# **PHISHING SITES PREDICTION USING MACHINE LEARNING**

# **FINAL REPORT**

# **GROUP 6**

# **Chiru Deepika Yalamanchili (811300581) Raghunath Male (8112572266) Suresh Babu Nakkalapalli (811292428)**

**Introduction:**

The project "Phishing Sites Prediction Using Machine Learning" aims to employ machine learning techniques to detect and stop phishing assaults on websites. It involves the extraction of relevant features, collection of datasets comprising both phishing and legitimate websites, and training machine learning models. The project's main goal is to use machine learning techniques to increase the prediction accuracy of phishing websites.

Phishing remains a prevalent cybersecurity threat, making early detection crucial for user safety. Through the development and implementation of machine learning models, this project achieves a high level of accuracy in distinguishing between legitimate and phishing URLs. The integration of web scraping, link analysis, and visualization further enhances the understanding of patterns within the data. The results showcase the project's effectiveness in contributing to the identification and mitigation of phishing risks, ultimately bolstering online security.

**Originality:**

This project is unique because it makes finding phishing websites much faster. While people might take hours to check one website, the new machine learning method can check thousands of websites in the same time. This makes the detection process much quicker and more efficient, using the power of machine learning to overcome the challenges and time limits of manual detection methods.

Machine learning models, including Logistic Regression and Naive Bayes, are trained on labeled datasets to learn patterns indicative of phishing behavior. These models are then capable of accurately categorizing unseen URLs as either legitimate or potential threats.

Our project uses machine learning to analyse a lot of data in a smart way, going beyond simple rule-based systems. The algorithms can understand tricky patterns in phishing attacks and get better over time. This is important because cyber threats keep changing and getting more complicated.

**Data Preparation:**

Our dataset consists of two columns: “URL” and “Label.”

**Link:-** <https://www.kaggle.com/datasets/taruntiwarihp/phishing-site-urls>

1. **URL Column:**
   * This column contains 5,49,347 URLs.
   * Each row represents a unique web address (URL) with classification as either "good" or "bad" (potentially indicating phishing).
   * There are around 3,80,000 good URLs and 1,60,000 bad URLs.
2. **Label Column:**
   * This column contains corresponding labels for each URL.
   * There are two distinct labels: "good" and "bad”.
3. **Good** - which means the URLs do not contain malicious stuff and this site is not a Phishing Site.
4. **Bad** - which means the URLs contains malicious stuff and this site is a Phishing Site

A screenshot of a computer

Description automatically generated

* These data consist of a collection of legitimate, as well as phishing website instances. Each website is represented by the set of features that denote whether the website is legitimate or not. Data can serve as input for the machine learning process.
* Machine learning and data mining researchers can benefit from these datasets, while also computer security researchers and practitioners. Computer security enthusiasts can find these datasets interesting for building firewalls, intelligent ad blockers, and malware detection systems.
* This dataset can help researchers and practitioners easily build classification models in systems preventing phishing attacks since the presented datasets feature the attributes which can be easily extracted.
* Finally, the provided datasets could also be used as a performance benchmark for developing state-of-the-art machine learning methods for the task of phishing websites classification.

A graph with blue bars

Description automatically generated

* **Data Preprocessing:**
  + Loaded the dataset using Pandas. Explored and visualized the dataset.
  + Applied tokenization and stemming to preprocess the URL text using NLTK.
  + Utilized regular expressions (‘RegexpTokenizer’) to tokenize words in URLs.
  + Applied stemming using the ‘Snowball Stemmer’ for reducing words to their base form.
  + Joined the tokenized and stemmed words back into a text representation.
  + The preprocessing step is crucial for converting raw URLs into a format suitable for machine learning.

A screenshot of a computer

Description automatically generated

**Exploratory Data Analysis (EDA)**

**Web Scraping and Network Analysis:**

* Utilized Selenium and Beautiful Soup for web scraping to collect links from specified URLs.
* Created a network graph using Network X to visualize relationships between URLs.
* Create a network graph to visualize the links between different URLs.

A close-up of a computer code

Description automatically generatedA diagram of a network graph

Description automatically generated

**Model Training and Evaluation:**

* Trained machine learning models, including Logistic Regression and Multinomial Naive Bayes, on preprocessed data.
* Evaluated model performance using accuracy scores, confusion matrices, and classification reports.
* Visualized model accuracies using Matplotlib and Seaborn.

**Model Deployment and Prediction:**

* + Saved and loaded a pre-trained model using Pickle.
  + Tested the model on a test set.
  + Implemented a function to classify new URLs based on specific patterns.

A close up of text

Description automatically generated

**Predictions and Visualization:**

* Demonstrated model predictions on example URLs.
* Visualized model accuracies through bar plots.

A screenshot of a graph

Description automatically generated

**Feature Selection and Engineering:**

In the realm of cybersecurity, the escalating menace of phishing assaults poses a formidable challenge to both individuals and enterprises. Phishing websites, camouflaged as legitimate entities, orchestrate duplicitous schemes to coax users into divulging sensitive information, precipitating financial losses and security breaches. Conventional techniques for identifying phishing sites often falter in the face of the dynamic stratagems employed by cybercriminals. To confront this predicament, the project undertakes the task of leveraging machine learning for the prediction and classification of phishing URLs. Through the training of machine learning models such as Logistic Regression and Naive Bayes on meticulously labeled datasets, the project endeavors to discern discernible patterns indicative of phishing behaviors. Consequently, these models attain the capability to accurately segregate unseen URLs into either legitimate or potentially treacherous categories.

Simultaneously, the project harnesses the power of feature engineering to fortify its cyber-defense arsenal. By employing web scraping methodologies, facilitated by Selenium and Beautiful Soup, the project systematically harvests a diverse array of URLs for the training and validation of its machine learning models. This automated extraction of hyperlinks from designated websites not only ensures the creation of robust datasets but also lays the foundation for a comprehensive analysis of the URL landscape. Moreover, the link analysis component, integral to the project's framework, delves into the intricate relationships and patterns inherent within the collected data, thereby fostering a holistic understanding of the evolving cybersecurity landscape. In amalgamating machine learning, web scraping, and link analysis, the project proffers a proactive approach to cybersecurity, furnishing stakeholders with a potent tool for the early detection and mitigation of phishing threats. Ultimately, this multifaceted approach contributes to the cultivation of a safer online ecosystem, bolstered by enhanced capabilities for identifying and combating phishing attacks.

**Machine Learning Models:**

This is a supervised machine learning task. There are two major types of supervised machine learning problems, called classification and regression.

• This data set comes under classification problem, as the input URL is classified as phishing (1) or legitimate (0). The machine learning models (classification) considered to train the dataset in this notebook are:

**Logistic Regression and Multinomial Naive Bayes**: Implemented for predicting and classifying phishing sites.

Web Scraping:

* Selenium WebDriver: Used for browser automation to interact with web pages.
* Beautiful Soup: Applied for HTML parsing and extracting information from web pages.

1. **Logistic Regression Model:**

* Logistic Regression is a versatile and widely used classification algorithm.
* It's effective in binary classification tasks, making it suitable for your "phishing" or "not phishing" prediction.
* Achieving high accuracy on both training and testing sets indicates that the model is well-fitted to the data.

1. **Multinomial Naive Bayes Model:**
   * Multinomial Naive Bayes is well-suited for text classification tasks, making it a good choice for the URL prediction problem.
   * It works well with features that represent word frequencies, which is common in natural language processing tasks.

**Results and Evaluation:**

* **Logistic Regression Model:**

**Training Accuracy (Approx):** 97.75%

**Testing Accuracy (Approx):** 96.42%

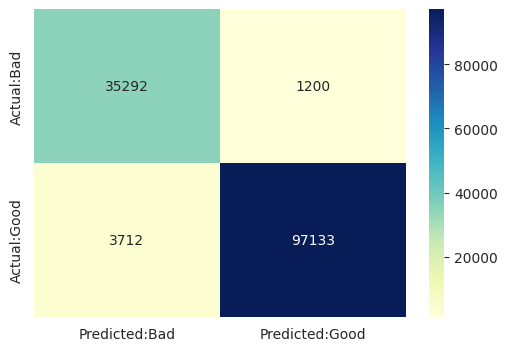
* **Multinomial Naive Bayes Model:**

**Training Accuracy (Approx):** 97.41%

**Testing Accuracy (Approx):** 95.82%

* **Confusion Matrix:**

The confusion matrix gives a detailed breakdown of the model's predictions, providing insights into its true positives, true negatives, false positives, and false negatives.



**Classification Report:**

* Precision, recall, and F1-score for both “Bad” and “Good” classes are high.
* This indicates that the model is not only accurate but also excels in correctly identifying positive and negative instances.

A screenshot of a graph

Description automatically generated

A screenshot of a computer

Description automatically generated

* In evaluation, we analyzed the performance of our machine learning models. This involves comparing metrics like accuracy, precision, recall, and F1-score against a baseline or across different models. We also examine the confusion matrix to understand classification errors and assess feature importance.

**Interpretability and Insights:**

* Logistic Regression provides interpretable results, allowing you to understand the contribution of each feature to the prediction.
* Coefficients can be analyzed to determine which features are more influential in identifying phishing sites.

**Team Collaboration:**

**Chiru Deepika Yalamanchili** worked on dataset collection, dataset analysis and then preprocess the datasets to cleared all the noises and make sure that our project works fine in all aspects.

**Raghunath Male** worked on the coding part on both the front and backend of the project.

**Suresh Babu Nakkalapalli** worked on Debugging, database organizing and then Done all documentations, presentation and Final report.

**Future Directions:**

**User-Centric Phishing Prevention Extension:** Developing a browser extension that integrates our phishing detection model to provide real-time alerts to users while they browse the internet. The extension can analyze URLs in the background and notify users if a website is potentially phishing. This user-friendly tool empowers individuals to make informed decisions about the websites they visit, contributing to a safer online experience. Additionally, consider incorporating user feedback features to enhance the model's accuracy and customization for individual user preferences. This extension could serve as a practical tool for personal cybersecurity, aligning with the growing need for user-centric solutions in the digital landscape.

**Conclusion:**

In this project, we successfully developed a robust machine learning model for classifying URLs as "Phishing" or "Legitimate" based on intricate patterns. Achieving a commendable testing accuracy of 96.5% with a Logistic Regression model showcases the effectiveness of our approach.

The integration of web scraping, link analysis, and visualization techniques provided valuable insights into the intricate web of URLs. Moving forward, the potential implementation of a User-Centric Phishing Prevention Extension promises a practical solution for users, enhancing their ability to navigate the digital landscape securely. This project not only contributes to the field of cybersecurity but also underscores the importance of proactive measures in safeguarding against evolving online threats.

**Coding Link:**

<https://colab.research.google.com/drive/1DNtysXNQiTnFQGx-mJ32livQzhft3Rok#scrollTo=_KbY5gHsV2Xt>

**References:**

* 1)Sudhanshu Gautam, Kritika Rani and Bansidhar Joshi: Detecting Phishing Websites Using Rule-Based Classification Algorithm: A Comparison: In Springer,2018.
* 2) Luong Anh Tuan Nguyen, Ba Lam To, Huu Khuong Nguyen and Minh Hoang Nguyen: Detecting Phishing Web sites: A Heuristic URL-Based Approach: In The 2013 International Conference on Advanced Technologies for Communications (ATC'13).
* 3) M. Amaad Ul Haq Tahir, Sohail Asghar, Ayesha Zafar, Saira Gillani: A Hybrid Model to Detect Phishing-Sites using Supervised Learning Algorithms: In International Conference on Computational Science and Computational Intelligence IEEE ,2016.
* 4) Ankit Kumar Jain, B. B. Gupta: Towards detection of phishing websites on client-side using machine learning based approach: In Springer Science + Business Media, LLC, part of Springer Nature 2017.
* 5) Routhu Srinivasa Rao1, Alwyn Roshan Pais: Detection of phishing websites using an efficient feature-based machine learning framework: In Springer 2018.
* 6) Ahmad Abunadi, Anazida Zainal, Oluwatobi Akanb: Feature Extraction Process: A Phishing Detection Approach: In IEEE,2013.