# **URL-BASED PHISHING DETECTION**

Detailed Description of the Features

1. **`having\_IP`**: Checks for the presence of an IP address in the URL. Phishers may use IP addresses instead of domain names to deceive users. If an IP address is used, the value is set to -1 (phishing), otherwise, it's 1 (legitimate).

2. **`having\_IP\_Address`**: Binary feature derived from `having\_IP`, indicates the presence (1) or absence (-1) of an IP address in the URL.

3. **`URL\_Length`**: Measures the length of the URL. Short URLs are considered more legitimate, so if the length is less than 54 characters, the value is 1 (legitimate), otherwise, 0 (phishing).

4. **`Shortining\_Service`**: Detects the use of URL shortening services. Phishers might use these services to hide the true destination. If a shortening service is detected, the value is -1 (phishing), otherwise, 1 (legitimate).

5. **`having\_At\_Symbol**`: Checks for the presence of '@' symbol in the URL. Phishers might use it to hide the real address, making it -1 (phishing) if present, 1 (legitimate) otherwise.

6. **`double\_slash\_redirecting**`: Detects multiple "//" in the URL, which could indicate redirection. If found, the value is -1 (phishing), otherwise, 1 (legitimate).

7. `**Prefix\_Suffix**`: Checks for the presence of '-' in the domain part of the URL. Phishers may add prefixes/suffixes to mimic legitimate websites, so -1 (phishing) if found, 1 (legitimate) otherwise.

8. `**having\_Sub\_Domain**`: Binary feature indicating the presence (1) or absence (-1) of subdomains in the URL.

9. `**SSLfinal\_State**`: Checks for "https" in the domain part of the URL. If found, the value is 1 (legitimate), otherwise, -1 (phishing).

10. `**Domain\_registeration\_length**`: Measures the age of the domain. If the domain age is more than 6 months, the value is 1 (legitimate), otherwise, -1 (phishing).

11. `**Favicon**`: Detects the presence (1) or absence (-1) of a favicon in the URL.

12. `**port**`: Detects if a port is specified in the URL. If yes, the value is -1 (phishing), otherwise, 1 (legitimate).

13. `**HTTPS\_token**`: Checks for "https" in the URL. If present, the value is 1 (legitimate), otherwise, -1 (phishing).

14. `**Request\_URL**`: Binary feature indicating the presence (1) or absence (-1) of a request URL.

15. `**URL\_of\_Anchor**`: Binary feature indicating the presence (1) or absence (-1) of URLs in the anchor tags.

16. `**Links\_in\_tags**`: Binary feature indicating the presence (1) or absence (-1) of domain links in anchor tags.

17. `**SFH**`: Checks if the website uses a Server Form Handler. If yes, the value is -1 (phishing), otherwise, 1 (legitimate).

18. `**Submitting\_to\_email**`: Checks if any form submits to an email. If yes, the value is -1 (phishing), otherwise, 1 (legitimate).

19. `**Abnormal\_URL**`: Detects if the URL contains excessive hyphens or non-standard characters. If found, the value is -1 (phishing), otherwise, 1 (legitimate).

20. `**Redirect**`: Checks the number of forwardings in the URL. If it exceeds 2, the value is 0 (phishing), otherwise, 1 (legitimate).

21. `**on\_mouseover**`: Checks if there is an event "onMouseOver" in the webpage source code. If yes, the value is -1 (phishing), otherwise, 1 (legitimate).

22. `**RightClick**`: Checks the status of the right-click attribute. If it's disabled, the value is -1 (phishing), otherwise, 1 (legitimate).

23. `**popUpWidnow**`: Detects the use of pop-up windows. If found, the value is -1 (phishing), otherwise, 1 (legitimate).

24. `**Iframe**`: Detects the use of iframes. If found, the value is -1 (phishing), otherwise, 1 (legitimate).

25. `**age\_of\_domain**`: Measures the age of the domain. If the domain is older than 6 months, the value is 1 (legitimate), otherwise, -1 (phishing).

26. `**DNSRecord**`: Checks the existence of DNS records. If none are found, the value is -1 (phishing), otherwise, 1 (legitimate).

27. `**web\_traffic**`: Measures web traffic of the URL using Alexa rank. If the rank is below 100,000, the value is 0 (phishing), otherwise, 1 (legitimate).

28. `**Page\_Rank**`: Uses PageRank algorithm to measure URL importance. If the PageRank is greater than or equal to a threshold, the value is 1 (legitimate), otherwise, -1 (phishing).

29. `**Google\_Index**`: Checks if the URL is indexed by Google. If yes, the value

is 1 (legitimate), otherwise, -1 (phishing).

30. `**Links\_pointing\_to\_page**`: Measures the number of links pointing to the page. If there are many such links, the value is 1 (legitimate), otherwise, -1 (phishing).

Machine Learning Models Used

**1. Decision Tree:**

- Decision Tree is a popular classification algorithm that works by recursively splitting the dataset into subsets based on features.

- It creates a tree-like structure where each node represents a feature, each branch a decision, and each leaf node a class label (phishing or legitimate).

- Decision Trees are easy to interpret, handle categorical features well, but they can overfit the training data.

**2. XG Boost:**

- XG Boost stands for Extreme Gradient Boosting, which is an advanced and efficient implementation of the Gradient Boosting algorithm.

- It builds multiple weak learners (typically decision trees) sequentially, where each tree corrects the errors of the previous one.

- XG Boost is computationally efficient, handles missing data, and often provides better performance than traditional Gradient Boosting.

**3. SVM Classifier:**

- SVM stands for Support Vector Machine, a powerful supervised learning algorithm used for classification and regression tasks.

- SVM finds a hyperplane that best divides the dataset into classes. In the case of a binary classifier, it finds the optimal line that separates the two classes.

- SVM is effective in high-dimensional spaces, memory-efficient, and versatile due to its various kernel functions.

**4. Logistic Regression:**

- Logistic Regression is a simple and widely used classification algorithm for binary and multiclass problems.

- It models the probability of a certain class using the logistic function, which maps the output to the range [0, 1].

- Logistic Regression is easy to implement, interpret, and works well for linearly separable data.

**5. KNN (K-Nearest Neighbors):**

- KNN is a non-parametric and instance-based classification algorithm.

- It assigns a class label to an unclassified instance based on the majority class of its k nearest neighbors in the feature space.

- KNN is simple, lazy (no explicit training phase), but can be computationally expensive for large datasets.

**6. Random Forest:**

- Random Forest is an ensemble learning method that creates multiple decision trees during training and combines their predictions.

- Each tree in the forest is built using a random subset of the data and random subset of features, making the model robust and reducing overfitting.

- Random Forest is known for its accuracy, ability to handle high-dimensional data, and resistance to overfitting.

**7. Naive Bayes:**

- Naive Bayes is a probabilistic classification algorithm based on Bayes' theorem and the assumption of feature independence.

- Despite its simplicity, Naive Bayes can perform well in many real-world situations, especially for text classification tasks.

- It is computationally efficient and requires a relatively small amount of training data.

These machine-learning models can be trained and evaluated using the feature-extracted dataset. Once the model is trained and its performance is satisfactory, it can be integrated into the Flask web application to provide real-time phishing detection for users. Keep in mind that the choice of the best model depends on the characteristics of the dataset and the specific requirements of the application. Regular model evaluation and updates with new data are essential to maintain the effectiveness of the phishing detection system.

**Dataset used** <https://drive.google.com/file/d/1VDSGbK78fOyeyThIN0pN_O6pmNo3UnSJ/view?usp=sharing>

**Table depicting the results of each machine learning model**

|  |  |
| --- | --- |
| Machine Learning Models used | Accuracy Attained |
| Decision Tree | 96.29 |
| XG Boost | 53.05 |
| SVM Classifier | 94.04 |
| Logistic Regression | 91.67 |
| KNN | 94.34 |
| Random Forest | 96.92 |
| Naïve Bayes | 61.51 |