**Tools Used**

**Python Libraries**

1) socket: This module is used to create and work with socket objects which allow for network communications. In my code, Using this to get the IP address of the domain.

2) re: The re module is used for working with regular expressions. used it for pattern matching in URLs.

3) datetime: This module provides classes for manipulating dates and times. using it to calculate the age of the domain.

4) urllib.parse: This module is used for breaking URLs into components or combining URL components into a URL string. Using this for parsing URLs into different components.

5) concurrent.futures: This module provides a high-level interface for asynchronously executing callables.using ThreadPoolExecutor to process URLs in parallel.

**External Libraries**

1) Pandas: Pandas, a powerful data manipulation library in Python, is used for loading and managing the phishing and legitimate URL datasets.

2) NumPy: This library is used for numerical operations, although it’s imported here, it’s not explicitly used in the provided script.

3) Scikit-learn: This is a machine learning library used for data splitting, preprocessing (scaling), model training, prediction, and evaluation.

4) Joblib: This is a library for saving and loading Python objects efficiently. Here, it’s used to save and reload the scalar object.

5) requests: This is a popular library used for making HTTP requests. Using it to retrieve the content of websites and check for redirects.

6) beautifulsoup4: This library is used for parsing HTML and XML documents and navigating, searching, and modifying the parse tree. Using it to parse the website’s HTML content.

7) whois: This library is used to extract domain information. Using it to get

information about domain registration.

8) tldextract: This library accurately separates the TLD from the registered domain and subdomains of a URL. Using it to count the number of subdomains.

**Techniques**

1) Regular expressions: These are used extensively to search for patterns in URLs.

2) Web scraping: Using the requests and BeautifulSoup libraries to scrape websites and extract information.

3) Parallel processing: Using the ThreadPoolExecutor to process multiple URLs

Simultaneously.

4) Logistic Regression: This is a machine learning algorithm used for binary classification problems. Here, it’s used to predict whether a given URL is a phishing URL (”bad”) or not.

5) Train-Test Split: This is a method used to evaluate the performance of a machine learning model. The dataset is split into a training set (used for training the model) and a test set (used for testing the model).

6) Standard Scaler: This is a preprocessing technique used to standardize the features by removing the mean and scaling to unit variance.

7) Accuracy, Precision, Recall, F1 Score: These are evaluation metrics used to measure the performance of the classification model. Accuracy is the proportion of correct predictions, precision is the proportion of true positive predictions among all positive predictions, recall (sensitivity) is the proportion of true positive predictions among all actual positives, and F1 score is the harmonic mean of precision and recall.