**Project Report Format**

1. **INTRODUCTION** 
   * Project Overview

A phishing website is a common social engineering method that mimics trustful uniform resource locators (URLs) and webpages. The objective of this project is to train machine learning models and deep neural nets on the dataset created to predict phishing websites. Both phishing and benign URLs of websites are gathered to form a dataset and from them required URL and website content-based features are extracted. The performance level of each model is measures and compared.

* + Purpose

The purpose of this project is to design an intelligent system for detecting phishing websites. Phishing is one of the social attack which aims in stealing sensitive information of the users such as login credentials, credit card numbers etc.

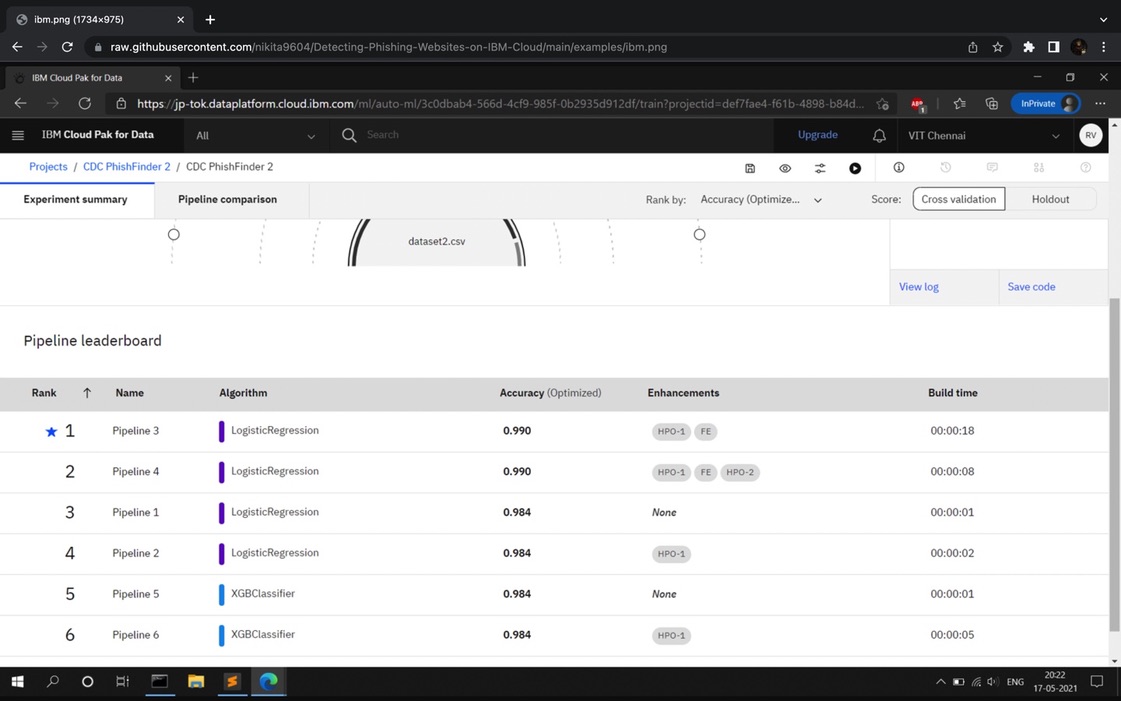
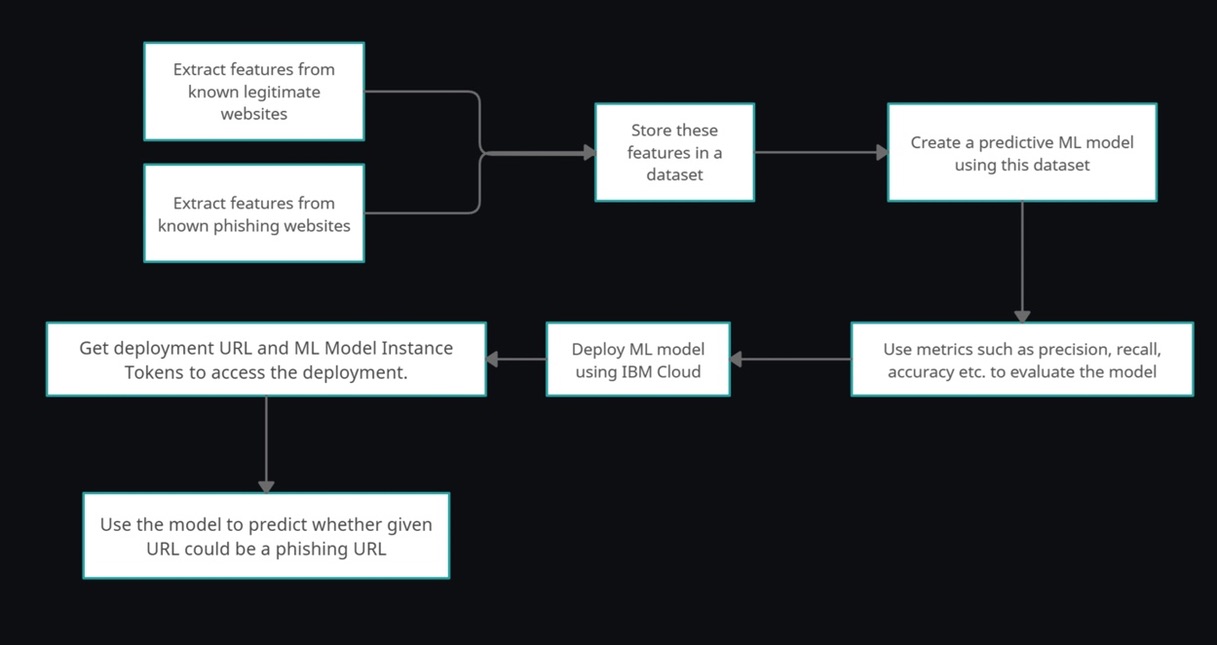
1. **LITERATURE SURVEY**
   * Existing problem
2. **Content-Based Approach:** Analyses text-based content of a page using copyright, null footer links, zero links of the body HTML, links with maximum frequency domains. Using only pure TF-IDF algorithm, 97% of phishing websites can be detected with 6% false positives.
3. **URL Based Approach:** Uses page rank and combines it with other metrics derived from URL based on a priori knowledge. This method can detect upto 97% phishing websites.
4. **Machine Learning Approach:** Uses different machine learning models trained over features like if URL contains @, if it has double slash redirecting, pagerank of the URL, number of external links embedded on the webpage, etc. This approach could get upto 92% true positive rate and 0.4% false positive rate.
   * References

<https://www.researchgate.net/publication/277476345_Phishing_Websites_Features>

* + Problem Statement Definition
  + Internet has dominated the world by dragging half of the world’s population exponentially into the cyber world. With the booming of internet transactions, cybercrimes rapidly increased and with anonymity presented by the internet, Hackers attempt to trap the end-users through various forms such as phishing, SQL injection, malware, man-in-the-middle, domain name system tunnelling, ransomware, web trojan, and so on. Among all these attacks, phishing reports to be the most deceiving attack. Our main aim of this paper is classification of a phishing website with the aid of various machine learning techniques to achieve maximum accuracy and concise model.

1. **IDEATION & PROPOSED SOLUTION**

Detection and prevention of phishing websites endure measure continuously a major space for analysis. There are different types of phishing techniques that offer torrential and essential ways that offer attackers to penetrate the data of people and organizations. Uniform resource locator URLs sometimes are also referred to as “Weblinks” play a vital role in a phishing attack. Uniform resource locator has a vulnerability of redirecting the pages i.e., through the hyperlink; which could redirect to the legitimate website or the phishing site. Different techniques in making phishing sites are emerging day by day. This actually motivated several researchers to put up their concentrate on finding the phishing sites.

1. **REQUIREMENT ANALYSIS**
   * Functional requirement
   * Non-Functional requirements
2. **PROJECT DESIGN**
   * Data Flow DiagramsThe technique comprises of host based, page based and lexical feature extraction of collected websites. The primary step is the collection of phishing and benign websites. In the host-based approach, admiration based and lexical based attributes extractions are performed to form a database of attribute value. This database consists of knowledge mined that uses different machine learning techniques. On evaluating the algorithms, a selective classifier is opted and is implemented in Python
3. **PROJECT PLANNING & SCHEDULING**

## Feature Extraction

The below mentioned category of features are extracted from the URL data:

1. Address Bar based Features  
             In this category 9 features are extracted.
2. Domain based Features  
             In this category 4 features are extracted.
3. HTML & Javascript based Features  
             In this category 4 features are extracted.

The details pertaining to these features are mentioned in the [*URL Feature Extraction.ipynb.*](https://github.com/shreyagopal/Phishing-Website-Detection-by-Machine-Learning-Techniques/blob/master/URL%20Feature%20Extraction.ipynb)

So, all together 17 features are extracted from the 10,000 URL dataset and are stored in '[5.urldata.csv](https://github.com/shreyagopal/Phishing-Website-Detection-by-Machine-Learning-Techniques/blob/master/DataFiles/5.urldata.csv)' file in the DataFiles folder.  
The features are referenced from the <https://archive.ics.uci.edu/ml/datasets/Phishing+Websites>.

## Models & Training

Before stating the ML model training, the data is split into 80-20 i.e., 8000 training samples & 2000 testing samples. From the dataset, it is clear that this is a supervised machine learning task. There are two major types of supervised machine learning problems, called classification and regression.

This data set comes under classification problem, as the input URL is classified as phishing (1) or legitimate (0). The supervised machine learning models (classification) considered to train the dataset in this project are:

* Decision Tree
* Random Forest
* Multilayer Perceptrons
* XGBoost
* Autoencoder Neural Network
* Support Vector Machines

All these models are trained on the dataset and evaluation of the model is done with the test dataset. The elaborate details of the models & its training are mentioned in [Phishing Website Detection\_Models & Training.ipynb](https://github.com/shreyagopal/Phishing-Website-Detection-by-Machine-Learning-Techniques/blob/master/Phishing%20Website%20Detection_Models%20%26%20Training.ipynb)

## Presentation

The short video presentaion for this project is @ <https://youtu.be/I1refTZp-pg>.  
The slide presentaion used in this video is [Phishing Website Detection by Machine Learning Techniques Presentation.pdf](https://github.com/shreyagopal/Phishing-Website-Detection-by-Machine-Learning-Techniques/blob/master/Phishing%20Website%20Detection%20by%20Machine%20Learning%20Techniques%20Presentation.pdf)

## End Results

From the obtained results of the above models, XGBoost Classifier has highest model performance of 86.4%. So the model is saved to the file '[XGBoostClassifier.pickle.dat](https://github.com/shreyagopal/Phishing-Website-Detection-by-Machine-Learning-Techniques/blob/master/XGBoostClassifier.pickle.dat)'

Source Code :https://colab.research.google.com/drive/1ku3nPsNdHOiHn\_KKHp7CzTPxfC2-EHZR?usp=sharing

GitHub : <https://github.com/IBM-EPBL/IBM-Project-26802-1660038070>

BY

SAKTHIPRIYA PANDIYAN (1919106082)

BTECH FINAL YEAR