Phishing Website Detection by Machine Learning Techniques

By

Tanmoy Bhowmick (181001001096)   Nilanjan Tarafder (181001001102)   Ritesh Saha (181001001146)

OBJECTIVES

- The objective of this project is to train machine learning models on the dataset created to predict whether the website is legit or not.

METHOD

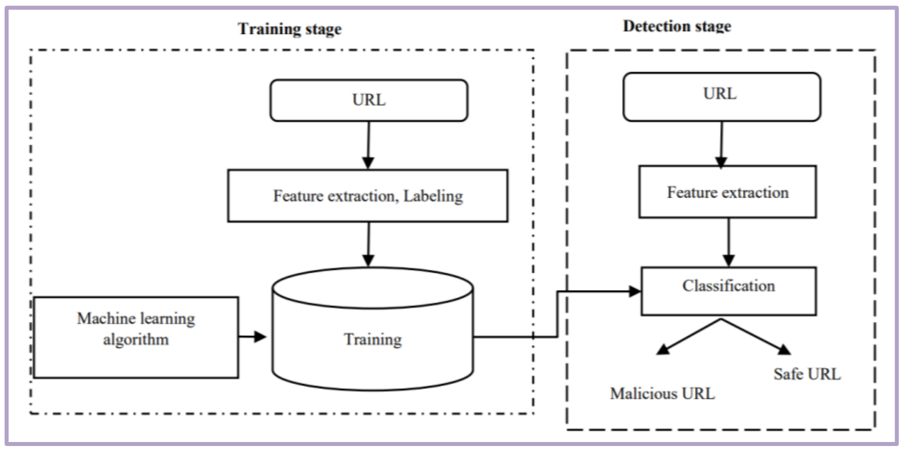
- **Data Collection**: A labelled dataset of Phishing and benign websites is collected from the PhishTank and New Brunswick University.
- **Data Cleaning and Extraction**: Pre-processing includes random 5000 URL from both phishing and legit csv and then extraction of features like Address Bar Based Features is done, add them to form a new csv called ***urldata.csv***.
- **Data Visualization**: Data are put into different graphs to understand the data more well.
- **Model Training**: Sklearn python library is used for training the model using different machine learning techniques such as Decision Tree (DT), Random Forest and Support Vector Machines (SVM) on 80% of the data.
- **Model Testing**: Trained model is tested on the remaining 20 % of the data. Hyperparameters are tuned to increase accuracy, precision, and recall.
- **Model Comparison**: The machine learning classification techniques are compared based on evaluation metrics.

FEATURE EXTRACTION

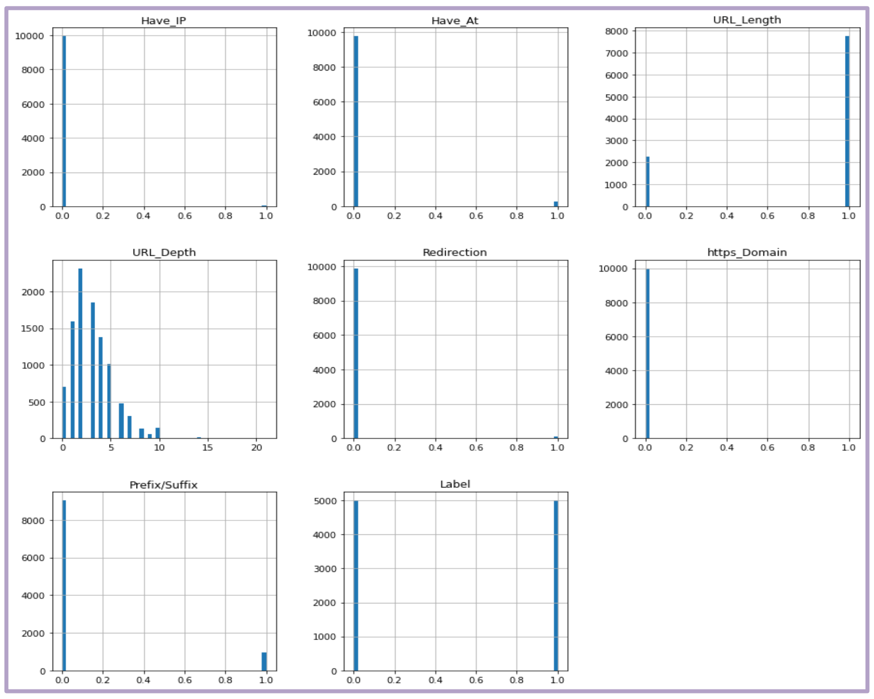
- **Presence of IP address in URL**
- **Presence of @ symbol in URL**
- **Length of URL**
- **Number of slashes in URL**
- **URL redirection**
- **HTTPS token in URL**
- **Prefix or Suffix separated by (-) to domain**

MODEL METHODOLOGIES

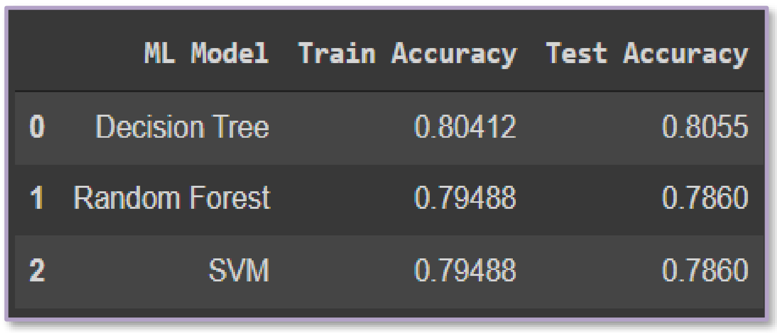
- Decision Tree
- Random Forest Algorithm
- Support Vector Machines



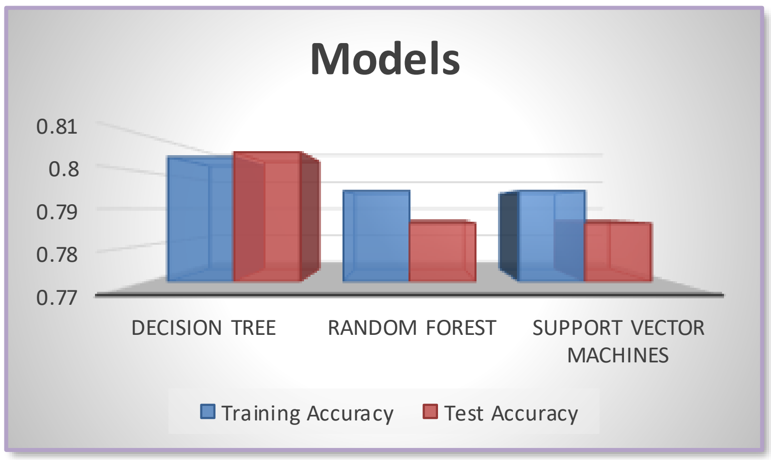
DATA VISUALIZATION



MODEL EVALUATION



	ML Model	Train Accuracy	Test Accuracy
0	Decision Tree	0.80412	0.8055
1	Random Forest	0.79488	0.7860
2	SVM	0.79488	0.7860



CONCLUSIONS

- Decision tree gives best result among all the 3 algorithms.
- The accuracy of decision tree is 80.55%