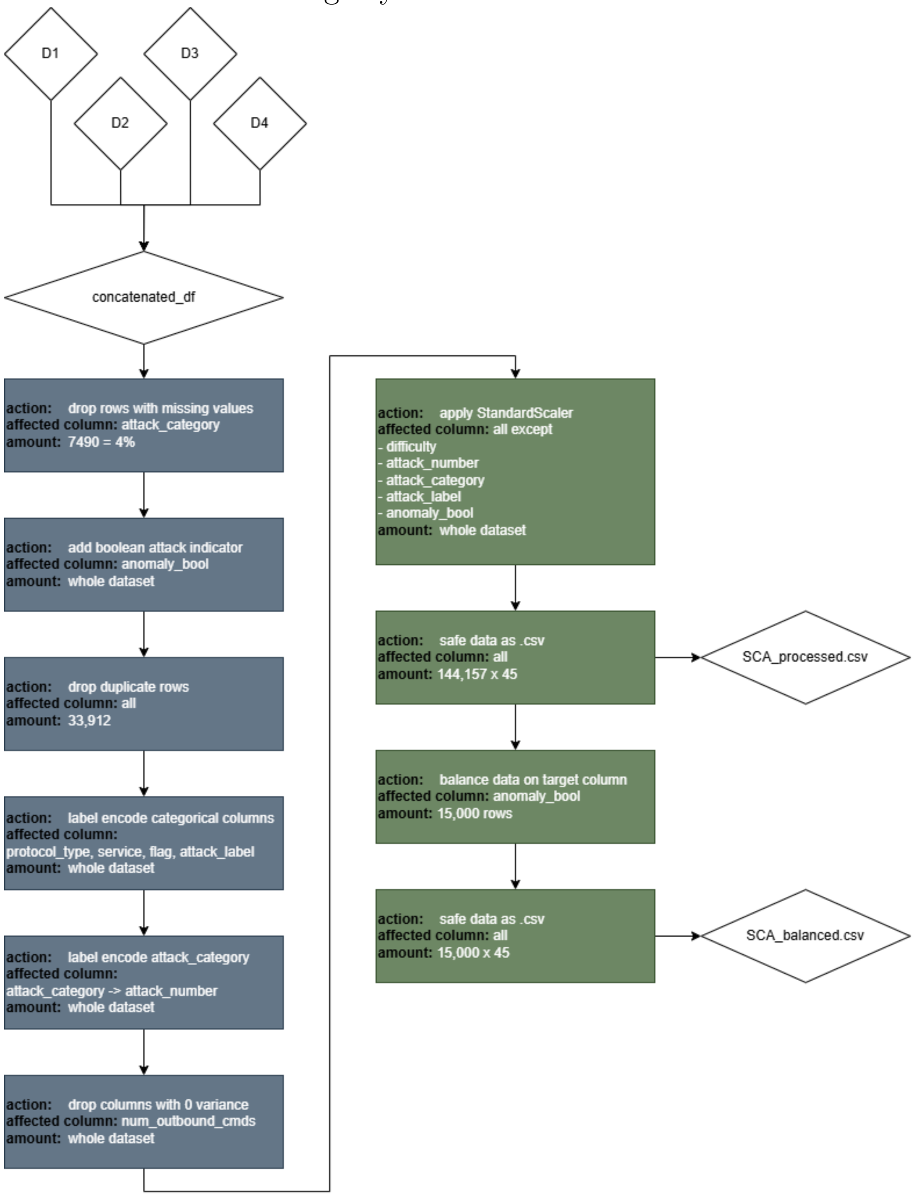

d1
d2
d3
d4
d5
d6
d7
d8
concatenated_df
action: drop duplicate rows
affected column: all
amount: 308,381
action: replace infinity values with NaN
affected column:
Flow Bytes/s, Flow Packets/s
amount: 2,775
action: add boolean attack indicator
affected column: anomaly_bool
amount: whole dataset
action: impute missing values with median
affected column:
Flow Bytes/s, Flow Packets/s
amount: 1,564 ; 1,564
action: impute missing values with median
affected column:
Flow Bytes/s, Flow Packets/s
amount: whole dataset
action: map attacks to categories
affected column:
Attack Type, Label
amount: whole dataset
action: drop not needed columns
affected column: Label
amount: whole dataset
action: label encode categorical columns
affected column:
Attack Type -> Attack Number
amount: whole dataset
action: drop columns with 0 variance
affected column:
- Bwd PSH Flags
- Bwd URG Flags
- Fwd Avg Bytes/Bulk
- Fwd Avg Packets/Bulk
- Fwd Avg Bulk Rate
- Bwd Avg Bytes/Bulk
- Bwd Avg Packets/Bulk
- Bwd Avg Bulk Rate
amount: whole dataset
action: apply StandardScaler
affected column: all except
- anomaly_bool
- Attack Number
- Attack Type
amount: whole dataset
action: safe data as .csv
affected column: all
amount: 2,522,362 x 73
SCA_processed.csv
action: pca transform scaled data
affected column: all except
- anomaly_bool
- Attack Number
- Attack Type
amount: 500 batches, feature/2 components
action: save scaler and PCA model
affected column: -
amount: -
scaler.pkl
ipca.pkl
action: safe data as .csv
affected column: all
amount: 2,522,362 x 38
PCA_processed.csv
action: balance data on target column
affected column: Attack Type
amount: 15,000 rows in SCA data
action: safe data as .csv
affected column: all
amount: 15,000 x 73
SCA_balanced.csv
action: balance data on target column
affected column: Attack Type
amount: 15,000 rows in PCA data
action: safe data as .csv
affected column: all
amount: 15,000 x 38
PCA_balanced.csv

Figure 2: Flowchart visualizing key actions executed on the dataset **NSL-KDD**


D1
D2
D3
D4
concatenated_df
action: drop rows with missing values
affected column: attack_category
amount: 7490 = 4%
action: add boolean attack indicator
affected column: anomaly_bool
amount: whole dataset
action: drop duplicate rows
affected column: all
amount: 33,912
action: label encode categorical columns
affected column:
protocol_type, service, flag, attack_label
amount: whole dataset
action: label encode attack_category
affected column:
attack_category -> attack_number
amount: whole dataset
action: drop columns with 0 variance
affected column: num_outbound_cmds
amount: whole dataset
action: apply StandardScaler
affected column: all except
- difficulty
- attack_number
- attack_category
- attack_label
- anomaly_bool
amount: whole dataset
action: safe data as .csv
affected column: all
amount: 144,157 x 45
SCA_processed.csv
action: balance data on target column
affected column: anomaly_bool
amount: 15,000 rows
action: safe data as .csv
affected column: all
amount: 15,000 x 45
SCA_balanced.csv