FUNCTIONAL FEATURES OF WEB PHISHING DETECTION

ALGORITHM:

Step 1: : Import the Dataset.

Step 2: Read the Dataset.

Step 3: Extract the data from the dataset for processing

Step 4: Make predictions for the test dataset.

Step 5: Applying Machine Learning algorithms to the dataset

Step 6: Predict the best and worst accuracy algorithms from ML algorithms.

MODULES :

- Data Collection
- Data preprocessing
- Feature Extraction
- Evaluation model

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 Pre-Processing :

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

D. Evaluation Model

The evaluation 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. To prevent overfitting, two very different methods require a test carried out to analyze the accuracy of the model. In evaluating the efficiency of each classification model, the median efficiency is employed. The final product will take the form that has been imagined. Information during classification is represented using graphical representations. Accuracy is measured by the proportion of predictions made using the testing dataset. It's easy to calculate by dividing the total number of forecasts even by correctly predicted guesses. We calculate accuracy as the difference between actual and

expected output calculate accuracy as

Where TP = True Positives
TN = True Negatives
FN = False Negatives
FP = False Positives

The factors for discovery and bracket of phishing websites are as follows

Html and Java script based features:

1.Redirect This number of cases a webpage is being rerouted seems to be the distinguishing factor between phishing and legitimate websites.

2. Right-Click disablement

JavaScript is used by phishers to block the right-click on a feature, preventing customers from accessing and purchasing website programs is written. This function is used at the same time that Using on Mouseover to Cover the Link is handled.

3. Making use of Pop-Up Window

It is really rare as for come across a malicious website that asks visitors to provide private details through a pop-up window.

4. Redirection of the IFrame

An IFrame is a type of HTML tag that allows you as for embed another website inside the one you're now viewing.

Domain-Based Features

1. Age of Domain

The WHOIS network may be adjusted to extract this function. The bulk of such phishing websites is only operational for a small period. We can observe from the data also that the qualifying area must be at least 6 months old.

2. DNS Record

In the case of phishing websites, either the declared identity cannot be verified using the WHOIS database, or there are no facts to back up the claim. The website is rated as Phishing if a DNS document is blank sometimes no longer
available; otherwise, this is categorized under Valid.

3. Web Traffic

The function calculates the number of visitors and the pages they visit to determine the overall reputation of a site.

4. Page Rank

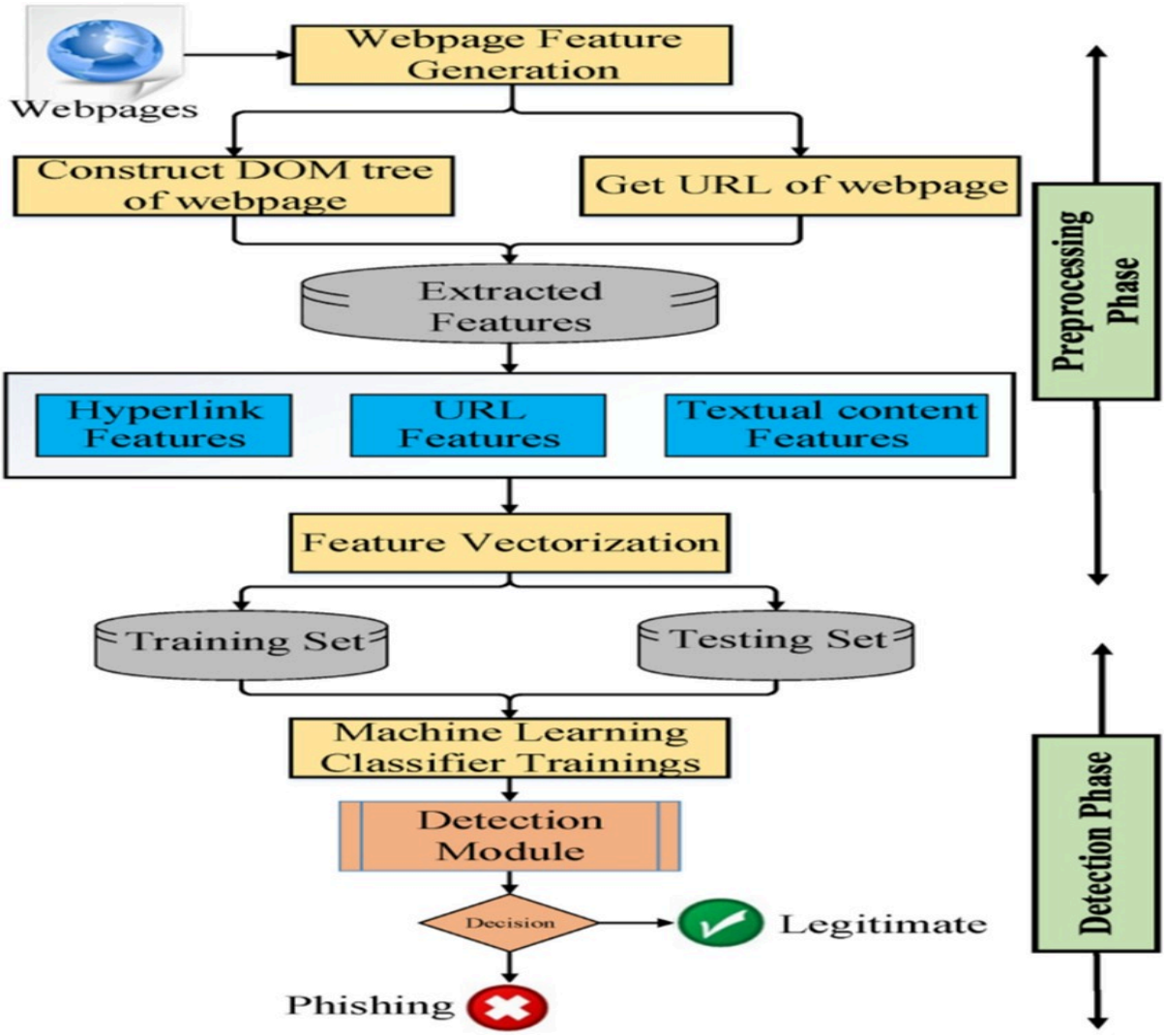
PageRank is a market price that goes between 0 and 1. PageRank seeks to determine a website's popularity on the Internet.

5. Google Index

The function determines not whether a website should be indexed by Google. Whenever a domain is registered with Google, it shows up on the list.

FUNCTIONAL WORKFLOW :

We propose a phishing detection approach, which extracts efficient features from the URL and HTML of the given webpage without relying on third-party services. Thus, it can be adaptable at the client side and specify better privacy. An efficient solution for phishing detection that extracts the features from website's URL and H



TML source code proposed.

Result :

Machine learning methods were imported using the Scikit-learn library. Each classification is performed using a training set, and the performance of the classifiers is evaluated using a testing set. The accuracy score of classifiers was calculated to assess their performance.

Feature HTML and JavaScript Based Abnormal Based Features