# **Exploratory data Analysis (Harshkumar Dharmendrabhai Patel)**

A graph of different colored bars

Description automatically generated

Classification frequencies are shown by this display of dataset labels. Seaborn.countplot displays labels (e.g., "malware," "attack-pattern") on the x-axis and frequencies on the y-axis. Malware is the most common label, followed by categories like "attack-pattern" and "TIME." "MD5" and "REGISTRYKEY," uncommon labels. This mismatch highlights the need for tactics like resampling to maintain equitable representation of classes in Machine Learning. Rotating labels aid reading.

A computer screen shot of words

Description automatically generated

Using a word cloud, this visualisation shows the clean\_text column of the dataset's most frequent words. Using a single string of preprocessed text data, the approach creates a WordCloud process. Larger words like "file," "attack," "malware," and "used" dominate the dataset. These terms clarify the text's primary themes, like cyberattacks. A word cloud is a simple way to display important terms for exploratory data analysis and text patterns.

A screen shot of a graph

Description automatically generated

The violin plot shows the dataset's label text lengths. The clean\_text column's len() function calculates preprocessed text length. Each label has one violin to represent text length density and dispersion. Wider pieces are more common than thinner ones. Labels like "SHA2" and "malware" exhibit wider distributions, indicating text length variety. This study finds cross-category text structure trends.

A screen shot of a graph

Description automatically generated

The word counts across dataset labels are shown in this boxplot. Splitting clean\_text into words and counting them yields the word\_count column. Each box labels the word count median, interquartile range, and outliers. Labels like "malware" and "attack-pattern" have varying word counts, indicating different text lengths. Outliers have high word counts. Identifying text complexity patterns across labels aids feature engineering and model development.

A screen shot of a computer

Description automatically generated

This pie chart illustrates the top five dataset labels by frequency and percentage. Using the value\_counts() function, the chart displays the most popular labels in pastel hues. The dataset's labels "malware" (26.6%), "location" (20.8%), "SOFTWARE" (18.1%), "attack-pattern" (17.5%), and "identity" (17.0%) dominate. Percentages show "malware" dominates the dataset. To train machine learning models properly, consider label imbalances while making decisions like resampling.

A computer screen shot of a code

Description automatically generated

A group of colorful bars

Description automatically generated with medium confidence

The code snippet above focuses on analyzing the most frequent words associated with specific labels in the dataset. It uses a **bar chart visualization** to present the top 10 most common words for the four most frequent labels: **malware**, **location**, **SOFTWARE**, and **attack-pattern**.

1. **Code Explanation**:
   * **Data Aggregation**: The top labels are identified using the .value\_counts() method, which ranks the labels by their frequency.
   * **Text Extraction**: For each top label, the corresponding clean\_text is aggregated into a single string.
   * **Word Frequency Calculation**: The Counter function is applied to the text, breaking it into individual words and counting the occurrences of each.
   * **Bar Chart Visualization**: Using bar charts and seaborn, the word frequencies are visualised. Every graphic shows word frequency for each label.
2. **Key Insights**:

* Common malware label terms include "rat," "attack," and "file," highlighting remote access tools and file-based malware.
* The location label encourages using terms like phishing, campaign, and email to discuss phishing websites and attack campaigns.
* Terms like "user," "security," and "android" dominate the SOFTWARE label, referring to software vulnerabilities.
* The attack-pattern label stresses email-based attacks and credential theft using terms like phishing and credential theft.

1. **Purpose of Visualization**:
   * Visualisation identifies common terms and interprets labels' context and meaning.
   * Discovering cybersecurity threat linguistic trends aids feature engineering and model interpretation.

Textual patterns are analysed to assist construct robust threat detection models and techniques.

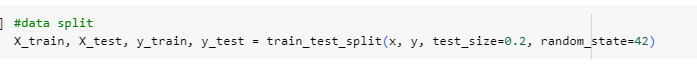
A graph with text in the middle

Description automatically generated with medium confidence

The bar plot shows each dataset label's average text length based on cleaned text characters. The x-axis shows labels (DOMAIN, FILEPATH, REGISTRYKEY) and the y-axis average text length.

Labels like REGISTRYKEY and SHA1 have the longest average text length. Low-verbose labels like DOMAIN and EMAIL have shorter average text lengths. Other labels, like infection and campaign, have intermediate text length descriptions and brief text length. This study shows how labels distribute verbose or succinct text preparation and model training descriptions.

# Train-Test Split - (Harshkumar Dharmendrabhai Patel)



This code divides the dataset into training and testing using train\_test\_split. Features and labels are separated into 80% training data and 20% testing data.