High Level Design (HLD)

Air Quality Index Prediction

Revision Number: 1.0

Last date of revision: 9 August 2024

Pranav Ghorpade

**Table of Content**

Abstract ………………………………………………………………………………………… 3

Introduction ………………………………………………………………………………….. 4

1. Why this High-Level Document? ……………………………………………. 4

2. Scope …………………………….………………………………………………………. 4

3. Definition ………………………………………………………………………………. 4

General Description ………………………………………………………………………. 5

1. Product Perspective ……………………………………………………….………. 5

2. Problem Statement ………………………………………………………..………. 5

3. Problem Solution ……………………………………………………………………. 5

4. Further Improvements ……………………………………………………………. 5

5. Data Required ……………………………………………………...…………………. 5

6. Tools Used ………………………………………………………………………………. 5-6

7. Constraints …………………………………………………………………..…………. 6

8. Assumptions ……………………………………………………………………………. 6

Design Details ……………………………………………………………..…………………. 7

1. Process Workflow ………………………………………………………..…………. 7

Performance …………………………………………………………………………….……. 8

1. Reusability ……………………………………………………………..………………. 8

2. Application compatibility ………………………………………..………………. 8

3. Deployment ……………………………………………………………………..…..….8

Conclusion ………………………………………………………………………….……..…….9

Reference ………………………………………………………………………………..………10

**Abstract**

Phishing attacks pose a significant threat to online security, leading to financial losses and compromised personal information. This project focuses on the detection of phishing URLs using machine learning techniques. Initially, logistic regression was employed but proved inadequate in capturing the complex patterns of phishing attempts. Consequently, a random forest model was developed and implemented, demonstrating superior performance in identifying malicious URLs. The model was trained and evaluated on a dataset comprising various features extracted from URLs, including lexical, host-based, and content-based attributes. The random forest model achieved a high accuracy rate, showcasing its efficacy in distinguishing phishing URLs from legitimate ones. This project highlights the potential of machine learning in enhancing cybersecurity measures and provides valuable insights into the practical application of random forest classifiers in phishing detection.

**Introduction**

**1. Why this High-Level Design Document?**

The purpose of this High-Level Design (HLD) Document is to add the necessary detail to the current project description to represent a suitable model for coding. This document is also intended to help detect contradictions prior to coding and can be used as a reference manual for how the modules interact at a high level.

The HLD will:

* Present all the design aspects and define them in detail
* Describe the user interface being implemented
* Describe the hardware and software interfaces
* Describe the performance requirements
* Include design features and the architecture of the project

**2. Scope**

The HLD documentation presents the structure of the system, such as the database architecture, application architecture (layers), application flow (Navigation), and technology architecture. The HLD uses non-technical to mildly technical terms which should be understandable to the administrators of the system.

**3. Definition**

The terms used in the projects are:

* Phishing url: The link designed to trick user into revealing sensitive information.
* Labels: Good or Bad

**General Description**

**1. Product Perspective**

The phishing URL detection system is designed to seamlessly integrate into existing security infrastructures, providing real-time analysis and identification of malicious URLs. Its user-friendly interface allows for easy deployment and monitoring, while the robust random forest model ensures high accuracy and reliability. This tool empowers organizations to proactively protect their digital assets and users from phishing threats, enhancing overall cybersecurity.

**2. Problem Statement**

To create the machine learning based solution to detect phishing urls.

**3. Problem Solution**

To address this, we developed a phishing URL detection system utilizing a random forest model. This machine learning approach captures complex patterns in URL features, significantly improving detection accuracy. The model is trained on a comprehensive dataset of lexical, host-based, and content-based features, enabling it to effectively distinguish between legitimate and malicious URLs. This solution enhances cybersecurity measures, providing reliable, real-time protection against phishing attacks.

**4. Further Improvement**

Future improvements could include incorporating deep learning techniques to further enhance detection accuracy and expanding the feature set with real-time threat intelligence data for even more robust protection.

**5. Data Required**

For training the model we need the data that consist of different phishing or non-phishing urls and their labels.

Data is completely depending upon our problem statement

**6. Tools Used**

* Python programming language and frameworks such as NumPy, Pandas, Scikit-learn, Matplotlib, Seaborn are used to build the whole model.
* PyCharm and Visual Studio Code is used as IDE.
* For visualization of the plots, Matplotlib and Seaborn are used.
* Front end development is done using Flask.
* GitHub is used as version control system.

**7. Constraints**

The phishing url prediction website should be user friendly.

**8. Assumptions**

The model assumes that the extracted URL features are comprehensive enough to differentiate phishing from legitimate URLs. Additionally, it presumes the random forest model's accuracy will generalize effectively to new, unseen data.

**Design Details**

**1. Process Workflow**

Model Training and Evaluation



**Performance**

**1. Reusability**

The phishing URL detection model is designed for reusability, allowing seamless integration into various cybersecurity systems with minimal adjustments. Its high performance, demonstrated by consistent accuracy across diverse datasets, ensures reliable detection in different operational environments.

**2. Application compatibility**

The phishing URL detection system is compatible with various applications, including web browsers, email filters, and security software. Its flexible API design ensures easy integration and interoperability across different platforms and operating systems.

**3. Deployment**

The code is deployed in GitHub.

**Conclusion**

In conclusion, the phishing URL detection system leverages the power of a random forest model to accurately identify malicious URLs, significantly enhancing cybersecurity measures. With its efficient resource utilization, broad application compatibility, and high performance, this solution offers a robust and scalable approach to protecting users from phishing attacks. Future improvements and deep learning integration promise even greater accuracy and resilience against evolving threats.

**References**

* <https://docs.streamlit.io/en/stable/>
* <https://scikit-learn.org/stable/user_guide.html>
* <https://numpy.org/doc/>
* <https://seaborn.pydata.org/examples/regression_marginals.html>
* <https://seaborn.pydata.org/examples/scatterplot_matrix.html>
* <https://matplotlib.org/>
* <https://pandas.pydata.org/docs/>
* <https://phishtank.org/>