

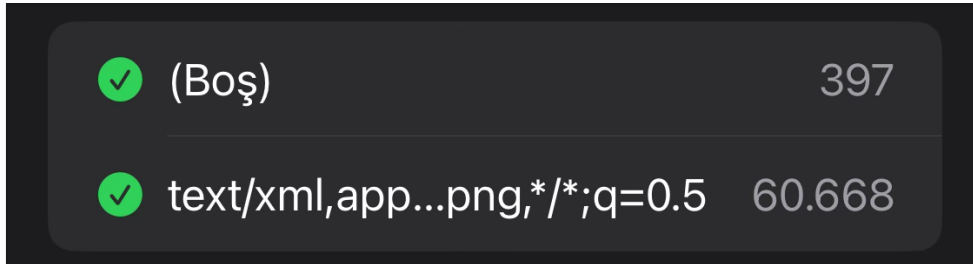
Data Preprocessing Report

CSIC 2010 Web Application Attacks Dataset

Data Cleaning Steps

1. Removal of Duplicate Columns:

- Several columns were found to contain identical values across all rows.
- These columns provided no variability and were removed to reduce redundancy and improve dataset clarity.



A screenshot of a data visualization tool, likely Tableau, showing two rows of data. Each row is preceded by a green checkmark icon. The first row shows '(Boş)' with a value of 397. The second row shows 'text/xml,app...png,*/*;q=0.5' with a value of 60.668.

✓ (Boş)	397
✓ text/xml,app...png,*/*;q=0.5	60.668

2. Column Splitting:

- The URL column, which contained composite data in URLs, was split into multiple features.
- This division aimed to enhance the granularity of the dataset and facilitate more detailed analysis and modeling.

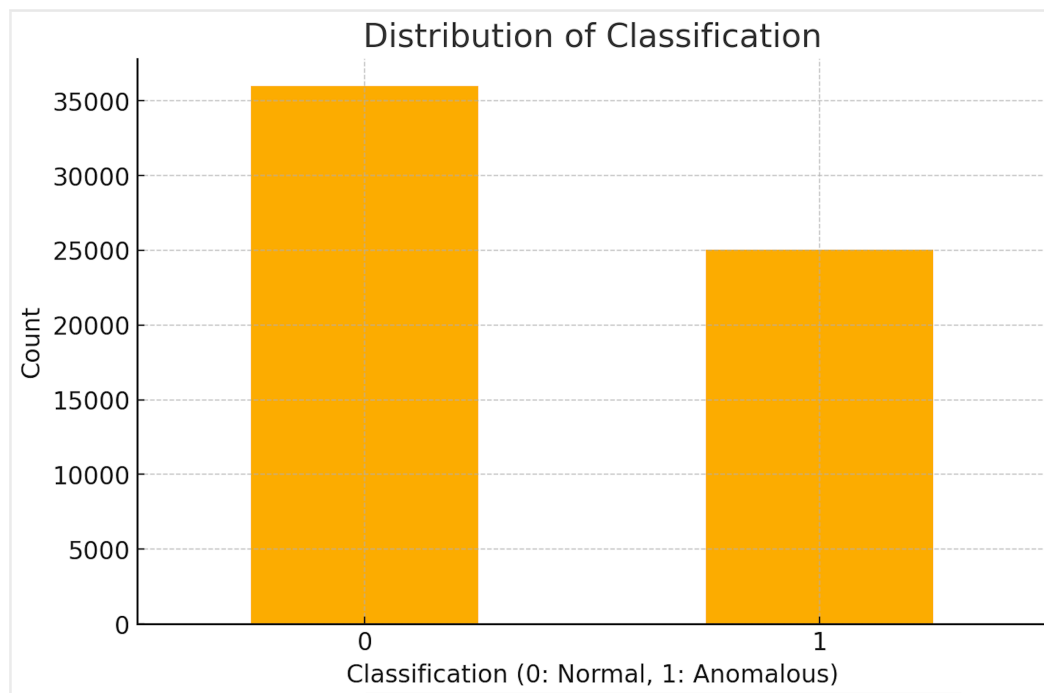
Exploratory Data Analysis (EDA) Findings

1. Potentially Useful Fields:

- Method, Accept, host, cookie, and URL may provide insight into request patterns.
- classification serves as the target variable.

2. Data Quality:

- Many null values in fields like content-type, lenght, and content.
- classification has no missing values and serves as the label for "Normal" or "Anomalous."



The dataset has two classifications:

- **Normal (0):** 36,000 entries
- **Anomalous (1):** 25,065 entries

The target variable distribution is imbalanced but still has a significant number of entries for each class, allowing for meaningful analysis.

Feature Selection and Engineering Rationale

1. Feature Selection:

- Uninformative features (e.g., constant or near-constant columns) were removed to streamline model training.
- Features with high missing values or low relevance were excluded after thorough analysis.

2. Feature Engineering:

- Split composite columns into individual sub-features to enhance interpretability.
- Created new features by combining or transforming existing ones to capture meaningful interactions and trends.

3. Scaling and Encoding:

- Standardized numerical features to ensure compatibility with algorithms sensitive to feature scales.
- Applied one-hot encoding and label encoding to categorical variables based on their cardinality and model requirements.

Future Implications

- The preprocessing steps undertaken aim to ensure a robust foundation for building accurate and interpretable models.
- By cleaning redundant data, enhancing feature granularity, and balancing feature distributions, we anticipate improved model performance and reduced overfitting risks.

Visualizations, along with detailed code and methodologies, are available upon request for further review.