

**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.**

**Dataset Overview:**

The PhiUSIIL Phishing URL dataset contains **235,795 URL records** with **54 features**, and is classified into two categories; - Category 1 is Phishing URLs.

Category 2 is Legitimate URLs.

The features include various characteristics of URLs, such as length, domain details, and specific derived features like CharContinuationRate, TLDLegitimateProb, etc.

**Questions before the data wrangling.**

1. What data do I need to predict whether a URL is a phishing URL or a legitimate URL?

**Answer**

I need data containing both phishing URLs and legitimate URLs

**Data wrangling.**

In data wrangling, I checked my dataset for missing values, fortunately my data had no missing values.

To understand the structure of the dataset, including types of features, statistics, and class distribution,

I used info() and describe(). The dataset had a balanced distribution of phishing (1) and legitimate (0) URLs, making it suitable for classification without major class imbalance handling.

**Exploratory Data Analysis (EDA)**

**Univariate Analysis**

Histograms were plotted for all the continuous features.

Phishing URLs typically had longer lengths compared to legitimate URLs, which is consistent with known phishing tactics where longer URLs may be used to disguise malicious intent.

Similar to URL length, phishing domains were often longer or had multiple subdomains, which contributed to suspicion.