# DETECTION OF PHISHING WEBSITES USING MACHINE LEARNING

**BY**

# UMAR GHALI MUKHTAR

# FCP/CSC/18/1032

**A PROJECT SUBMITTED TO THE DEPARTMENT OF COMPUTER SCIENCE,**

# FEDERAL UNIVERSITY DUTSE, IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE AWARD DEGREE OF THE DEGREE OF BACHELOR OF SCIENCE (B.Sc.) IN COMPUTER SCIENCE.

**JANUARY, 2024**

**DECLARATION**

I, Umar Ghali Mukhtar, declare that this project report entitled ”DETECTION OF PHISHING WEBSITES USING MACHINE LEARNING” submitted in partial fulfillment of the requirements for the degree of computer science in the department of computer science at federal university dutse, is my original work. It has not been submitted for any other degree or examination at any other university.

I also affirm that this project work complies with the guidelines and regulations of federal university dutse and that all sources used have been properly cited and referenced.

Date:

Signature: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**CERTIFICATION**

This is to certify that this project titled “Detection of phishing websites using machine learning” was carried out by Umar Ghali Mukhtar with the registration number FCP/CSC/18/1032 meet the regulations governing the award of degree of computer science at Federal University Dutse.

# ……………………………… ……………………………

Dr. Z.S. Iro Date

(Supervisor)

# ……………………………… ……………………………

Dr. Aminu Abdullahi Date

(Project Coordinator)

**DEDICATION**

To the Pillars of My Dreams, This project stands as a testament to the unwavering support, boundless love, and enduring belief you have showered upon me throughout my journey. It is with profound gratitude and immense joy that I dedicate this endeavor to my parents.

**ACKNOWLEDGEMENT**

No angle leap will fall without the knowledge of Almighty Allah. I testify that there is no deity worthy of worship except Allah and prophet Muhammad (SAW) is his messenger.

My first appreciation goes to my beloved parent for their parental care attended to me throughout the course of my study. I pray you live long and reap the point of your tireless labor on myself and my siblings.

My profound gratitude goes to my disciplinarian able supervisor, Dr. Zahraddeen Salele Iro without whose guidance, this project work would not have attained this highly esteemed standard, may Almighty Allah continue to protect you and your family.

In the same vein. I appreciate the prayers, love and support of my sibling as, I remain grateful.

I have no expression to qualify the immense contribution of Habibu Ibrahim, Kabir Kabir Dalha, Alamin Nasir Hanga, Ibrahim Suleiman and all my course mates to the success of my study.

May the Almighty Allah protect, guide and bless everybody who directly or indirectly contributed to the successful completion of the programme.

**ABSTRACT**

As the internet becomes an integral part of daily life, the threat of phishing attacks has surged, jeopardizing users' sensitive information. This project focuses on leveraging machine learning algorithms to detect phishing websites effectively. By employing a dataset comprising features extracted from the structure and content of known phishing and legitimate websites, our model aims to discern subtle patterns indicative of malicious intent. Through the application of supervised learning techniques, such as decision trees and ensemble methods, we achieve high accuracy in differentiating between phishing and legitimate websites. The proposed approach not only enhances current phishing detection capabilities but also contributes to the ongoing efforts to fortify cyber security in the ever-evolving digital landscape.

Contents

[DECLARATION ii](#_Toc155643171)

[CERTIFICATION iii](#_Toc155643172)

[DEDICATION. i](#_Toc155643173)v

[ACKNOWLEDGEMENT……………………………………………………………………………..](#_Toc155643175)v

ABSTRACT……………………………………………………………………………………………vi

CHAPTER ONE………………………………………………………………………………………..1

1.0 Overview of the study………………………………………………………………………………1

1.1 Background of the study…………………………………………………………………...1

1.2 Statement of the problem………………………………………………………………….2

1.3 Motivation of the study……………………………………………………………………2

1.4 Aim of the study…………………………………………………………………………..3

1.5 Objectives of the Study…………………………………………………………………..3

[CHAPTER TWO 5](#_Toc155643176)

[LITERATURE REVIEW 5](#_Toc155643177)

[2.1 Overview of the Study 5](#_Toc155643178)

[2.2 Theoretical Review 5](#_Toc155643179)

[2.2.1 Feature selection 6](#_Toc155643180)

[2.2.2 Feature extraction 6](#_Toc155643181)

[2.3 Conceptual Review 7](#_Toc155643182)

[2.3.1 Phishing mechanism 7](#_Toc155643183)

[2.3.2 Visual similarity-based phishing detection and filtering approaches 10](#_Toc155643184)

[2.4 Empirical Review 12](#_Toc155643185)

[CHAPTER 3 15](#_Toc155643186)

[3.1 Methodology 15](#_Toc155643187)

[3.2 Overview of System Analysis 15](#_Toc155643188)

[3.3 Analysis of Existing System 16](#_Toc155643189)

[3.4 Proposed System 16](#_Toc155643190)

[3.4.1 Benefits of the new system 16](#_Toc155643191)

[3.5 Model Development 17](#_Toc155643192)

[i. Data Collection 17](#_Toc155643193)

[ii. Preprocessing 18](#_Toc155643194)

[iii. Exploratory data analysis 18](#_Toc155643195)

[iv. Feature Extraction 18](#_Toc155643196)

[v. Model Training 20](#_Toc155643197)

[vi. Model Testing 20](#_Toc155643198)

[vii. Model Evaluation 20](#_Toc155643199)

[3.5 System Modeling 21](#_Toc155643200)

[3.5.1 Use a case diagram of the system 22](#_Toc155643201)

[3.5.2 Flowchart of the system 25](#_Toc155643202)

[CHAPTER FOUR 27](#_Toc155643203)

[SYSTEM IMPLEMENTATION AND RESULTS 27](#_Toc155643204)

[4.1 Installation Requirements 27](#_Toc155643205)

[4.1.1 Hardware requirements 27](#_Toc155643206)

[4.1.2 Software requirements 27](#_Toc155643207)

[4.2 Model Development 27](#_Toc155643208)

[4.2.1 Data collection 28](#_Toc155643209)

[4.1.1 Feature extraction on the datasets 30](#_Toc155643210)

[4.3 General Working of the System 30](#_Toc155643211)

[4.3.1 The home page 31](#_Toc155643212)

[4.3.2 The about page 31](#_Toc155643213)

[4.3.3 The use case page 31](#_Toc155643214)

[4.3.4 Resource page 32](#_Toc155643215)

[4.3.5 Web application Source Code 32](#_Toc155643216)

[Figure 4.18: Code for the web application 36](#_Toc155643217)

[CHAPTER FIVE 37](#_Toc155643218)

[SUMMARY, CONCLUSION, AND RECOMMENDATIONS 37](#_Toc155643219)

[5.1 Summary 37](#_Toc155643220)

[5.2 Contribution to Knowledge 37](#_Toc155643221)

[5.3 Conclusion 37](#_Toc155643222)

[5.4 Recommendation 37](#_Toc155643223)

[REFERENCE 38](#_Toc155643224)

[APPENDIX A: Web application Source code 42](#_Toc155643225)

[APPENDIX B: Source code for API integration to web app 51](#_Toc155643226)

**CHAPTER ONE**

**INTRODUCTION**

1. **Overview**

Phishing is a type of social engineering attack that attempts to trick users into revealing confidential information, such as passwords, credit card numbers, or Social Security numbers. Phishing attacks are typically carried out through email and websites, but they can also be launched through social media, text messages, or even phone calls.

Phishing attacks are often successful because they rely on human psychology. Phishing emails may use social engineering techniques to create a sense of urgency or fear, or they may try to impersonate a trusted organization such as a bank or government agency. Phishing websites may also be designed to look like legitimate websites, using similar logos, colours, and layouts.

Machine learning is a type of artificial intelligence (AI) that allows computers to learn without being explicitly programmed. Machine learning algorithms can be trained on data to identify patterns and make predictions. This makes them well-suited for detecting phishing websites, as they can learn to identify the common features that distinguish phishing websites from legitimate websites.

**1.1 Background of the study**

Detecting phishing websites using machine learning involves understanding the prevalence and sophistication of phishing attacks, the limitations of traditional rule-based detection methods, and the potential of machine learning techniques in improving detection accuracy. It is crucial to highlight the significance of addressing the growing threat of phishing attacks, which aim to deceive users into revealing sensitive information, such as login credentials, financial details, or personal data.

Phishing attacks have become increasingly sophisticated, often bypassing traditional security measures and exploiting human vulnerabilities. These attacks can lead to significant financial losses, identity theft, and damage to individuals and organizations. Existing rule-based detection systems often rely on predefined patterns and heuristics, making them less effective in identifying new and evolving phishing techniques.

In contrast, machine learning algorithms can offer a more adaptive and dynamic approach to detect phishing websites. By analysing large datasets of known phishing websites and legitimate web pages, machine learning models can learn complex patterns and features that distinguish between genuine and fraudulent websites. This approach allows for the detection of subtle anomalies and behavioural patterns that may not be apparent to rule-based systems.

Furthermore, the integration of machine learning in phishing detection systems can potentially enhance the speed and accuracy of identifying phishing websites, thereby reducing the risk of falling victim to fraudulent activities. However, it is essential to acknowledge the challenges associated with the implementation of machine learning, such as the need for high-quality training data, the potential for false positives and false negatives, and the requirement for continuous model updates to adapt to evolving phishing tactics.

Overall, the background of the study emphasizes the urgency of leveraging advanced technologies, such as machine learning, to bolster cyber security measures and mitigate the threats posed by phishing attacks. By addressing the limitations of traditional detection methods and harnessing the power of machine learning, researchers aim to develop more robust and effective solutions to safeguard individuals and organizations from the detrimental consequences of phishing scams.

**1.2 Statement of the problem**

Phishing attacks have emerged as a pervasive and evolving cyber security threat, targeting individuals and organizations worldwide. These attacks exploit human vulnerabilities to deceive users into divulging sensitive information, leading to financial losses, identity theft, and data breaches. While traditional rule-based methods for detecting phishing websites have been in use, they often struggle to keep pace with the continually changing tactics employed by cybercriminals. In this case, we would craft an artificially intelligent machine learning system to do this detection to near perfect accuracy.

**1.3 Motivation of the study**

The motivation for this study lies in the pressing need to combat the escalating threat of phishing attacks, which continue to jeopardize individuals, organizations, and government entities, leading to financial losses, identity theft, and data breaches. Traditional rule-based detection methods are increasingly ineffective in identifying sophisticated and evolving phishing tactics, necessitating the exploration of innovative and adaptive solutions facilitated by the advancement of machine learning and artificial intelligence technologies. Real-time protection against phishing attacks is a crucial goal, as traditional approaches often lack immediate detection capabilities, leaving users exposed for extended durations. The study is further motivated by the challenges of acquiring high-quality, balanced datasets for machine learning model development, as well as the recognition of the complementary role of user education and awareness programs in bolstering cyber security. Additionally, the motivation stems from the aspiration to create machine learning models that can generalize across diverse phishing tactics and adapt to emerging methods, all while addressing ethical and privacy considerations. Ultimately, the study aims to reduce cyber security risks and protect individuals, organizations, and sensitive data from the detrimental effects of phishing attacks.

**1.4 Aim of the study**

This project aims to detect phishing websites using machine learning and deep neural

Networks by developing a web application that allows users to check if a URL is

Phishing or legitimate and have access to resources to help tackle phishing attacks.

**1.5 Objectives of the Study**

To accomplish the project's purpose, the following particular objectives have been

Established:

i. dataset collection and pre-processing;

ii. machine-learning model selection and development;

iii. Development of a web-based application for detection;

iv. Integration of the developed model to a web application.

**1.6 Scope of this project**

This study explores data science and machine learning models that use datasets gotten from open-source platforms to analyse website links and distinguish between phishing and legitimate URL links.

The model will be integrated into a web application, allowing a user to predict if a URL link is legitimate or phishing. This online application is compatible with a variety of browsers.

**1.7 Significance of this study**

The study bears significance in the sense that it can help at least reduce instances of people falling for fake websites that they may give their information to. It really helps for people who do not have all the time in the world to carefully identify exact URLs to verify the addresses.

**1.8 Definition of terms**

Phishing: Phishing is a type of social engineering attack often used to steal user data, including login credentials and credit card numbers.

Website: a set of related web pages located under a single domain name, typically produced by a single person or organization.

Machine learning: Machine learning is a type of AI focused on building computer systems that learn from data, enabling software to improve its performance over time.

URL: URL stands for Uniform Resource Locator.

HTTPS: Hypertext transfer protocol secure is the secure version of HTTP, which is the primary protocol used to send data between a web browser and a website.

# CHAPTER TWO

# LITERATURE REVIEW

# Overview of the Study

This chapter offers an insight into various important studies conducted by excellent scholars from articles, books, and other sources relevant to the detection of phishing websites. It also provides the project with a theoretical review, conceptual review, and empirical review to demonstrate understanding of the project.

# Theoretical Review

Ankit and Gupta (2020) mentioned that Statistics show that according to Internet world stats ("Internet world stats usage and population statistics", 2018), the total numbers of Internet users worldwide are 4.27 billion in 2018; that is, more than 38% of the world population uses the Internet. Hackers take advantage of the insecure Internet system and can fool unaware users to fall for phishing scams. A phishing e-mail is used to defraud both individuals and financial organizations on the Internet. (“RSA Anti-Fraud Command Center”, n.d.) Said the Anti-Phishing Working Group (APWG) is an international consortium that is dedicated to promoting research, education, and law enforcement to eliminate online fraud and cyber-crime. In 2020, total phishing attacks increased by 160% over 2015, signifying a record year in phishing volumes. The total phishing attacks detected in 2015 were approximately 450,000 and led to financial losses of more than 5.9 billion dollars (“RSA Anti-Fraud Command Center”, n.d.). Total attack increases by 1% in 2020 as compared to 2015. The total number of phishing attacks noticed in Q1 (first quarter) of 2020 was 125,215, a 10.7 percent increase over Q4 (fourth quarter) of 2020. More than 55% of phishing websites contain the name of the target site in some form to fool users and 99.4% of phishing websites use port 80 ("Anti-Phishing Working Group (APWG)Phishing activity trends report first quarter",

2020). According to the APWG report in the first quarter of 2020, the second-highest number of phishing attacks ever recorded was between January and March 2020 ("Anti- Phishing Working Group (APWG) Phishing activity trends report first quarter", 2020) and payment services are the most targeted industry. During the second half of 2020, 123,972 unique phishing attacks were observed ("APWG report", 2020).

# Feature selection

Feature Selection is the method of reducing the input variable to your model by using only relevant data and getting rid of noise in data.

It is also the process of automatically choosing relevant features for your machine- learning model based on the type of problem you are trying to solve. We do this by including or excluding important features without changing them. It helps in cutting down the noise in our data and reducing the size of our input data. Figure 2.5 shows the feature selection process.

Feature selection models are of two types:

1. Supervised Models:

Supervised feature selection refers to the method which uses the output label class for feature selection. They use the target variables to identify the variables which can increase the efficiency of the model.

1. Unsupervised Models:

Unsupervised Feature selection refers to the method which does not need the output label class for feature selection. We use them for unlabeled data.

# Feature extraction

Feature extraction is a process of dimensionality reduction by which an initial set of raw data is reduced to more manageable groups for processing. A characteristic of these large data sets is a large number of variables that require a lot of computing resources to process. Feature extraction is the name for methods that select and or combine variables into features, effectively reducing the amount of data that must be processed, while still accurately and completely describing the original data set (deepAI, n.d.).

Why is Feature Extraction Useful? The process of feature extraction is useful when you need to reduce the number of resources needed for processing without losing important or relevant information. Feature extraction can also reduce the amount of redundant data for a given analysis. Also, the reduction of the data and the machine’s efforts in building variable combinations (features) facilitate the speed of learning and generalization steps in the machine learning process. (DeepAI, n.d.)

According to (Rami, Fadi & Lee, 2019), they have compounded important features that have proved to be sound and effective in predicting phishing websites. In addition, they have proposed some new features, experimentally assign new rules to some well-known features and update some other features.

These feature selections include:

1. Address Bar based Features
2. Abnormal Based Features
3. HTML and JavaScript-based Features
4. Domain-based Features

All the listed feature selection above consists of feature extractions which are guided by rules.

# 2.3 Conceptual Review

# 2.3.1 Phishing mechanism

Figure 2.8 below shows a fake website is the clone of a targeted genuine website, and it always contains some input fields (e.g., textbox). When the user submits his/her details, the information is transferred to the attacker. An attacker steals the credential of the innocent user by performing the following steps:

* + - 1. Construction of Phishing Site. In the first step, the attacker identifies the target as a well-known organization. Afterward, the attacker collects detailed information about the organization by visiting their website. The attacker then uses this information to construct the fake website.
      2. URL Sending. In this step, the attacker composes a bogus e-mail and sends it to thousands of users. The attacker attached the URL of the fake website in the bogus e-mail. In the case of a spear-phishing attack, an attacker sends the e-mail to selected users. An attacker can also spread the link of phishing websites with the help of blogs, forums, and so forth (Kruegel, Kirda, Mutz, Robertson, & Vigna, 2019).
      3. Stealing of the Credentials: when the user clicks on the attached URL, consequently, fake site is opened in the web browser. The fake website contains a fake login form that is used to take the credential of an innocent user. Furthermore, the attacker can access the information filled by the user.
      4. Identity Theft: attacker uses this credential for malicious purposes. For example, an attacker purchases something by using the credit card details of the user. (Ankit & Gupta, 2017).

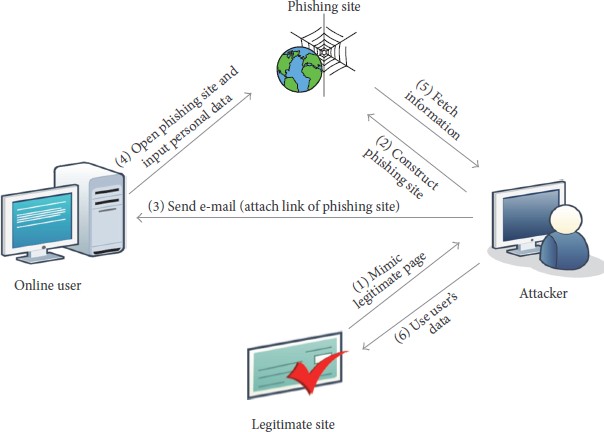


Figure 2.8 Phishing Mechanism (Source: Ankit & Gupta, 2017)

# Visual similarity-based phishing detection and filtering approaches

A user could become the victim of the phishing attack by looking at the high visual resemblance of phishing websites with the targeted legitimate site, such as page layouts, images, text content, font size, and font color. Take for example the fake and genuine web pages of PayPal as shown in Figure 2.11.

Both pages have the same visual appearance but different URLs. It is not always necessary that people carefully notice on URL and SSL (Secure Socket Layer) certificates of websites. If an attacker does not copy the visual appearance of a targeted website well, then the chances of inputting credentials by Internet users are very less. An attacker fools the user in the following ways:

1. Visual Appearance: The phishing website looks similar to its legitimate website.

Attackers used to copy the HTML source code of a genuine website to build the fake website.

1. Address Bar: Attackers also cover the address or URL bar of the website by script or image. The user would believe that they are inputting information on the right website. (Ankit & Gupta, 2017).

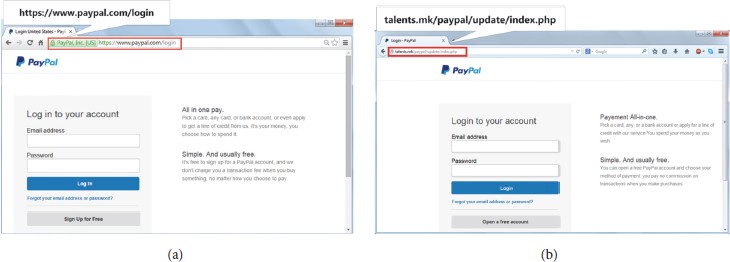


Figure 2.11 (a) Legitimate PayPal webpage and (b) phishing webpage of PayPal. (Source: Ankit & Gupta, 2017).

# Empirical Review

In 2017, the paper titled **A Theoretical Study on Different ways to identify the Phishing URL and Its Prevention Approaches** by V. Karamchand Gandhi and Dr.

M. Suriakala works on identifying the general format and procedure of phishing URLs by analyzing 200 phishing email archives provided by APWG. Based on the analysis, they find out the suitable approaches to prevent them from the fraudulent process.

The strengths of the paper are that the approaches, technical techniques, and classification of phishing hyperlinks in phishing e-mails enhance better results in the identification and prevention of phishing attacks.

The weakness of the paper is that the paper faces the challenge of inadequate data, as any machine learning algorithm can only give correct readings/predictions if it is applied to reliable data.

To address this weakness, this project will ensure proper and accurate dataset collection and pre-processing of updated phishing emails from the Anti-Phishing Working Group (APWG).

In 2015, the paper titled **Protecting users against phishing attacks with AntiPhish** by Engin Kirda and Christopher Kruegel presents AntiPhish, a browser extension that aims to protect inexperienced users against spoofed website-based phishing attacks. AntiPhish keeps track of the sensitive information of a user and generates warnings whenever sensitive information is typed into a form on a website that is considered untrusted. The tool has been implemented as a Mozilla Firefox plug-in and is free for public use.

The strengths of the paper give an overview of the different types of phishing attacks and discuss a real-world example.it also describes the solution of AntiPhish, and provides details about its implementation, as well as presents an example that shows how AntiPhish mitigates a typical phishing attempt.

The weakness of the paper is that despite that Antiphish is free for public use; the AntiPhish browser extension does not support other Web browsers such as Chrome, Brave, and internet Explorer but is limited to only the Mozilla browser.

To address this weakness, a cross-platform plug-in is developed for any browser to endure AntiPhish is not limited to a specific browser since it is free for public use.

In 2019, the paper titled **A Review Paper on Detection of Phishing Websites Using Machine Learning** by Ashritha Jain R, Chaithra Kulal, Deekshitha S, and Mrs. Mangala Kini proposed different algorithms (models) as well as different features of phishing attacks and techniques to detect phishing websites.

The strengths of the paper discuss works and different methods for phishing detection. It also introduces the proposed methodology that can be implemented to predict the phishing website accurately. The investigation gap provides more scope to study phishing detection.

The weakness of the paper is that it is limited to few feature selections of URL websites as well as four models for the prediction of phishing websites which might not produce the best accuracy of the model.

To address this weakness, several resources are sourced from updated published work to uncover different Machine learning models and features of a website to test and train the dataset for accuracy of the model. This will help find the best accuracy of the model for the detection of a phishing website.

In 2019, the paper titled **Phishing Websites Detection Using Machine Learning** by (Kiruthiga. R., & Akila, D.) Explained a novel approach to detect phishing websites using machine learning algorithms. They also compared the accuracy of five machine learning algorithms Decision Tree (DT), Random Forest (RF) (Shad & Sharma, 2018), Gradient Boosting (GBM), Generalized Linear Model (GLM), and Generalized Additive Model (GAM)(Shad & Sharma, 2018). Accuracy, Precision, and Recall evaluation methods were calculated for each algorithm and compared. Website attributes (30) are extracted with the help of Python and performance evaluation done with open-source programming language R. Top three algorithms namely Decision Tree, Random Forest, and GBM performance were compared in the table. From the tables of accuracy, recall, and performance, it is shown that the Random Forest algorithm has given the highest 98.4% accuracy, 98.59% recall, and 97.70% precision.

The strengths of the paper are that different authors proposed different approaches for feature selection algorithm, automatic feature extraction, and phishing detection model to detect phishing performance effectively by using different machine learning models.

In this paper, the authors Sönmez, Tuncer, Gökal, & Avci (2018) **proposes a classification model to classify phishing attack**s. This model comprises feature extraction from sites and the classification of websites. In feature extraction, 30 features have been taken from the UCI Irvine machine learning repository data set and phishing feature extraction rules have been clearly defined. To classification, these features, Support Vector Machine (SVM), Naïve Bayes (NB), and Extreme Learning Machine (ELM) (Sönmez et al., 2018) were used. In Extreme Learning Machine (ELM), six activation functions were used and achieved 95.34% accuracy than SVM and NB. The results were obtained with the help of MATLAB.

The Authors Peng, Harris, & Sawa, (2018) present an approach to **detect phishing email attacks using natural language processing and machine learning**. This is used to perform the semantic analysis of the text to detect malicious intent. A natural Language Processing (NLP) technique is used to parse each sentence and finds the semantic jobs of words in the sentence in connection to the predicate. In light of the job of each word in the sentence, this strategy recognizes whether the sentence is an inquiry or an order. Supervised machine learning (Peng et al., 2018) is used to generate the blacklist of malicious pairs. Authors defined the algorithm SEA Hound (Peng et al., 2018) for detecting phishing emails and Net-craft Anti-Phishing Toolbar is used to verify the validity of a URL. This algorithm is implemented with Python scripts and the dataset Nazario phishing email set is used. Results of Net-craft and SEA Hound (Peng et al., 2018) are compared and obtained precision 98% and 95% respectively.

The weakness of the paper is that it is limited to a few feature selections and extraction of URL websites.

To address this weakness, new techniques and features will be added or replaced to help detect new phishing URL attacks.

# CHAPTER 3

**METHODOLOGY AND SYSTEMANALYSIS**

# 3.1 Methodology

An extensive review was done on related topics and existing documented materials such as journals, e-books, and websites containing related information gathered which was examined and reviewed to retrieve essential data to better understand and know how to help improve the system.

The methodology used to achieve the earlier stated objectives is explained below.

The dataset collection consists of phishing and legitimate URLs which were obtained from open-source platforms. The dataset was then pre-processed that is cleaned up from any abnormality such as missing data to avoid data imbalance. Afterward, expository data analysis was done on the dataset to explore and summarize the dataset. Once the dataset was free from all anomalies, website content-based features were extracted from the dataset to get accurate features to train and test the model. An extensive review was done on existing works of literature and machine learning models on detecting phishing websites to best decide the classification models to solve the problem of detecting phishing websites. Hence, Series of these machine learning classification models such as Decision Tree, Support Vector Machine, XGBooster, Multilayer perceptions, Auto encoder Neural Network and Random Forest was deployed on the dataset to distinguish between phishing and legitimate URLs. The best model with high training accuracy out of all the deployed models was selected then integrated into a developed web application. Thus, a user can enter a URL link on the web application to predict if it is phishing or legitimate.

# Overview of System Analysis

This chapter describes the various process, methods, and procedures adopted by the researcher to achieve the set aim and objectives and the conceptual structure within which the research was conducted.

The methodology of any research work refers to the research approach adopted by the researcher to tackle the stated problem. Since the efficiency and maintainability of any application are solely dependent on how designs are prepared. This chapter provides detailed descriptions of methods employed to proffer solutions to the stated objectives of the research work.

According to the Merriam-Webster dictionary (11th.Ed), system analysis is "the process of studying a procedure or business to identify its goals and purposes and create systems and procedures that will efficiently achieve them". It is also the act, process, or profession of studying an activity (such as a procedure, a business, or a physiological function) typically by mathematical means to define its goals or purposes and to discover operations and procedures for accomplishing them most efficiently. System analysis is used in every field where the development of something is done. Before planning and development, you need to understand the old systems thoroughly and use the knowledge to determine how best your new system can work.

# Analysis of Existing System

The existing system of phishing detection techniques suffers low detection accuracy and high false alarm especially when different phishing approaches are introduced. Above and beyond, the most common technique used is the blacklist-based method which is inefficient in responding to emanating phishing attacks since registering a new domain has become easier, no comprehensive blacklist can ensure a perfect up-to-date database for phishing detection.

# Proposed System

The proposed phishing detection system utilizes machine learning models and deep neural networks. The system comprises two major parts, which are the machine learning models and a web application. These models consist of Decision Tree, Support Vector Machine, XGBooster, Multilayer Perceptions, Auto Encoder Neural Network, and Random Forest.

These models are selected after different comparison-based performances of multiple machine learning algorithms. Each of these models is trained and tested on a website content-based feature, extracted from both phishing and legitimate dataset.

Hence, the model with the highest accuracy is selected and integrated into a web application that will enable a user to predict if a URL link is phishing or legitimate.

# Benefits of the new system

* 1. Will be able to differentiate between phishing (0) and legitimate (1) URLs.
  2. It will help reduce phishing data breaches for an organization.
  3. It will be helpful to individuals and organizations.
  4. It is easy to use.

# Model Development

The model development method takes several models, tests them, and adds them to an iterative process until a model that meets the required requirements is developed. Figure

* 1. Shows the steps used in the development of machine learning models using both supervised and unsupervised learning.

The following are the stages to machine learning model development for phishing detection systems:

# Data Collection

The data used to generate the datasets on which the models are trained are gotten from different open-source platforms. The dataset collection consists of phishing and legitimate URL dataset.

The set of phishing URLs are collected from an open-source service called Phish Tank. This service provides a set of phishing URLs in multiple formats like CSV, JSON and so on that gets updated hourly. This dataset is accessible from the phishtank.com website. From this dataset, over 5000 random phishing URLs are collected to train the ML models.

The set of legitimate URLs are obtained from the open datasets of the University of New Brunswick, This dataset is accessible on the university website. This dataset has a collection of benign, spam, phishing, malware & defacement URLs. Out of all these types, the benign URL dataset is considered for this project. From this dataset, Over 5000 random legitimate URLs are collected to train the ML models.

# Preprocessing

Data preprocessing is the first and crucial step after data collection. The raw dataset obtained for phishing detection was prepared by removing redundant and irregular data and also encoded using the One-Hot Encoding technique into a useful and efficient format suitable for the machine learning model.

# Exploratory data analysis

Exploratory data analysis (EDA) technique was used on the dataset after series of data cleaning. The data visualization method was employed to analyze, explore and summarize the dataset. This visualization consist of heat-map, histograms, box plots, scatter plots, and pair plots to uncover patterns and insights within data.

# Feature Extraction

Feature Extraction aims to reduce the number of features in a dataset by creating new features from the existing ones. Thus, Website content-based features were extracted from phishing and legitimate datasets such as the Address bar-based feature which consists of 9 features, Domain-based feature which consists of 4 features, and HTML & JavaScript-based Feature which consists of 4 features. So, altogether 17 features were extracted for phishing detection.

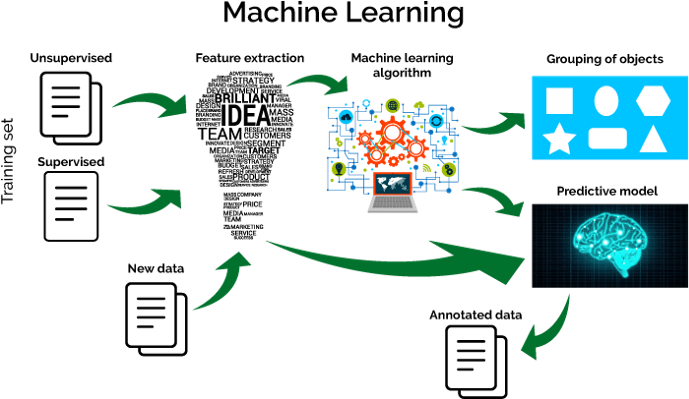


Figure 3.1 Machine Learning development process (Source: Ayush, 2019)

# Model Training

Model Training involves feeding Machine learning algorithms with data to help identify and learn good attributes of the dataset.

This research problem is a product of supervised learning, which falls under the classification problem. The algorithms used for phishing detection consist of supervised machine learning models (4) and deep neural network

(2) Which was used to train the dataset. These algorithms include Decision Tree, Random Forest, Support Vector Machines, XGBooster, Multilayer Perceptron, and Auto-encoder Neural Network. All these models were trained on the dataset. Thus, the dataset is spitted into training and testing set. The training model consists of 80% of the dataset to enable the machine learning models to learn more about the data and be able to distinguish between phishing and legitimate URLs.

# Model Testing

Model Testing involves the process where the performance of a fully trained model is evaluated on a testing set.

Thus, after 80% of data has been trained, 20% of the dataset is used to evaluate the trained dataset to see the performance of the models.

# Model Evaluation

Model Evaluation involves estimating the generalization accuracy of models and deciding whether the model performs better or not.

Thus, Scikit-learn (sklearn matrics) module was used to implements several score and utility functions to measure the classification performance to properly evaluate the models deployed for phishing detection.

# System Modeling

System modeling involves the process of developing an abstract model of a system, with each model presenting a different view or perspective of the system. It is the process of representing a system using various graphical notations that shows how users will interact with the system and how certain parts of the system function. The proposed system was modeled using the following diagrams:

* + 1. Architectural diagram
    2. Use case diagram
    3. Flowcharts

The proposed system will be implemented using Python Programming language along with different machine learning models and libraries such as pandas, scikit-learn python who-is, beautiful-Soup, NumPy, seaborn, and matplotlib. Etc.

**3.6 System architecture**

Architectural design is concerned with understanding how a system should be organized and designing the overall structure of that system, it shows how different components of the system work together to achieve its main objectives. It is the process for identifying sub-systems making up a system and the framework for sub-system control and communication. The diagram below represents a graphical overview of the architectural design of the proposed system.

Figure 3.1 shows the architecture view of the proposed phishing detection system such that a user enters a URL link and the link moves through different trained machine learning and deep neural network models and the best model with the highest accuracy is selected. Thus, the selected model is deployed as an API (Application Programming Interface) which is then integrated into a web application. Hence, a user interacts with the web application which is accessible across different display devices such as computers, tablets, and mobile devices.

# Use a case diagram of the system

The Use Case diagram describes the functionality of the system as designed from the requirements; it summarizes the details of a system and the users within the system. It is a behavior diagram and visualizes the observable interactions between actors and the system under development. The Use case diagram consists of the system, the related use cases, and actors and relates to each other.

Figure 3.2 shows the Use case scenarios that a user can carry out on the phishing detection system.

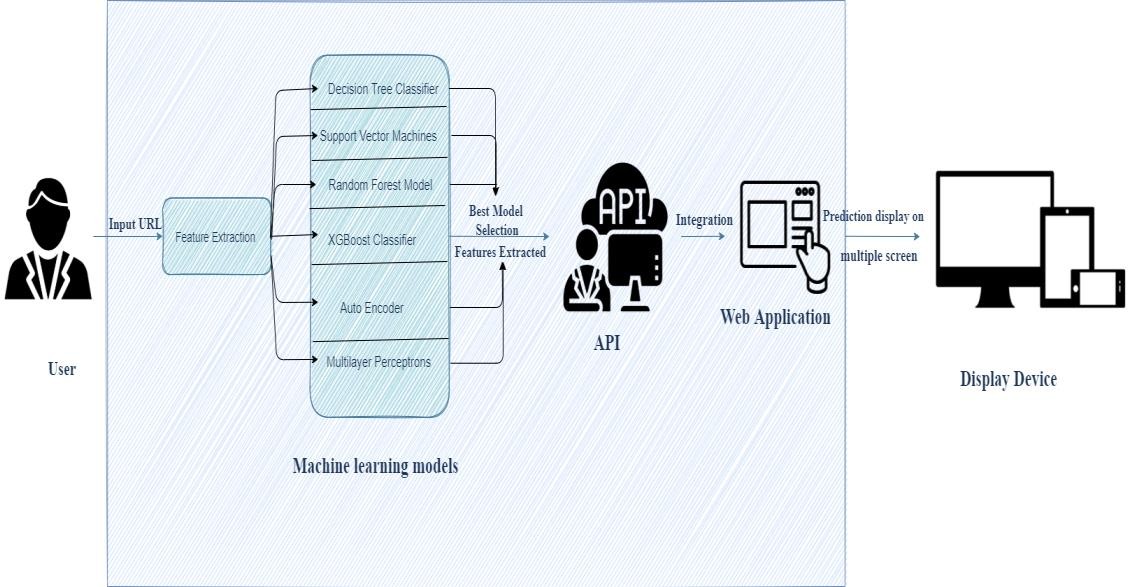


Figure 3.2: Architectural Design of the Proposed System

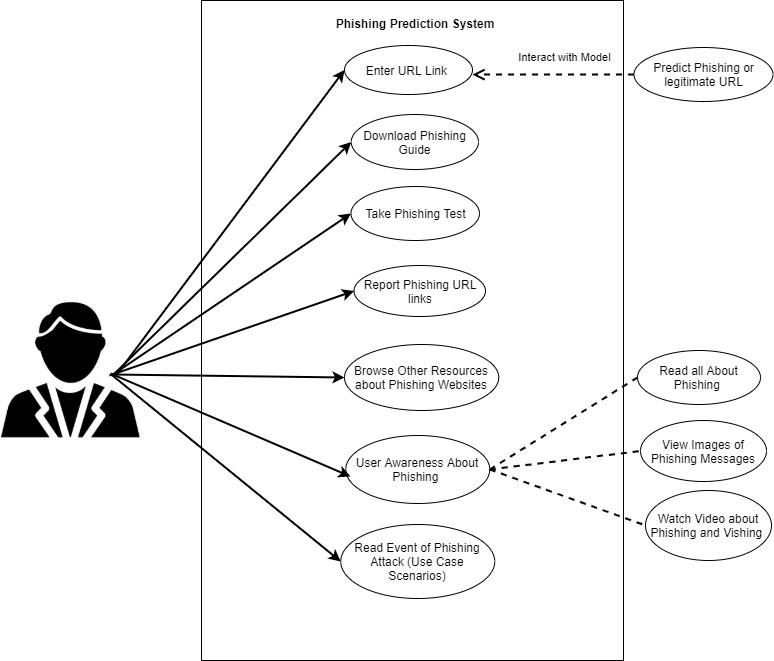


Figure 3.3 Use Case diagram for Proposed System

# Flowchart of the system

A flowchart is a diagram that depicts a process, system, or computer algorithm. It is a graphical representation of the steps that are to be performed in a system; it shows the steps in sequential order. It is used in presenting the flow of algorithms and to communicate complex processes in clear, easy-to-understand diagrams.

Figure 3.3 shows the flow of phishing detection systems using the machine learning process.

Figure 3.4 shows the phishing detection web interface system. The user inputs a URL link and the website validates the format of the URL and then predicts if the link is phishing or legitimate.

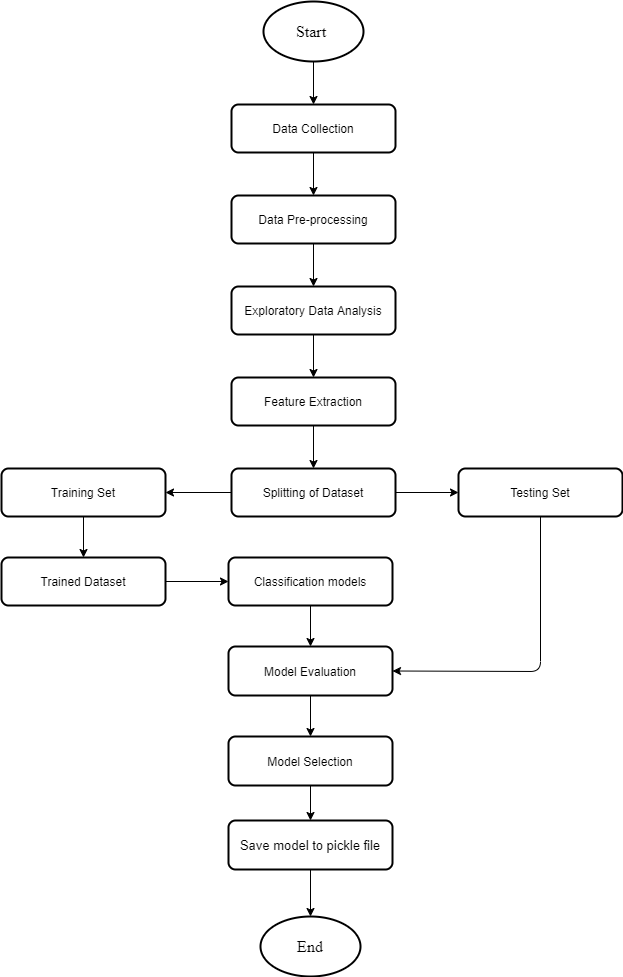


Figure 3.4 Flowchart of the proposed System

# CHAPTER FOUR

# SYSTEM IMPLEMENTATION AND RESULTS

This chapter deals with the implementation of multiple machine learning models for the detection of underlying diseases and illnesses as earlier designed in the preceding chapters.

The implementation is concerned with all the activities that took place to put up the newly developed system into operation (using the approach that was stated in the methodology (e.g. architectural diagram, flowchart, uses case, etc.).) to achieve the objectives of the project to convert the theoretical design into a working system. The components of the system were also tested and evaluated.

# Installation Requirements

The hardware (physical components of a computer system that can be seen, touched, or felt) and software (both system software and the application software installed and used in the system development) tools needed to satisfy these objectives highlighted below:

# Hardware requirements

The hardware requirement includes:

* + - 1. A laptop or desktop computer (Preferably 64bit)
      2. Random Access Memory (RAM): 8 Gigabytes Minimum
      3. Processor: Intel Core i5, 2.4 GHz Minimum

# Software requirements

The software requirements for the development of this system include:

* + - 1. Windows Operating System (8/10/11)
      2. Anaconda Navigator (Jupyter Notebook)
      3. PyCharm Community edition
      4. Web browser (Preferably Chrome)
      5. Visual Studio Code
      6. Postman

# Model Development

The model for detecting phishing URL websites was built using a python programming language with over six (6) machine learning models and deep neural network algorithms altogether and the most accurate test score on the tested 5,000 datasets were used.

# Data collection

The dataset used for the classification was sourced from was gotten from multiple sources listed in the earlier stated methodology.

The dataset used for classifying the dataset into phishing and legitimate URLs was sourced from open source websites, samples of which are shown below in figure 4.1 and 4.2 respectively.

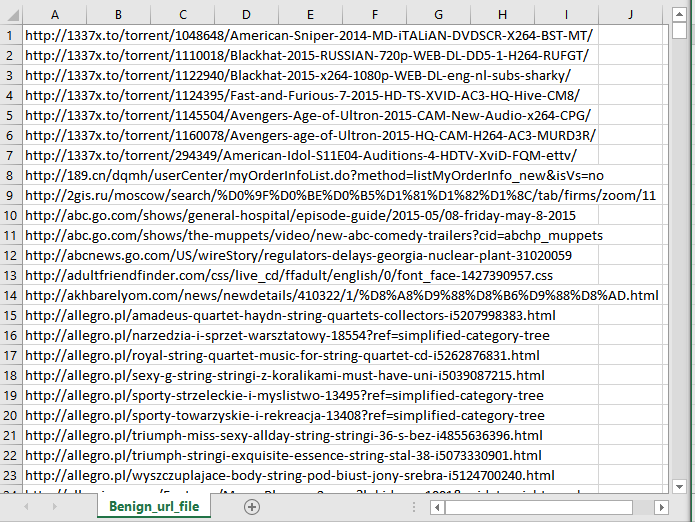


Figure 4.1 Dataset of Phishing URLs

Source: The Dataset is collected from an open-source service called Phish-Tank. This dataset consists of 5,000 random phishing URLs which are collected to train the ML models.

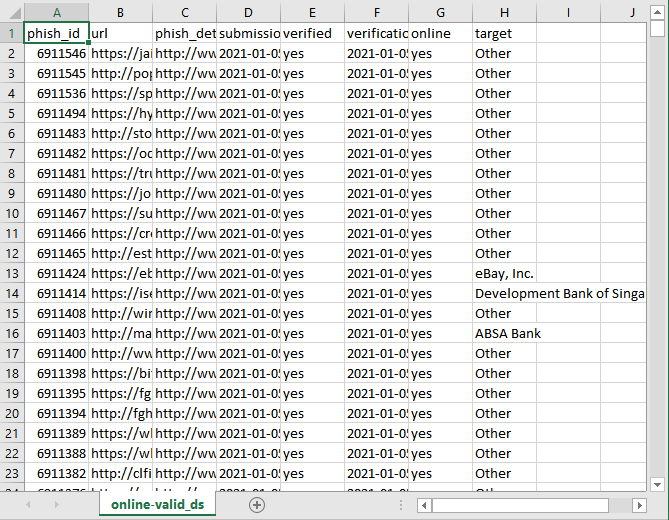


Figure 4.2 Dataset of Legitimate URLs

Source: The Dataset were obtained from the open datasets of the University of New Brunswick, The dataset consists of collections of benign, spam, phishing, malware & defacement URLs. Out of all these types, the benign URL dataset is considered for this project. This dataset consists of 5,000 random legitimate URLs which are collected to train the ML models.

# Feature extraction on the datasets

The features extraction used on the dataset are categorized into:

* + - 1. Address bar based features
      2. Domain-based features
      3. Html & java-script based features

# General Working of the System

A one-page phishing detection web application called “Phish-Buster” has been developed to run on any browser. The application was developed using programming languages such as HTML, CSS, PHP, and JavaScript.

The phishing detection web application has the following pages:

# The home page

The home page contains a session for a user to enter a URL and predict if it is phishing or legitimate. It predicts the state of the URL base on the feature selection as shown in figure 4.14.

The purpose of this page is to help its users validate a URL link and also provide various resources on phishing attacks. The User can also take a Google phishing test to help understand how to detect phishing messages and URLs. Also, users can download a book that contains information and other resources on phishing.

# The about page

The about page contains details about the application and also information about the author of the project as shown in figure 4.15

# The use case page

The use case page contains different case scenarios of phishing attacks that happened to various companies in previous years back. Also, it contains images of phishing attacks messages sent by phishers as shown in figure 4.16

# Resource page

The resource page it contains different resources regarding phishing such as definition, types, and techniques of a phishing attack as well as reference links to the source in which the content where retrieved from Also, it contains two (2) sub-session link: the first session reports phishing case and the second session consists of a phishing website.

The First Session: The Report phishing Case link redirects users to submit any case of phishing attack be it URL or a phishing email to Google Support and Google safe browsing.

The Second Session: Phishing website link redirects users to two (2) websites that contain resourceful information and phishing test. These sites consist of phishing box and intradyn as shown in Figure 4.17.

# Web application Source Code

As shown from figure 4.18, consists of pages of source code of the web application running on visual studio code, other source codes are shown in Appendix A



Figure 4.14 (a): The Home page

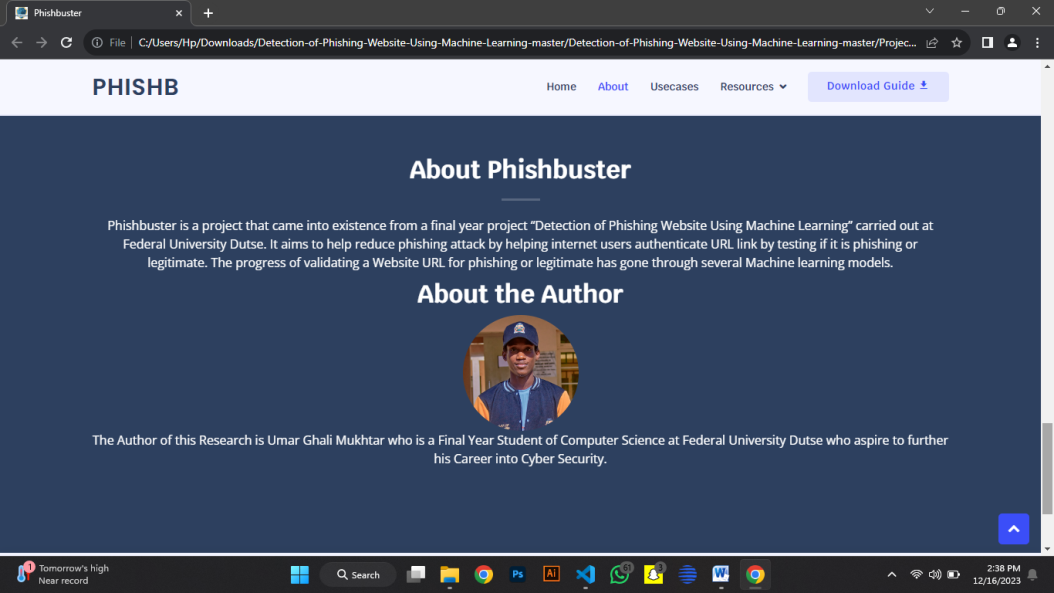


Figure 4.15: The About page



Figure 4.16 (a): The Use-case page

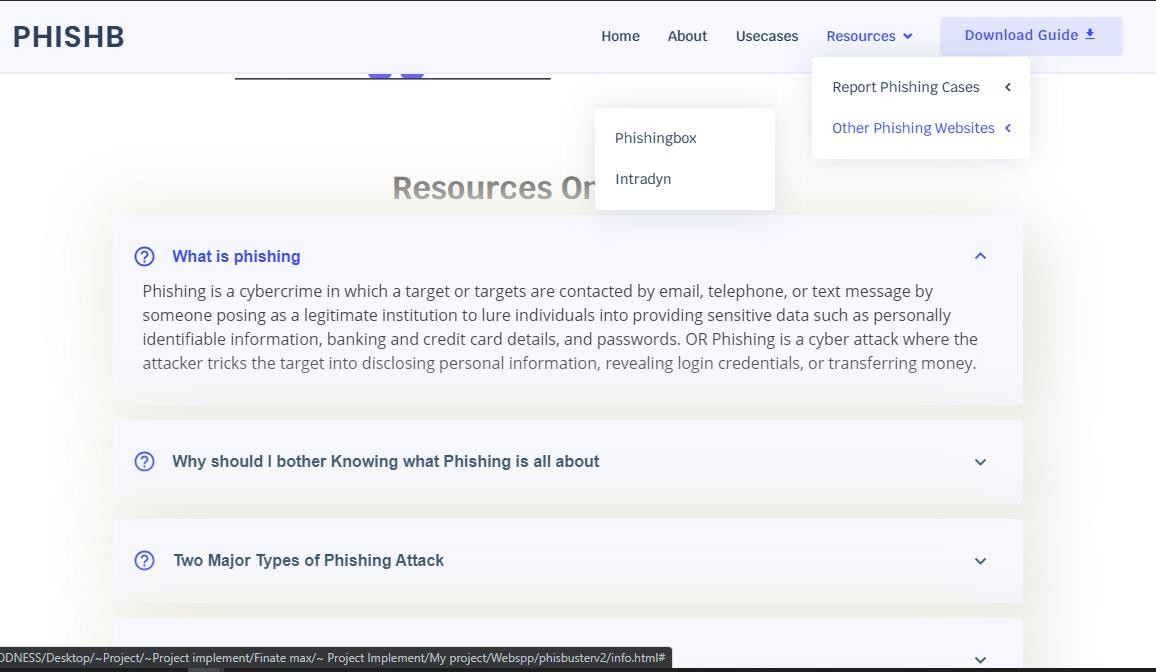
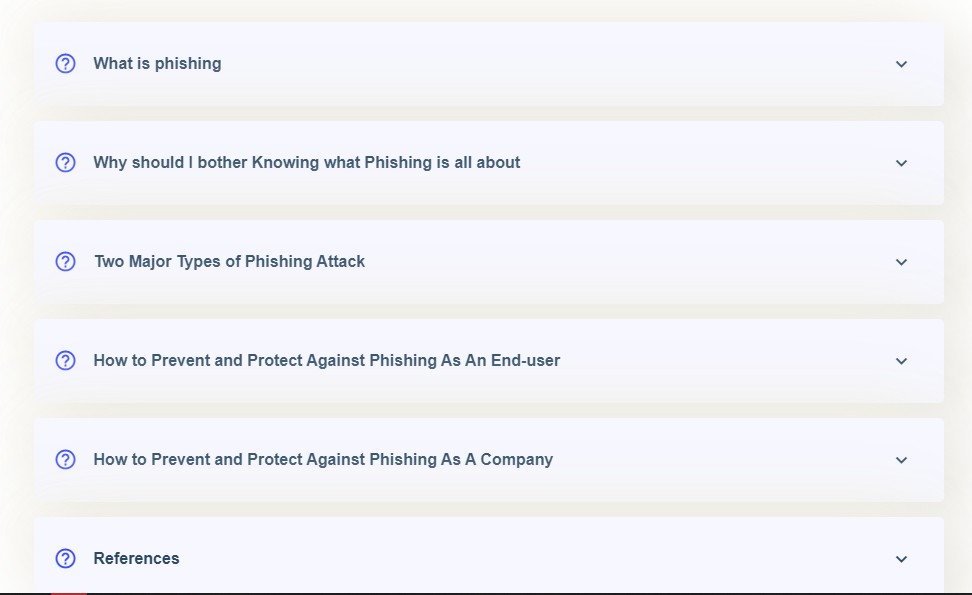
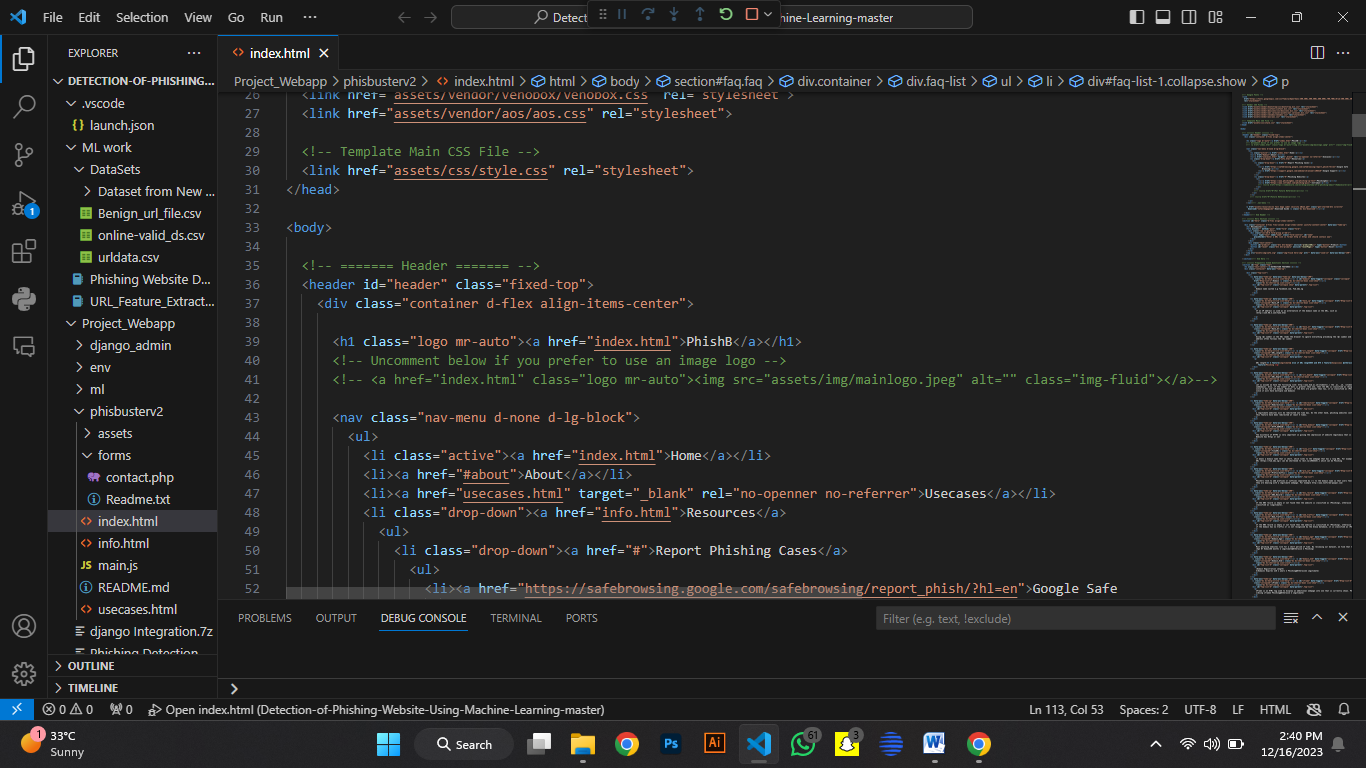


Figure 4.17: The Resource page



# Figure 4.18: Code for the web application

# CHAPTER FIVE

# SUMMARY, CONCLUSION, AND RECOMMENDATIONS

# Summary

Phishing attacks are a rapidly expanding threat in the cyber world, costing internet users billions of dollars each year. It involves the use of a variety of social engineering tactics to obtain sensitive information from users. Hence, Phishing techniques can be detected using a variety of types of communication, including email, instant chats, pop-up messages, and web pages.

This project was able to categorize and recognize how phishers carry out phishing attacks and the different ways in which researchers have helped to solve phishing detection. Hence, the proposed system of this project worked with different feature selection and machine learning and deep neural networks such as Decision Tree, Support Vector Machine, XGBooster, Multilayer Perceptions, Auto Encoder Neural Network, and Random Forest to identify patterns in which URL links can be detected easily.

The Model with the highest accuracy based on the feature extraction algorithm used to identify phishing URL from legitimate URL links was integrated to a web application where users can input website URL links to detect if it is legitimate or phishing.

# Contribution to Knowledge

This project provides a new and faster way to help users detect if a URL link is phishing or legitimate and also provide them access to educational resources about phishing attacks.

# Conclusion

The system developed detects if a URL link is phishing or legitimate by using machine learning models and deep neural network algorithms. The feature extraction and the models used on the dataset helped to uniquely identify phishing URLs and also the performance accuracy of the models used.

# Recommendation

Through this project, one can know a lot about phishing attacks and how to prevent them. This project can be taken further by creating a browser extension that can be installed on any web browser to detect phishing URL Links.

# REFERENCE

Abdelhamid, N., Thabtah F., & Abdel-Jaber, H. Phishing detection: A recent intelligent machine learning comparison based on models’ content and features," 2017 IEEE International Conference on Intelligence and Security Informatics (ISI), Beijing, 2017, pp. 72-77, DOI: 10.1109/ISI.2017.8004877.

Anjum N. S., Antesar M. S., & Hossain M.A. (2016). A Literature Review on Phishing Crime, Prevention Review and Investigation of Gaps. Proceedings of the 10th International Conference on Software, Knowledge, Information Management & Applications (SKIMA), Chengdu, China, 2016, pp. 9-15, DOI: 10.1109/SKIMA.2016.7916190.

Almomani, A., Gupta, B. B., Atawneh, S., Meulenberg, A., & Almomani, E. (2013). A survey of phishing email filtering techniques, Proceedings of IEEE Communications Surveys and Tutorials, vol. 15, no. 4, pp. 2070–2090.

Ashritha, J. R., Chaithra, K., Mangala, K., & Deekshitha, S. (2019). A Review Paper on Detection of Phishing Websites using Machine Learning.Proceedings of International Journal of Engineering Research & Technology (IJERT), 7, 2. Retrieved from [www.ijert.org.](http://www.ijert.org/)

Anti-Phishing Working Group (APWG) Phishing activity trends report the first quarter. (2014) Retrieved from [http://docs.apwg.org/reports/apwg trends report q1](http://docs.apwg.org/reports/apwg%20trends%20report%20q1%202014.pdf)

[2014.pdf](http://docs.apwg.org/reports/apwg%20trends%20report%20q1%202014.pdf)

APWG report. (2014). Retrieved from <http://apwg.org/download/document/245/>APWG Global Phishing Report 2H

2014.pdf

Ayush, P. (2019). Workflow of a Machine Learning project. Retrieved from [https://towardsdatascience.com/workflow-of-a-machine-learning-project-](https://towardsdatascience.com/workflow-of-a-machine-learning-project-ec1dba419b94)

[ec1dba419b94](https://towardsdatascience.com/workflow-of-a-machine-learning-project-ec1dba419b94)

Camp W. (2001). Formulating and evaluating theoretical frameworks for career and technical education research. Journal of Vocational Education Research, 26(1), 4- 25.

DeepAI (n.d.). About clinical psychology. Retrieved from [https://deepai.org/machine-](https://deepai.org/machine-learning-glossary-and-terms/feature-extraction)

[learning-glossary-and-terms/feature-extraction](https://deepai.org/machine-learning-glossary-and-terms/feature-extraction)

Engine K., & Christopher K. (2005). Protecting Users Against Phishing Attacks. Proceedings of the Oxford University Press on behalf of The British Computer Society, Oxford University, 0, 2005, Retrieved from: <https://sites.cs.ucsb.edu/~chris/research/doc/cj06_phish.pdf>

Gandhi, V. (2017). A Theoretical Study on Different ways to identify the Phishing URL and Its Prevention Approaches: presented at International Conference on Cyber Criminology, Digital Forensics and Information Security at DRBCCC Hindu College, Chennai. Retrieved from [https://www.researchgate.net/publication/319006943\_A\_Theoretical\_Study\_on\_](https://www.researchgate.net/publication/319006943_A_Theoretical_Study_on_Different_ways_to_Identify_the_Phishing_URL_and_Its_Prevention_Approaches)

[Different\_ways\_to\_Identify\_the\_Phishing\_URL\_and\_Its\_Prevention\_Approaches](https://www.researchgate.net/publication/319006943_A_Theoretical_Study_on_Different_ways_to_Identify_the_Phishing_URL_and_Its_Prevention_Approaches)

Gupta, B. B., Tewari, A., Jain, A. K., & Agrawal, D. P. (2016). Fighting against phishing attacks: state of the art and future challenges, Neural Computing and Applications.

Internet world stats usage and population statistics. (2014). Retrieved from <http://www.internetworldstats.com/stats.htm>

Kiruthiga, R., Akila, D. (2019, September). Phishing Websites Detection Using Machine Learning. Retrieved from [https://www.researchgate.net/publication/337049054 Phishing Websites Detection](https://www.researchgate.net/publication/337049054%20Phishing%20Websites%20Detection%20Using_Machine_Learning)

[Using\_Machine\_Learning](https://www.researchgate.net/publication/337049054%20Phishing%20Websites%20Detection%20Using_Machine_Learning).

KnowBe4 (2021). Phishing Techniques. Retrieved from <https://www.phishing.org/phishing-techniques>

Kondeti, P. S., Konka, R. C., & Kavishree, S. (2021). Phishing Websites Detection using Machine Learning Techniques. International Research Journal of Engineering and Technology, 08(4), Page 1471-1473. Retrieved from <https://www.irjet.net/archives/V8/i4/IRJET-V8I4274.pdf>

Noel, B. (2016). Support Vector Machines: A Simple Explanation. Retrieved from [https://www.kdnuggets.com/2016/07/support-vector-machines-simple-](https://www.kdnuggets.com/2016/07/support-vector-machines-simple-explanation.html)

[explanation.html](https://www.kdnuggets.com/2016/07/support-vector-machines-simple-explanation.html)

Osanloo, A., & Grant, C. (2016). Understanding, selecting, and integrating a theoretical framework in dissertation research: creating the blueprint for your “house”. Administrative issues journal: connecting education, practice and research 4(2), 7.

Peng, T., Harris, I., & Sawa, I. (2018). Detecting Phishing Attacks Using Natural Language Processing and Machine Learning. Proc. - 12th IEEE Int. Conf. Semant. Comput. ICSC 2018, vol. 2018–Janua, pp. 300–301.

Rami, M. M., Fadi, T., & Lee, M. (2015). Phishing Websites Features. Retrieved from <https://eprints.hud.ac.uk/id/eprint/24330/6/MohammadPhishing14July2015.pdf>

Rishikesh, M., & Irfan, S. (2018a). Phishing Website Detection using Machine Learning Algorithms. International Journal of Computer Applications, 23, 45. doi:10.5120/ijca2018918026

Rishikesh, M., & Irfan, S. (2018b). Phishing Website Detection using Machine Learning Algorithms. International Journal of Computer Applications, 23, 45-46. doi:10.5120/ijca2018918026

Rahul, S. (2017). How the decision tree algorithm works. Retrieved from <https://dataaspirant.com/how-decision-tree-algorithm-works/>

Rishikesh, M., & Irfan, S. (2018c). Phishing Website Detection using Machine Learning Algorithms. International Journal of Computer Applications, 23, 46-47. doi:10.5120/ijca2018918026

Saimadhu, P. (2017). How the random forest algorithm works in machine learning. Retrieved from [https://dataaspirant.com/random-forest-algorithm-machine-](https://dataaspirant.com/random-forest-algorithm-machine-learing/)

[learing/](https://dataaspirant.com/random-forest-algorithm-machine-learing/)

Shaikh, A.N., Shabut, A.M., Hossain, M.A. (2016, December 15-17). A literature review on phishing crime, prevention review, and investigation of gaps. Paper presented at the Tenth International Conference on Software, Knowledge,

Shreya, G. (2020). Phishing website detection by machine learning techniques. Retrieved from [https://github.com/shreyagopal/Phishing-Website-Detection-by-](https://github.com/shreyagopal/Phishing-Website-Detection-by-Machine-Learning-Techniques)

[Machine-Learning-Techniques](https://github.com/shreyagopal/Phishing-Website-Detection-by-Machine-Learning-Techniques)

Sönmez, Y., Tuncer, T., Gökal, H., & Avci, E. (2018). Phishing web sites features classification based on extreme learning machine. 6th Int. Symp. Digit. Forensics Secure. ISDFS 2018 - Proceeding, vol. 2018–Janua, pp. 1–5.

RSA Anti-Fraud Command Center (n.d.) Retrieved from <https://www.emc.com/collateral/fraud-report/rsa-online-fraud-report-012014.pdf>.

Shad, J., & Sharma, S. (2018). A Novel Machine Learning Approach to Detect Phishing Websites Jaypee Institute of Information Technology, pp. 425–430.

The RSA Current State of Cybercrime. (n.d.). Retrieved from [https://www.rsa.com/en-](https://www.rsa.com/en-us/perspectives/industry/online-fraud)

[us/perspectives/industry/online-fraud.](https://www.rsa.com/en-us/perspectives/industry/online-fraud)

Tewari, A., Jain, A. K, & Gupta, B. B. (2016). A recent survey of various defense mechanisms against phishing attacks. Journal of Information Privacy and Security, vol. 12, no. 1, pp. 3–13.

Joachim, S. (n.d). Missing Value Imputation (Statistics) – How to Impute Incomplete Data. Retrieved from [https://statisticsglobe.com/missing-data-imputation-](https://statisticsglobe.com/missing-data-imputation-statistics)

[statistics](https://statisticsglobe.com/missing-data-imputation-statistics)

Kartik, M. (2021). Everything You Need to Know About Feature Selection in Machine Learning. Retrieved from [https://www.simplilearn.com/tutorials/machine-](https://www.simplilearn.com/tutorials/machine-learning-tutorial/feature-selection-in-machine-learning)

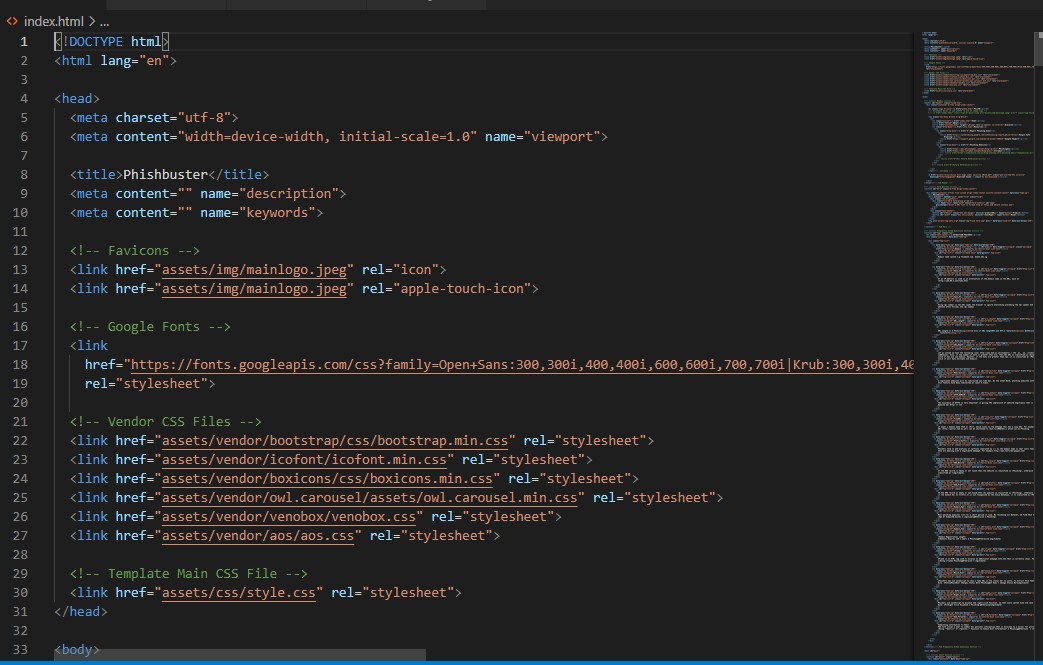
[learning-tutorial/feature-selection-in-machine-learning](https://www.simplilearn.com/tutorials/machine-learning-tutorial/feature-selection-in-machine-learning)

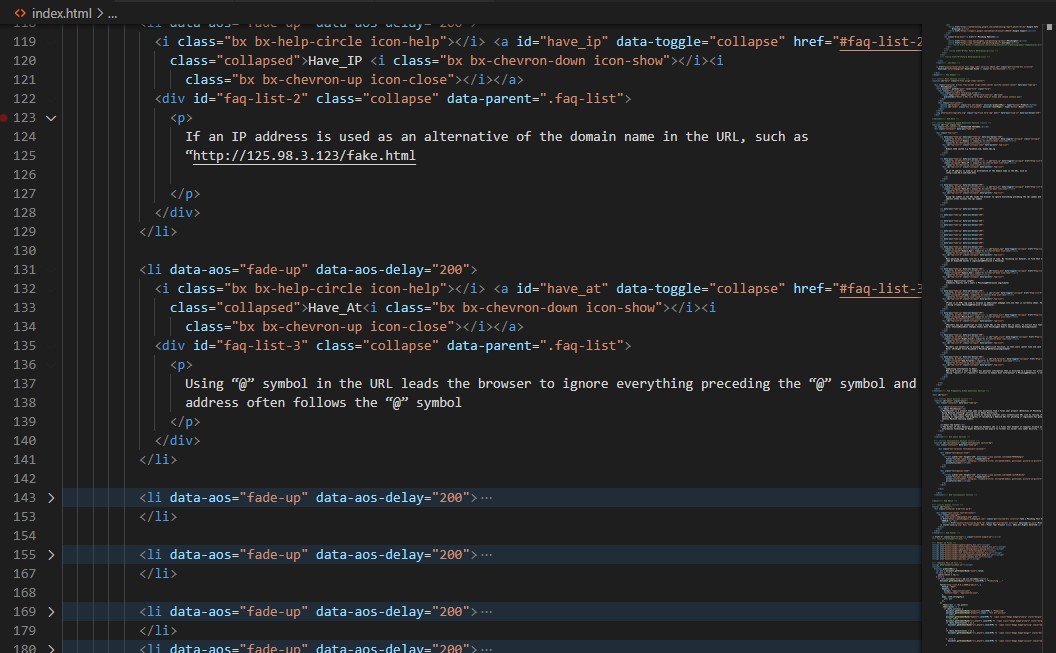
Will, B. (2019). 6 Different Ways to Compensate for Missing Values In a Dataset (Data Imputation with examples). Retrieved from [https://towardsdatascience.com/6-](https://towardsdatascience.com/6-different-ways-to-compensate-for-missing-values-data-imputation-with-examples-6022d9ca0779)

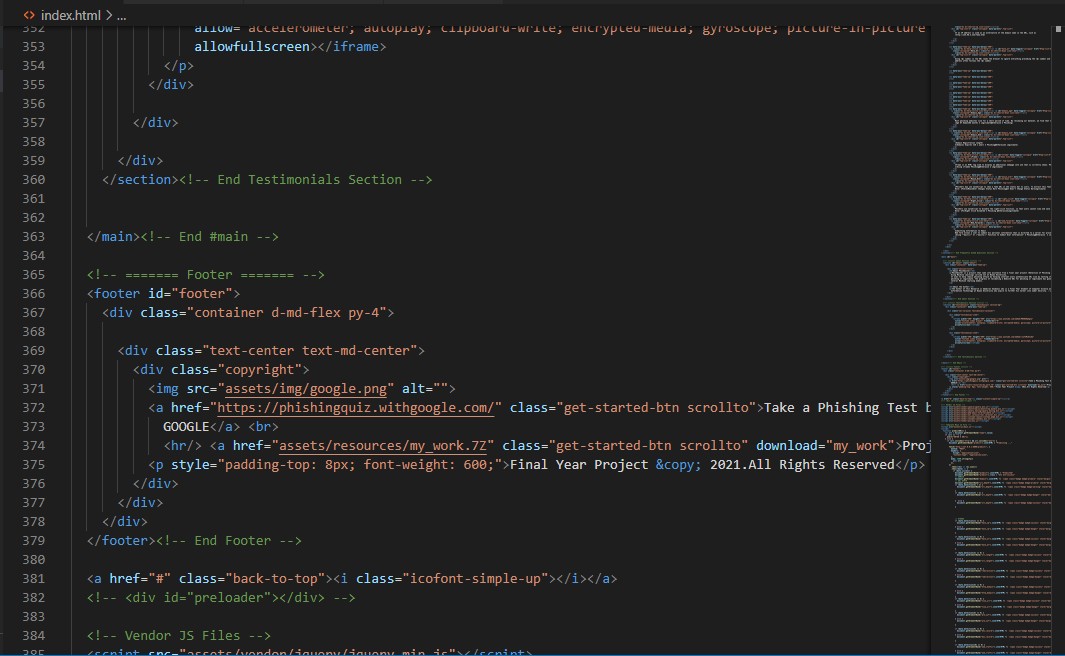
[different-ways-to-compensate-for-missing-values-data-imputation-with-](https://towardsdatascience.com/6-different-ways-to-compensate-for-missing-values-data-imputation-with-examples-6022d9ca0779)

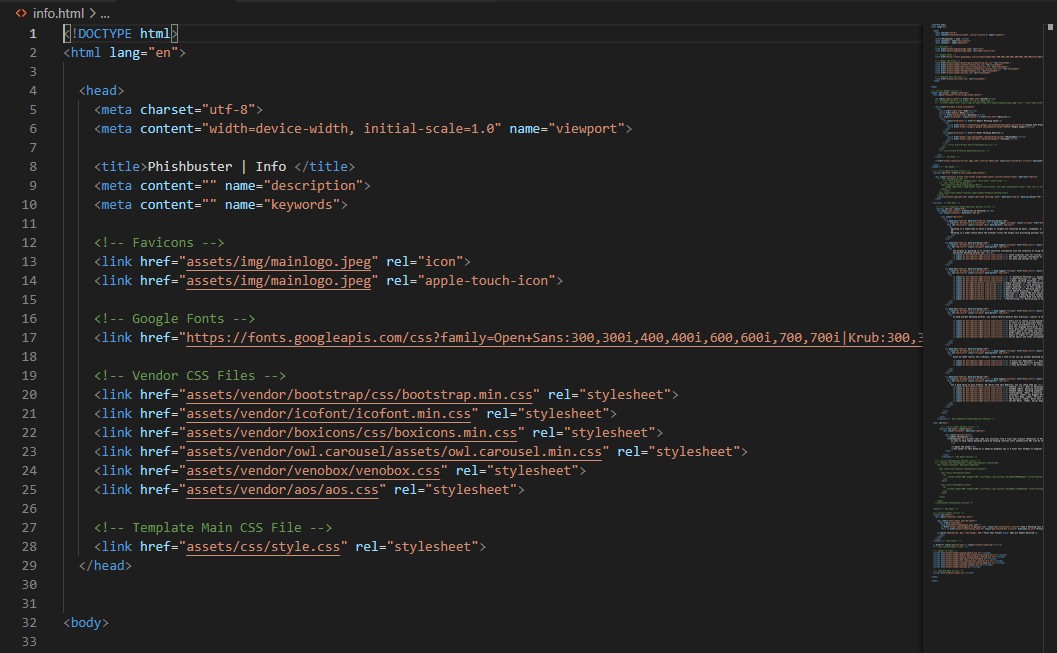
[examples-6022d9ca0779](https://towardsdatascience.com/6-different-ways-to-compensate-for-missing-values-data-imputation-with-examples-6022d9ca0779)

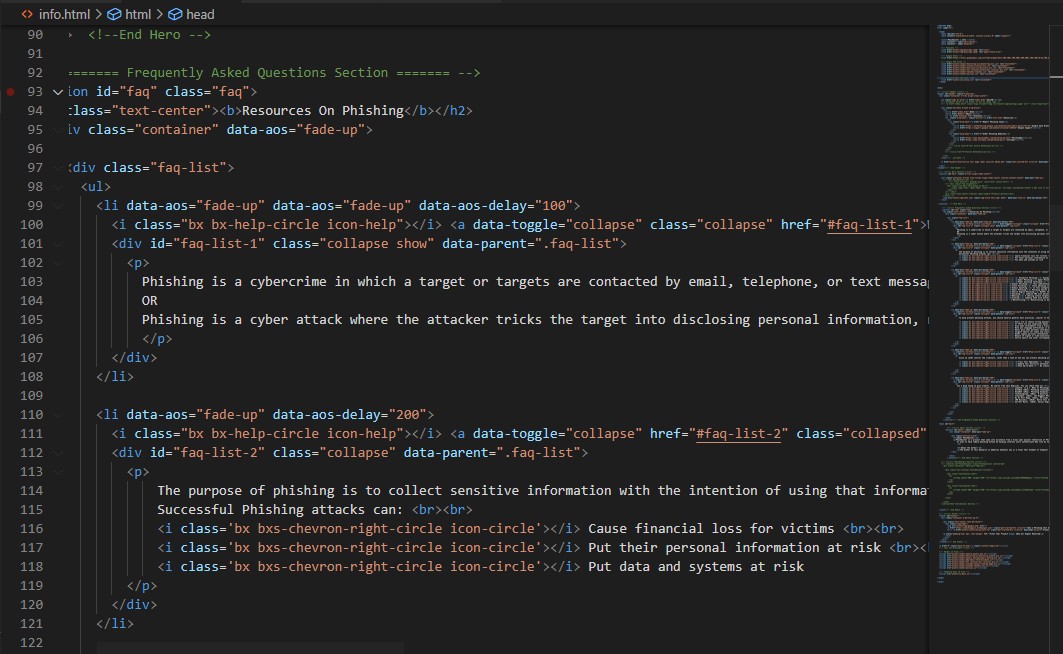
# APPENDIX A: Web application Source code

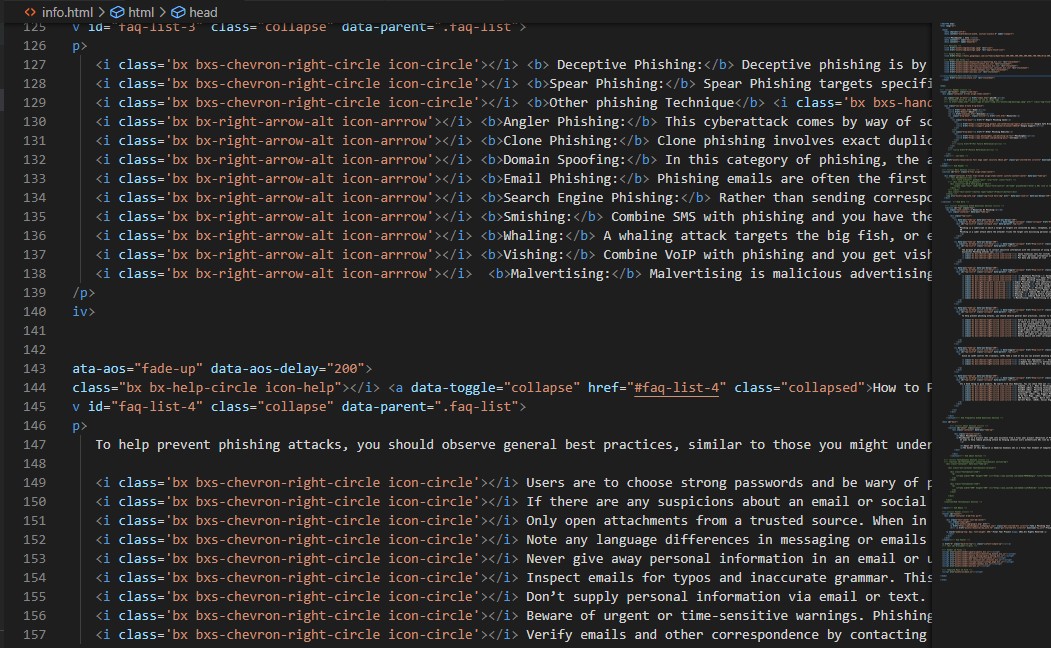


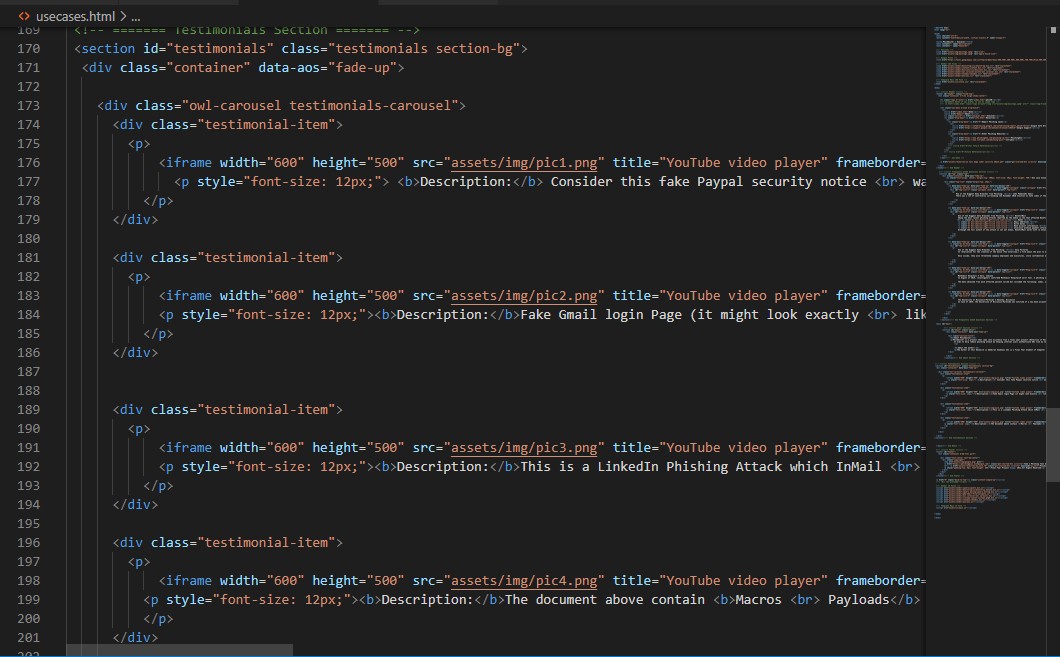


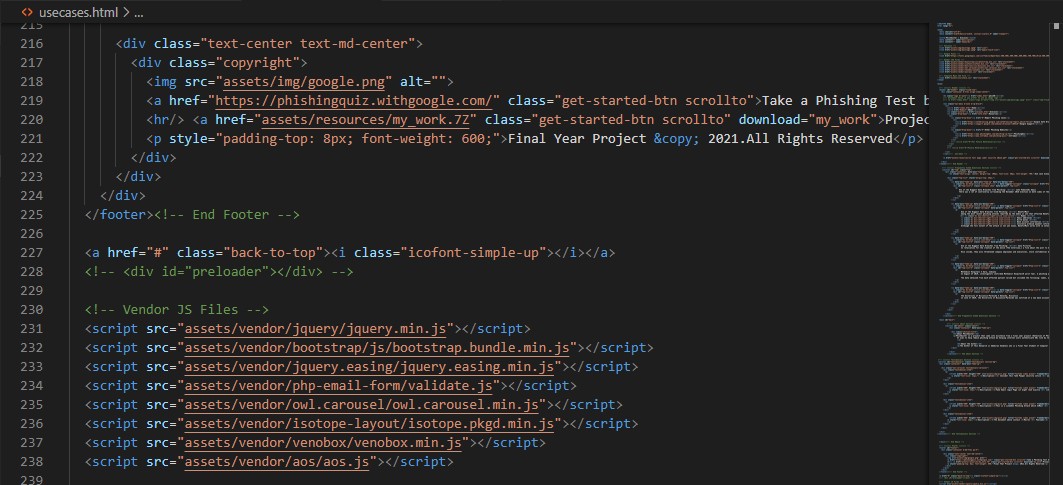


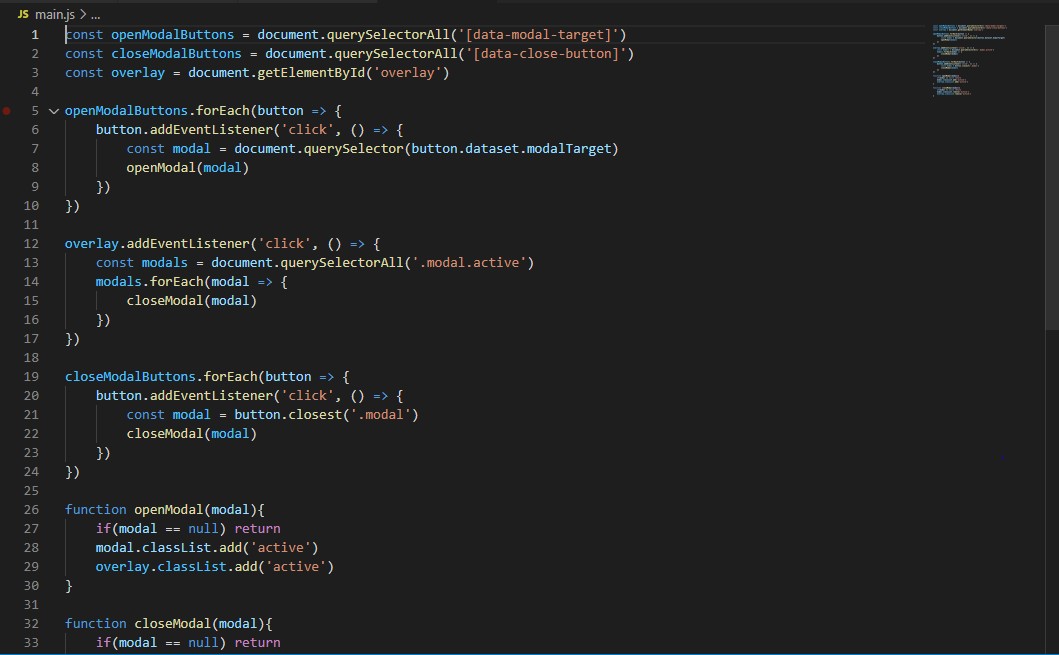




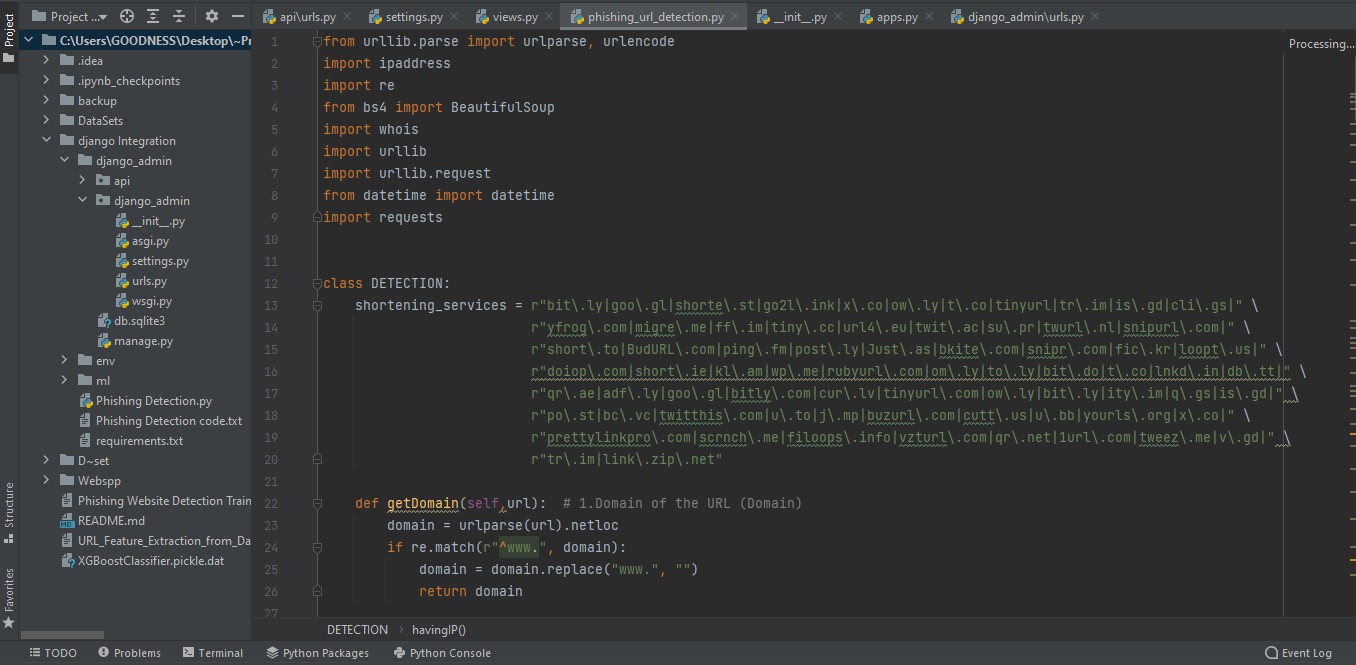


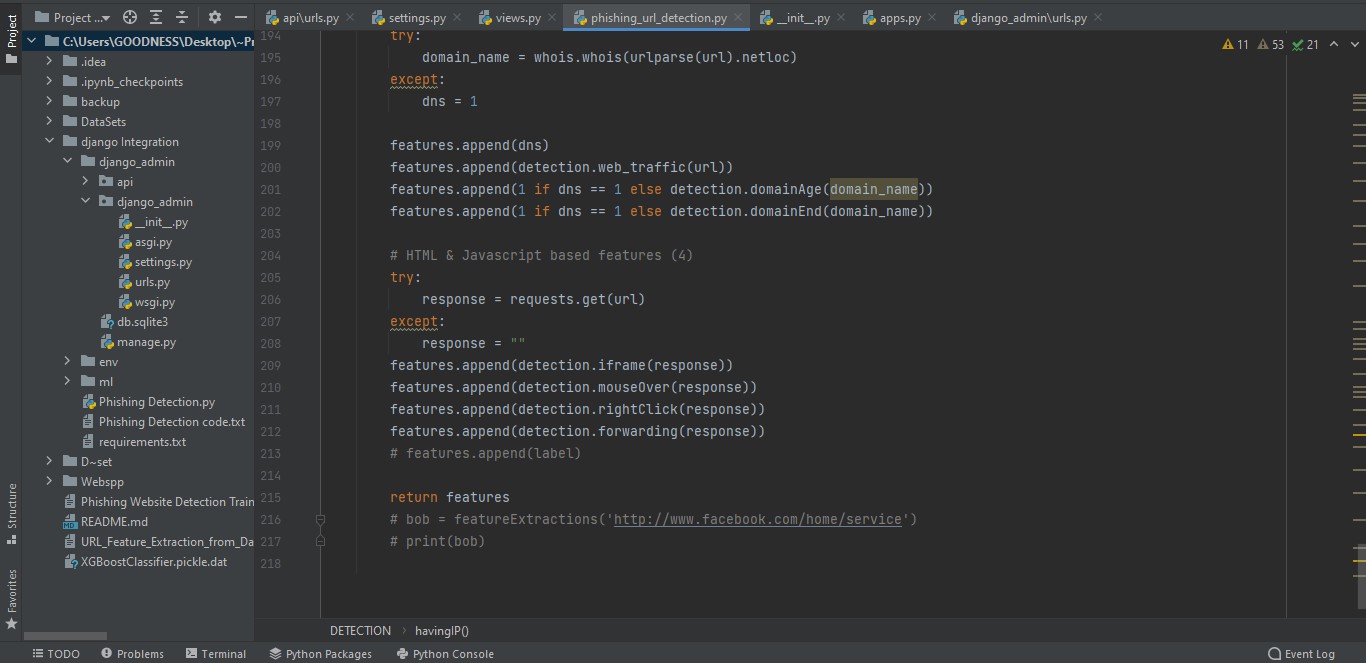


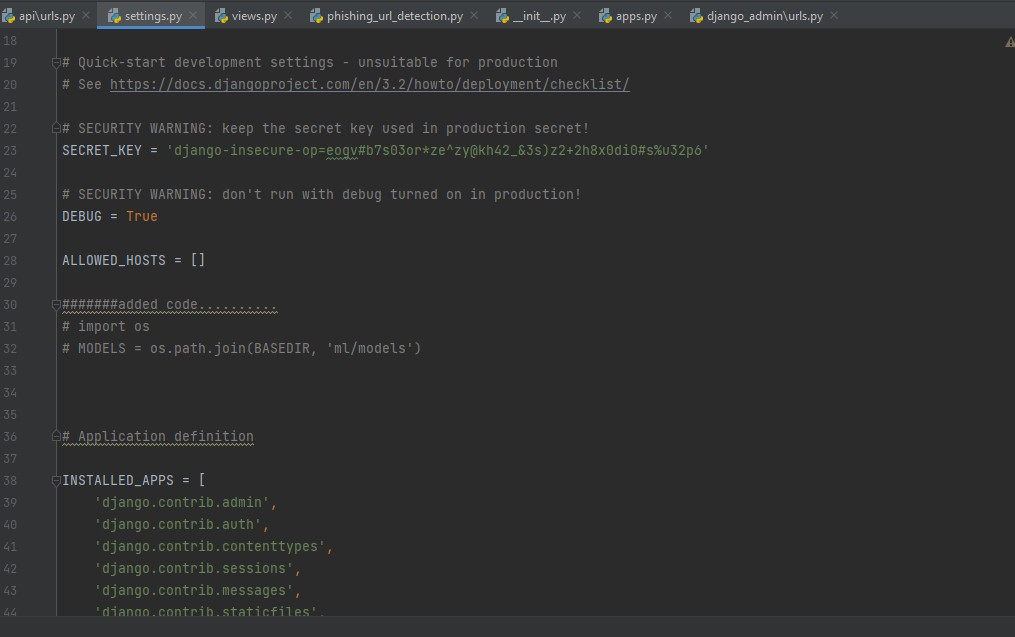




# APPENDIX B: Source code for API integration to web app







# 