Prediction of Phishing Websites using Machine Learning

BY

Shubham Chaudhary (Enrolment No. 18/11/EC/016)
Jatin Verma (Enrolment No. 18/11/EC/002)
Arun Pratap Singh (Enrolment No. 18/11/EC/025)
Shivam Jha (Enrolment No. 18/11/EC/037)

under the guidance of

Prof. TV Vijay Kumar, School of Engineering JNU, Delhi
Dr. Manju Khari, Other School, JNU, Delhi

in the partial fulfilment of the requirements
for the award of the degree of

Bachelor of Technology
(a part of Five-Year Dual Degree Course)

School of Engineering
Jawaharlal Nehru University, Delhi
Jan, 2022

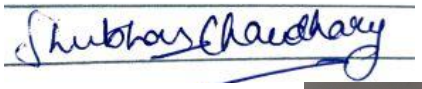
JAWAHARLAL NEHRU UNIVERSITY
SCHOOL OF ENGINEERING

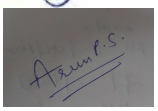
DECLARATION

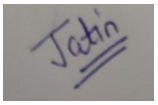
We declare that the project work entitled **"Prediction of Phishing Websites using Machine Learning"** which is submitted by me/us in partial fulfillment of the requirement for the award of degree B.Tech. (a part of Dual-Degree Programme) to School of Engineering, Jawaharlal Nehru University, Delhi comprises only my/our original work and due acknowledgement has been made in the text to all other material used.

Shivam Jha 

Shubham Chaudhary



Arun Pratap Singh 

Jatin Verma 

(Full Name and Sign of all Group Members)

JAWAHARLAL NEHRU UNIVERSITY
SCHOOL OF ENGINEERING

CERTIFICATE

This is to certify that the project work entitled **"Prediction Of Phishing Website Using Machine Learning"** being submitted by **Mr. Shivam Jha** (Enrolment No.- 18/11/EC/037) in fulfilment of the requirements for the award of the **Bachelor of Technology** (part of Five-Year Dual Degree Course) in **Computer Science and Engineering**, will be carried out by him under my supervision.

In my opinion, this work fulfills all the requirements of an Engineering Degree in respective stream as per the regulations of the School of Engineering, Jawaharlal Nehru University, Delhi. This thesis does not contain any work, which has been previously submitted for the award of any other degree.

Prof. TV Vijay Kumar
(Supervisor)
Professor
School of Engineering
.
Jawaharlal Nehru University, Delhi

Dr. Manju Khari
(Co-Supervisor)
Professor
School of Computational and .
System Sciences,
Jawaharlal Nehru University, Delhi

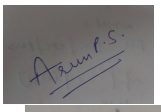
JAWAHARLAL NEHRU UNIVERSITY
SCHOOL OF ENGINEERING

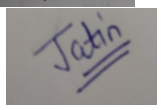
ACKNOWLEDGMENT

We would like to thank Prof. TV Vijay Kumar and Dr. Manju Khari for their unwavering support throughout our B.Tech Major Project. We want to express our profound gratitude to Professor TV Vijay Kumar, our supervisor, and Dr. Manju Khari, our co-supervisor, for their excitement, patience, insightful remarks, valuable information, practical guidance, and never-ending ideas, which has greatly aided us during our study and preparation for the project. Their vast knowledge, extensive experience, and professional competence in Machine Learning have been extremely beneficial to us throughout the project.

Shivam Jha 

Shubham Chaudhary

Arun Pratap Singh 

Jatin Verma 

(Full Name and Sign of all Group Members)

ABSTRACT

Phishing is the most straightforward method of obtaining sensitive information from unsuspecting consumers. The goal of phishers is to obtain sensitive information such as usernames, passwords, and bank account numbers. People working in cyber security are now looking for reliable and consistent detection strategies for phishing websites. The purpose of this work is to use machine learning to detect phishing URLs by extracting and evaluating various aspects of authentic and phishing URLs. Phishing websites are detected using Logistic Regression, Decision Tree, Random Forest, Support Vector Machine and Bagging and Boosting algorithms. The goal of the study is to detect phishing URLs and to narrow down the best machine learning method by analysing each algorithm's accuracy rate, false positive rate, and false negative rate.

LIST OF CONTENTS

Content

Page No.

LIST OF FIGURES

Details of Figure **Page No.**

CHAPTER-1

INTRODUCTION

1.1. INTRODUCTION

Phishing has become a major source of concern for security professionals in recent years since it is relatively easy to develop a phoney website that appears to be identical to a legitimate website

Although experts can recognise bogus websites, not all users can, and as a result, some users become victims of phishing attacks. The attacker's main goal is to steal bank account credentials. Businesses in the United States lose $2 billion each year as a result of phishing attacks. According to the 3rd Microsoft Computing Safer Index Report, which was released in February 2014, the yearly global impact of phishing might be as high as $5 billion. Because of a lack of user awareness, phishing assaults are becoming more successful.

Because of a lack of user awareness, phishing assaults are becoming more successful. Because phishing attacks take advantage of user flaws, it's tough to counteract them, but it's critical to improve phishing detection techniques.

Machine learning technology is made up of a variety of algorithms that use historical data to produce predictions about future data. The algorithm will use this technique to examine numerous banned and valid URLs and their attributes in order to accurately detect phishing websites, including zero-hour phishing websites.

1.2. THESIS OBJECTIVE

The primary objective of this thesis is to :

To predict and classify phishing websites using supervised learning techniques.

CHAPTER-2

LITERATURE SURVEY

2.1. INTRODUCTION

We studied about the different components of URL including the domain/subdomain names and the different tags present on the website that may be used maliciously by the phishers to trap the user.

2.2 Different Features of a URL:

2.2.1. HTTPS (Hyper Text Transfer Protocol with Secure Sockets Layer)

The existence of HTTPS is very important in giving the impression of website legitimacy. The difference between the two protocols i.e. HTTP and HTTPS is that HTTPS uses TLS (SSL) to encrypt normal HTTP requests and responses. As a result, HTTPS is far more secure than HTTP

2.2.2. Website Traffic

This feature measures the popularity of the website by determining the number of visitors and the number of pages they visit. However, since phishing websites live for a short period of time, they may not be recognized by the Alexa database (Alexa the Web Information Company., 1996). By reviewing our dataset, we find that in worst scenarios, legitimate websites ranked among the top 100,000. Furthermore, if the domain has no traffic or is not recognized by the Alexa database, it is classified as “Phishing”. Otherwise, it is classified as “Suspicious”.

2.2.3. Sub Domain and Multi Sub Domains:

A domain name might include the country-code top-level domains (ccTLD), which in our example is “uk”. The “ac” part is shorthand for “academic”, the combined “ac.uk” is called a second-level domain (SLD). If a URL has more than one subdomain, the website may be deemed as suspicious

2.2.4. Adding Prefix or Suffix Separated by (-) to the Domain:

The dash symbol is rarely used in legitimate URLs. Phishers tend to add prefixes or suffixes separated by (-) to the domain name so that users feel that they are dealing with a legitimate webpage.

2.2.5 Links in Script and Link tags
It is common for legitimate websites to use "< Meta >" tags to offer metadata about the HTML document; "< script >" tags to create a client side script; and "< Link >" tags to retrieve other web resources.It is expected that these tags are linked to the same domain of the webpage.

2.2.6 URL of Anchor
An anchor is an element defined by the tag <a>. This feature is treated exactly as “Request URL”. However, for this feature we examine: If the tags and the website have different domain names. This is similar to request URL feature.

2.2.7 Request URL
Request URL examines whether the external objects contained within a webpage such as images, videos and sounds are loaded from another domain. In legitimate webpages, the webpage address and most of objects embedded within the webpage are sharing the same domain.

2.2.8 Server Form Handler (SFH):
SFHs that contain an empty string or “about:blank” are considered doubtful because an action should be taken upon the submitted information. In addition, if the domain name in SFHs is different from the domain name of the webpage, this reveals that the webpage is suspicious because the submitted information is rarely handled by external domains.

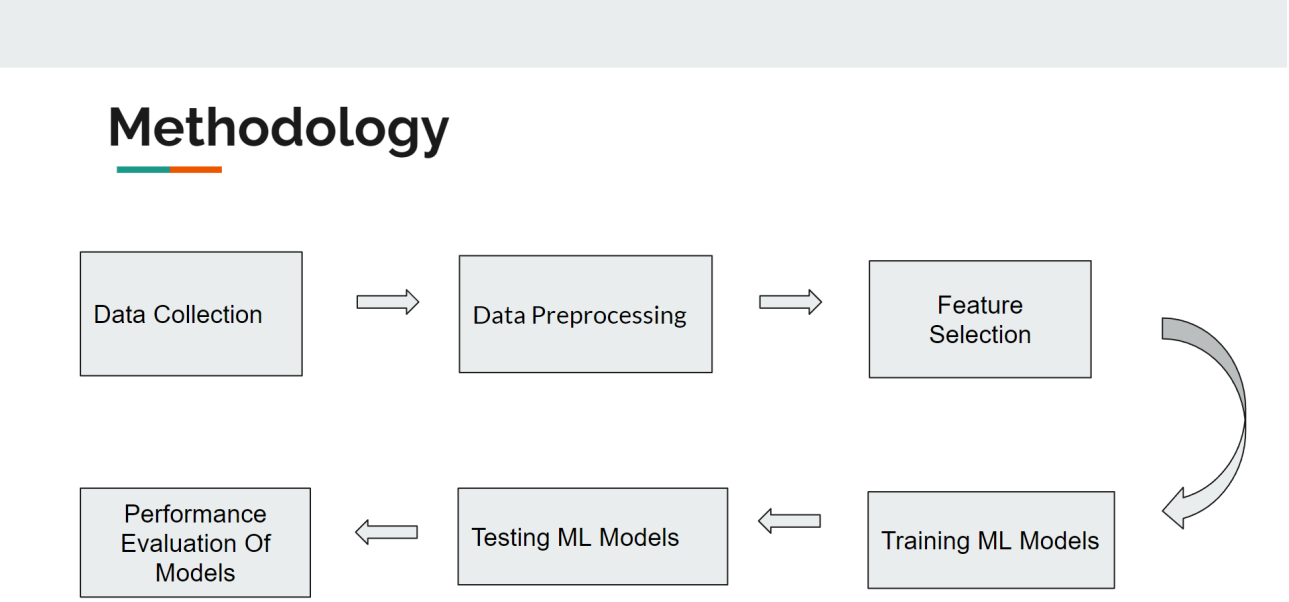
2.2.9 Domain Registration Length:
Based on the fact that a phishing website lives for a short period of time, we believe that trustworthy domains are regularly paid for several years in advance. In our dataset, we find that the longest fraudulent domains have been used for one year only.

2.2.10 Age of Domain:
This feature can be extracted from the WHOIS database. Most phishing websites live for a short period of time. By reviewing our dataset, we find that the minimum age of the legitimate domain is 6 months.

2.2.11. Using the IP Address:
If an IP address is used as an alternative of the domain name in the URL, such as “http://125.98.3.123/fake.html”, users can be sure that someone is trying to steal their personal information. Sometimes, the IP address is even transformed into hexadecimal code as shown in the following link “http://0x58.0xCC.0xCA.0x62/2/paypal.ca/index.html”.

CHAPTER-3

PROPOSED WORK AND METHODOLOGY



3.1. INTRODUCTION

Step 1) Data Collection: Dataset Used: Phishing website Detector Kaggle
The dataset consists of a collection of website URLs for 11000+ websites. Each sample originally had 32 parameters but after feature importance ranking we are left with 12 most prominent website parameters and a class label identifying it as a phishing website or not (1 or -1).

Step 2) Data Preprocessing:
Data preprocessing can refer to manipulation or dropping of data before it is used in order to ensure or enhance performance, and is an important step in the data mining process.

Step 3) Feature Selection:
Feature Selection is one of the core concepts of Machine Learning which hugely impacts the performance of the model. Irrelevant or partially relevant features can negatively impact model performance.
We have used two methods:-

1. Feature Importance method
2. Correlation matrix with heat map

Step 4) Training Machine Learning Models:

We used a Phishing Website Detection dataset for designing classifiers using machine learning techniques like Decision trees, Logistic regression, K-nearest neighbours, Random Forest, Gaussian Naive Bayes, Support Vector Machines and Bagging and Boosting.
We have used the following Machine Learning Algorithms to create our models:

1. Logistic Regression :
2. Decision Tree
3. KNN
4. SVM :
5. Naive Bayes
6. Random Forest
7. ADA Boost
8. XGBoost

Step 5) Testing Machine Learning Models:
In machine learning, model testing is referred to as the process where the performance of a fully trained model is evaluated on a testing set.
We divided the dataset into a 70 to 30 ratio of training set and testing set. Then we used this testing dataset to evaluate the accuracy and other metrics of our models.

Step 6) Performance Evaluation Of Models:
The three main metrics used to evaluate a classification model are accuracy, precision, and recall. Accuracy is defined as the percentage of correct predictions for the test data. It can be calculated easily by dividing the number of correct predictions by the number of total predictions.

3.2. MODELS

3.2.1. Logistic Regression : It is a powerful supervised machine learning approach for binary classification issues is logistic regression (when target is categorical). The best way to think of logistic regression is as a type of linear regression that is used to solve classification difficulties. The logistic function defined below is used to model a binary output variable in logistic regression (Tolles & Meurer, 2016).The main distinction between linear and logistic regression is that the range of logistic regression is limited to 0 and 1. Furthermore, logistic regression does not require a linear relationship between input and output variables, unlike linear regression.This is due to applying a nonlinear log transformation to the odds ratio (will be defined shortly).

$$\Delta \equiv \frac{1}{1+e^{-x}}$$

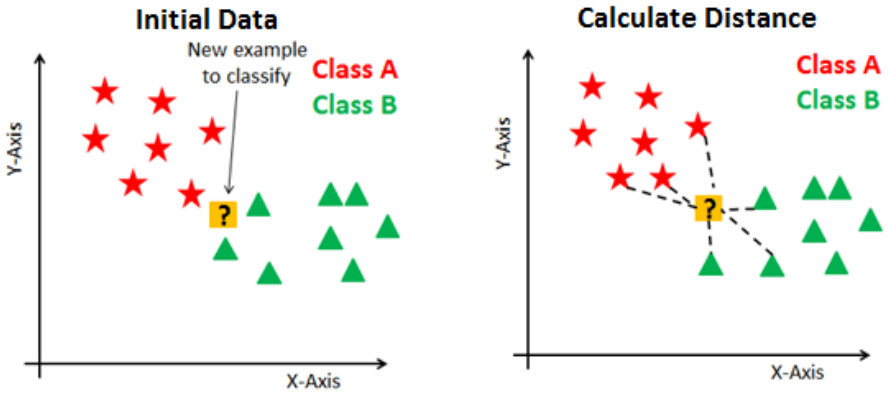
3.2.2. Decision Tree : One of the most extensively used algorithms in the field of machine learning. The decision tree algorithm is simple to comprehend and apply. The decision tree starts by selecting the best splitter from the available qualities for categorization, which is referred to as the tree's root.The algorithm keeps building the tree until it reaches the leaf node.In tree representation, each internal node of the tree corresponds to attribute and each leaf node of the tree belongs to class

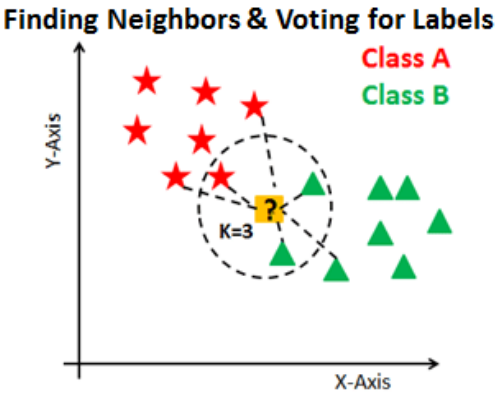
label, which is used to anticipate target value or class. The gini index and information gain approaches are employed in the decision tree algorithm to determine these nodes.

3.2.3. K Nearest Neighbour : The KNN algorithm is a sort of supervised machine learning method that may be used to solve both classification and regression predicting problems. However, in industry, it is mostly used to solve classification and prediction problems. The following two characteristics would be a good way to describe KNN:

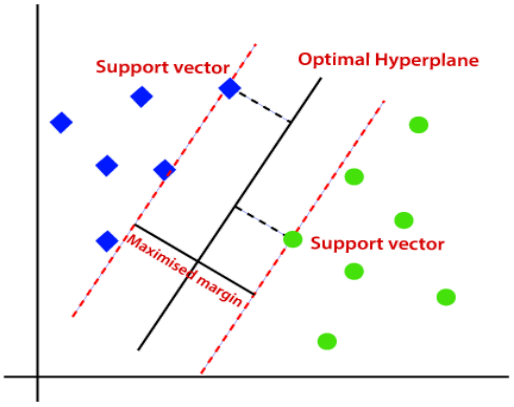
KNN is a lazy learning algorithm because it doesn't have a dedicated training phase and instead uses all of the data for training and classification.

KNN is also a non-parametric learning algorithm because it makes no assumptions about the underlying data.





3.2.4. Support Vector Machine : Another useful approach in machine learning is the support vector machine. Each data item in the support vector machine method is plotted as a point in n-dimensional space, and the support vector machine algorithm creates a separating line for classification of two classes, which is known as a hyperplane.The closest points, known as support vectors, are sought by the support vector machine, which then creates a line linking them. The support vector machine then creates a separation line that is perpendicular to the connecting line and bisects it. The margin should be as large as possible in order to accurately classify data. The margin is the distance between the hyperplane and support vectors in this case. Because it is impossible to segregate complicated and nonlinear data in the real world, the support vector



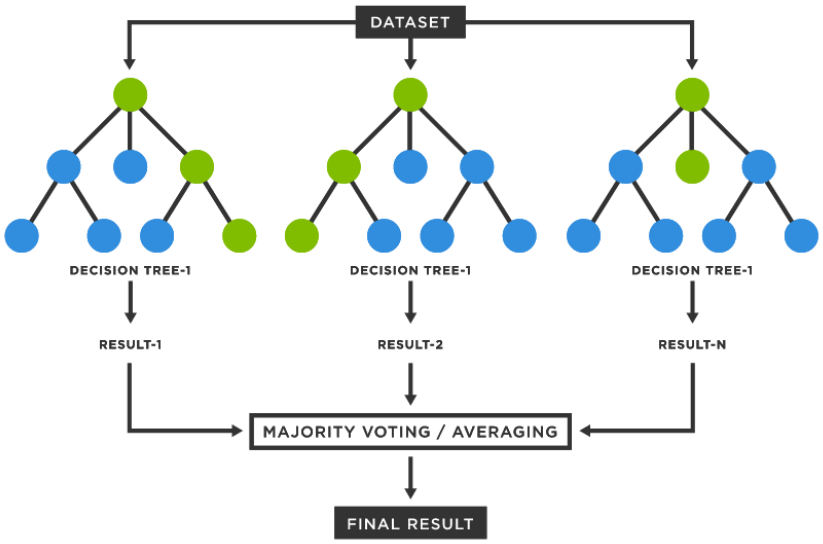
machine employs a kernel method that converts lower dimensions space to higher dimensional space to tackle this difficulty.

3.2.5. Naive Bayes :A probabilistic learning model is what the NB classifier is called. It is based on Bayes' theorem and is used to solve classification difficulties. The independence of the features or qualities involved in prediction is a significant underlying assumption of this approach. This algorithm computes the class prior probability, the likelihood to a given class, the posterior probability, and the predictor prior probability to make predictions about any instance to a particular class, as shown below:

$$P(D|X) = P(x|d)P(d)/P(x)$$

- where,
- $P(D \mid X)$ is the posterior probability for class D, predictor given as (x, attributes)
- $P(x \mid d)$ is the likelihood for a particular label class
- $P(d)$ is the prior probability for a label class
- $p(x)$ is the prior probability for the predictor

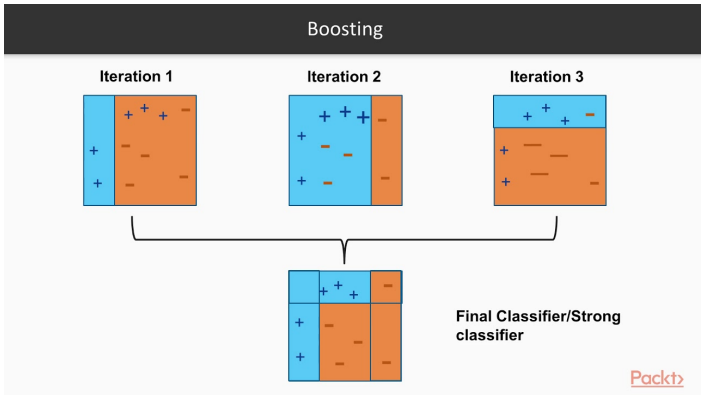
3.2.6. Random Forest : The random forest algorithm, which is based on the notion of the decision tree algorithm, is one of the most powerful algorithms in machine learning technology. The random forest algorithm generates a forest with a large number of decision trees. A large number of trees results in a high level of detection accuracy. The bootstrap method is used to create trees. To generate a single tree, the bootstrap approach selects characteristics and samples from the dataset at random with replacement. Random forest algorithm, like decision tree algorithm, chooses the best splitter for classification from randomly picked features. Random forest algorithm also uses gini index and information gain methods to determine the best splitter. This technique will be repeated until the random forest has produced n trees. Each tree in the forest predicts the target value, and the



algorithm then calculates the votes for each target value predicted. Finally, the random forest algorithm considers the predicted target with the most votes as the final prediction.

3.3 BOOSTING ALGORITHM BASED MODELS

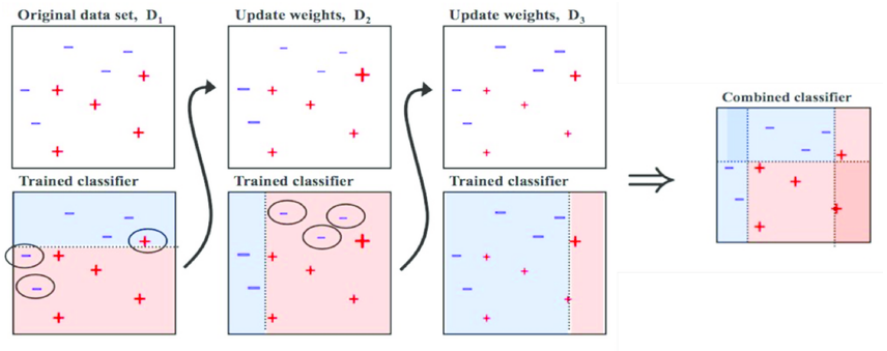
Boosting Algorithms :



Boosting is an ensemble modelling strategy that aims to create a strong classifier out of a large number of weak ones. It is accomplished by constructing a model from a sequence of weak models. To begin, a model is created using the training data. The second model is then created, which attempts to correct the faults in the first model. This approach is repeated until either the entire training data set is properly predicted or the maximum number of models has been added.

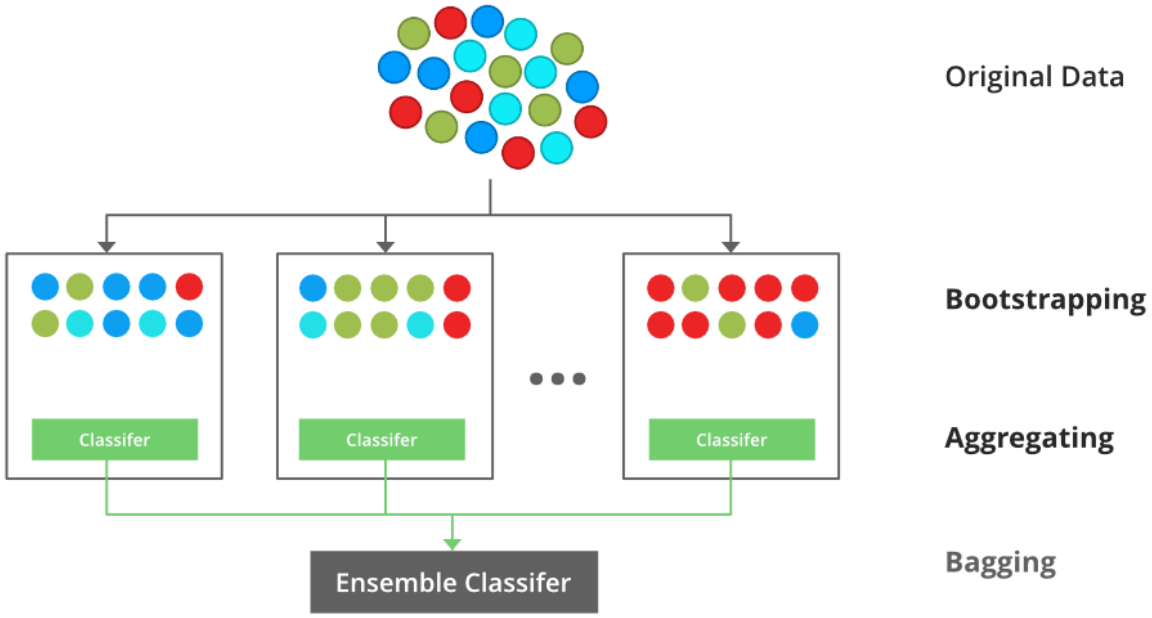
3.3.1. ADA Boost :

Freund and Schapire proposed the AdaBoost algorithm in 1997, and it is a highly valuable method. Because of its great speed, low complexity, and outstanding interoperability, it is frequently used. Viola and Jones were the first to use the AdaBoost algorithm to solve feature selection difficulties in face recognition. For each feature, this method builds a simple weak classifier. Because the weight of correctly classified samples will be appropriately reduced and the weight of misclassified samples will be appropriately increased in the iterative process, the original classifier does not require a high accuracy if the accuracy is higher than that of random classification. The distribution



of samples is shifted in this way. A strong classifier with greater performance is created by merging the weak classifiers produced from each cycle.

3.3.2. XGBoost :



Gradient Boosted decision trees are implemented in XGBoost. C++ was used to create this library. It is a type of software library that was created with the goal of improving model speed and performance. In recent years, it has dominated in the field of applied machine learning. Many Kaggle competitions are dominated by XGBoost models.

Decision trees are constructed sequentially in this approach. In XGBoost, weights are very significant. All of the independent variables are given weights, which are subsequently fed into the decision tree, which predicts outcomes. The weight of variables that the tree predicted incorrectly is increased, and the variables are then fed into the second decision tree. These various classifiers/predictors are then combined to create a more powerful and precise model. It has the potential to help with regression

CHAPTER-4

RESULT DISCUSSION

4.1. INTRODUCTION

Dataset was divided into training sets and testing sets in 70:30 ratios respectively. Each classifier is trained using a training set and a testing set is used to evaluate performance of classifiers. Performance of classifiers has been evaluated by calculating the classifier's accuracy score, recall score, Precision score and F1 score.

We have used the following Machine Learning Algorithms to create our models:

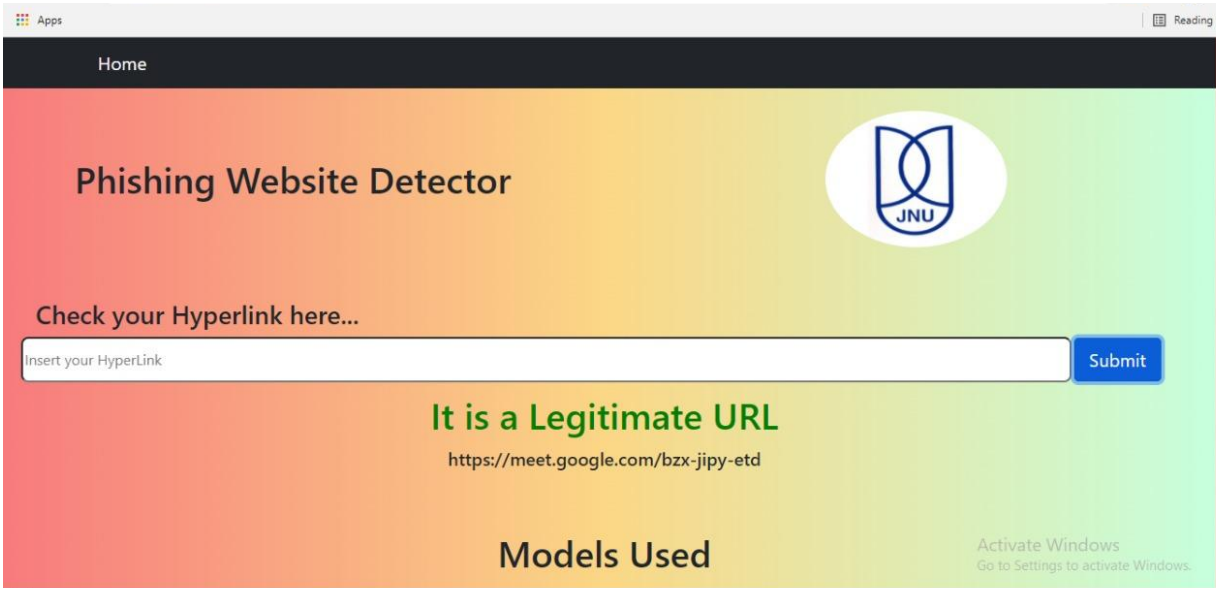
1. Logistic Regression
2. Decision Tree
3. KNN
4. SVM
5. Naive Bayes
6. Random Forest
7. ADA Boost
8. XGBoost

By comparing and analysing all the models we find that the Random Forest model outperforms every other model in almost all the metrics. We have also ignored the Naive Bayes model because it has performed at a subpar level.

	Accuracy	Precision	Recall	F1-score
Logistic Regression	0.92	0.88	0.93	0.91
Decision Tree	0.95	0.95	0.94	0.94
KNN	0.93	0.9	0.94	0.92
Random Forest	**0.95**	**0.95**	**0.95**	**0.95**
SVM	0.93	0.9	0.95	0.92
Naive Bayes	0.59	**1**	0.52	0.69
ADA Boost	0.93	0.9	0.93	0.92
XG Boost	0.94	0.91	0.94	0.93

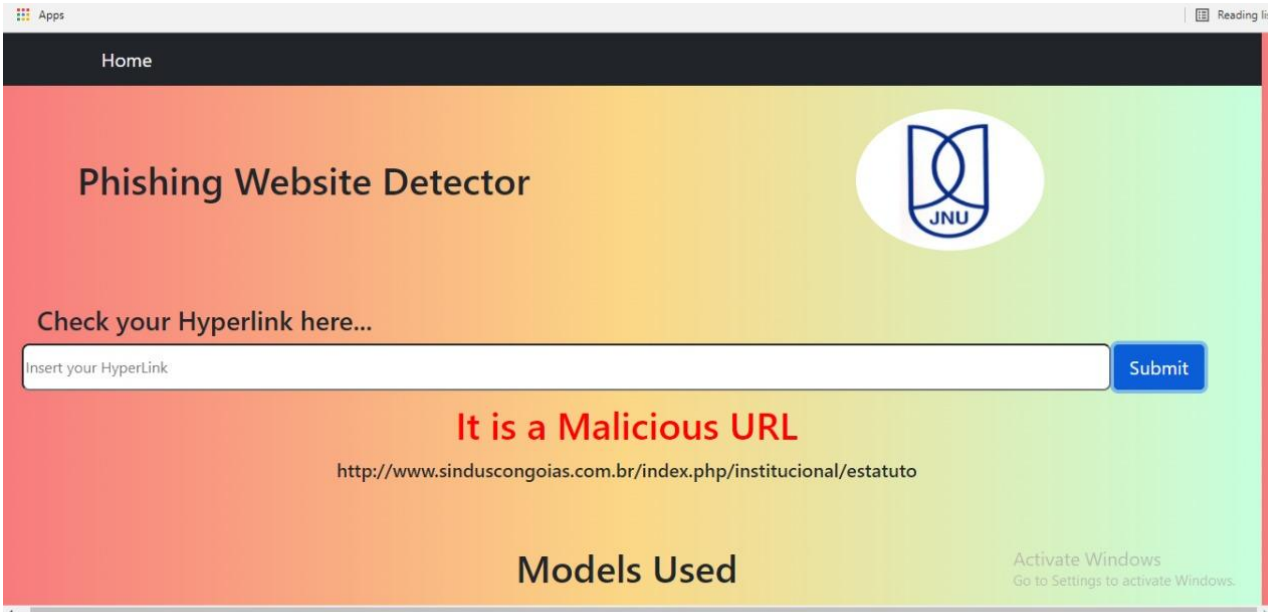
After deploying the model on live server, here are the results on the following URLs :-

1. https://meet.google.com/bzx-jipy-etd (legitimateURL)



---

2. http://www.sinduscongoias.com.br/index.php/institucional/estatuto (phishy URL)



CHAPTER-5

CONCLUSION AND FUTURE SCOPE

5.1. CONCLUSION

Using machine learning technologies, this paper tries to improve the detection mechanism for phishing websites. Using a random forest method with the lowest false positive rate, we were able to obtain 95 % detection accuracy. In addition, we have selected 12 most important features out of 31 using Feature Importance and Correlation Matrix with HeatMap.

5.2. FUTURE SCOPE

The project has come to a stage where it may be implemented at the backend of any web app or web extensions that will prompt a warning message whenever the user encounters any URL that has been classified as phishy by the proposed model.

BIBLIOGRAPHY

[1] Rishikesh Mahajan, “Phishing Website Detection using Machine Learning Algorithms 2018”

[2] https://www.alexa.com/login/

[3] https://in.godaddy.com/offers/whois-b?

[4] Rami M. Mohammed and Fadi Thabtah “Phishing Websites Features”