SMARTBRIDGE INTERNSHIP

DETECTING PHISHING WEBSITES

Submitted by

Abirami. S

Nithyashree. S

1. **INTRODUCTION**

1.1 OVERVIEW

In the last decades, the web and online services have revolutionized the modern world. However, by increasing our dependence on online services, as a result, online security threats are also increasing rapidly. One of the most common online security threats is a so-called Phishing attack, the purpose of which is to mimic a legitimate website such as online banking, e-commerce or social networking website in order to obtain sensitive data such as user-names, passwords, financial and health-related information from potential victims.

AI Machine Learning is the scientific study of algorithms and statistical models that computer systems use in order to perform a specific task effectively without using explicit instructions, relying on patterns and inference instead. It is seen as a subset of artificial intelligence. Machine learning algorithms build a mathematical model based on sample data, known as “training data”, in order to make predictions or decisions without being explicitly programmed to perform the task.

Machine learning tasks are classified into several broad categories. In supervised learning, the algorithms build a mathematical model from a set of data that contains both the inputs and desired outputs. Classification algorithms and regression algorithms are types of supervised learning. Classification algorithms are used when the outputs, are restricted to a limited set of values. Regression algorithms are named for their continuous outputs, meaning they may have any value within a range. In unsupervised learning, the algorithm builds a mathematical model from a set of data which contains only inputs and no desired output labels. Unsupervised learning can discover patterns in the data, and can group the inputs into categories.

1.2 PURPOSE

Phishing is a criminal mechanism employing both social

engineering and technical tricks to steal consumers’ personal

identity data and financial account credentials.

Phishing is a criminal mechanism employing both social

engineering and technical tricks to steal consumers’ personal

identity data and financial account credentials.

Phishing is a criminal mechanism employing both social

engineering and technical tricks to steal consumers’ personal

identity data and financial account credentials.

Phishing is a criminal mechanism employing both social

engineering and technical tricks to steal consumers’ personal

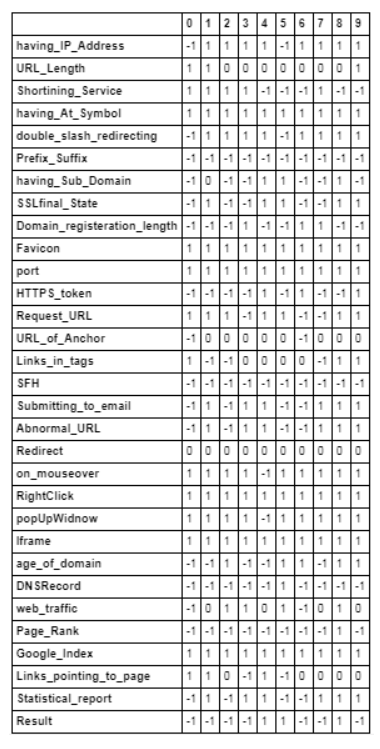
identity data and financial account credentials.

Phishing is a criminal mechanism employing both social engineering and technical tricks to steal consumers’ personal identity data and financial account credentials. Phishing continues to be one of the rapidly growing classes of identity theft scams on the internet that is causing both short term and long-term economic damage. There have been nearly 33,000 phishing attacks globally each month in the year 2012, totaling a loss of $687 million.

Therefore, developing a machine that will enhance the detection of phishing websites will be of a great advantage in improving security. We propose a machine learning based approach to classifying Web sites into 2 classes: Legitimate and Phishing website. Our mechanism only analyses the Uniform Resource Locator (URL) itself without accessing the content of Web sites. Thus, it eliminates the run-time latency and the possibility of exposing users to the browser-based vulnerabilities.

1. **LITERATURE SURVEY**

2.1 EXISTING PROBLEM

The dataset has a total of **30** predictors like the above and the label is saved in the feature named **Result**. A value of -**1** in the Result column denotes that the corresponding website is a phishing website and a value of **1** denotes that the corresponding website is a normal one. Each row (total **11055** rows) in the dataset represents a website by means of the quantified metrics. 

This dataset seems to be a classic example of supervised learning. We have been provided with a fixed number of features for each data point, and our aim will be to train a variety of Supervised algorithms on this data, so that, when a new data point arises, our best performing classifier can be used to categorize the data point as a legitimate or phishing website.

2.2 PROPOSED SOLUTION

The main objective of this research is to use classification algorithms to identify the phishing websites from legitimate ones. This project also aims to compare the classification algorithms based on their performance factors.

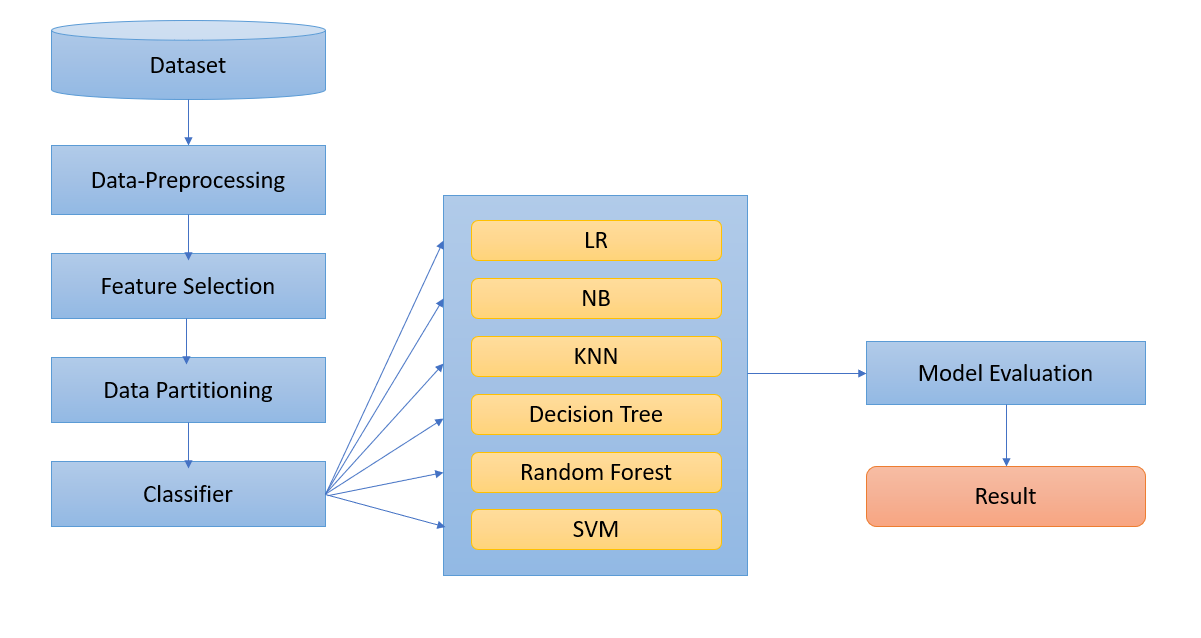
To help the people in the identification of phishing websites on the internet, a graphical user interface is developed. The GUI can be readily utilized for quicker results.

The following algorithms are trained for this purpose, to choose the best performing classifier.

* Logistic Regression
* Naïve Bayes
* K-Nearest Neighbors
* SVM
* Decision Tree Classifier
* Random Forest

1. **THEORITICAL ANALYSIS**

3.1 BLOCK DIAGRAM



3.2 HARDWARE/SOFTWARE DESIGNING

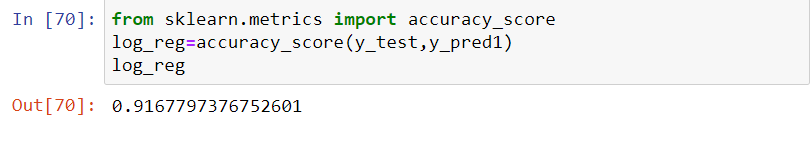
The steps followed in developing the model:

* Data Collection: The dataset was downloaded from the UCI ML Repository.
* Data Analysis: Evaluating cleanliness of the dataset by looking for any irrelevant data and handling missing data.
* Search for any trends, relations and correlations.
* Developing a GUI where a website can be identified to be legitimate or not.

1. **EXPERIMENTAL INVESTIGATIONS**
2. Logistic Regression

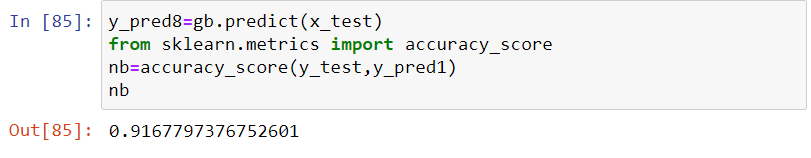
Since the outcome is binary and we have a reasonable number of examples at our disposal compared to number of features, this approach seems suitable. Since for this data, it already knows the output beforehand, it continuously adjusts the weights such that when these weights summed up with their features are introduced in the logistic function, the results are as near as possible to the actual ones.

Once presented with a test value, it again inserts the value into our logistic function and returns the output as a number between 0 and 1, which represents the probability of that test value being in a particular class.



1. Naïve Bayes

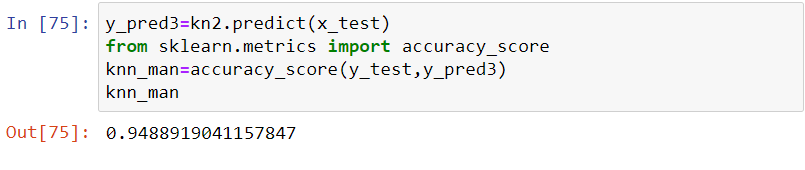
In [statistics](https://en.wikipedia.org/wiki/Statistics), **Naïve Bayes classifiers** are a family of simple "[probabilistic classifiers](https://en.wikipedia.org/wiki/Probabilistic_classification)" based on applying [Bayes' theorem](https://en.wikipedia.org/wiki/Bayes%27_theorem) with strong (naïve) [independence](https://en.wikipedia.org/wiki/Statistical_independence) assumptions between the features. They are among the simplest [Bayesian network](https://en.wikipedia.org/wiki/Bayesian_network) models. Naïve Bayes classifiers are highly scalable, requiring a number of parameters linear in the number of variables in a learning problem.



1. K-Nearest Neighbors

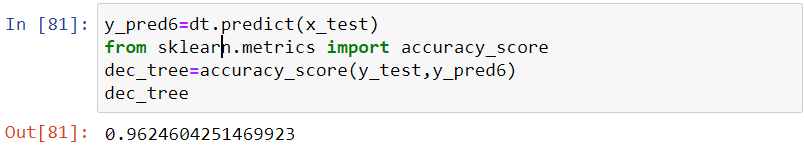
In [pattern recognition](https://en.wikipedia.org/wiki/Pattern_recognition), the **k-nearest neighbors algorithm** (***k*-NN**) is a [non-parametric](https://en.wikipedia.org/wiki/Non-parametric_statistics) method proposed by Thomas Cover used for [classification](https://en.wikipedia.org/wiki/Statistical_classification) and [regression](https://en.wikipedia.org/wiki/Regression_analysis). In both cases, the input consists of the *k* closest training examples in the [feature space](https://en.wikipedia.org/wiki/Feature_space).

In k-NN classification, the output is a class membership. An object is classified by a plurality vote of its neighbors, with the object being assigned to the class most common among its *k* nearest neighbors (*k* is a positive [integer](https://en.wikipedia.org/wiki/Integer), typically small). If *k* = 1, then the object is simply assigned to the class of that single nearest neighbor.



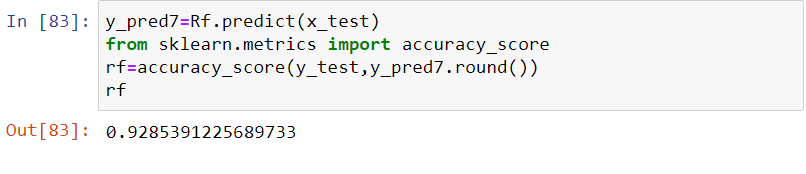
1. Decision Tree Classifier

**Decision Trees (DTs)** are a non-parametric supervised learning method used for [classification](https://scikit-learn.org/stable/modules/tree.html#tree-classification) and [regression](https://scikit-learn.org/stable/modules/tree.html#tree-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.



1. Random Forest

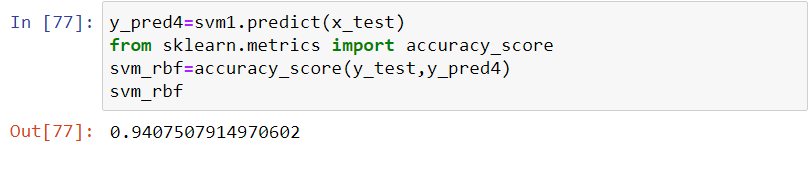
**Random forests** or **random decision forests** are an [ensemble learning](https://en.wikipedia.org/wiki/Ensemble_learning) method for [classification](https://en.wikipedia.org/wiki/Statistical_classification), [regression](https://en.wikipedia.org/wiki/Regression_analysis) and other tasks that operate by constructing a multitude of [decision trees](https://en.wikipedia.org/wiki/Decision_tree_learning) at training time and outputting the class that is the [mode](https://en.wikipedia.org/wiki/Mode_(statistics)) of the classes (classification) or mean prediction (regression) of the individual trees. Random decision forests correct for decision trees' habit of [overfitting](https://en.wikipedia.org/wiki/Overfitting) to their [training set](https://en.wikipedia.org/wiki/Test_set).



1. SVM

In [machine learning](https://en.wikipedia.org/wiki/Machine_learning), **support-vector machines** (**SVMs**, also **support-vector networks**) are [supervised learning](https://en.wikipedia.org/wiki/Supervised_learning) models with associated learning [algorithms](https://en.wikipedia.org/wiki/Algorithm) that analyze data used for [classification](https://en.wikipedia.org/wiki/Statistical_classification) and [regression analysis](https://en.wikipedia.org/wiki/Regression_analysis).

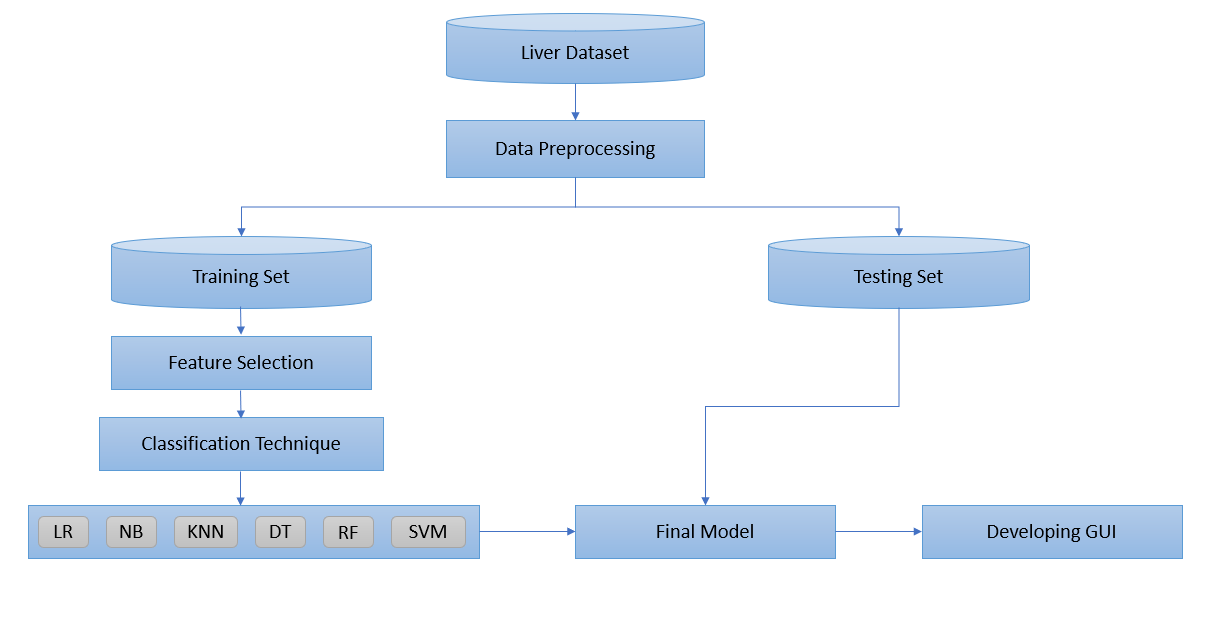
The Support Vector Machine (SVM) algorithm is a popular machine learning tool that offers solutions for both classification and regression problems.



SUMMARY

|  |  |
| --- | --- |
| **Model** | **Accuracy Score** |
| Logistic Regression | 0.916780 |
| SVM | 0.940751 |
| Decision Tree Classifier | 0.962460 |
| Random Forest | 0.928539 |
| KNN | 0.948892 |
| Naïve Bayes | 0.916780 |

1. **FLOWCHART**



1. **RESULT**

Based on the inputs entered by the user, the model predicts whether the website is a phishing website or not.

1. **ADVANTAGES AND DISADVANTAGES**

The benefits of this model are:

* Easy interface
* Straight forward results
* Accurate performance calculations
* This system can be used by many E-commerce or other websites in order to have good customer relationship.
* User can make online payment securely.
* With the help of this system user can also purchase products online without any hesitation.

**Disadvantages:**

As our dataset is small, it’s training dataset is similar to test dataset. So, it is difficult for the model to predict accurately for larger dataset.

1. **APPLICATIONS**

This project makes it easier to predict whether the website is legitimate or not.

1. **CONCLUSION**

In this project, we have proposed methods detecting phishing website using machine learning techniques. The six machine learning techniques that were used include SVM, Logistic Regression, KNN, Naïve Bayes, Decision Tree Classifier and Random Forest. The system was implemented using all the models and their performance was evaluated. Performance evaluation was based on certain performance metrics. A GUI, which can be used as a tool by professionals or common people was implemented.

1. **FUTURE SCOPE**

Database should be expanded on which the system will be tested much better.

1. **BIBLIOGRAPHY**

* <https://machinelearningmastery.com/a-tour-of-machine-learning-algorithms/>
* <https://medium.com/intel-software-innovators/detecting-phishing-websites-using-machine-learning-de723bf2f946>