Safeguarding the Web: Machine learning techniques for Phishing Detection

॥वसुधैव कुटुम्बकम्॥

THESIS SUBMITTED TO
Symbiosis Institute of Geoinformatics

FOR PARTIAL FULFILLMENT OF THE M. Sc.
DEGREE

By

Swachchha Roy

(Batch 2023-25 / PRN 23070243054)

Symbiosis Institute of Geoinformatics
5th Floor, Atur Centre, Gokhale Cross Road, Model Colony,
Pune - 411016

CERTIFICATE

Certified that this thesis titled **Safeguarding the Web: Machine learning techniques for Phishing Detection** is a bonafide work done by **Mr. Swachchha Roy**, at Symbiosis Institute of Geoinformatics, under our supervision.

Supervisor, Internal
Dr. Rajesh Dhumal
Symbiosis Institute of Geoinformatics

Index

LIST OF FIGURES

Acknowledgement

I would like to pay my respect and profound regard to my internal guide of this study, Dr Rajesh Dhumal sir for extending his support and constant encouragement during the conduction of the research. The knowledge and experience he has shared in the project have been instrumental in coming up with this research work.

I would also like to express my appreciation to Symbiosis Institute of Geoinformatics and for offering me the necessary means and a conducive setting in which I could perform my study. The knowledge and experience gained during my time have been instrumental in shaping this project.

The project would have not been completed without the immense help and worthy experience. I was able to research and analyze a lot of topics during this project. It helped me expand my knowledge and skill.

Additionally, I would like to extend my thanks to my colleagues for their constant support and feedback which helped me in refining my work.

List of Abbreviations

Abbreviations	Full Forms
DT	Decision Tree
RF	Random Forest
SVM	Support Vector Machine
XGB	Extreme Gradient Boosting
CNN	Convolutional Neural Network
ANN	Artificial Neural Network
NLP	Natural Language Processing
DNN	Deep Neural Network
RNN	Recurrent Neural Network
LSTM	Long Short-Term Memory
GANs	Generative Adversarial Network
HTML	Hyper Text Markup Language
URL	Uniform Resource Locator
ML	Machine Learning
DL	Deep Learning
CSV	Comma-separated Values
JSON	Java Script Object Notation

Preface

Advancement in the internet uses has seen a very massive growth in convenience and interconnection but has caused emergence of new types of hostile activities most of which are cyber-attacks and the most common type is Phishing. Phishing websites are those websites which are actually fake websites, created mainly to trap people into revealing their personal details like account details, passwords etc. These activities are dangerous as they can lead to monetary losses, identity theft and loss of confidence in online communication.

To this concern, my project titled as "**Safeguarding the Web: Machine learning techniques for Phishing Detection**" aims to provide a useful prediction of fraudulent websites. It was the feature extraction in which properties including domain, IP address, use of @ symbol, use of tiny or short URLs along with web traffic data were collected and transformed into a dataset. Using that dataset, the process of machine learning is carried out.

My reason for choosing such topic is rooted in personal concern of cybercrimes and increased vulnerability in the online web. The effectiveness of modern phishing attacks means that the task of constructing the corresponding defense systems is complex and requires development of new ideas. My goal in this project is to help evolve the field of security enough that it can effectively protect people when they go online and make the internet as safe as it can be.

Introduction

Cybercriminal activities including phishing are commonly practiced where the aim of the criminals is to obtain personal details from the users. It is regarded as a social engineering attack, a technique that targets people's behavior in order to compromise organizational security controls. Being one of the typical cyberattacks employed by hackers, phishing most frequently aims at making people provide accounts' information, including credit card numbers, user names, and passwords. Sometimes, phishing becomes a tool for massive distribution of malware in a network, which adds to the problems of protecting the networks.

Thus, the adversaries continue to popularize phishing that can be classified as spoofing, malware-based phishing, DNS-based phishing, data theft, email/spam, web-based delivery, and phone. All these attacks harness several forms of communication including but not limited to email, instant messages, QR codes, and social media. Web users are tricked by pretending to be trustworthy like banking, credit card services, or popular online shopping stores, into entering their login credentials on fake websites. The ramification of such attacks is serious, which can result in access to bona fide accounts and massive losses.

List based methods or Heuristic-based detection are some techniques previously used to fight against phishing, but those techniques are not very efficient. The techniques fail in detecting the fake sites since there are new methods in phishing that have evolved in the market. Therefore, there is demanding a need for more advanced methods for the identification of phishing websites.

A solution to this problem is provided by Machine Learning. It is for this reason that through the use of big data and algorithms ML models can be able to make sense of and discern the features that set the phishing websites from the original ones. The purpose of this project is to investigate and develop methods for using machine learning that would help identify phishing sites, which concretize the protection of cyberspace.

Literature Review

After deciding the objectives of my study, I reviewed various research publications where authors have followed various algorithms with a variety of datasets. The following table give a brief review of the researches of different authors, along with the description of the datasets used and also the accuracies obtained for each of them.

Review Table

Sr. No.	Author	Dataset	Methods	Result	Publish Year
1	Asadullah Safi, Satwinder Singh	PhishTank website, Alexa website	Machine Learning techniques, Heuristic techniques, Visual Similarity, CNN.	Convolutional Neural Network (CNN) achieved the highest accuracy of 99.98% in detecting phishing websites.	11th January 2023
2	Shouq Alnemari, Majid Alshammar	Phishing dataset from UCI machine learning repository	Artificial Neural networks (ANN), SVM, Decision Tree classifier, Random Forest.	Random forest achieved 97%.	2023
3	Aniket Garje, Namrata Tanwani, Sammed Kandale, Twinkle Zope, Prof. Sandeep Gore	Phishing dataset from UCI machine learning repository	Decision Tree, Generalized Linear Model, Gradient Boosting, Generalized Additive Model, and Random Forest	The Random Forest algorithm showed the highest accuracy of 98.4%, 98.59% recall, and precision of 97.70%	September 2023

4	Mahmoud Khonji, Youssef Iraqi, Senior Member, IEEE, and Andrew Jones	Kaggle	machine learning and clustering algorithms (k-NN, C4.5, SVM, k-means, DBSCAN), similarity-based detection, Bayesian approach, hybrid features, Google Safe Browsing API, Bayesian classification, NLP techniques	Automatic Detection of Phishing Target from Phishing Webpage: FP rate of 3.4% and FN rate of 8.56%. Detecting DNS-poisoning-based phishing attacks from their network performance characteristics: FP rate of 0.7% and FN rate of 0.6%. Textual and Visual Content-Based Anti-Phishing: A Bayesian Approach: FP rate of 0-0.02% and FN rate of 0-1.95%.	
5	Cagatay Catal, Görkem Giray, Bedir Tekinerdogan, Sandeep Kumar, Suyash Shukla	18 different datasets for deep learning-based phishing detection, with PhishTank being the most used dataset	Deep Neural Networks (DNN), Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), Long Short-Term Memory Networks (LSTM), Autoencoders Generative Adversarial Networks (GANs)	Achieved an accuracy ranging from 94% to 99.34%.	2020
6	Ali Aljofey, Qingshan Jiang, Abdur Rasool, Hui Chen, Wenyin Liu, Qiang Qu, Yang Wang	The dataset used in the study includes 60,252 webpages, with 32,972 benign webpages and 27,280 phishing webpages.	URL character sequence features, various hyperlink information, and textual content of the webpage to train the XGBoost classifier. The features are extracted from the HTML source code and URL without relying on third-party services.	The proposed approach achieved an accuracy of 96.76% with a false-positive rate of 1.39% on the custom dataset.	2022

7	Ashit Kumar Dutta	PhishTank and Crawler dataset	Utilizes a combination of URL and HTML features evaluated through various machine learning classifiers.	The proposed method achieved accuracy of 98.48%.	2022
8	P. Amba Bhavani, Chalamala Madhumitha, Pinnam Sree Likhitha, Chanda Pranav Sai	Kaggle	CNN LSTM, Logistic regression	CNN LSTM 57.85% and Logistic Regression 91.89%	

The document titled **"A systematic literature review on phishing website detection techniques"** is a systematic literature review on phishing website detection techniques. It compares dissimilar approaches such as list based, visual similarity, heuristic, machine learning and deep learning. This review was performed on 80 scientific papers published in the last 5 years including algorithms, datasets and research questions related to phishing websites detection. Machine learning techniques had the most applications with 57 studies using them. The Phish-tank website was the primary source for phishing website which were explored in 53 studies while Alexa's site was used for downloading legitimate datasets in 29 studies. In terms of ML methods, RF was used by 31 articles. CNN achieved the highest accuracy at a rate of 99.98%.

"Detecting phishing domains with Machine Learning" is an article that focuses on the menace of phishing and suggests employing machine learning to identify them. The researchers used the UCI Phishing domains to evaluate 4 ML models: SVM, ANN, RF and DT. The RF model has the highest accuracy of 97% of all other works in this area. It also hints on previous researches conducted on phishing domains such as Ensemble learning and analysis of HTML page source codes.

ML techniques demonstrated in **“Detecting phishing websites using Machine Learning”** as a means of discovering phishing websites. It refers to the phishing as a type of cyber-security threat which often involves theft of personal information like passwords and credit card numbers. Phishing website detection uses ML algo like KNN, Naïve Bayes, Gradient Boosting, DT. The paper additionally contains an overview of various research papers that have been dedicated to this issue. The study also contains tables that show the confusion matrices for different algorithms used. To conclude, f1 score is best with DT being concluded as better for phishing website detection.

The document **“Phishing Detection: a Literature Survey”** includes topic such as expanding user awareness and using extra software that are used to phishing attempts. It includes topics such as expanding user awareness and using extra software that are used to detect phishing attempts. In addition to this, it goes into list-based detection methods particularly through the use of whitelist and blacklist techniques; machine learning based detection approaches. The proposed approach improves accuracy of phishing detection by extracting and analyzing various characteristics on which suspects’ websites can be identified. For example, it introduces eight new features combined with existing ones to generate a feature vector for each webpage. During its detection phase, XGB classifier has been used for building strong classifiers for phishing detection purposes. The authors have also assessed how different features perform under the same classifiers thereby illustrating why their approach works better than others do within this context. The work ends by stating where the method does not work very well at present as well as possible future studies on identifying phishing attacks.

The document **“Applications of Deep Learning for Phishing Detection: a systematic literature review”** contains a systematics literature review (SLR) for the deep learning-based phishing detection techniques. References are categorized under regular or primary references. Regular references are derived from the basic research areas covering the DLS (Deep Learning Techniques based detection of Phishing). They are broadly classified as follows:

Deep Learning, Machine Learning, Convolutional Neural Networks, Recurrent Neural Networks, Feature Engineering, Detection Models Primary Studies: These studies covered the area of DLS based phishing Detection. They are pertaining to the above mentioned four regular references. Research Objectives: This part details out the research objectives of the SLR. This presents the method followed in this SLR.

The document titled **as "An effective detection approach for phishing websites using URL and HTML features"** gives an overview of phishing detection techniques which are increasing awareness of the user and additional software. Also, list-based detection methods, as long as whitelist or blacklist methods, are discussed here and also machine learning-based detection techniques. Thus, the proposed technique suggests that in order for there to be better identification rates with regard to cases such as website forgery there should be some extraction and examination done on particular elements constituting those pages minus warning signs among others. To do this job well requires extraction of various features that together make up suspected webpages including those that have not yet surfaced formerly called novel document features belonging into each candidate set liveness text in form field-based Image-only approach VKDZ distinctiveness. Introduce eight new features, combining them with those that already exist.

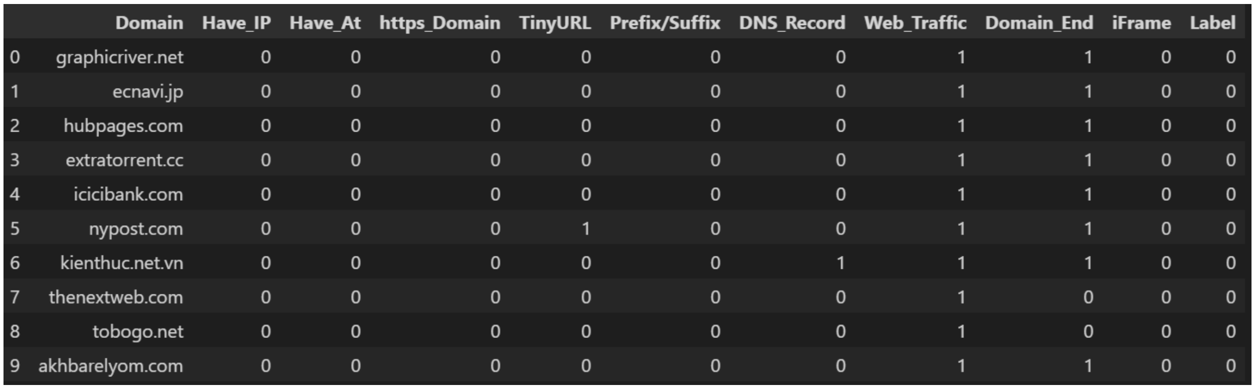
The article **"Detecting Phishing websites using Machine Learning technique"** considers a research study for detecting either malicious or legitimate URLs using Machine Learning methods, specifically the RNN-LSTM approach. The proposed technique is called LURL. Compared with other state-of-the-art URL detectors, it had better accuracy and F1-score. Details are provided in this document pertaining to methodology with equations and algorithms used in data collection, preprocessing, training, and testing. It also covers the history of phishing attacks, classification of phishing attack techniques, and related research works. At the end, the paper is concluded by presenting the need for ML-based anti-phishing techniques and specifying the future directions. Also, one can find the reference list on Phishing Detection and Machine Learning with respect to URL-based phishing detection, machine learning algorithms, feature selection techniques, and use of deep learning for malicious URL detection.

The document **“Phishing Website detection using Machine Learning”** find the phishing URL by comparing different machine learning algorithms in terms of accuracy, false positive and false negative. The document further includes a literature survey, detailing how the respective methods used till now i.e., online toolbars, data mining algorithms and deep learning approaches. Results section will provide the accuracy of CNN LSTM, CNN Bi-LSTM, Logistic regression and XGB machine learning models. This discussion will shed a light on shortcomings of previous methods and benefit using machine learning-based approaches.

Methodology

The first and foremost step is data collection. The set of phishing URLs are collected from opensource service that exist as PhishTank in web. This service delivers a combination of phishing URLs in CSV, JSON and any format of file that is updated hourly. Thus, from this dataset, 5000 random samples are extracted.

The legitimate URLs are collected from open datasets of University of Brunswick. This dataset has collection of benign, spam, phishing, malware and defacement URLs. Out of all these types, the 5000 benign URL dataset is collected for this project to train the ML models on them.



	Domain	Have_IP	Have_At	https_Domain	TinyURL	Prefix/Suffix	DNS_Record	Web_Traffic	Domain_End	iFrame	Label
0	graphicriver.net	0	0	0	0	0	0	1	1	0	0
1	ecnavi.jp	0	0	0	0	0	0	1	1	0	0
2	hubpages.com	0	0	0	0	0	0	1	1	0	0
3	extratorrent.cc	0	0	0	0	0	0	1	1	0	0
4	icicibank.com	0	0	0	0	0	0	1	1	0	0
5	nypost.com	0	0	0	1	0	0	1	1	0	0
6	kienthuc.net.vn	0	0	0	0	0	1	1	1	0	0
7	thenextweb.com	0	0	0	0	0	0	1	0	0	0
8	tobogo.net	0	0	0	0	0	0	1	0	0	0
9	akhbarelyom.com	0	0	0	0	0	0	1	1	0	0

The final dataset, which contains 10000 URLs of which 5000 are phishing and 5000 are legit. This dataset is formed by concatenating the two datasets and shuffled.

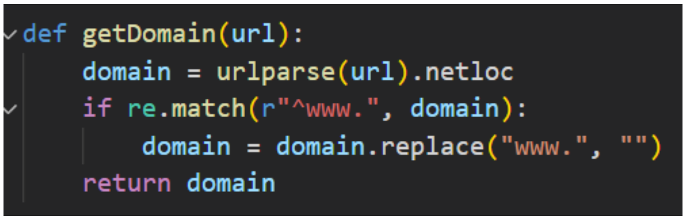
The second step involves feature extraction. Feature extraction is a very essential step in the detection process where the characteristics of the URLs are mainly categorized into 3 main groups:

1. Address Bar based features
2. Domain based features
3. HTML based features

Address Bar Based Features

The following features were pulled from the address bar of the given URL:

- Domain: The first part of the URL which tells the legitimacy of the website.



```python
def getDomain(url):
    domain = urlparse(url).netloc
    if re.match(r"^www.", domain):
        domain = domain.replace("www.", "")
    return domain
```

- IP address: URL contains usually the domain name but if it contains an IP address, then there is a possibility of phishing attack.
- "@" symbol: The case in which "@" symbol is placed within the URL can make the browser go to a different address, which can be a phishing attempt.
- "https" in Domain: Official or legal sites employ "https" for protected communication with the client.
- URL shortening service: The actual link usually disguises itself and in doing so, URL shortening services are usually employed by phishers.
- Prefix or suffix "-" in domain: Hyphens in the domain name can also be considered as a sign of phishing due to the tendency of phishers to use sub-domains to replicate definite websites.

Domain Based Features

The following domain-based features were considered:

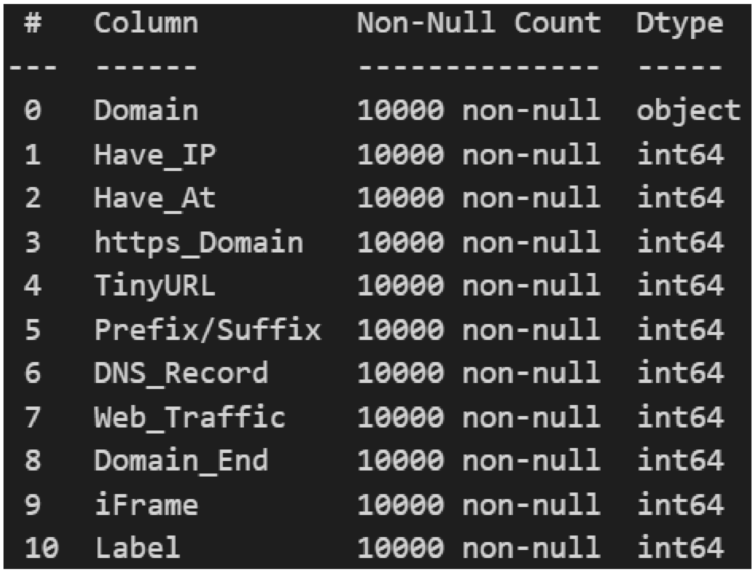
- Website Traffic: It means traffic or popularity of the website in receiving traffic. A well-established website is generally more popular as compared to the fake one.
- End period of Domain: The time period for which this domain has been registered. Cybercriminals registering the various phishing domains usually do so for shorter durations.

HTML based features

HTML based features are the characteristics of the page that are extracted by analyzing the HTML code that describes the content of the page.

- IFrame Redirection: Similar to the case with referential metadata tags, the utilization of the IFrame tags with a view of inserting an outside content can be exploited for unethical motives with intention of concealing the genuine identity of a webpage.

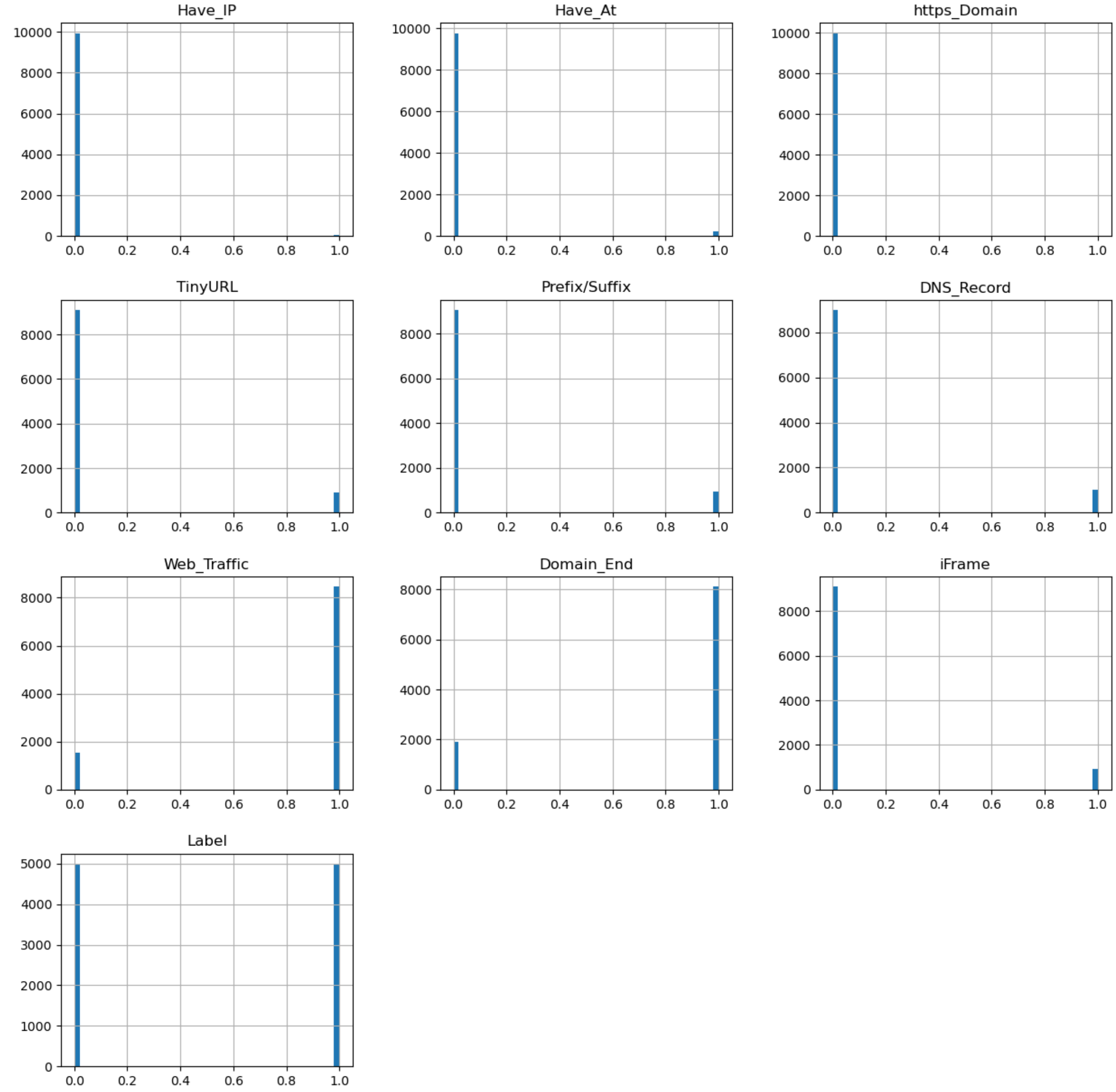
Now comes the Data Preprocessing, where raw data is transformed into clean and usable format before using it in machine learning models.



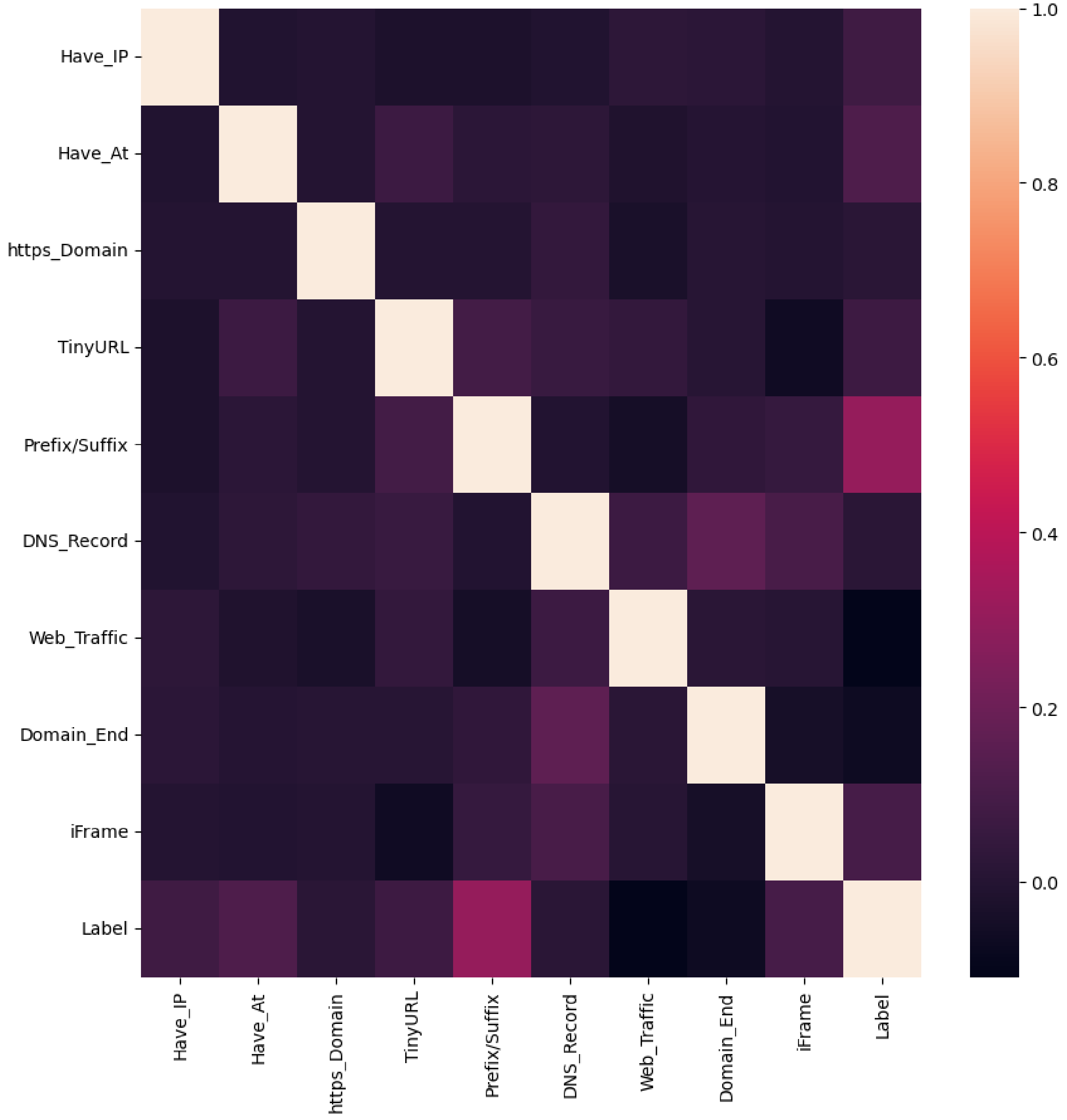
```
 #   Column         Non-Null Count  Dtype 
---  ------         --------------  ----- 
 0   Domain         10000 non-null  object
 1   Have_IP        10000 non-null  int64 
 2   Have_At        10000 non-null  int64 
 3   https_Domain   10000 non-null  int64 
 4   TinyURL        10000 non-null  int64 
 5   Prefix/Suffix  10000 non-null  int64 
 6   DNS_Record     10000 non-null  int64 
 7   Web_Traffic    10000 non-null  int64 
 8   Domain_End     10000 non-null  int64 
 9   iFrame         10000 non-null  int64 
 10  Label          10000 non-null  int64 
```

Information regarding the data

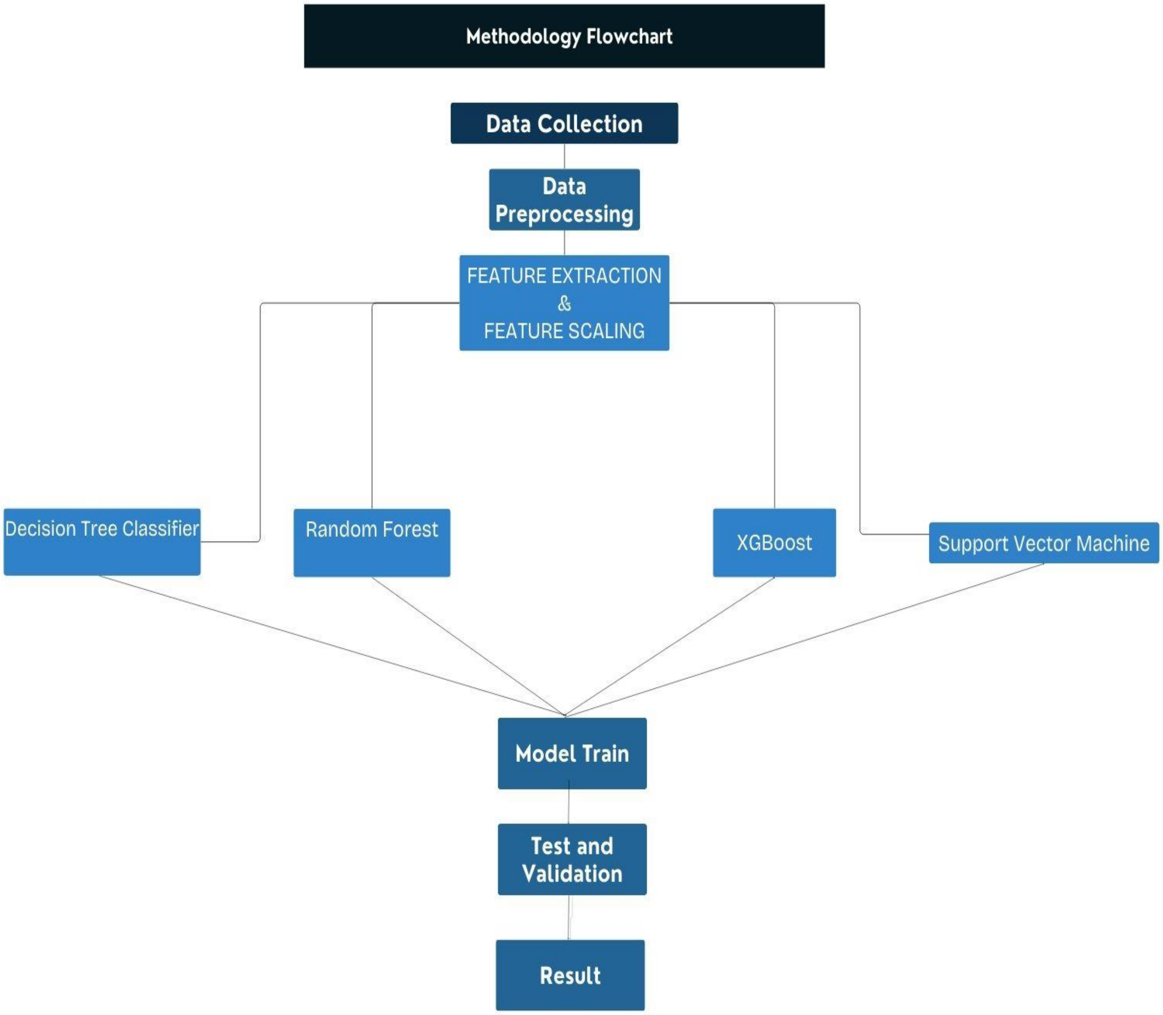
This is the histogram which represent the frequency of data in the class 0 and 1.



Then comes the Correlation heatmap which represents the correlation between the features of the data. Correlation is the relation between the features of the data or it defines how the features of the data are correlated with each other.



After the preprocessing, the data is split into two ratios: first is 80% - 20% as training and testing and secondly 70% - 30% as training and testing respectively. Then that trained data is passed into the machine learning algorithms like Decision Tree Classifier, Random Forest, XGB and Support Vector Machine. The motive behind selecting the above algorithms is that these algorithms are beneficial for classification problems under supervised machine learning. Models are evaluated using confusion matrix and accuracy of the training and testing data. Following is the architectural diagram for the methodology:



Decision Tree is one of the categories of Supervise learning that is applied in the solution of both classification and regression problems. It partitions the data by making a decision regarding the values of the input features which takes it to look like a decision tree.

Random forest is another ensemble learning technique that build up and provide median of those trees' prediction for the classification problem. RF is more stable as it averages multiple trees from which it reduces variance, less prone to overfitting and good for large datasets.

XGBoost is a specialized gradient boosting framework, enhanced to be efficient, versatile and portable. It consumes less memory, fast robust in dealing with NA values.

SVM is a learning technique categorized under supervised learning and is applied to classify as well as regression analysis. It looks for the hyperplane that best centralizes the classes in a multidimensional space while maximizing the distance between them.

By using these algorithms, we trained the model with the data which we split it into 80% – 20% and 70% - 30% respectively and perform the relevant operations.

Results

In the result section, the findings on the use of different algorithms in identifying the phishing website are discussed. This section contains he comparative analysis that compares the efficiency of the utilized algorithms and the corresponding visualizations. The approaches involved in this project are DT, RF, XGB and SVM and the model is trained in two ways:

- 80-20 split for training and testing
- 70-30 split for training and testing

Accuracy of the model is the percentage of the number of instances correctly classified out of the total number of instances.

Confusion matrix is a matrix which assess the model of classification based on various measures of performance.

DECISION TREE CLASSIFIER (80 - 20 split)

- Accuracy: 1.0 (on training data), 1.0 (on testing data)
- Confusion matrix: [[1615 0], [0 385]]

DECISION TREE CLASSIFIER (70 - 30 split)

- Accuracy: 1.0 (on training data), 1.0 (on testing data)
- Confusion matrix: [[2365 0], [0 635]]

RANDOM FOREST (80 – 20 split)

- Accuracy: 0.99525 (on training data), 0.998 (on testing data)
- Confusion Matrix: [[1615 0], [4 381]]

RANDOM FOREST (70 – 30 split)

- Accuracy: 0.99571 (on training data), 0.996 (on testing data)
- Confusion matrix: [[2365 0], [12 623]]

XGBOOST (80 – 20 split)

- Accuracy: Accuracy: 1.0 (on training data), 1.0 (on testing data)
- Confusion matrix: [[1615 0], [0 385]]

XGBOOST (70 – 30 split)

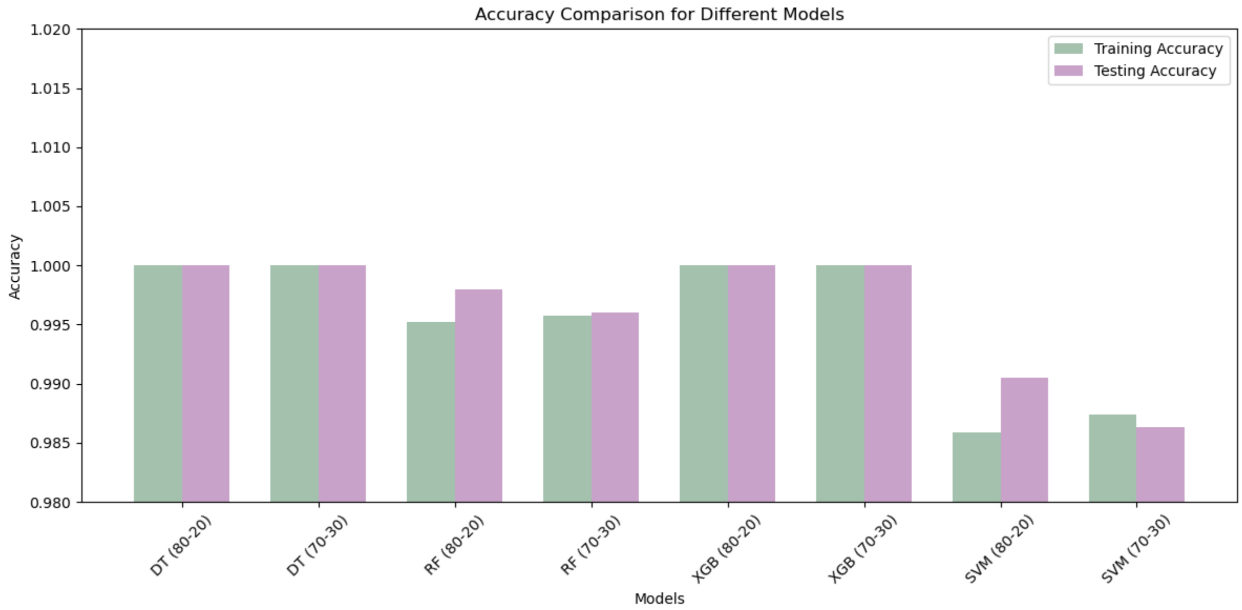
- Accuracy: Accuracy: 1.0 (on training data), 1.0 (on testing data)
- Confusion matrix: [[2365 0], [0 635]]

SUPPORT VECTOR MACHINE (80 – 20 split)

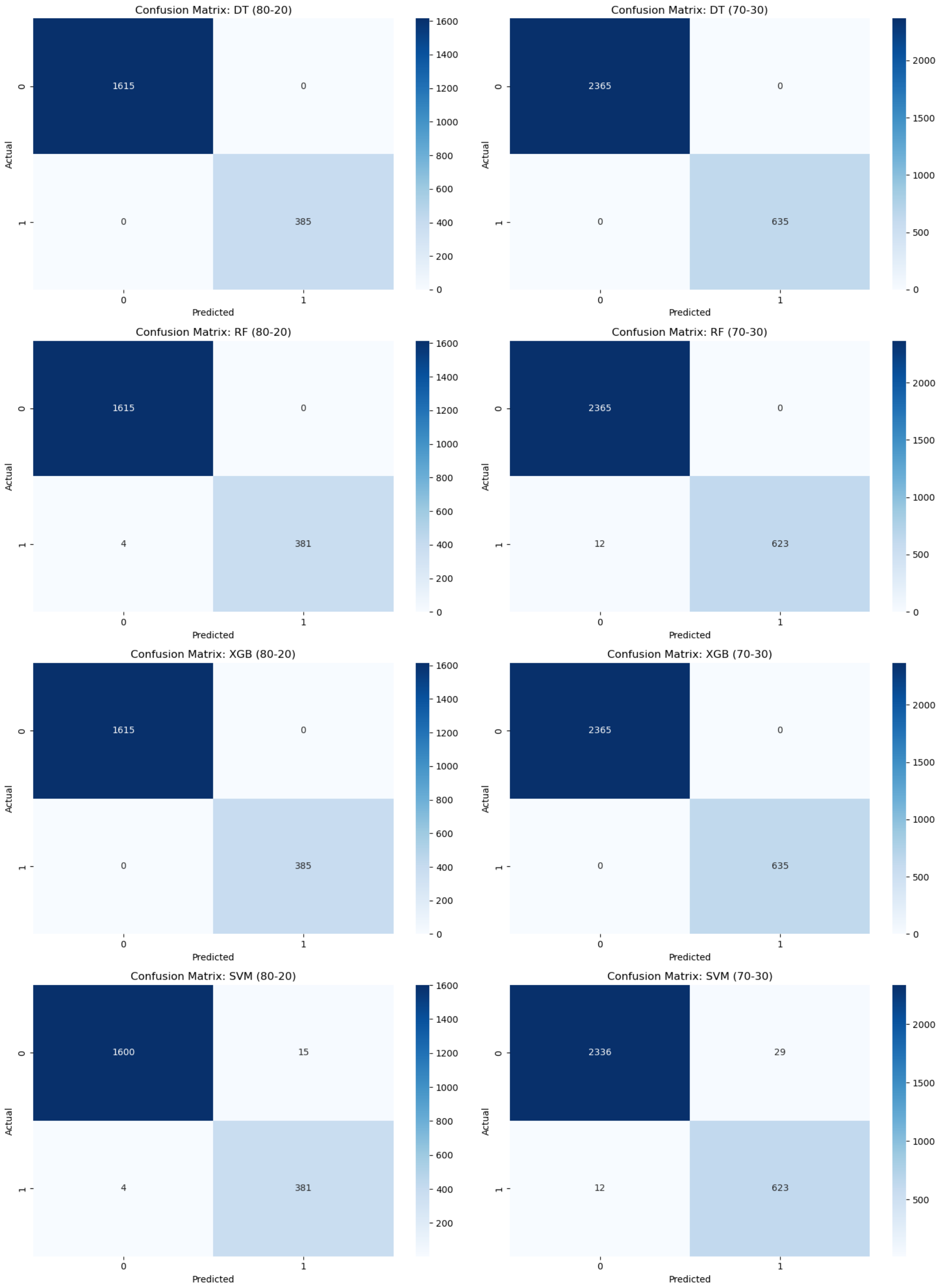
- Accuracy: Accuracy: 0.985875 (on training data), 0.9905 (on testing data)
- Confusion matrix: [[1600 15], [4 381]]

SUPPORT VECTOR MACHINE (70 – 30 split)

- Accuracy: Accuracy: 0.98742 (on training data), 0.98633 (on testing data)
- Confusion matrix: [[2336 29], [12 623]]



This bar chart provides us a clear comparison of the accuracy of the different models on both training and testing data for two different data splits. From the chart we can conclude that DT and XGBoost are much more accurate than other models.



The visualization in the previous page is a confusion matrix heatmap which shows the number of true positives, false positives, true negative and false negatives. They help in understanding the performance of the classification models.

Conclusion

The project "Safeguarding the Web: Machine Learning techniques for Phishing Detection" has successfully illustrated how various machine learning techniques can be used and how they are useful in the identification of phishing websites and the differentiation of these sites from genuine ones. The process started with amassing a broad set of phishing URLs from the PhishTank website and a host of legitimate URLs from the University of Brunswick. The dataset with 10,000 URLs is clean and comprehensively it was prepared in an optimal format for the model.

There was a feature extraction step where URLs were divided under Address based, Domain based and HTML based. From these features, the machine learning algorithms were able to learn from and detect the presence of phishing attempts.

After employing two kinds of data splitting (80-20 and 70-30), the fundamental set of classifiers included the following: DT, RF, XGB and SVM. The performance achieved by these models was exemplary and both the DT and the XGB scored 100% with respect to training and testing data sets. The SVM that featured slightly lower accuracy was also very reliable in classification.

The work done in this project shows that feature extraction is a very crucial step. It must be pointed out that according to the presented accuracy and confusion matrix results, ML techniques could be rather useful in identifying phishing websites and thus minimize fraud and protect the users from scams. These work not only helps in expanding the field of cybersecurity but also gives groundwork for the subsequent study and advancement for the identification of all other kinds of threats through machine learning on the internet.

References

1. Safi, A., & Singh, S. (2023). A systematic literature review on phishing website detection techniques. *Journal of King Saud University - Computer and Information Sciences*, *35*(2), 590–611. https://doi.org/10.1016/j.jksuci.2023.01.004

2. Alnemari, S., & Alshammari, M. (2023). Detecting phishing domains using machine learning. *Applied Sciences*, *13*(8), 4649. https://doi.org/10.3390/app13084649

3. Garje, A., Tanwani, N., Kandale, S., Zope, T., & Gore, S. (2021). Detecting phishing websites using machine learning. *PloS One*, *9*(2320-2882)

4. Khonji, M., Iraqi, Y., & Jones, A. (2013). Phishing Detection: A Literature Survey. *IEEE Communications Surveys & Tutorials*, *15*(4), 2091–2121. https://doi.org/10.1109/surv.2013.032213.00009

5. Mahajan, R., & Siddavatam, I. (2018). Phishing Website Detection using Machine Learning Algorithms. *International Journal of Computer Applications*, *181*(23), 45–47. https://doi.org/10.5120/ijca2018918026

6. Catal, C., Giray, G., Tekinerdogan, B., Kumar, S., & Shukla, S. (2022). Applications of deep learning for phishing detection: a systematic literature review. Knowledge and Information Systems, 64(6), 1457–1500. https://doi.org/10.1007/s10115-022-01672-x

7. Aljofey, A., Jiang, Q., Rasool, A., Chen, H., Liu, W., Qu, Q., & Wang, Y. (2022). An effective detection approach for phishing websites using URL and HTML features. *Scientific Reports*, *12*(1). https://doi.org/10.1038/s41598-022-10841-5

8. Dutta, A. K. (2021). Detecting phishing websites using machine learning technique. *PLoS ONE*, *16*(10), e0258361. https://doi.org/10.1371/journal.pone.0258361

9. Bhavani, P. A., Chalamala, M., Likhitha, P. S., & Sai, C. P. S. (2022). Phishing websites detection using machine learning. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4208185