

PHISHING URL DETECTION A COMPARATIVE STUDY USING MACHINE LEARNING MODELS

A Project Report

*submitted in partial fulfilment of the
requirements for the Award of Degree of*

Bachelor of Technology

in

COMPUTER SCIENCE & ENGINEERING

by

Purnima Painkra

Roll No.: 20115072

Under the guidance of

Dr. Dilip Singh Sisodia

Associate Professor



DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

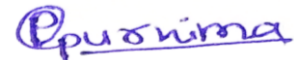
National Institute of Technology Raipur, CG(INDIA)

December 2023

DECLARATION

I, hereby declare that the work described in this report, entitled “**Phishing URL Detection Using Machine Learning Technique**” which is being submitted by me in partial fulfillment of the award of the degree of **Bachelor of Technology in Computer Science and engineering** to the department of CSE, National Institute of Technology, Raipur is the result of investigations carried out by us under the guidance of **Dr. Dilip Singh Sisodia**.

The work is original and has not been submitted for any Degree/Diploma of this or any other Institute/university.



Signature

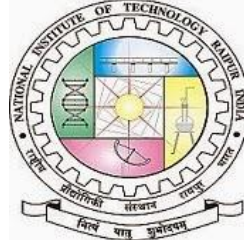
Purnima Painkra

Roll No.: 20115072

Place : Raipur

Date : 8 Dec 2023

**DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING
NATIONAL INSTITUTE OF TECHNOLOGY RAIPUR**



CERTIFICATE

This is to certify that the project entitled “**Phishing URL Detection a Comparative study Using Machine Learning Models**”, that is being submitted by **Purnima Painkra (Roll No. 20115072)** in partial fulfillment of the requirements for the award of the degree of **Bachelor of Technology in Computer Science & Engineering** to National Institute of Technology Raipur is a record of bonafide work carried out by them under my guidance and supervision.

The matter presented in this project document has not been submitted by them for the award of any other degree/diploma elsewhere.

Signature of Supervisor

Dr. Dilip Singh Sisodia

Assistant Professor

Department of Computer Science & Engineering

National Institute of Technology, Raipur (CG.)

Signature of Project Coordinator

Dr. K. Jairam Naik

Department of C.S.E

NIT Raipur (CG.)

Signature of H.O.D.
21/12/2018
H.O.D.

Dr. Pradeep Singh

Department of C.S.E

NIT Raipur (CG.)

ACKNOWLEDGEMENT

We would like to acknowledge our college **National Institute of Technology, Raipur** for providing a holistic environment that nurtures creativity and research-based activities.

We express our sincere thanks to **Dr. Dilip Singh Sisodia, Assistant Professor CSE Department, NIT Raipur**, the supervision of the project for guiding and correcting throughout the process with attention and care. He has frequently suggested us creative ideas and guided through the major hurdles that occurred during the duration of the project.

We would also thank **Dr. Pradeep Singh, Head of the department** and all the faculty members without whom this project would be a distant reality. We also extend our heartfelt thanks to my family and friends who supported us.

Thank You!!

Purnima Painkra

Roll No. 20115072

CSE – 7th sem

ABSTRACT

Due to the quick advancement of technology, a growing number of people now work remotely and use social media more frequently. There is a rise in mobile phishing campaigns and attacks because the majority of internet-connected devices do not have sufficient security safeguards. The quantity and variety of web services available on the Web have increased dramatically over the past few years. As people carry out daily tasks, web services like social networking, online gaming, and banking have quickly developed. Consequently, a substantial volume of data is posted to the Internet every day. These online services create more opportunities for people to interact, but they also give criminals new avenues to operate.

The simplest method of obtaining personal information from careless users is through phishing attacks. Phishers attempt to obtain personal data, including bank account details, usernames, and passwords. Phishing causes large losses for internet users annually. This is an incredible technique for obtaining personal and financial information from Internet users. Cybersecurity experts are currently looking for reliable and robust methods to detect phishing websites. Phishers often create URLs that look real, such as those from reputable banks or online stores. Clicking on these links takes users to fake websites that look like real websites. User data entered on these bogus websites can be obtained by fraudsters, who can then use it in their own fraudulent schemes. To identify phishing URLs, this article describes a machine learning method that extracts and identifies legitimate phishing URLs. Support vector machine algorithms, random forests, and decision trees are examples of phishing site identification techniques. To find the best learning algorithm, this article evaluates the false positive, false negative, and accuracy rates of each algorithm for phishing URL detection.

CONTENT

DESCRIPTION

Page No.

Contents

DECLARATION	ii
<i>CERTIFICATE</i>	iii
ABSTRACT	v
CONTENT	vi
LIST OF FIGURES	viii
List of Tables.....	ix
1. INTRODUCTION	1
1.1 Overview.....	1
Summary of Literature Review	6
3. PROPOSED METHODOLOGY	8
3.1 Block diagram of various stages of this project:.....	8
3.2 Overview of problem	8
3.3 Data Collection.....	9
3.4 Data Cleaning and Data preprocessing	9
3.5 Extraction of Features.....	9
3.6 Address Bar based Feature.....	10
3.7 Domain based Features	12
<i>Age of Domain</i>	12
3.8 Html and JavaScript based feature	13
Website Forwarding	14
3.9 Models and Training	14
4. IMPLEMENTATION	15
4.1 Dataset	15

4.2 Data Preprocessing	15
<i>Phishing Count in pie chart</i>	<i>17</i>
4.3 Data Splitting	18
4.4 Model building and training.....	18
<i>4.5 Logistic Regression</i>	<i>19</i>
<i>4.6 K-Nearest Neighbors</i>	<i>20</i>
<i>4.7 Support Vector Classifier</i>	<i>21</i>
<i>4.8 Naive Bayes Classifier</i>	<i>21</i>
<i>4.9 Decision Trees Classifier</i>	<i>23</i>
<i>4.10 Random Forest Classifier</i>	<i>24</i>
<i>4.11 Gradient Boosting Classifier</i>	<i>25</i>
<i>4.12 Multi-classifier Perceptron</i>	<i>26</i>
5. EXPERMENT AND RESULT	27
5.1 Performance Evaluation Matrix	27
5.2 Comparison of Models	28
6. CONCLUSION & DISCUSSION	29
References	31

LIST OF FIGURES

Figure. No.	Title	Page No.
3.1	Block Diagram of proposed work	8
3.2	Feature importance using permutation on full model	14
4.1	Frequency of legitimate and phishing URLs	16
4.2	Correlation heat map	16
4.3	pair plot for particular feature	17
4.4	Phishing and Legitimate count in pie chart	18
5.1	Confusion Matrix of logistic regression	19
5.2	Confusion Matrix of KNN	20
5.3	Confusion matrix of Support Vector Classifier	21
5.4	Confusion matrix of Naïve Bayes Classifier	22
5.5	Confusion matrix of Decision Tree Classifier	23
5.6	Confusion matrix of Random Forest Classifier	24
5.7	Confusion matrix of Gradient Boosting Classifier	25
5.8	Confusion matrix of Multi Classifier Perceptron	26

List of Tables

Table. No.	Title	Page No.
2.1	Overview of existing work	5
3.1	Tabular representation of Extracted features	10
5.9	Comparison of models	25

1. INTRODUCTION

1.1 Overview

Internet usage has increased over the past few years, more people are using the Internet as a platform for online business, information sharing, and e-commerce. The expansion of the internet has led to the rise of cybercrime[1]. Cybercriminals use a variety of techniques to steal information, most frequently phishing. Phishing is a kind of online fraud in which perpetrators send unsolicited emails with malicious links in an attempt to fool the recipient into downloading malware or clicking on links to phony websites. Email has always been the primary method of sending messages, but other options include text, social media, and phone. In 2022, phishing will continue to be the most prevalent form of cybercrime affecting businesses in the UK; 83% of these businesses report that they were the target of a phishing attack. 323,972 internet users globally were the target of phishing attacks in 2021. Phishing attacks cost victims \$136 on average, and in 2021, cybercriminals made off with \$44.2 million in stolen funds. Phishing attacks often use email to target their victims. Every 100 internet users globally will receive 16.5 emails on average by 2021 [2]. Cybercriminals can purchase these related files and utilize them in phishing attacks because they are available for purchase on the dark web's illicit market. By 2021, 1 in 5 internet users will have their emails compromised, amounting to approximately 1 billion emails. This could clarify some one of the more prevalent phishing scams.

Over 850 million users in more than 200 countries and territories use LinkedIn. The platform is a prime target for email phishing attacks because of its large user base. Phishing emails with LinkedIn cover photos were the most frequently spotted by social media in 2021 (42%), surpassing Facebook (20%) and Twitter (9%). A prime target are recent hires who update their job status on LinkedIn. Fraudsters pose as business executives to steal personal information. Some will request that employees buy itunes gift cards or give them a call to go over important job requirements. Since 2021, LinkedIn has been a top target for cybercriminals. Reports state that in the first quarter of 2022, LinkedIn was the most copied brand worldwide, accounting for 52% of all detected phishing attacks[3].

Phishing has become a major problem that affects individuals, companies, and even whole countries. The availability of numerous industries, such as social networking, software downloads, online banking, entertainment, and education, has led to the rapid growth of the web in recent years. As a result, the amount of data being uploaded and downloaded to the Internet never stops. Phishing emails in which users pretend to be reputable companies and enterprises Social engineering tactics include assuming the identity of trustworthy websites and deceiving users into disclosing passwords, usernames, and financial information[4]. Technological trickery installation of malicious software, even though the system usually prevents user names and passwords from being accessed from online accounts, on devices where credentials are likely to be stolen.

Markup for Hypertext The architecture and structure of languages are what propel the web development industry's rapid expansion (HTTP). The HTTP protocol is so useful that it makes it easy for people to copy webpages and images. After making a copy of the original email, the phisher will then distribute the duplicate to more users who are most likely involved in this fraudulent attempt. When users click and open these emails, they are taken to a phishing website that mimics the real website[5].

Types of Phishing

Email phishing

This uses fake hyperlinks or attachments to lure email recipients into clicking on them and revealing their data.[6]

Spear phishing

This targets specific individuals by using information about their jobs or social lives to make the emails more convincing.[6]

Whale Watching

This type of spear phishing goes after VIPs or high-ranking employees.[6]

Smishing

This uses text messages or SMS to send phishing links or requests.[6]

Vishing

This is the result of phishers tricking people into visiting phony websites. Another tool that attackers use in their illegal activities is secure browser connections. Businesses in this sector neglect to give their staff proper training because they lack the resources to handle phishing attacks, which in turn

causes phishing attacks to rise. Using encryption as a defense against phishing attacks is one of the most popular ways to fall victim to them. Companies provide security updates for all of their systems and train staff members on how to spot fraudulent phishing scams. Phishing websites mimic reliable internet companies. [6]

1.2 Motivation

The primary objective of the system's development is to safeguard clients' credit card information when they make online purchases. A variety of security measures are provided by banks and other financial institutions to lessen the likelihood of unauthorized access to their accounts. Since today's online transactions through a variety of applications depend entirely on online banking, it is critical that this activity be secure.

1.3 Objectives and Importance of the Project

The "Identify Phishing URLs Using Machine Learning Techniques" project aims to create a system that can recognize and categorize URLs as either phishing or legitimate. It is intended to imitate trustworthy websites and trick users into disclosing private information, like credit card numbers or passwords. Even though they are correctly classified as phishing or legitimate, machine learning techniques can be used to analyze URL content, including domain names, URL lengths, and the presence of suspicious keywords.

1.3.1 Data Collection and Preprocessing

Create a pipeline for semantically segmenting satellite imagery dedicated to land cover mapping in a self-supervised manner. This involves utilizing the abundance of available unlabeled satellite imagery and utilizing contrastive self-supervised techniques to learn meaningful representations.

1.3.2 Feature Extraction

Extract relevant attributes from URLs, such as domain name length, presence of dangerous keywords, and URL redirects. These objects will serve as inputs for machine learning models.

1.3.3 Model Selection and Training

Compare and contrast models such as Support Vector Machines (SVMs), Random Forests, and Neural Networks to find the optimal machine learning algorithm for phishing URL detection. Train the selected model with the preprocessed data.

1.3.4 Model Evaluation

Analyze the trained model's output using a different test dataset. Evaluate the model's recall, accuracy, precision, and F1-score to see how well it classifies URLs.

1.3.5 Real-time Implementation

Integrate the trained model into a practical system that can be deployed to detect phishing URLs in real-time. This may involve developing a web application or API that can analyze URLs and provide classification results

Applying machine learning techniques to the detection of phishing URLs has many advantages, both socially and personally. Phishing attacks pose a significant risk to individuals and companies alike, as they have the potential to cause financial losses, identity theft, and reputational damage. Machine learning offers a promising defense against these threats by making it easier to identify phishing URLs accurately and automatically.

1.4 Scope

Using machine learning for phishing URL detection has a wide range of applications and has a significant impact on mitigating the growing threat posed by phishing attacks. Feature extraction from URLs, supervised model training, behavioral analysis, real-time detection, adaptability to changing threats, ensemble approaches for increased accuracy, security system integration, false positive mitigation, and user education are important components. By proactively detecting and blocking malicious URLs, this method improves cybersecurity by insulating users and organizations from the dangers of phishing attacks. In the ever-changing field of cybersecurity, machine learning models must be continuously monitored and updated to remain effective.

2. LITRATURE REVIEW

2.1 Overview of existing Work

Numerous researchers have examined the statistics of phishing URLs. Our technology integrates important components from earlier studies. We review previous studies on URL features for phishing website detection, which supports our approach.

table 2. 1 Overview Of existing Work

Author	Title	Models	Summary
Qasem Abu Al-Haijalet.al[2]	An intelligent identification and classification system for malicious uniform resource locators (URLs)	The proposed system makes use of four ensembles supervised machine learning approaches, the ensemble of bagging trees (En_Bag) , the ensemble of k-nearest neighbor (En_kNN) , the ensemble of boosted decision trees (En_Bos) , and the ensemble of subspace discriminator (En_Dsc).	This paper discusses a new detection and classification system to identify and categorize the malicious uniform resource locators (URLs), which is developed, trained, validated, evaluated, and discussed.
Ashish Kumar Jha et.al[1]	Intelligent phishing website detection using machine learning	Linear Regression and MultinomialNB are used as the prime methods for the classification apart from other techniques viz. Random Forest, Artificial Neural Network and Support Vector Machine.	The designed pipelined model using Logistic regression, achieved an accuracy of around 98%.
Anitha R[4]	Detecting URL Phishing Attacks Using Machine Learning & NLP Techniques	SVM (Support Vector Machine) is used to identify the phishing or safe status of the given URL.	In this paper, the structure of URL is identified and features are extracted using NLP techniques and svm is trained to perform classification to detect phishing.

Mehmet Korkmaz[7]	Detection of Phishing Websites by Using Machine Learning-Based URL Analysis	In the machine learning based system, 8 different algorithms were run in the experiment. These are: Logistic Regression (LR), K-Nearest Neighborhood (KNN), Support Vector Machine (SVM), Decision Tree (DT), Naive Bayes (NB), XGBoost, Random Forest (RF) and Artificial Neural Network (ANN). The models created with these algorithms trained by using the Sklearn library in the Python programming language.	we aimed to implement a phishing detection system by using some machine learning algorithms. The proposed systems are tested with some recent datasets in the literature and reached results are compared with the newest works in the literature.
Ashit Kumar DuttaI et.al[8]	Detecting phishing websites using machine learning technique	LSTM, RNN	Recurrent neural network (RNN) based method is proposed for detecting phishing links. The projected study highlights the phishing techniques in the context of classification. In this study, 7900 malicious URLs were detected with the help of the proposed method

Summary of Literature Review

Phishing is a prevalent form of cyberattack wherein malevolent emails or messages are sent to the intended recipients with the intention of tricking them into visiting fraudulent websites and providing personal data, including usernames, passwords, and credit card details, in exchange for money. The first stage in the phishing lifecycle, as shown in the figure, is the development of a fake

website that closely resembles the real one. Cybercriminals commonly use techniques like misspelled URLs and spoofing text and logos to trick users into thinking they are visiting a reliable website. Attackers typically target websites such as payment, login, and search pages that allow users to submit forms and request sensitive data. The next step in the phishing process involves sending emails or messages that force the recipient to click on a link to a phony website. website. Cybercriminals can use a range of methods, such as SMS, voice messages, QR codes, and fake mobile apps, to spread false information and trick people into clicking on links. When someone clicks on the link, they are directed to a phony website where identity thieves are collecting personal information. Because fake and real websites often have similar content, user interfaces, and logos, it can be difficult for users to tell the difference between them. To acquire sensitive information, hackers typically target websites for payment, password reset, login, and personal data recovery. In the final phase, hackers steal money from the user's account by using their real information.

3. PROPOSED METHODOLOGY

3.1 Block diagram of various stages of this project:

In Figure 3. 1 shows block diagram of various stages of this project is shown and described in further paragraph.

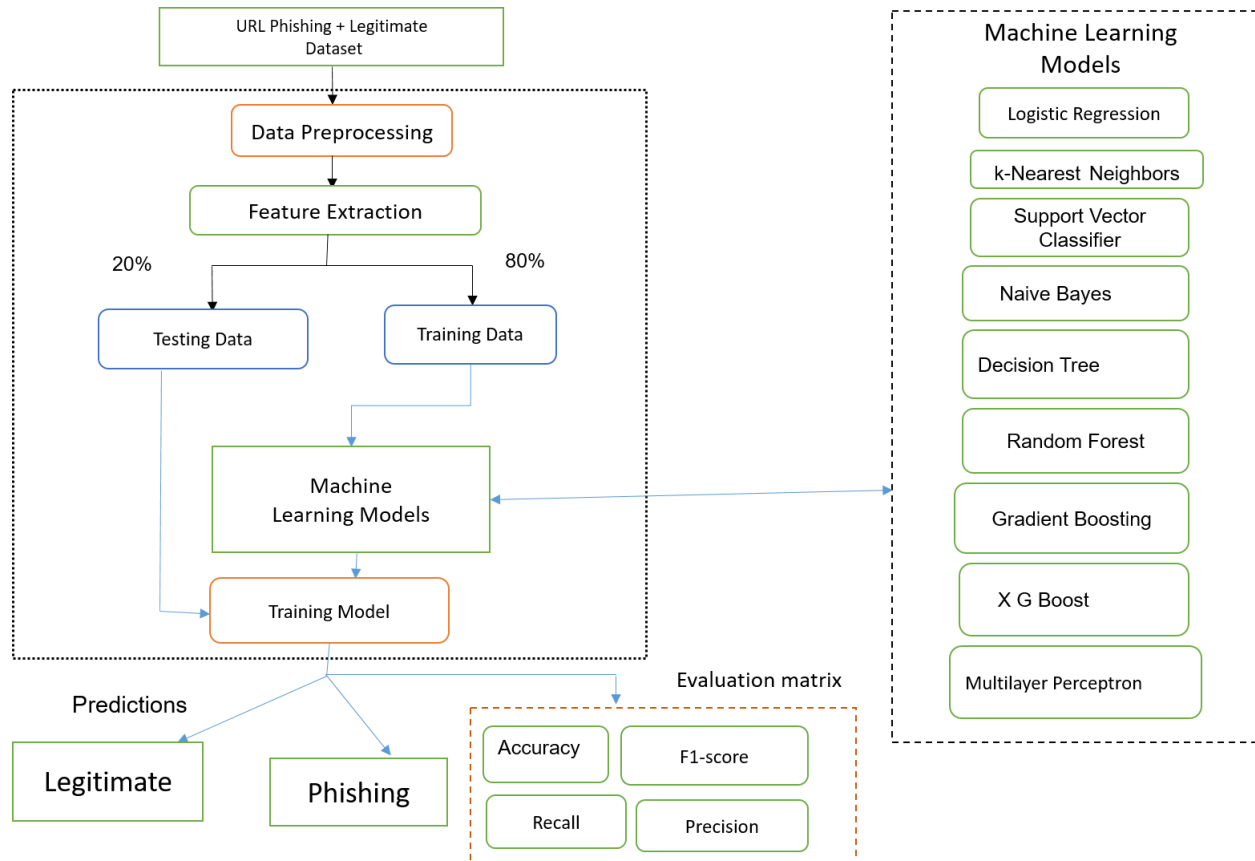


Figure 3. 1 Block Diagram of proposed work

3.2 Overview of problem

URL, also known as "web link", is the most important thing that helps us find information on the internet. Our goal is to choose the best classifier that can distinguish between URLs and legitimate phishing websites. Classifier selection is based on analysis of the keyword, host and page characteristics of the URL. We use the Python language to analyze machine learning algorithms.

3.3 Data Collection

The open-source resource Phish Tank is the source of the list of phishing URLs. This service provides a collection of phishing URLs that is updated every hour in multiple formats, including csv, json, and others[9]. This dataset contains 5000 randomly chosen phishing URLs that are used to train machine learning models.

Valid URLs can be obtained from the University of New Brunswick Open Datasets [10] A range of malware, phishing, spam, defacement, and safe URLs are included in this dataset. Of all these types, the benign URL dataset is considered for this project. This dataset contains 5,000 randomly selected, valid URLs that are used to train machine learning models.

The design of our system is build up as shown in Figure. First, dataset of phishing and legitimate URLs are collected.

3.4 Data Cleaning and Data preprocessing

To clean up the data, add missing values, smooth out crooked data, find and remove outliers, and fix anomalies.

Data preprocessing is a process to convert raw data to meaningful data using different techniques. This frequently entails a variety of mathematical procedures and methods to purify, alter, and enhance the quality of the data.

3.5 Extraction of Features

Numerous algorithms and data formats are available in the literature and in commercial products for identifying phishing URLs. Certain characteristics shared by phishing URLs and the websites they link to can help differentiate them from malicious URLs. An attacker might create a long, complex name, for example, in order to conceal the true domain name. Machine learning algorithms in academic research use a variety of features in their recognition processes. The following features were extracted from academic research on machine learning-based phishing domain detection.

Table 3.1 Tabular representation of Extracted features

Address Bar based Features	<ul style="list-style-type: none">• Domain of the URL• IP address in the URL• “@” Symbol in URL• Length of URL• Depth of URL• Redirection “//” in URL• Http/Https in Domain name• Using URL Shortening Services• Prefix or Suffix “-“ in Domain
Domain based Features	<ul style="list-style-type: none">• DNS Record• Web Traffic• Age of Domain• End Period of Domain
HTML & JavaScript based Features	<ul style="list-style-type: none">• IFRAME redirection• Status Bar customization• Disabling Right Click• Website Forwarding

In this project, there are three features extracted from the data. This functionality is based on address bar, domains, HTML, and JavaScript based features. In **Table 3.1** extracted features are shown.

3.6 Address Bar based Feature

Domain of URL

All we need to do here is extract the domains present in the URL. This feature is not very important for training. During model training, the model may sometimes crash.[11]

IP Address in URL

When an IP address is present in the domain portion of the URL, this attribute is assigned a value of 1 (phishing) or 0 (legitimate) model.[11]

"@" Symbol in URL

The real address frequently comes after the "@" symbol in a URL, and the browser ignores anything that comes before it when you use it. The function is assigned a value of 1 (phishing) if the URL contains the symbol "@," and 0 (legitimate) otherwise.

Length of URL

Determine the URL's length. Long URLs can be used by scammers to conceal questionable content in the address bar. For the purposes of this project, a URL is considered phishing if it contains 54 characters or more; if not, it is considered legitimate. The function's assigned value is 1 (phishing) if the URL's length is 54 characters or more, and 0 (legitimate).

Depth of URL

Determine the depth of the URL. This function uses '/' to determine how many subpages are present in the provided URL. Based on the URL, the function has a numerical value.

Redirection "/" in URL

A "/" in the URL path will direct the user to an alternate website. As you can see, if the URL starts with "HTTP," "/" will appear in her sixth position. However, if the URL uses "HTTPS," "/" needs to appear at the seventh position. Anywhere in the URL other than after the protocol that "/" appears, this function receives a value of 0 (legitimate) or 1 (phishing).

"http/https" in Domain name

Verify whether the domain portion of the URL contains "http/https". By appending a "HTTPS" token to the URL's domain part, phishers can deceive users. If "http/https" appears in the domain portion of the URL, this function's value is 1 (phishing); otherwise, it is 0 (legitimate).

Using URL Shortening Services "Tiny URL"

A shortened URL is a type of "World Wide Web" that, provided the URL is sufficiently short, enables you to visit a desired website. This is accomplished momentarily by sending a link to a web page with a lengthy URL via a "HTTP redirect" of the domain name. This attribute is assigned a value of 1 (phishing) or 0 (allowed) when using a URL shortening service.

Prefix or Suffix "-" in Domain

In order to give the impression that a user is interacting with a genuine website, phishers frequently append prefixes or suffixes to domain names, separating them with a (-). This attribute takes on the values 1 (phishing) or 0 (enabled) if the domain part of the URL contains the "-" character.

3.7 Domain based Features

DNS Record

The WHOIS database is either unaware of the purported identity of phishing websites or lacks a verified record for the host name. The value assigned to this attribute is 1 (phishing) or, alternatively, 0 (enabled) if the DNS record is empty or cannot be found. The WHOIS database is either unaware of the purported identity of phishing websites or lacks a verified record for the host name. The value assigned to this attribute is 1 (phishing) or, alternatively, 0 (enabled) if the DNS record is empty or cannot be found.[11]

Web Traffic

This feature calculates the quantity of visits and pages viewed in order to assess a website's popularity. However, because phishing websites are fleeting, the Alexa database is unable to detect them (Alexa the Web Information Company, 1996). In the worst-case scenarios, relevant websites rank in the top million, according to our analysis of our data sets. Furthermore, if a domain is not recognized by the Alexa database or receives no traffic, it is flagged as "Phishing". In the event that the domain rank is less than 100,000, this property's value is 1 (phishing), and otherwise it is 0 (legal).[11]

Age of Domain

The WHOIS database may be cleared of this characteristic. The majority of phishing websites are temporary. For the sake of this project, a site must be at least 12 months old.

End Period of domain

The WHOIS database may no longer contain this feature. The object's remaining domain time can be found by subtracting the current time from the elapsed time. This project must be completed in the best possible location in no more than six months. If the domain expiration date is less than six months away, the attribute's value is 1 (phishing); if not, it is 0 (legitimate).

3.8 Html and JavaScript based feature

I Frame Redirection

An additional website can be displayed within an active one using an HTML tag known as an I Frame. Phishers can make their content invisible or without frame boundaries by using the "iframe" tag. Phishers exploit this by telling the browser to display a visible border through the use of the "frame Border" feature.

Status Bar Customization

Phishers can display fake URLs in the status bar by using JavaScript. To remove this feature, we need to look at the web browser's source code, specifically the "onMouseOver" event, to see if the status bar is changed in any way. If the response is either blank or onmouseover, the attribute's value is either 0 (valid) or 1 (phishing).

Disabling Right Click

An additional website can be displayed within an active one using an HTML tag known as an IFrame. Phishers can make their content invisible or without frame boundaries by using the "iframe" tag. Phishers exploit this by telling the browser to display a visible border through the use of the "frameBorder" feature. When the answer is not consistently detected or the iframe is empty, this characteristic is given a value of 0 (legal) or 1 (phishing)[11].

Website Forwarding

The thin line separating legitimate websites from phishing ones is the quantity of times the website has been activated. We find that in our data set, trustworthy websites are redirected no more than once. Phishing websites with this feature, however, always have four redirects or more[11].

Figure 3. 2 Shows features importance of the Dataset

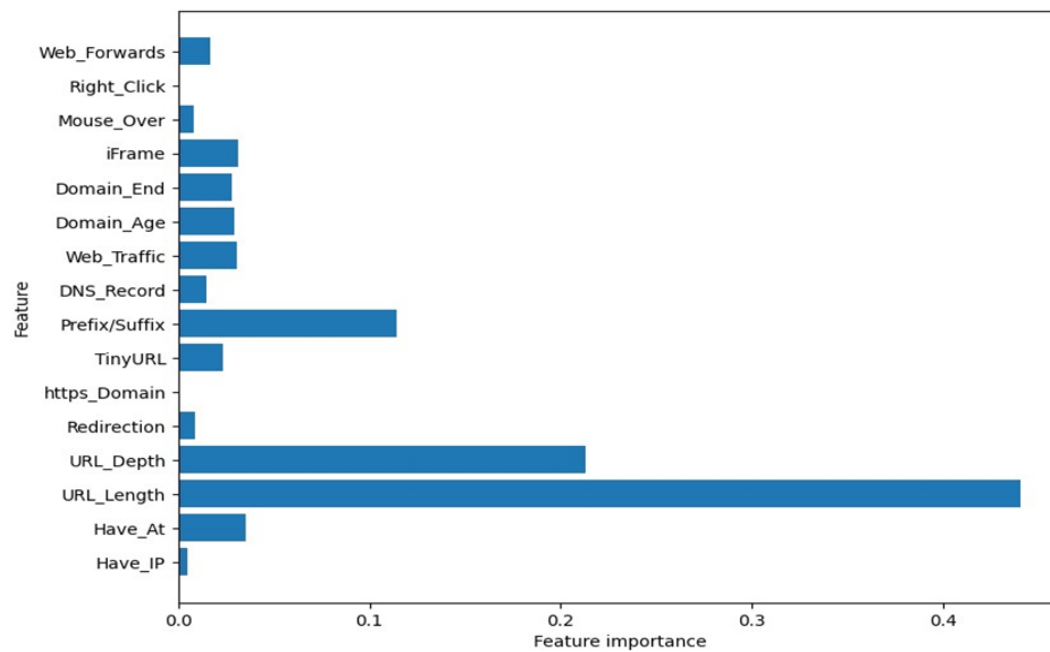


Figure 3. 2 Feature importance using permutation on full model

3.9 Models and Training

Before the ML model is trained, the data is split into 8,000 training samples and 2,000 testing samples. The dataset clearly indicates that this is a problem involving supervised machine learning. Classification and regression are her two main types of supervised machine learning problems. There is a classification problem since the input URLs in this dataset are labeled as either phishing (1) or legitimate (0). The following supervised machine learning models were investigated for dataset training in the context of this project: K-Nearest Neighbors logistic regression is used.

- Support Vector classifier

- Naïve Bayes
- Decision Tree
- Random Forest
- Gradient Boosting
- X g-Boost
- Multilayer Perceptron

4. IMPLEMENTATION

4.1 Dataset

We gathered data sets from Phishing Tank, an open-source platform. The dataset for storage was in CSV format. We used the data preprocessing method to transform the dataset, which had eighteen characters in it. We employed a number of data frame approaches to find features in the data. To show and understand the distribution of the data and the relationships between the factors, a few plots and graphs are included. The machine learning model is not affected by the domain column during training. 16 items and value columns are present now. Without undergoing any cycles, the retrieved attributes of the authentic and phishing URL data sets were simply combined into the excluded attribute file. To balance the classification, we must separate the classifier into training and test sets and sort the data. Additionally, this removes the possibility of overfitting during model training.

4.2 Data Preprocessing

Any data analysis pipeline must include data preprocessing as a crucial step to convert raw data into a format that can be processed and analyzed further. This frequently entails a variety of mathematical procedures and methods to purify, alter, and enhance the quality of the data.

Frequency of Legitimate and Phishing URL

Figure 4. 1 Shows The provided code snippet creates a countplot to visualize the frequency of legitimate and phishing URLs in a dataset. It utilizes the Seaborn library's `sns.countplot` function to generate the plot and applies various customizations, including adding value annotations, setting

plot title and labels, and adjusting font sizes. The resulting plot effectively summarizes the distribution of URL labels (legitimate vs. phishing) and provides a clear representation of the data.

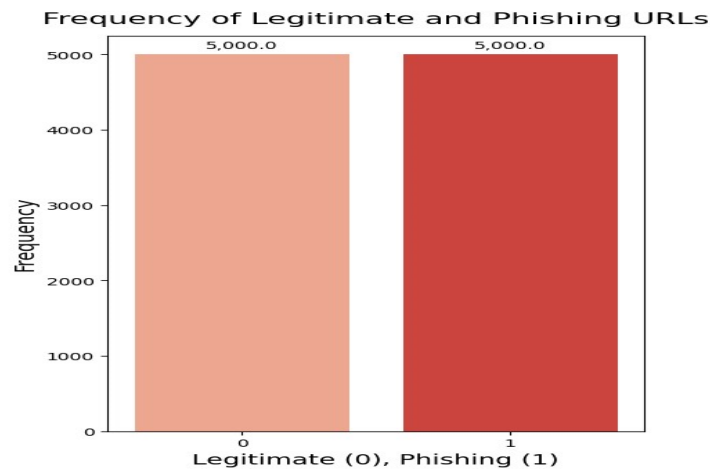


Figure 4. 1 Frequency of legitimate and phishing URLs

Correlation Heatmap

Figure 4. 2 shows the resulting heatmap effectively summarizes the relationships between the variables and provides insights into the underlying structure of the data

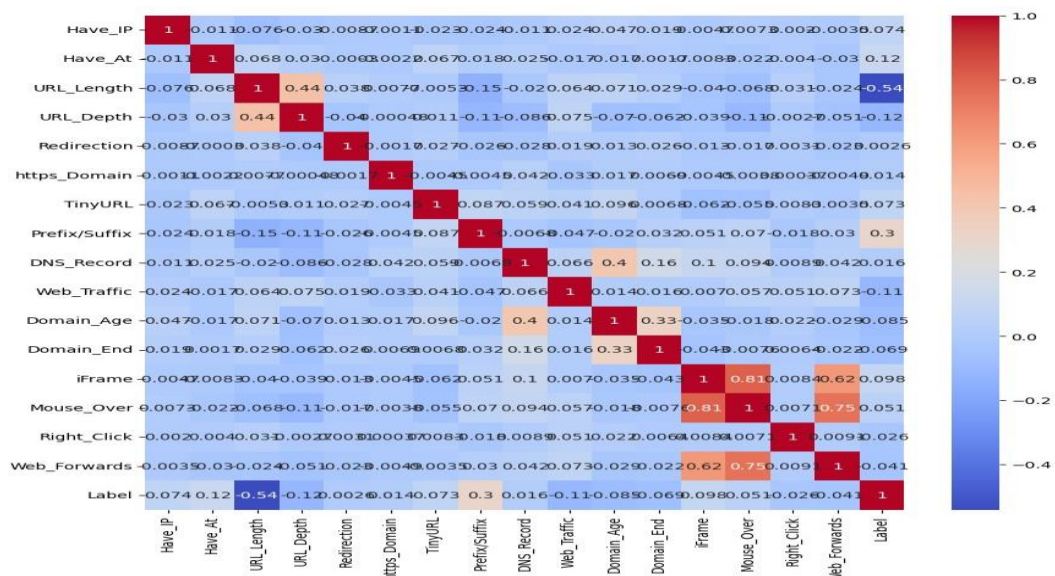


Figure 4. 2 Correlation Heatmap

Pairplot for particular features

Figure 4. 3 shows the pairplot provides a comprehensive overview of the relationships between the selected variables, allowing for the identification of linear, non-linear, and correlational patterns. The color-coding based on Label helps distinguish the characteristics of legitimate and phishing URLs. The marginal histograms provide additional insights into the distribution of each variable.

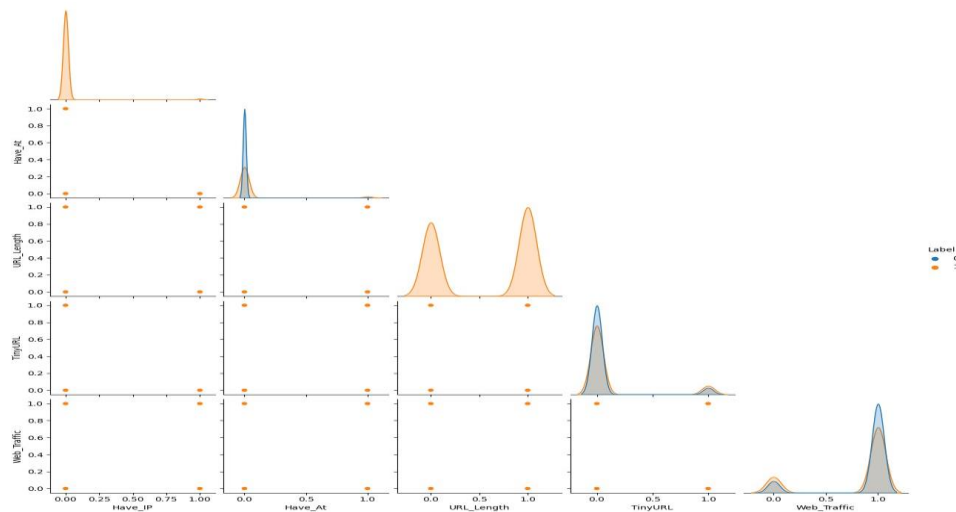


Figure 4. 3 Pair plot for particular features

Phishing Count in pie chart

Figure 4. 4 shows using a dataset, this code generates a pie chart that displays the proportion of phishing and non-phishing emails. "Phishing Count" is the label on the pie chart.

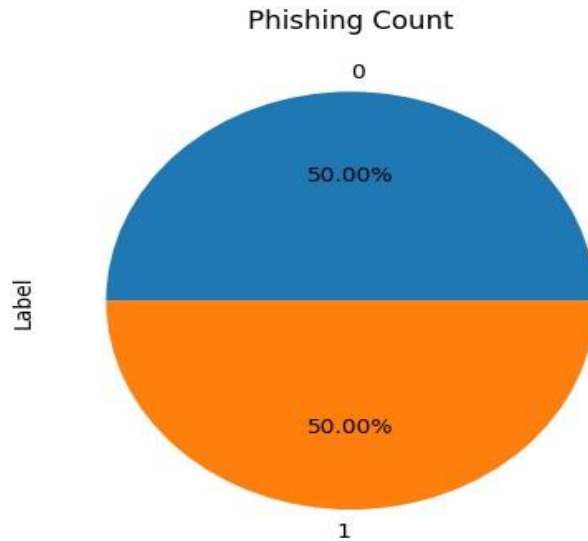


Figure 4. 4 Phishing and legitimate count in pie chart

4.3 Data Splitting

The data is divided into tests and trains 80-20. In machine learning, especially in supervised learning projects, dividing data into trains and experiments is standard practice. Creating training and test sets from the available data is the goal. The machine learning model is developed using the training set, and its performance on untested data is evaluated using the testing set.

We call the data segment that is assigned to each group an 80-20 division. In this case, 80% of the data is used for training and 20% is used for testing. Because it provides enough data for training and stores enough data for insightful analysis, this formula is frequently used.

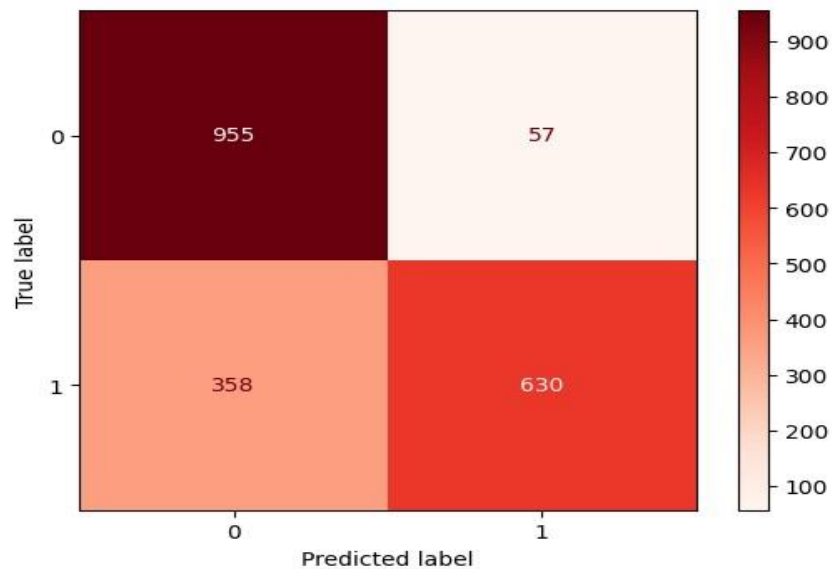
4.4 Model building and training

Supervised learning is one of the most well-liked and successful types of machine learning. Using examples of feature-label pairs, supervised learning can be used to predict a particular label or outcome from a given set of features. These feature-label pairs make up our training set, which we use to build a machine learning model. Our goal is to produce precise predictions for new, unseen data.

Supervised machine learning problems fall into two main categories: classification and regression. We can categorize the suicide rate prediction in our data set as a regression problem because it is a continuous number, or, as programmers would say, a floating-point number. The dataset in this notebook was trained using a supervised machine learning model called regression.

4.5 Logistic Regression

Logistic regression is used to predict a categorical dependent variable's result. Consequently, a discrete or category value must be the result. With the exception of application, logistic regression and linear regression are very similar. Logistic regression is used to solve problems involving classification, while linear regression is used to solve regression problems.[12]



5. 1 Confusion Matrix of Logistic Regression

[[955 57]

[358 630]]

Accuracy: 0.7925

Misclassification Rate: 0.2075

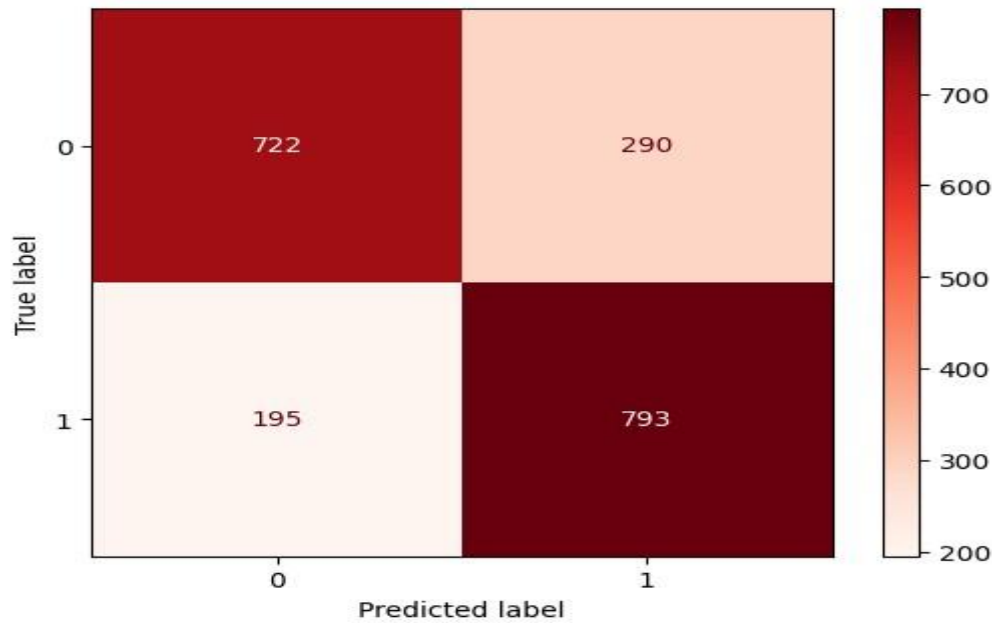
Recall: 0.6376518218623481

Specificity: 0.9436758893280632

Precision: 0.9170305676855895

4.6 K-Nearest Neighbors

One of the most basic machine learning algorithms, K-Nearest Neighbor, is based on the supervised learning approach. The K-NN algorithm places the new case in the category most similar to the existing categories on the assumption that the new case and its data are similar to the cases that are already available.[13]



5. 2 Confusion matrix of KNN

[[722 290]

[195 793]]

Accuracy: 0.7575

Misclassification Rate: 0.2425

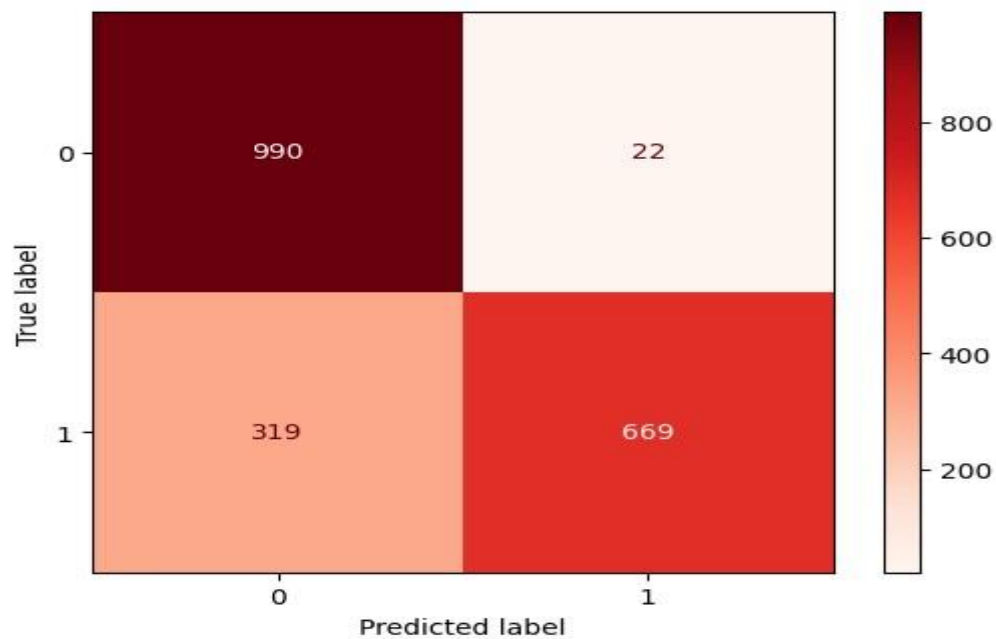
Recall: 0.8026315789473685

Specificity: 0.7134387351778656

Precision: 0.7322253000923361

4.7 Support Vector Classifier

One of the most widely used supervised learning algorithms for both regression and classification problems is support vector machine, or SVM. To make it simpler to categorize new data points in the future, the SVM algorithm searches for the optimal line or decision boundary that can divide n-dimensional space into classes.[14]



5. 3 Confusion matrix of Support Vector machine

[[990 22]

[319 669]]

Accuracy: 0.8295

Misclassification Rate: 0.1705

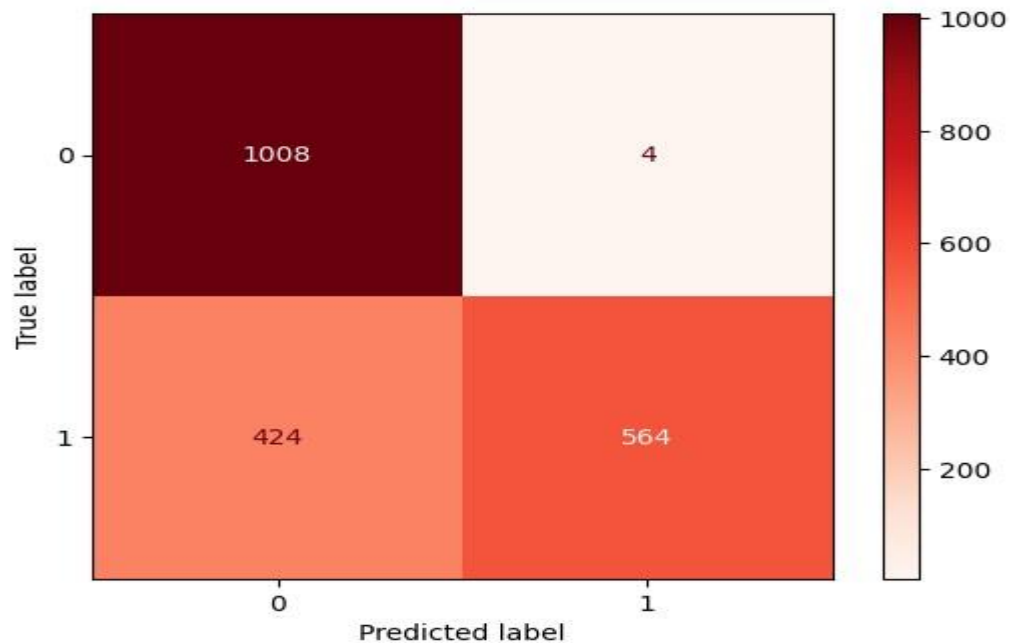
Recall: 0.6771255060728745

Specificity: 0.9782608695652174

Precision: 0.9681620839363242

4.8 Naive Bayes Classifier

A supervised learning method for classifying data is the Naïve Bayes algorithm. The Bayes theorem forms its foundation. It uses a high-dimensional training dataset for text and image classification, which is its main use. The Naïve Bayes classifier, one of the simplest and most effective classification algorithms, facilitates the quick creation of machine learning models with quick prediction capabilities.



5. 4 Confusion matrix of Naïve Bayes Classifier

```
[[1008  4]
```

```
 [ 424 564]]
```

Accuracy: 0.786

Misclassification Rate: 0.214

Recall: 0.5708502024291497

Specificity: 0.9960474308300395

Precision: 0.9929577464788732

4.9 Decision Trees Classifier

Despite being a supervised learning method, decision trees are mostly employed in the resolution of classification issues. Regression problems can also be resolved with them, though. With internal nodes representing dataset features, branches representing decision rules, and leaf nodes representing the outcomes, this classifier has a tree-structure.[2]

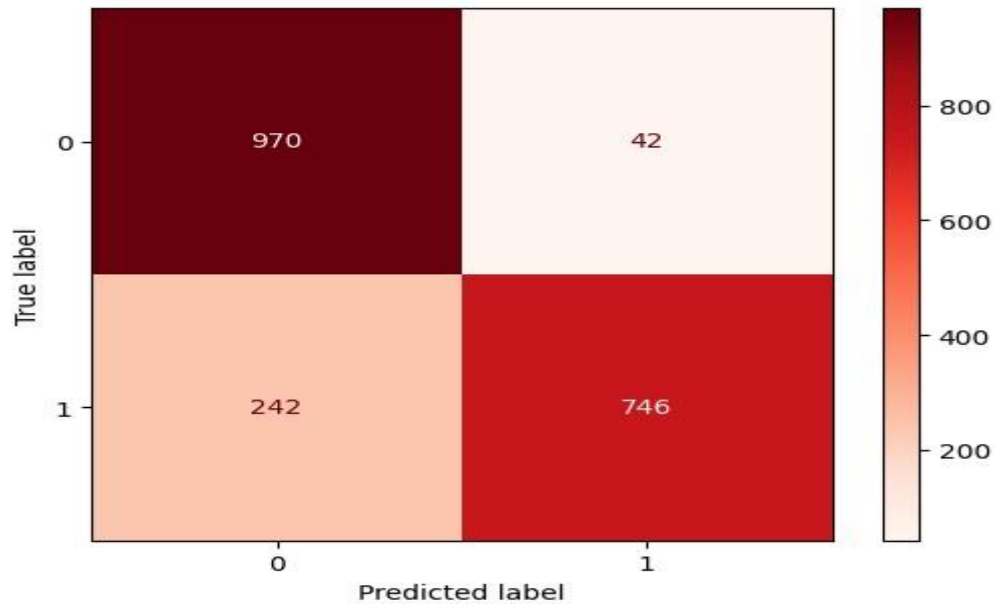


Figure 5.5 Confusion matrix of Decision Tree Classifier

[[970 42]

[242 746]]

Accuracy: 0.858

Misclassification Rate: 0.142

Recall: 0.7550607287449392

Specificity: 0.958498023715415

Precision: 0.9467005076142132

4.10 Random Forest Classifier

Random Forest is a well-liked machine learning algorithm that is applied to supervised learning methods. Regression and classification-based machine learning problems can be resolved with it. Its foundation is the idea of ensemble learning, which is the process of merging several classifiers to enhance the model's performance and resolve a challenging issue.[14]

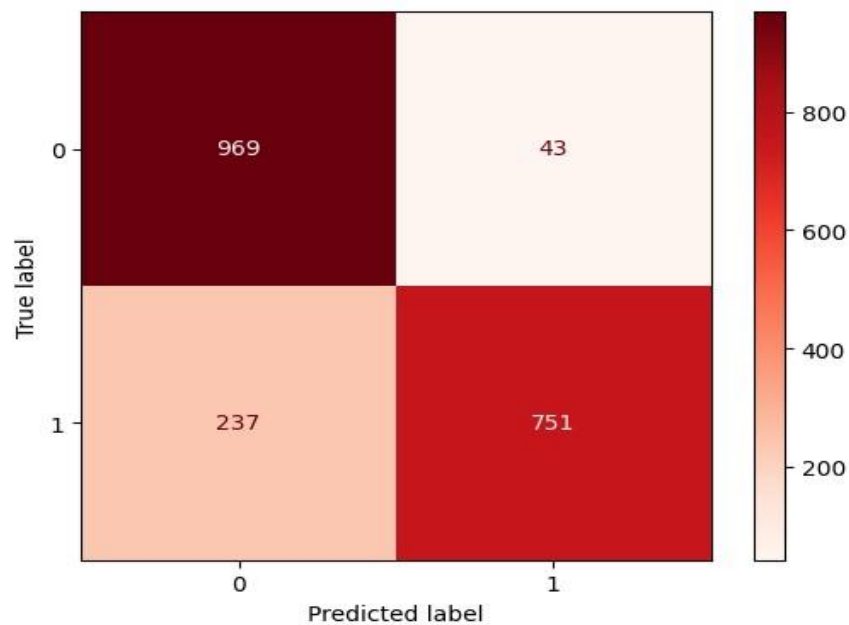


Figure 5.6 Confusion matrix of Random forest classifier

[[969 43]

[237 751]]

Accuracy: 0.86

Misclassification Rate: 0.14

Recall: 0.7601214574898786

Specificity: 0.9575098814229249

Precision: 0.9458438287153652

4.11 Gradient Boosting Classifier

Gradient boosting classifiers are a class of machine learning algorithms that create a single, potent predictive model by combining multiple weak learning models. In gradient boosting, decision trees are typically used. For the purpose of managing the bias-variance trade-off, boosting algorithms are indispensable. Unlike bagging algorithms, which only take high variance into account, boosting regulates a model's bias and variance. It is therefore believed to be more effective.[15]

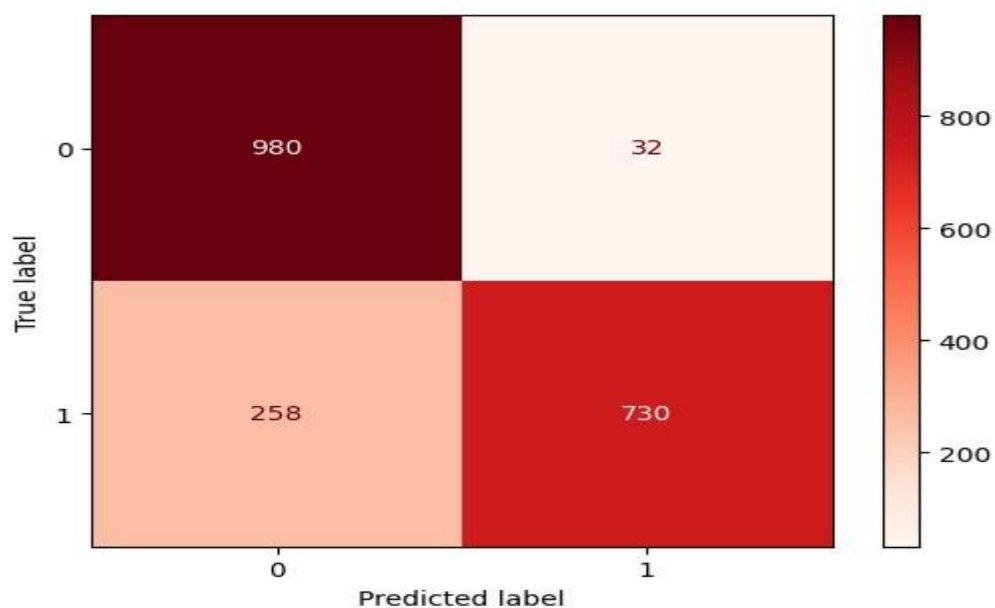


Figure 5.7 Confusion matrix of Gradient Boosting Classifier

[[980 32]

[258 730]]

Accuracy: 0.855

Misclassification Rate: 0.145

Recall: 0.7388663967611336

Specificity: 0.9683794466403162

Precision: 0.958005249343832

4.12 Multi-classifier Perceptron

MLP classifier, short for Multilayer Perceptron classifier, gets its name from artificial neural networks. To accomplish the classification task, the MLP Classifier depends on an underlying Neural Network, in contrast to other algorithms such as Support Vectors or Naive Bayes Classifier.[15]

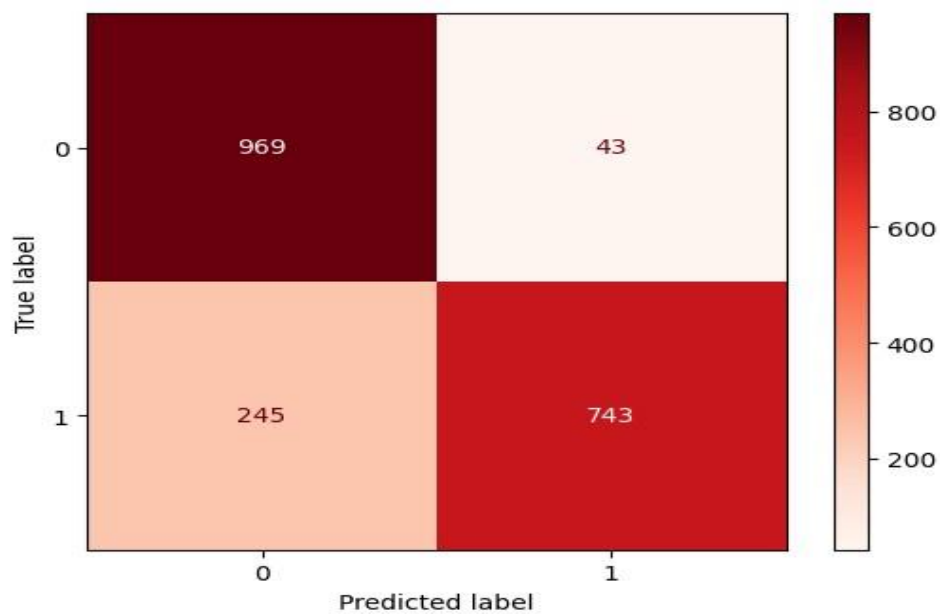


Figure 5.8 Confusion Matrix of Multi-Classifier Perceptron

[[969 43]
[245 743]]
Accuracy: 0.856
Misclassification Rate: 0.144
Recall: 0.7520242914979757
Specificity: 0.9575098814229249
Precision: 0.94529262086514

5. EXPERIMENT AND RESULT

5.1 Performance Evaluation Matrix

To evaluate the efficiency of a system, we use certain parameters. For each machine learning model, we compute the Accuracy, Precision, Recall, F1 Score to decide its exhibition. Every one of these measurements is determined dependent on True Positive(TP), True Negative(TN), False Negative(FN).

On account of URL classification, True positive (TP) is the quality of phishing URLs that are accurately named phishing. True Negative(TN) is the quantity of real URLs that are accurately authentic. False positive (FP) is the quantity of genuine URLs that are named phishing. False Negative (FN) is the quantity of phishing URLs that are named genuine. These qualities are summed up in the table called Confusion Matrix.

Table 5. 1 Table of Confusion Matrix For Phishing Detection

	Predicted Phishing	Predicted Legitimate
Actual Phishing	TP	FN
Actual Legitimate	FP	TN

Precision is the quantity of URLs that are phishing out of the multitude of URLs anticipated as phishing. It estimates the classifier's precision. The recipe to work out precision is given by Equation (1) beneath.

$$\text{Precision} = \frac{TP}{(TP+FP)} * 100\% \quad (1)$$

Recall is the quantity of URLs that the classifier recognized as phishing out of the relative multitude of URLs that are phishing. It is likewise called sensitivity or True positive rate. It is a significant measure and ought to be pretty much as high as could be expected. The formula to compute Recall is given by Equation (2) beneath.

$$\text{Recall} = \frac{TN}{(TN+FP)} * 100\% \quad (2)$$

F1-Score is the weighted normal of accuracy and recall. It is utilized to quantify accuracy and recall simultaneously. The formula to compute F1-Score is given by Equation (3) beneath.

$$F1 \text{ Score} = 2 * \frac{\text{Recall} * \text{Precision}}{(\text{Recall} + \text{Precision})} \quad (3)$$

Accuracy is the quantity of cases that were accurately ordered out of the relative multitude of cases in the test information. The recipe to ascertain is given by Equation (4) beneath.

$$\text{Accuracy} = \frac{(TP+TN)}{(TP+TN+FP+FN)} \quad (4)$$

Specificity indicates the percentage of real negatives out of all negatives.

5.2 Comparison of Models

Table 5. 2 Comparison of Models

No.	ML Model	Accuracy	Precision	Recall	F1_Score
1.	Random forest	0.860	0.932	0.797	0.843
2.	Decision Tree	0.858	0.933	0.795	0.840
3.	Multilayer perceptron	0.856	0.922	0.781	0.838

4.	Gradient Boosting Classifier	0.855	0.913	0.781	0.834
5.	Support Vector Machine	0.830	0.963	0.694	0.797
6.	Logistic Regression	0.792	0.929	0.666	0.752
7.	Naïve Bayes Classifier	0.786	0.981	0.600	0.725
8.	K-nearest neighbors	0.758	0.753	0.856	0.766

Gradient Boosting Classifier correctly classify URL upto 86.4% respective classes and hence reduces the chance of malicious attachments.

6. CONCLUSION & DISCUSSION

The ultimate goal of the project is to analyze machine learning models, conduct research on phishing datasets, and understand the properties of this data. Working on this project, I learned a lot about The factors that affect the model's ability to determine whether a URL is safe or not. I also learned how to open patterns and influence their performance. The final conclusion of this paper is that certain attributes (such as "URL_Length", "URL_Depth" and "Prefix/Suffix") are more important for classifying URLs than determining whether they are phishing or not. The current gradient boosting classifier classifies approximately 86.4% of URLs, reducing the risk of false positive. A user friendly, secure environment can detect the presence of illegal activities. Phishing can be used to track these anonymous individuals and intercept their information using website URLs. It allows users to learn about some of the current phishing websites and phishing website affecting our system. This article presents a machine learning technique for comparative URL prediction. Securing and protecting users' access to their sensitive information is an important goal. The authenticity of the website can be determined by machine learning algorithms. In this study, we found that random

forest distribution with 16 features has the highest accuracy compared to other models. This functionality can be added by creating browser extensions and including graphical user interfaces. With this example we can determine whether the given URL is phishing or legitimate.

In this paper, we aimed to implement a phishing detection system by using some machine learning algorithms. The proposed systems are tested with some recent datasets in the literature and reached results are compared with the newest works in the literature. The comparison results show that the proposed systems enhance the efficiency of phishing detection and reach very good accuracy rates. As future works, firstly, it is aimed to create a new and huge dataset for URL based Phishing Detection Systems. With the use of this dataset, we plan to enhance our system by using some hybrid algorithms, and also deep learning models.

References

- [1] A. K. Jha, R. Muthalagu, and P. M. Pawar, “Intelligent phishing website detection using machine learning,” *Multimed. Tools Appl.*, vol. 82, no. 19, pp. 29431–29456, Aug. 2023, doi: 10.1007/s11042-023-14731-4.
- [2] Q. Abu Al-Haija and M. Al-Fayoumi, “An intelligent identification and classification system for malicious uniform resource locators (URLs),” *Neural Comput. Appl.*, vol. 35, no. 23, pp. 16995–17011, Aug. 2023, doi: 10.1007/s00521-023-08592-z.
- [3] “Phishing Statistics.” [Online]. Available: <https://www.geeksforgeeks.org/types-of-phishing-attacks-and-how-to-identify-them/>
- [4] R. Anitha, S. Swathi, R. Vasuhi, and P. Thenmozhi, “Detecting URL Phishing Attacks Using Machine Learning & NLP Techniques,” *tacks Using Mach. Learn. NLP Tech. Int. J. Sci. Res. Comput. Sci. Eng. Inf. Technol.*, vol. 5, no. 2, pp. 53–56, 2019, [Online]. Available: <https://ijsrceit.com/paper/CSEIT19522.pdf>
- [5] K. Vanitha, “Detection of Phishing Web Pages Based on Features Vector and Prevention using Multi Layered Authentication,” *Int. J. Pure Appl. ...*, no. March, 2018, [Online]. Available: https://www.researchgate.net/profile/Vanitha-Muthuraman/publication/326345682_Detection_of_phishing_web_pages_based_on_features_vector_and_prevention_using_multi_layered_authentication/links/5c85fa72458515831f9aaf9b/Detection-of-phishing-web-pages-based-o
- [6] “Types of Phishing Attacks and How to Identify them.”
- [7] M. Korkmaz, O. K. Sahingoz, and B. Dİri, “Detection of Phishing Websites by Using Machine Learning-Based URL Analysis,” *2020 11th Int. Conf. Comput. Commun. Netw. Technol. ICCCNT 2020*, no. June, 2020, doi: 10.1109/ICCCNT49239.2020.9225561.
- [8] A. K. Dutta, “Detecting phishing websites using machine learning technique,” *PLoS One*, vol. 16, no. 10 October, pp. 1–17, 2021, doi: 10.1371/journal.pone.0258361.
- [9] “PhishTank.” [Online]. Available: <https://www.phishtank.com/index.php>
- [10] “URL dataset (ISCX-URL2016).” [Online]. Available: <https://www.unb.ca/cic/datasets/url-2016.html>
- [11] R. M. Mohammad, F. Thabtah, and L. Mccluskey, “Phishing Websites Features,” *Ieee*, pp. 1–7, 2013, [Online]. Available: papers3://publication/uuid/6A553382-D05D-48FA-97AA-

382C3203BB1F

- [12] “PhishingProject.” [Online]. Available: https://eforensicsmag.com/2017_cybercrime/
- [13] J. Rashid, T. Mahmood, M. W. Nisar, and T. Nazir, “Phishing Detection Using Machine Learning Technique,” *Proc. - 2020 1st Int. Conf. Smart Syst. Emerg. Technol. SMART-TECH 2020*, pp. 43–46, 2020, doi: 10.1109/SMART-TECH49988.2020.00026.
- [14] T. Shrivastava, S. Bhatt, N. R. Roy, and A. Bhasney, “Phishing URL Detection Using Machine Learning: A Survey,” *Proc. - 2022 4th Int. Conf. Adv. Comput. Commun. Control Networking, ICAC3N 2022*, pp. 1992–1997, 2022, doi: 10.1109/ICAC3N56670.2022.10074337.
- [15] S. S. Ravindra, S. J. Sanjay, S. N. A. Gulzar, and K. Pallavi, “Phishing Website Detection Based on URL,” *Int. J. Sci. Res. Comput. Sci. Eng. Inf. Technol.*, vol. 3307, pp. 589–594, 2021, doi: 10.32628/cseit2173124.