**PHISHING WEBSITE DETECTION USING MACHINE LEARNING**

**Pallavi Male Shruti Pargaonkar Ashish Sasatte**

**Prof. Mr. Shantanu Pawar**

Department of Information Technology, P.E.S. Modern College of Engineering, Pune, Maharashtra, INDIA

**Abstract**

In Today`s world almost all people are dependent on internet. Nowadays Phishing has become the main area of concern for security researchers because it is not difficult to create the fake website which looks so close to legitimate website. So, the internet devices are more prone to cybercrimes. Also, the most popular cybercrime that has been performed in back 2 years is Phishing. Phishing is the way to obtain sensitive information from innocent users. Aim of the phishers is to acquire critical information like username, password and bank account details.This project will determine whether the URL is legitimate or phishing. Cyber security people are now looking for trustworthy and steady detection techniques for phishing website detection. This system deals with machine learning technology for detection of phishing URL`s by extracting and analyzing various features of legitimate and phishing URL`s. Decision tree, Support Vector Machine, Random Forest such algorithms are used to detect phishing websites. The aim is to make user aware of the phishing website easily and protect them from getting hoodwinked and losing their sensitive information.

1. **INTRODUCTION**

This is the phishing website detection chrome extension. As we all know, amounts of users getting phished are increasing, the cybercrime has increased rapidly since pandemic. Unknowingly after clicking on the trapped website, user can easily get hacked. Phishing is when users are tricked into giving their sensitive details like login credentials to the hacker. The phishing links looks very similar to the legitimate links, this hoodwinks the user. Considering these, points the machine learning algorithm is used to check whether the website is phishing or legitimate, and this will be indicated through the chrome extension.

1. **MOTIVATION AND OBJECTIVES**

**Motivation:**

1. To detect phishing website.
2. To make user secure from the phishing website
3. To make people aware about the phishing attack performed through URLs.

**Objectives:**

1. To train the Machine Learning algorithm such that they will able to detect the legitimate and phishing website effectively.
2. The End user will able to recognize legitimate / phishing website by an alert message.
3. By using the extension method, it will be able to save the user`s time.
4. **PROBLEM STATEMENT**

Detection of phishing website through chrome extension using machine learning.

There are many techniques to detect the phishing websites such as, whitelist, blacklist, fuzzy rule, gradient boosting and XGBoost and many more, but looking at the accuracy criteria we are using machine learning algorithms to train the models which will detect whether the website is legitimate or not. The most giving accuracy model is finalized for the further implementation, The Algorithms we are going to check are Logistic Regression, Decision Tree, Random Forest, Naïve Bayes, Support Vector Machine (SVM). The key advantage of machine learning is the ability to create flexible models for specific tasks like phishing detection.

1. **SCOPE**

Traditional method of detecting phishing website somewhere lags in the accuracy because phishing website can often evade detection by using sophisticated techniques like social engineering, domain spoofing and hence is risky, to overcome this limitation we had used Machine Learning in order to get better accuracy compared to Traditional method.

Machine Learning will help system to adapt and learn from new data, making it more effective.

1. **METHODOLOGY**

There are various methods available for the detection of phishing websites like, whitelist, blacklist, fuzzy rule, gradient boosting and XGBoost, web scraping and more. But, to gain more efficiency, compared to other methods Machine Learning can be used as a powerful tool. Following are the machine learning tools we have used to check the accuracy:

1. **Logistic Regression:**

Logistic Regression is a supervised learning used in binary classification. The logistic function maps the weighted sum to a value between 0 and 1, which represents the probability of the website being a phishing website.1 is Phishing and 0 is Legitimate.

1. **Decision Tree:**

Decision tree algorithm recursively partitions the feature into smaller and smaller subset having informative feature on each node. The partitioning starts from the root and further divided into sub nodes until the leaf node is acquired indicating whether the website is phishing or not.

1. **Random Forest:**

Random tree is a combination of decision trees. During training, each decision tree in the forest is trained on a random subset of the data and a random subset of the features, which helps to reduce overfitting and improve generalization.

1. **Nave Bayes:**

It is a probabilistic machine learning algorithm. we first need to collect a dataset of labelled examples (i.e., phishing or legitimate). Next, we need to select a set of relevant features that can help to distinguish between the two classes. Naïve is called because it makes the assumptions that all attributes are independent of each other. Bayes depends upon the Bayes Theorem, following is the formula

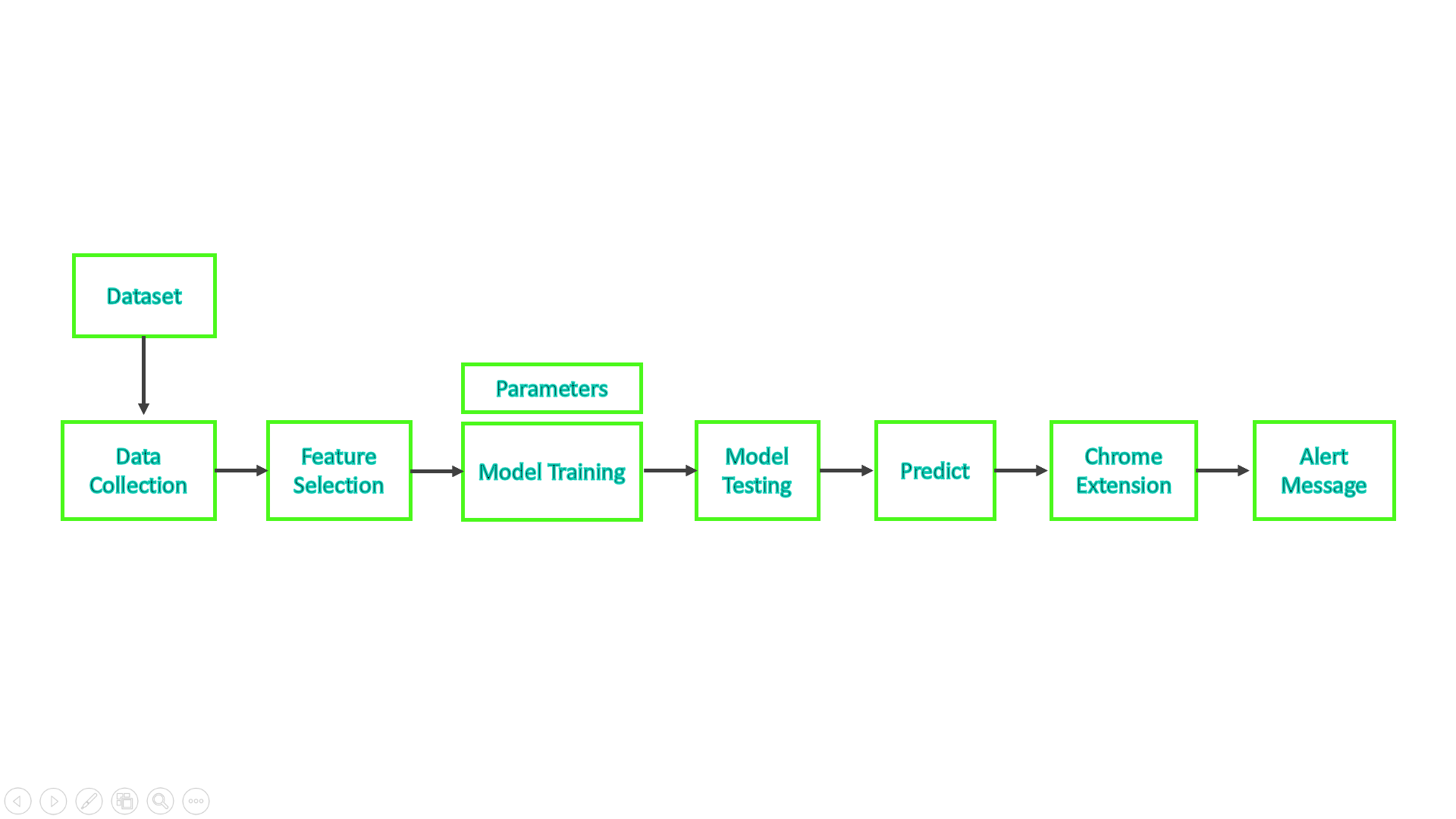
P(Class | Features) = P(Class) \* P(Features | Class) / P(Features) P(Class) = Prior probability

P(Features) = probability of features P(Features | Class) = likelihood

1. **Support Vector Machine:**

It is also a supervised learning algorithm based on binary classification, it`s aim is to draw a hyperplane by maximally segregating the two classes. In SVM, the data points that are closest to the hyperplane are called support vectors, and they are used to define the hyperplane. Websites can be classified by mapping them into the same feature space and determining which side of the hyperplane they fall on. If a website falls on one side of the hyperplane, it is classified as legitimate; if it falls on the other side, it is classified as phishing.

1. **WORKFLOW**



**Figure 1:** Workflow Diagram

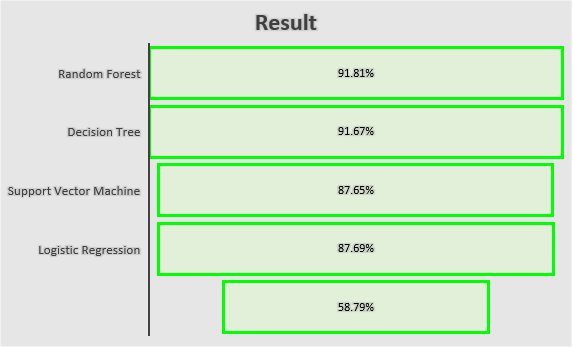
As seen in the above flowchart, data collection procedure is applied on the dataset, here we are using the dataset installed and extracted from the Kaggle website. Then, with the help of suitable algorithm we have trained and tested the model, the higher accuracy we got is for Random Forest Algorithm, After, training the model the model will predict whether the website is phishing or not, this message will be conveyed to the user through the chrome extension, after activation of the chrome extension.

Sample size or demographics: The size of the dataset used is of 95910 rows, which are further divided into training set and testing set, training set has 76728 of data whereas testing set has 19182 of data from the dataset.

Data collection techniques: Data is collected from the dataset, we have used the dataset which is downloaded from the Kaggle website, the dataset has final column as a label column which indicates whether the website is legitimate or phishing, if label value is 1, then the result is phishing, if value is 0, then the result is legitimate.

1. **RESULT**

We Obtained accuracy for the Machine Learning Model as follows:

****

Naïve Bayes

**Figure 2:** Result

1. **TECHNOLOGY REQUIREMENTS**
2. **IDE –** Jupyter Notebook, Visual Studio
3. **Language –** HTML, CSS, python, json,
4. **Dataset –** Kaggle
5. **LIMITATIONS AND FUTURE WORK**

**Limitations:**

1. There can be some phishing website having features like legitimate website, this may create problem to accurately distinguish between the two.
2. Hackers can try to make the same website link as the legitimate link, and can hoodwink the user, for this the extra security measures should be taken.
3. The output of the model will depend upon the size of dataset

**Future work:**

1. Implementing the model with Convolutional Neural Network (CNN).
2. Applying more security measures.
3. **CONCLUSION**

Overall, the research paper concludes that the project made will have maximum accuracy as accuracy of different algorithms are checked, the website will make user aware of the phishing website through the chrome extension. And it will help user from getting trapped and losing their information.

1. **REFERENCES**
2. Shahrivari, V., Darabi, M. M., & Izadi, M. (2020).

Phishing detection using machine learning techniques. arXiv preprint arXiv: 2009.11116.

1. Mahajan, R., & Siddavatam, I. (2018).

Phishing website detection using machine learning algorithms. International Journal of Computer Applications, 181(23), 45-47

1. Kalaharsha, P., & Mehtre, B. M. (2021).

Detecting Phishing Sites--An Overview. arXiv preprint arXiv: 2103.12739.

1. Arathi Krishna V, Anusree A, Blessy Jose, Karthika Anilkumar, Ojus Thomas Lee, (2021).

Phishing Detection using Machine Learning based URL Analysis: A Survey,

INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH & TECHNOLOGY

(IJERT) NCREIS – 2021 (Volume 09 – Issue 13)