Icon

Description automatically generated with low confidence

**Phishing Domain Detection (Classification [Machine Learning]).**

**High Level Design**

**Project Member: Aman Gupta.**

**Introduction :**

Phishing is the most common method to steal users information (Card details, passwords, and etc.) which can cause a huge damage to the user or group.

**Problem Statement:**

Phishing is a type of fraud in which an attacker impersonates a reputable company or person in order to get sensitive information such as login credentials or account information via email or other communication channels. Phishing is popular among attackers because it is easier to persuade someone to click a malicious link that appears to be authentic than it is to break through a computer's protection measures.

The main goal is to predict whether the domains are real or malicious.

**Approach:**

The classical machine learning tasks like Data Exploration, Data Cleaning, Feature Engineering, Model Building and Model Testing. Has been done on the project. Tried with different machine learning algorithms such as Logistic Regression, Support Vector Classifier, Decision Tree Classifier, Random Forest Classifier, Adaptive Boosting, Gradient Boosted Tree.

**Data-Set:**

This data set consist of 88,647 websites labelled as legitimate or phishing and allow the researchers to train their classification models, build phishing detection systems.

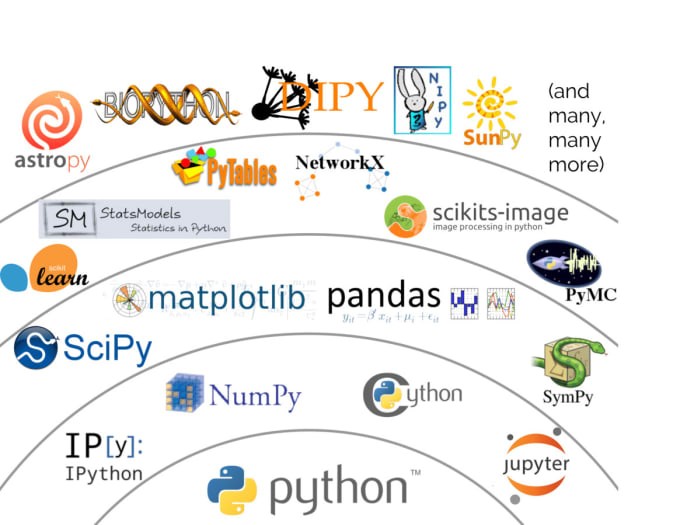
For more details of the dataset visit:

[Datasets for phishing websites detection - ScienceDirect](https://www.sciencedirect.com/science/article/pii/S2352340920313202)

[Datasets for phishing websites detection - ScienceDirect](https://www.sciencedirect.com/science/article/pii/S2352340920313202)

**Tools Used:**

Python Programming language with some packages like NumPy, Pandas, Scikit learn, Pickle, Flask, HTML, CSS, JS



**Design Flow:-**

Model Evaluation

Prediction

Saving the Models

Models

Testing

Training

Clusters

Dataset

**Conclusion :**

It turns out model is performing well with a recall score of 78% in the cluster one and in the cluster two it is giving recall of 93% but there are some False Positives which will effect.