

**SEMESTER ONE 2024/2025 ACADEMIC YEAR**

**SCHOOL COMPUTING AND IMFORMATICS TECHNOLOGY**

**DEPARTMENT OF COMPUTER SCIENCE**

**MASTER OF SCIENCE IN COMPUTER SCIENCE**

**MCS 7103**

**MACHINE LEARNING**

**AMPEIRE EDGAR**

**2024/HD05/21915U**

**2400721915**

**EDA Report on PhiUSIIL Phishing URL Dataset**

**Summary**

The PhiUSIIL Phishing URL dataset contains 235,795 URL records with 54 features. These URLs are categorized into two classes:

1. Phishing URLs
2. Legitimate URLs

The features in the dataset include characteristics such as URL length, domain details. The dataset aims to classify URLs as either phishing or legitimate based on these features

**Dataset Overview**

* **Total Records**: 235,795 URLs
* **Number of Features**: 54
* **Categories**:
  + **Category 1**: Phishing URLs
  + **Category 2**: Legitimate URLs

The dataset captures various aspects of URLs, such as:

* URL length
* Domain-specific details (e.g., top-level domain, subdomains)
* Other characteristics like character continuity, legitimate probability of the top-level domain (TLD), and more.

**Questions Before Data Wrangling**

1. **What data do I need to predict whether a URL is phishing or legitimate?**
   * **Answer**: To predict whether a URL is phishing or legitimate, I need data that includes both phishing URLs and legitimate URLs. The dataset must have clearly labeled examples of both classes to enable supervised learning. Additionally, relevant features such as URL structure, domain patterns, and statistical attributes will help in prediction.

**Data Wrangling**

During the data wrangling process, I checked the dataset for inconsistencies, missing values, and other potential data quality issues. Fortunately, there were no missing values in the dataset.

To better understand the structure and nature of the data, I performed an analysis using the following functions:

* **info()**: This provided a summary of the data types and the presence of null values, confirming no missing values.
* **describe()**: This gave a statistical overview of the features (e.g., mean, standard deviation, min, and max values).

**Class Distribution**: The dataset is well-balanced with an equal or near-equal number of phishing URLs (label 1) and legitimate URLs (label 0). This balance is beneficial for classification tasks, as there is no need for special handling to address class imbalance issues.

**Exploratory Data Analysis (EDA)**

**Univariate Analysis**

To gain initial insights into the dataset, I performed a univariate analysis, focusing on individual feature distributions. I plotted histograms for the continuous features such as URL length, domain length, and other numeric characteristics.

* **URL Length**: Phishing URLs were observed to be generally longer compared to legitimate URLs. This observation aligns with known phishing strategies, where attackers often use longer URLs to obscure malicious content or make the URL appear more legitimate.
* **Domain Structure**: Phishing URLs tend to have longer domains, often with multiple subdomains. These features were also more common in phishing URLs, contributing to a heightened suspicion. Attackers commonly use complex domain structures to imitate legitimate websites or hide the true nature of the domain.

Further exploration of the remaining features would reveal additional patterns and correlations that could assist in differentiating between phishing and legitimate URLs.

**Conclusion**

The PhiUSIIL Phishing URL dataset provides a comprehensive set of features for analyzing and predicting phishing URLs. The data is clean, well-balanced, and offers a range of URL characteristics for building a robust classification model. Initial EDA shows that phishing URLs tend to be longer and have more complex domain structures, which are typical of phishing attempts.