# Phishing Website Detection

Dhyey Savaliya
*Department of Electronics and Communication Engineering*
*Indian Institute of Information Technology Surat*
Surat, India
ui21ec60@iiitsurat.ac.in

***Abstract*—To develop a Machine Learning Model to detect Malicious URLs and Phishing Attempts. The ML Model is trained and tested on a Dataset with 11470 urls with 50% Phishing and 50% Legitimate URLs , The Model is an Ensemble Model based on several ML Algorithms such as K-Nearest Neighbours (KNN) , Decision Tree (DT) , Logistic Regression (LR) with soft voting . The document gives the complete idea about phishing URLs Detection and Working of the Model.**

***Index Terms*—Phishing , K-Nearest Neighbours, Decision Tree , Logistic Regression , Ensemble Model.**

## I. Introduction

Phishing is the most common form of cyber crime, with an estimated 3.4 billion spam emails sent every day. Phishing is a form of social engineering and scam where attackers deceive people into revealing sensitive information or installing malware such as ransomware. According to the FBI the Most Common Cybercrime is Phishing . The key problem with this is that there is a lack of knowledge among the common people about such scams and everyday a lot of people fall into such scams. The Scammers usually use Link Manipulation , Social Engineering or Filter Evasion .

Link Manipulation : The misspelling of URLs with some small error in the Spelling makes the untrained eye believe that the URL is a genuine URL . Filter Evasion : The use of Images instead of text in the Webpages for avoiding the Various filters used in Web .

Social Engineering : The use of prompts with some warning or some reward making the user believe or making him open the url by playing with Mindset of Winning or Losing something making a decision in a short time.

The project falls under Cyber Security utilizing Machine Learning, specifically designed for detecting Malicious URLs. The Malicious URL Detection operates on Feature Engineering and Supervised Learning. There are 87 features extracted from the URLs, such as Length, Length of Subdomain, Prefix, Suffix, Number of Symbols (@, $, #, etc.), Shortened URLs, Malicious Domain, Google Index, Number of Redirections, DNS Record, etc.

## II. Existing Solutions

There are several methods currently used in Cybersecurity , such as Blacklisting , Checking the no of Special Characters , Content Analysis , Collaboration and Threat Intelligence Sharing , Several Machine Learning Models and Artificial Intelligence , etc.

## III. Proposed Model

### A. Correlation

Correlation measures the strength and direction of a linear relationship between two variables. The Pearson correlation coefficient (often denoted as $r$) is a commonly used measure of correlation. It ranges from -1 to 1, where:

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n \sum x^2 - (\sum x)^2][n \sum y^2 - (\sum y)^2]}} \tag{1}$$

where:

$n$ : number of data points

$x$ : values of the first variable

$y$ : values of the second variable

$\sum$ : sum over all data points

The correlation coefficient $r$ indicates the following:

- $r > 0$: Positive correlation (as one variable increases, the other tends to increase).
- $r < 0$: Negative correlation (as one variable increases, the other tends to decrease).
- $r = 0$: No linear correlation (the variables are not linearly related).

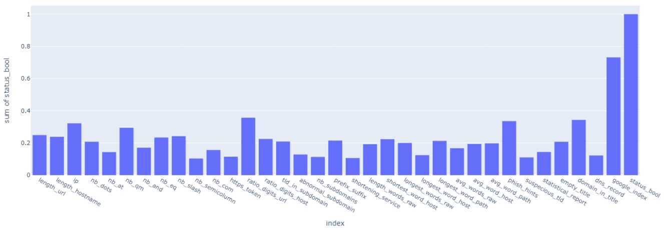

Fig. 1. Correlation between features and output variable

The ML Model Proposed is a Supervised Model based on the Ensemble Model of 3 Models . The Models used in the ensemble are KNN , DT and Logistic Regression . The dataset used for the Training of the Model has 11470 URLs and 87 Features extracted from the URLs , The features are numerical so the Correlation is calculated between the Status (Phishing(1) or Legitimate(0)). The features with less correlation (¡ 0.1) are dropped , 0.1 was taken as a reference for removing features as it is very small compared to other features , the best accuracy was found with 0.1 as a reference. The selection of KNN , DTs and Logistic Regression is because these models are very effective against large no of features and are not affected much if any value is missing in the data extracted from the new URL.

### B. KNN

K Nearest Neighbors-based classification is a type of distance-based learning or non-generalizing learning: it does not attempt to construct a general internal model, but simply stores values of the training data. Classification is computed from a simple majority vote of the k nearest neighbors of each point: a query point is assigned the data class which has the most representatives within the nearest neighbors of the point. The k-neighbors classification in K Neighbors Classifier is the most commonly used technique. The choice of the value of K is highly data-dependent: in general a larger suppresses the effects of noise (Empty Data , Null Values), but makes the classification boundaries less distinct.

### C. Decision Tree

Decision Trees (DTs) are a non-parametric supervised learning method used for classification and regression. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features. A tree can be seen as a piece wise constant approximation.

### D. Logistic Regression

Predictive analytics and classification frequently use this kind of machine learning regression model, also referred to as a logit model. Depending on the given dataset of independent features, the logistic regression model calculates the probability that an event will occur, such as voting or not voting. Given that the result is a probability of an event happening, the dependent feature's range is 0 to 1.

In the logistic regression model, the odds of winning the probability of success of an event divided by the probability of failure-are transformed using the logit formula.

### E. Voting Classifier

A voting classifier is a machine learning model that gains experience by training on a collection of several models and forecasts an output (class) based on the class with the highest likelihood of becoming the output. To forecast the output class based on the largest majority of votes, it averages the results of each classifier provided into the voting classifier. The concept is to build a single model that learns from various models and predicts output based on their aggregate majority of votes for each output class, rather than building separate specialized models and determining the accuracy for each of them. There are two types of Voting Classifiers : Hard and Soft. Hard Voting: In hard voting, the predicted output class is a class with the highest majority of votes, i.e., the class with the highest probability of being predicted by each classifier. For example, let's say classifiers predicted the output classes as (Cat, Dog, Dog). As the classifiers predicted class "dog" a maximum number of times, we will proceed with Dog as our final prediction. Soft Voting: In this, the average probabilities of the classes determine which one will be the final prediction. For example, let's say the probabilities of the class being a "dog" is (0.30, 0.47, 0.53) and a "cat" is (0.20, 0.32, 0.40). So, the average for a class dog is 0.4333, and the cat is 0.3067, from this, we can confirm our final prediction to be a dog as it has the highest average probability.

## IV. Flowchart

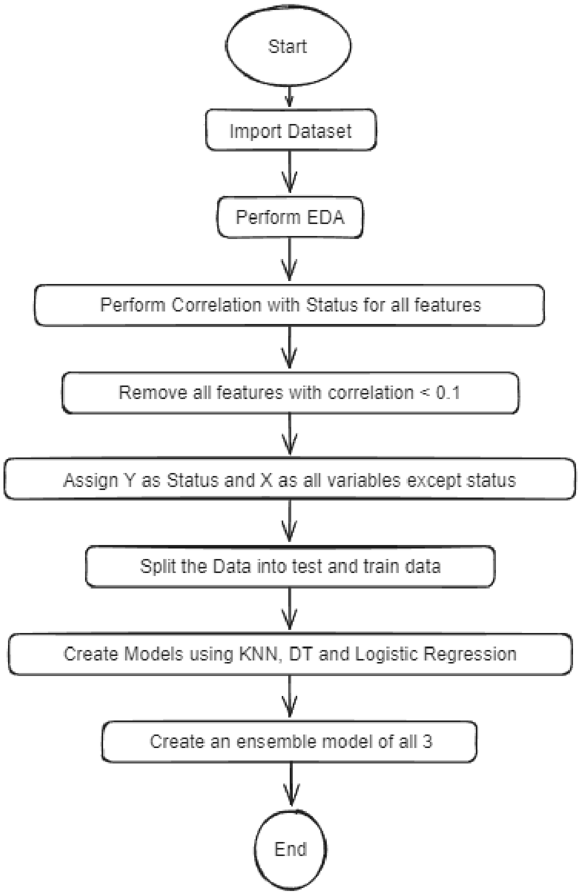
Start
Import Dataset
Perform EDA
Perform Correlation with Status for all features
Remove all features with correlation < 0.1
Assign Y as Status and X as all variables except status
Split the Data into test and train data
Create Models using KNN, DT and Logistic Regression
Create an ensemble model of all 3
End

Fig. 2. Flowchart for Model

## V. Results

### A. KNN

The optimal value of $k$ was determined to be 15, which resulted in the highest achieved accuracy of 81.85%.

TABLE I
Confusion Matrix for Phishing vs Legitimate for KNN

| | Phishing | Legitimate |
|---|---|---|
| Phishing | 1,000 | 143 |
| Legitimate | 272 | 871 |

### B. Descision Tree

The Decision Tree model achieved an accuracy of 89.85

TABLE II
CONFUSION MATRIX FOR PHISHING VS LEGITIMATE FOR DT

| | Phishing | Legitimate |
|---|---|---|
| Phishing | 1,031 | 112 |
| Legitimate | 120 | 1,023 |

### C. Logistic Regression

The Logistic Regression model achieved an accuracy of 90.90

TABLE III
CONFUSION MATRIX FOR PHISHING VS LEGITIMATE FOR LOGISTIC REGRESSION

| | Phishing | Legitimate |
|---|---|---|
| Phishing | 1,036 | 107 |
| Legitimate | 101 | 1,042 |

### D. Voting Classifier

The classifier used is soft classifier as the accuracy was more compared to that of Hard classifier. The Accuracy obtained is 91.51

TABLE IV
CONFUSION MATRIX FOR PHISHING VS LEGITIMATE FOR VOTING CLASSIFIER

| | Phishing | Legitimate |
|---|---|---|
| Phishing | 1,054 | 86 |
| Legitimate | 105 | 1,038 |

## VI. CONCLUSION

The ML model operates effectively with numerous URLs. However, it has a limitation in detecting misspelled URLs used by scammers. This issue can potentially be addressed by implementing a system similar to autocorrect. Such a system would require the creation of a dictionary containing legitimate website names. However, this task poses challenges, particularly with unique names deliberately misspelled by many companies.

## REFERENCES

1) Mishra, P., Kumar, M., & Poonia, P. (2021). Artificial Intelligence and Machine Learning in Healthcare: Current Trends and Future Directions. *Journal of Research in Medical Sciences: The Official Journal of Isfahan University of Medical Sciences*, 26(1), 106.
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7921422/
2) Phishing. (n.d.). *Wikipedia*. Retrieved from Phishing - Wikipedia.
https://en.wikipedia.org/wiki/Phishing
3) Scikit-learn. (n.d.). Nearest Neighbors Classification. *Scikit-learn Documentation*. 1.6. Nearest Neighbors — scikit-learn 1.4.2 documentation.
https://scikit-learn.org/stable/modules/neighbors.html
4) Scikit-learn. (n.d.). Decision Trees. *Scikit-learn Documentation*. 1.10. Decision Trees — scikit-learn 1.4.2 documentation.
https://scikit-learn.org/stable/modules/tree.html
5) JavaTpoint. (n.d.). sklearn - Logistic Regression. *JavaTpoint*. Sklearn Logistic Regression - Javatpoint.
https://www.javatpoint.com/sklearn-logistic-regression
6) Soni, A. (2021). Voting Classifier in Machine Learning. *GeeksforGeeks*. Voting Classifier - GeeksforGeeks.
https://www.geeksforgeeks.org/ml-voting-classifier-using-sklearn/
7) Scribbr. (n.d.). Correlation coefficient. Retrieved from https://www.scribbr.com/statistics/correlation-coefficient/