**Cyber Security**

Project 3

DESCRIPTION

**Problem Statement:**

Book-My-Show will enable the ads on their website, but they are also very cautious about their user privacy and information who visit their website. Some ads URL could contain a malicious link that can trick any recipient and lead to a malware installation, freezing the system as part of a ransomware attack or revealing sensitive information. Book-My-Show now wants to analyze that whether the particular URL is prone to phishing (malicious) or not.

**Dataset Details:**

The input dataset contains an 11k sample corresponding to the 11k URL. Each sample contains 32 features that give a different and unique description of URL ranging from -1,0,1.

 1: Phishing

 0: Suspicious

 1: Legitimate

The sample could be either legitimate or phishing.

**Project Task: Week 1**

**Exploratory Data Analysis:**

1. Each sample has 32 features ranging from -1,0,1. Explore the data using histogram, heatmaps.
2. Determine the number of samples present in the data, unique elements in all the features.
3. Check if there is any null value in any features.

**Project Task: Week 2**

**Correlation of features and feature selection:**

Next, we have to find if there are any correlated features present in the data. Remove the feature which might be correlated with some threshold.

**Project Task: Week 3 and 4**

**Building Classification Model:**

Finally, build a robust classification system that classifies whether the URL sample is a phishing site or not.

1. Build classification models using a binary classifier to detect malicious or phishing URLs.
2. Illustrate the diagnostic ability of this binary classifier by plotting the ROC curve.
3. Validate the accuracy of data by the K-Fold cross-validation technique.
4. The final output consists of the model, which will give maximum accuracy on the validation dataset with selected attributes.