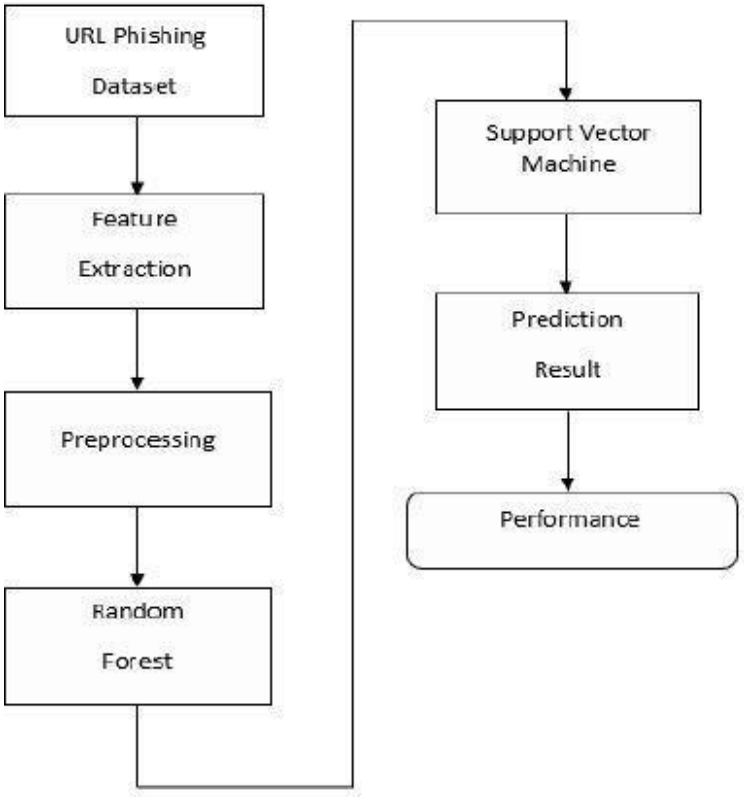
PROJECT DESIGN PHASE-1

SOLUTION ARCHITECTURE

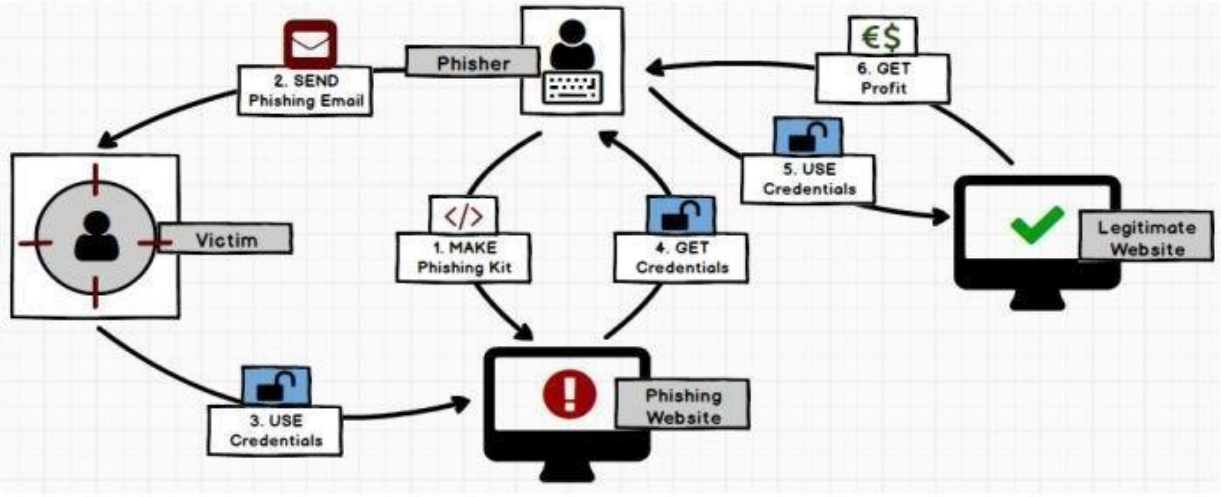
Team ID	**PNT2022TMID10245**
Project Name	**Web Phishing Detection**
Date	**20-10-2022**
Team Members	**1. Dhinesh M** **2. Gunasekaran R** **3. Jefril Angelan R** **4. Gokula Krishna K.S**



PROJECT DESCRIPTION:

A web service is one of the most important Internet communications software services. Using fraudulent methods to get personal information is becoming increasingly widespread these days. However, it makes our lives easier, it leads to numerous security vulnerabilities to the Internet's private structure. Web phishing is just one of the many security risks that web services face. Phishing assaults are usually detected by experienced users however, security is a primary concern for system users who are unaware of such situations. Phishing is the act of portraying malicious web runners as genuine web runners to obtain sensitive information from the end-user. Phishing is currently regarded as one of the most dangerous threats to web security. Vicious Web sites significantly encourage Internet criminal activity and inhibit the growth of Web services. As a result, there has been a tremendous push to build a comprehensive solution to prevent users from accessing such websites. We suggest a literacy-based strategy to categorize Web sites into three categories: benign, spam, and malicious. Our technology merely examines the Uniform Resource Locator (URL) itself, not the content of Web pages. As a result, it removes run-time stillness and the risk of drug users being exposed to cyber surfer-based vulnerabilities. When compared to a blacklisting service, our approach performs better on generality and content since it uses learning techniques.

SOLUTION ARCHITECTURE:



DATASET:

The kaggle dataset includes 11430 URLs with 87 extracted features. The dataset is designed to be used as benchmarks for machine learning-based phishing detection systems. Features are from three different classes: 56 extracted from the structure and syntax of URLs, 24 extracted from the content of their correspondent pages, and 7 are extracted by querying external services. The dataset is balanced, it contains exactly 50% phishing and 50% legitimate URLs.

APPROACH:

This project will be approached using Decision Tree , Random Forest & SVM Algorithms

- A decision tree is a non-parametric supervised learning algorithm, which is utilized for both classification and regression tasks. It has a hierarchical, tree structure, which consists of a root node, branches, internal nodes and leaf nodes.
- Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes or mean prediction of the individual trees.
- SVM is a supervised machine learning algorithm that is commonly used for classification and regression challenges. Common applications of the SVM algorithm are Intrusion Detection System, Handwriting Recognition, Protein Structure Prediction, Detecting Steganography in digital images, etc.

METHODOLOGY:

A. Data Collection

The data for this project is a collection of records. This stage includes choosing a sample of all available information on which to work. Data, especially as the huge quantity of data whereby the target output has been established, is the starting point for machine learning challenges. Data that has been labeled contains information for which we have an answer.

B. Data Pre-Processing

Organize the data we've chosen by formatting, cleaning, and sampling it. Three common data pre-processing steps are:

▪ **Formatting:** That information we've chosen will not be in an easy-to-work-with format. The data could have been in a relational database which we'd like to export to a flat file, or it could have been in a unique file format that you'd like to export to a relational database or a text description.

▪ **Cleaning:** Clearing information includes eliminating and replacing data that isn't present. There could be a situation when data is missing or imperfect, and we don't have all of the information we need to solve the problem. It is indeed likely that all these circumstances have to be removed. Moreover, a few of the characteristics might be sensitive data, which must be cleared or completely removed from the information.

▪ **Sampling:** There could be a lot more well-chosen data accessible than we need. Increased method execution durations and larger computational and storage requirements result from more information. We can choose a shorter sample size of the data sample before reviewing the complete dataset, which will allow us both to explore and develop ideas much faster.

C. Feature Extraction

The following stage is feature extraction, and that's an attribute extension that allows us to create more columns from URLs. Finally, we use a classifier algorithm to train our models. They take advantage of the obtained classified dataset. The remainder of our classified data would be used to validate the models. ML algorithms have been used to identify pre-processed data. That classifier utilized had been Random Forest.

WORKING:

The deployment of a model is a key step in its development. It helps us to figure out which model perfectly describes the data but also how this might perform as in years ahead. We build the trained model and publish it on Static Web Page using IBM Cloud facility. Here, We use Python Flask which is an API of Python that allows us to build up web-applications. Flask’s framework is more explicit than Django’s framework and is also easier to learn because it has less base code to implement a simple web-Application. Database integration is also very simple and easy using Flask**.**