Faculty of Engineering
Cairo University

Big Data and Cloud Computing

#CMP4011

Project Proposal

Team 10

Submitted to:

Eng. Omar Samir

Submitted By:

NAME	SEC	BN	ID
Sarah Mohamed Hossam	1	29	9202618
Abdelrahman Fathy	2	3	9202846
Yasmine Ashraf Ghanem	2	37	9203707
Yasmin Abdullah Nasser	2	38	9203717

Idea: PhiUSIIL Phishing URL (Website)

The dataset is designed for the detection and analysis of phishing URLs. Phishing attacks involve malicious websites pretending to be legitimate with the aim of deceiving individuals into proclaiming personal and sensitive information. It contains various features of URLs, including the lexical characteristics of the web addresses, website content features, and host-based features. This can give us a comprehensive framework for distinguishing between phishing and legitimate websites.

Dataset

- *Link:*

  PhiUSIIL Phishing URL (Website) - UCI Machine Learning Repository

- *Description:*

  The dataset contains the URLs and their corresponding webpages with 54 features including (URL, URLLength, Domain, Domain Length, … etc) and 235,795 instances. Possible classifications for each instance are label 0 which corresponds to a legitimate URL, and label 1 to a phishing URL

Planned approach / Proposed solution

Data Preprocessing: to handle any missing values and normalize any features that need normalization.

Exploratory Data Analysis (EDA): Visualize the dataset to understand the distribution of legitimate vs. phishing URLs , correlation among features, and any patterns that can help in classifying whether the URL is legitimate or not.

Feature Engineering & analytical techniques: Given the high dimensionality of the dataset, we will assess the importance of each feature in predicting fraudulent transactions and consider dimensionality reduction techniques if necessary to improve model performance.
Use more than 1 technique to be able to get useful information to provide answers and approaches for the given problem.
This is achieved by using at least 1 descriptive analysis method (Association) and several predictive ones (Classifiers, etc…)

Planned Approach: To handle the dataset's volume and perform distributed computing, we will explore the implementation of the chosen algorithms using MapReduce in Hadoop or Spark. Different methods will be implemented and tested, such as KNN or a shallow neural network.