Developing a Machine Learning System for Detecting and Mitigating Phishing Attacks

Preparing Your Literature Review:

Introduction:

This project is part of the fight against one of the most widespread and insidious cyber threats of our time: phishing. This advanced form of social engineering primarily aims to improperly appropriate sensitive user data, including but not limited to login credentials and credit card numbers **[1]**. At the heart of this initiative is a thematic emphasis on cybersecurity, enhanced by the integration of advanced machine learning algorithms. The main objective is twofold: firstly, to develop an innovative solution capable of precisely identifying and effectively neutralizing phishing attacks, in order to protect users' crucial information **[2]**. Secondly, it is just as imperative to raise awareness and educate users about the risks associated with phishing. By improving users' understanding of the threats, the project aims to foster a more security-conscious online environment **[3]**.

In essence, this project revolves around a technological and educational approach, aiming to strengthen digital security and user awareness in the face of the constantly evolving landscape of cyber threats.

Organization:

My literature review thematically or chronologically is going to be like that:

I. Introduction
II. Background
III. Methodology
IV. Discussions
V. Conclusion

Summary and Synthesis:

Here is a brief concise summary for my each papers:
“ https://ieeexplore.ieee.org/document/9214225 “
“ https://ieeexplore.ieee.org/document/9404714?denied “
“ https://arxiv.org/pdf/2201.10752.pdf#:~:text=,and%20building%20a%20large%20dataset “

Summary: The paper proposes a machine learning (ML)-based system for detecting phishing attacks. It utilized standard datasets from kaggle.com to input into ML algorithms. Two prevalent algorithms, Decision Tree (DT) and Random Forest (RF), were employed for classifying and detecting phishing websites. Principal Component Analysis (PCA) was used to identify and classify dataset components. The study found that the RF algorithm had less variance and managed the over-fitting issue better than DT **[1]**.

Synthesis: The paper contributes to the field by applying and comparing two different ML algorithms to detect phishing websites, addressing the need for robust detection systems in the face of evolving cyber threats. The use of PCA for feature selection and the comparison of DT and RF in terms of performance and overfitting provide valuable insights into the efficiency and accuracy of ML applications in cybersecurity. The high accuracy achieved by the RF algorithm, in particular, underscores the potential of advanced ML techniques in improving phishing detection systems. The comparison highlights that while both algorithms can be effective, RF may be more reliable in handling complex data without overfitting, which is crucial for adapting to the sophisticated and changing tactics used in phishing attacks **[2]**.

Conclusion:

The literature analysis concludes by highlighting how crucial it is to create sophisticated machine learning algorithms in order to identify and counteract phishing attempts, which are among the craftiest cyberthreats of the modern day. This research is important because it has the ability to effectively identify and neutralize phishing attempts, hence improving the security of sensitive user data **[4]**.

Additionally, by offering a unique solution that combines state-of-the-art machine learning techniques with user education to address both the technological and human elements of cybersecurity, this project will add to the body of knowledge already in existence. Through the development of more advanced detection techniques and raising user knowledge and comprehension of phishing hazards, the initiative will contribute to a safer online environment. It attempts to improve cybersecurity technology while also encouraging internet users to adopt a vigilante mindset, which is crucial given the constant evolution of cyberthreats. This two-pronged strategy guarantees that the initiative will have an immediate technological impact as well as long-lasting educational effects.

Reference

[1] *Dada, E. G., Bassi, J. S., Chiroma, H., Abdulhamid, S. M., Adetunmbi, A. O., & Ajibuwa, O. E. (2019). Machine learning for email spam filtering: review, approaches and open research problems. Heliyon, 5(6), e01802. https://doi.org/10.1016/j.heliyon.2019.e01802*

[2] *Phishing Attacks Detection using Machine Learning Approach. (2020, August 1). IEEE Conference Publication | IEEE Xplore. https://ieeexplore.ieee.org/document/9214225*

[3] *Survey on Detection and Prevention of Phishing Websites using Machine Learning. (2021, March 4). IEEE Conference Publication | IEEE Xplore. https://ieeexplore.ieee.org/document/9404714?denied=*

[4] *A. S. N. Aski A, «Proposed efficient algorithm to filter spam using machine learning techniques,» Pacific Science Review A: Natural Science and Engineering, vol. 18, pp. 145-149, 2016.*

Preparing Your Data Research:

My dataset: https://www.kaggle.com/datasets/shashwatwork/phishing-dataset-for-machine-learning/data

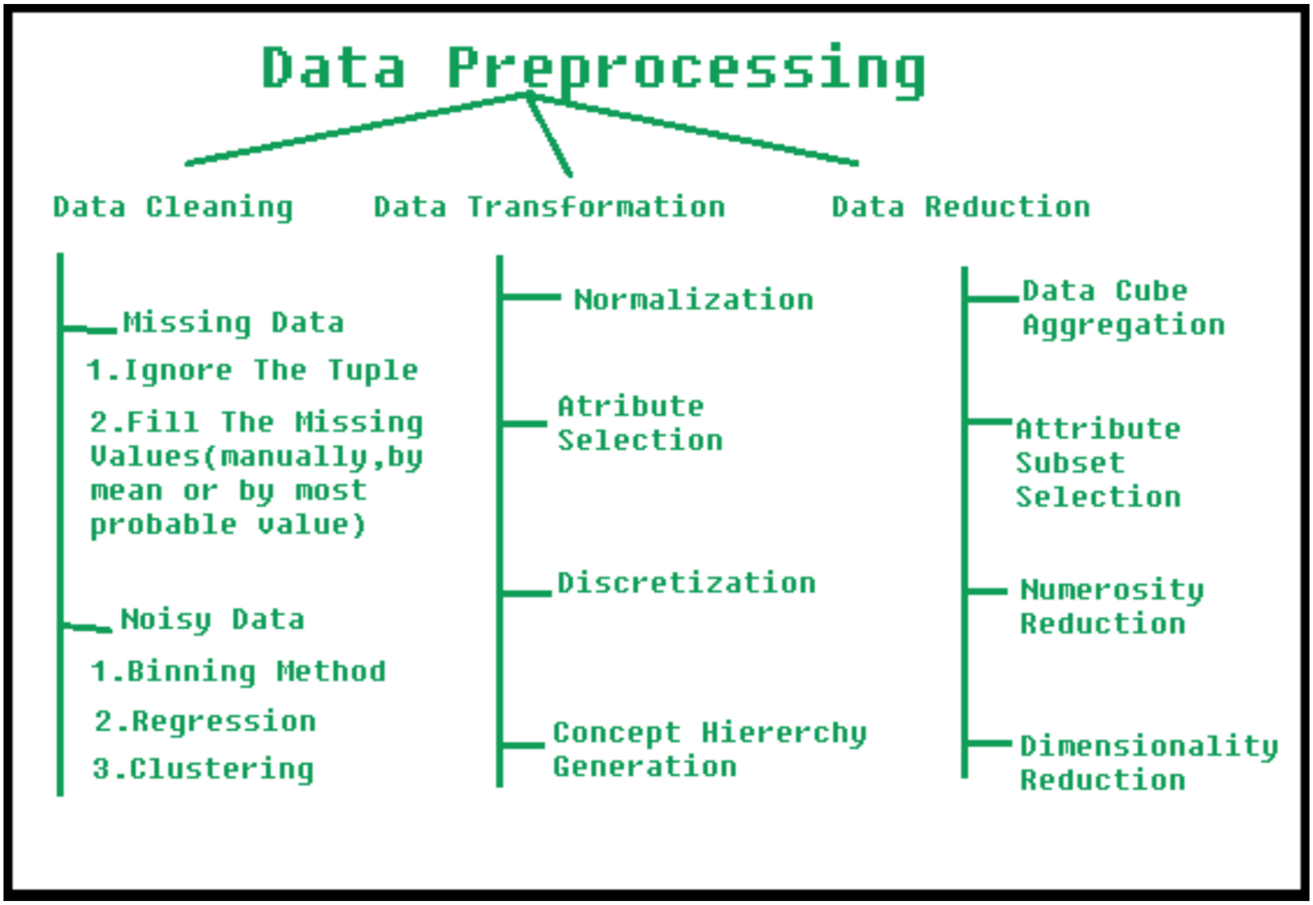
Introduction:

This data research project delves into the critical issue of phishing, a major cyber threat that exploits personal and financial data. The significance of our research questions is paramount; we aim to develop methods for precise phishing detection and prevention.
A detailed data analysis is vital, it informs the development of sophisticated machine learning models, ensuring their effectiveness against complex and evolving threats. By thoroughly examining the data, we aim to enhance cybersecurity measures and user awareness, providing robust defenses in an increasingly digital world. This foundation is essential for advancing online safety and privacy.

Organization:

The chronological structure of my dataset is as follows:

→ Preparation of Datasets
→ Data Preprocessing
→ Feature Selection
→ Machine learning modelling
→ Validation, Testing, and Information



My dataset appears to be related to website characteristics, likely used for identifying phishing websites. Each row in the dataset represents a different website, characterized by various features. Here are some insights based on the columns:

ID: Unique identifier for each record.

Numerical Features: Most of the columns like “NumDots”, “SubdomainLevel”, “PathLevel”, “UrlLength”, “NumDash”, etc., represent numerical features, possibly indicating different aspects of a website's URL or its structure.

Binary Features: Some columns, such as “AtSymbol”, “TildeSymbol”, “NumUnderscore”, etc., are binary (0 or 1), likely indicating the presence or absence of certain characteristics in the URL or the website content.

CLASS_LABEL: This column seems to be a target variable, indicating whether a website is phishing (1) or legitimate (0).

Data Source, Format, and Size:

Data Source: İt’s a dataset that comes from Kaggle.

Data Format: The data is in CSV format.

Data Size: The dataset contains 10,000 rows and 50 columns.

I selected this dataset for its perfect fit with my project 'Developing a Machine Learning System for Detecting and Mitigating Phishing Attacks'. Its detailed website characteristics are ideal for training models to distinguish between phishing and legitimate sites, aligning closely with my cybersecurity goals.

Data Analysis and Insights from the Dataset:

Descriptive Statistics Summary:

ID: Ranges from 1 to 10,000, indicating unique identifiers.

Numerical Features (e.g., “NumDots”, “UrlLength”): Vary considerably in range, with features like “NumDots” averaging around 2.45 with a standard deviation of 1.35, and “UrlLength” averaging around 70.26 with a standard deviation of 33.37.

Binary Features (e.g., “AtSymbol”, “IframeOrFrame”): Most binary features show a prevalence of one category over the other. For example, “AtSymbol” is almost always 0, whereas “IframeOrFrame” has a more balanced distribution.

CLASS_LABEL: The target variable is evenly distributed with a mean of 0.5, indicating a balanced dataset in terms of phishing and legitimate website classifications.

Trends and Patterns:

The data contains a mix of both binary and continuous variables, which suggests a need for different analysis and preprocessing approaches for each type.

The presence of certain features like “NumDash”, “NumUnderscore”, or “UrlLength” could be significant in distinguishing phishing websites from legitimate ones. Higher values in these features may indicate potential phishing activity.

Binary features like “IframeOrFrame” or “MissingTitle” may indicate specific tactics used by phishing websites.

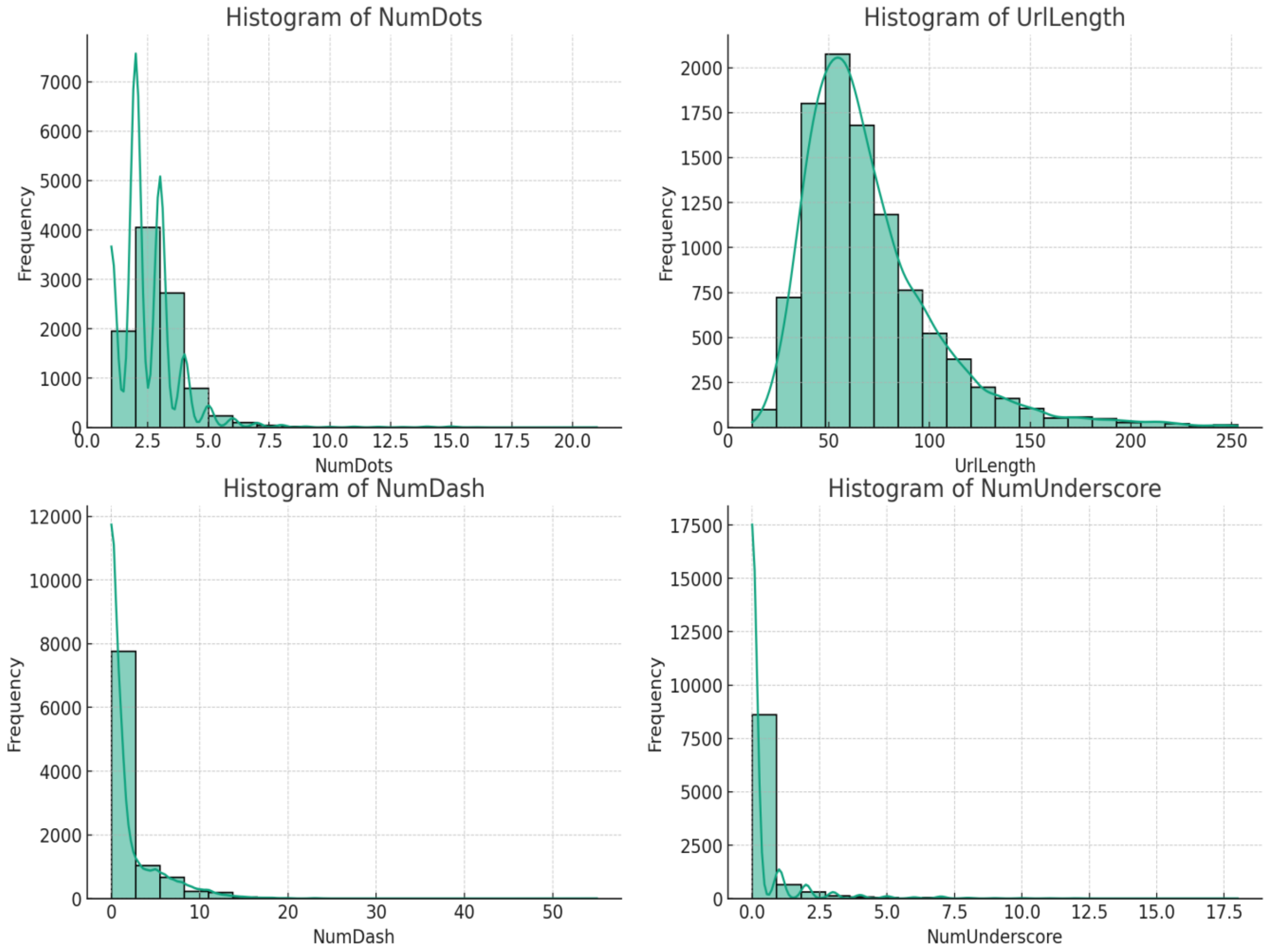
Visualizations:

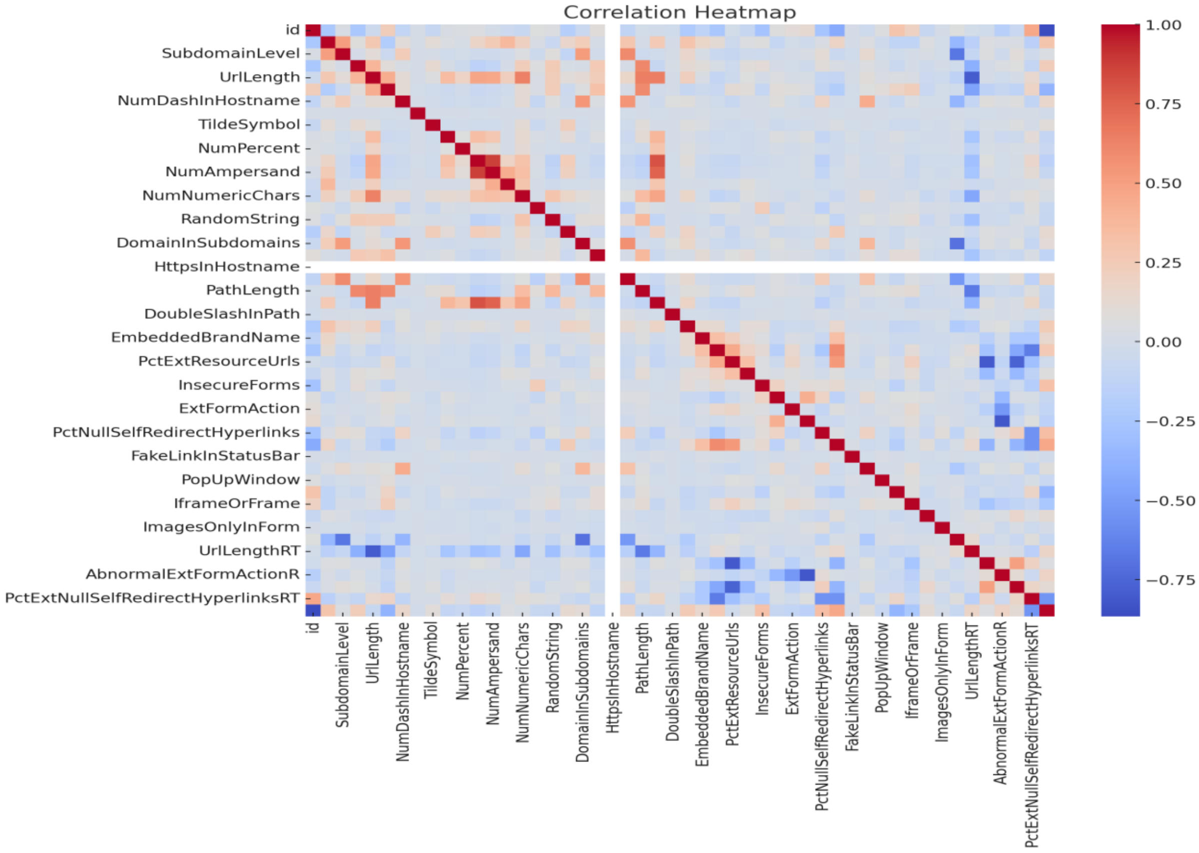
Next, I'll create a couple of visualizations to better illustrate these patterns:

Histograms for select numerical features to understand their distributions.

Correlation Heatmap to identify any notable correlations between features, especially in relation to the “CLASS_LABEL”.

Let's start with the visualizations.





Histograms of Selected Features:

NumDots: Most websites have a low number of dots in their URLs, with a few outliers having significantly more.

UrlLength: The length of URLs is skewed towards shorter lengths, which could be a distinguishing factor for phishing sites.

NumDash and NumUnderscore: These features also show a skew towards lower values with some exceptions.

Correlation Heatmap:

The heatmap illustrates the relationships between different features. While the detailed correlations are not visible in this overview, it provides a general idea of how features might be related to each other and to the “**CLASS_LABEL”**.

Features with strong correlations (either positive or negative) to the **“CLASS_LABEL”** are of particular interest as they could be key indicators of phishing activity.

Concluding Insights:

The dataset shows a variety of features that could potentially distinguish phishing websites from legitimate ones.

The balance in the CLASS_LABEL suggests that the dataset is well-suited for training classification models.

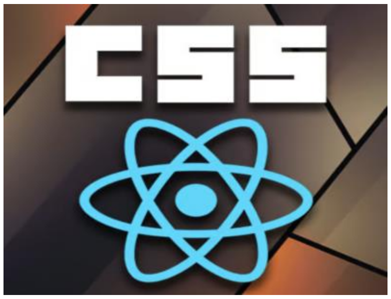
Careful feature selection based on these insights could improve the effectiveness of a machine learning model designed for phishing detection.

These visual and statistical analyses provide a foundational understanding of the dataset, crucial for any further modeling or deeper analysis in my project.

Preparing Your Technology Review:

Technology Requirements:

Python programming language and frameworks such as NumPy, Pandas, Scikit-learn will be used to build the whole model.

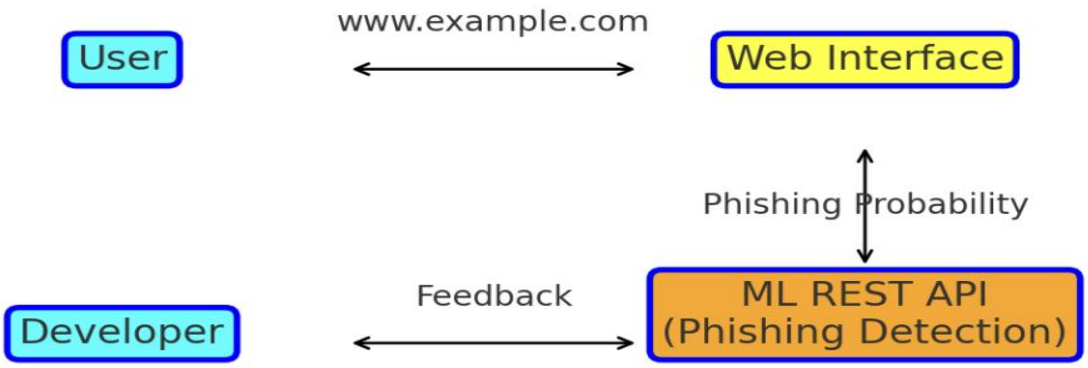


- PyCharm is used as an IDE.
- For visualization of the plots, Matplotlib,Seaborn and Plotly are used.
- Heroku is used for deployment of the model.
- Front end development is done using React & Css
- Python FastAPI is used for backend development.
- GitHub is used as a version control system.

Model Deployment:

Deployment is a key step in an organization gaining operational value from machine learning. The simplest way to deploy a machine learning model is to create a Web Microservice like an REST API. But to make it more acessable,such that even Non-technical people can interact with the model,we'll also be creating a Web Interface for it.

Below diagram describes the architecture of the system after deployment,it describes how both the developer and a user exchange data with the ML model.



Deployment Tools

Below are the following tools we'll be using for this project:

- ML RestAPI-FlaskorFastAPI Python Libraries
- WebInterface-ReactJS Javascript WebFramework
- CloudService-Heroku

Deployment Services (Cost : 10-15$):

For deployment of our model we'll be using cloud services which will host our models. At the start since we don't have many users,we may not be charged for hosting our services, but as our users grow our costs will also rise per usage. We'll be using Heroku or AWS cloud services for our project.