explain the flow of the project

Sure, here's a step-by-step explanation of the project flow:

Data Collection:

Gather a dataset containing URLs labeled as either phishing or legitimate. This dataset will be used for training and testing the machine learning models.

Feature Extraction:

Use the inputScript.py to extract features from each URL in the dataset. This script includes functions to analyze various aspects of a URL, such as:

Whether the URL contains an IP address (url\_having\_ip function).

The length of the URL (url\_length function).

Other characteristics like presence of '@' symbol, number of subdomains, etc.

Data Preprocessing:

Clean and preprocess the extracted features to prepare them for model training. This may involve:

Handling missing values.

Encoding categorical variables.

Normalizing numerical features.

Model Training:

Train different machine learning models using the preprocessed dataset. Common models used for this task include:

Logistic Regression

Random Forest

Support Vector Machine (SVM)

Each model will learn to classify URLs based on the extracted features.

Model Evaluation:

Evaluate the performance of each trained model using metrics such as:

Accuracy

Precision

Recall

F1-score

This helps in determining which model performs best on the dataset.

Model Selection:

Select the best-performing model based on the evaluation metrics. This model will be used for predicting phishing websites.

Deployment:

Deploy the selected model to a production environment where it can be used to classify new URLs in real-time. This may involve:

Creating a web service or API.

Integrating the model into an existing system.

Prediction:

Use the deployed model to predict whether new URLs are phishing or legitimate. The model will analyze the features of the URL and provide a classification result.

Monitoring and Maintenance:

Continuously monitor the performance of the deployed model and update it as needed. This may involve:

Retraining the model with new data to improve its accuracy.

Adapting to new phishing techniques

The key packages used in this project include:

regex: For advanced regular expression operations.

tldextract: For extracting the top-level domain (TLD) from URLs.

ssl: For handling SSL certificates.

socket: For network connections.

BeautifulSoup: For parsing HTML and XML documents.

urllib.request: For opening and reading URLs.

whois: For querying WHOIS information about domains.

datetime: For manipulating dates and times.

urllib.parse: For parsing URLs.

requests: For making HTTP requests.

re: For regular expression operations.

Additionally, for machine learning and data processing, the following packages are commonly used:

pandas: For data manipulation and analysis.

numpy: For numerical operations.

scikit-learn: For machine learning algorithms and tools.

matplotlib: For plotting and visualization.

+-------------------+

| Start |

+-------------------+

|

v

+-------------------+

| Data Collection |

+-------------------+

|

v

+-------------------+

| Feature Extraction|

+-------------------+

|

v

+-------------------+

| Data Preprocessing|

+-------------------+

|

v

+-------------------+

| Model Training |

+-------------------+

|

v

+-------------------+

| Model Evaluation |

+-------------------+

|

v

+-------------------+

| Model Selection |

+-------------------+

|

v

+-------------------+

| Deployment |

+-------------------+

|

v

+-------------------+

| Prediction |

+-------------------+

|

v

+-------------------+

|Monitoring & Maint.|

+-------------------+

|

v

+-------------------+

| End |

+-------------------+

ABOUT DATASETS:

The phishing.csv file contains a list of domains, which are likely used for training or testing a machine learning model to detect phishing websites. Each entry in the file represents a domain that may be classified as either legitimate or malicious.

Summary of the Datasets Folder

phishing.csv:

Content: A list of domains, which may include both phishing and legitimate sites.

Purpose: This dataset is likely used to train machine learning models to identify phishing websites based on their domain names.

phishcoop.arff:

This file is in ARFF (Attribute-Relation File Format), which is commonly used with Weka, a machine learning software. It likely contains similar data to phishing.csv but may include additional attributes or features relevant for machine learning tasks.

phishcoop.csv:

Similar to phishing.csv, this file may contain a different set of phishing-related data, possibly with additional features or a different format.

TITLE MEAN

Malicious: Refers to something that is harmful, intended to damage, or cause unauthorized access. In the context of the web, it often involves activities like hacking, phishing, or spreading malware.

Web: This refers to the World Wide Web, which is the collection of websites, web pages, and online content that can be accessed through the internet using browsers.

Deception: In cybersecurity, deception refers to methods used to mislead attackers, trick them into revealing their strategies, or deterring them from performing malicious actions.

Analysis: The process of examining something in detail in order to understand it better or detect any issues or anomalies. In this context, it refers to examining web traffic, user behavior, or website activities to detect signs of malicious activity.

# Random forest:

<https://youtu.be/5ks4WFmjL20?si=xoDvtcsUTXfLDsHK>

<https://youtu.be/fxE4nlU4UjI?si=v5dUFA1YeojW52SK>

decision tree

<https://youtu.be/WTjK0OUr9us?si=yN4YxsbEVIb393EN> 6min

Svm algorithm

<https://youtu.be/ERcEAkAISP8?si=B17hX4Q5uArNRqpe>

<https://youtu.be/cCpCjPI7Bnc?si=v4dMYVUrF5UPHtnW>

logistric regression

<https://youtu.be/g4WTafYlp5U?si=X47xh2NVI3ukZQRh>

dataset

convert html

flask

website

1. http://amaz0n-security-verify.tk/login
2. http://paypa1-account-verify.xyz/secure
3. http://g00gle-drive.gq/document
4. http://apple-id-confirm.top/verify
5. http://facebook-security-check.ml/login
6. http://netf1ix-account.xyz/payment
7. http://bank0famerica.tk/secure
8. http://micros0ft-365.gq/login