Mini Project Synopsis
On
“FAKE WEBSITE DETECTION”

GITAM
(DEEMED TO BE UNIVERSITY)
VISAKHAPATNAM HYDERABAD BENGALURU

SUBMITTED BY

321910307015 – G. Mahendra Reddy

321910307004 – B. Harsha Vardhan Reddy

321910307039 – A. Jagadeesh

321910307059 – K. Balu Reddy

Department of Computer Science and Engineering

SESSION: 2019 - 2023

Under the guidance of

Kavya G

Assistant Professor

DEPARTMENT OF Computer Science and Engineering

DEPT. OF CSE, SCHOOL OF TECHNOLOGY (GST),
GITAM UNIVERSITY (DEEMED TO BE UNIVERSITY)
BENGALURU, KARNATAKA, INDIA

Aim of the Project

“Creating a fake website link using a phishing tool and detecting that link using Machine Learning”

Abstract

Phishing is a common attack on credulous people by making them to disclose their unique information using counterfeit websites. The objective of phishing website URLs is to purloin the personal information like user name, passwords and online banking transactions. Phishers use the websites which are visually and semantically similar to those real websites. As technology continues to grow, phishing techniques started to progress rapidly and this needs to be prevented by using anti-phishing mechanisms to detect phishing. Machine learning OR Any Programming is a powerful tool used to strive against phishing attacks. since it is easier to trick someone into clicking a malicious link which seems legitimate than trying to break through a computer’s defences systems. The malicious links within the body of the message are designed to make it appear that they go to the spoofed organization using that organization’s logos and other legitimate contents. Here, we explain phishing domain (or Fraudulent Domain) characteristics, the features that distinguish them from legitimate domains, why it is important to detect these domains, and how they can be detected using Machine Learning techniques.

INTRODUCTION

Nowadays Phishing becomes a main area of concern for security researchers because it is not difficult to create the fake website which looks so close to legitimate website. Experts can identify fake websites but not all the users can identify the fake website and such users become the victim of phishing attack. Main aim of the attacker is to steal banks account credentials. Phishing attacks are becoming successful because lack of user awareness. Since phishing attack exploits the weaknesses found in users, it is very difficult to mitigate them but it is very important to enhance phishing detection techniques. Phishing may be a style of broad extortion that happens once a pernicious web site act sort of a real one memory that the last word objective to accumulate unstable info, as an example, passwords, account focal points, or MasterCard numbers. all the same, the means that there square measure some of contrary to phishing programming and techniques for recognizing potential phishing tries in messages and characteristic phishing substance on locales, phishes think about new and crossbreed procedures to bypass the open programming and frameworks. Phishing may be a fraud framework that uses a mixture of social designing what is additional, advancement to sensitive and personal data, as an example, passwords associate degree open-end credit unpretentious elements by presumptuous the highlights of a reliable individual or business in electronic correspondence. Phishing makes use of parody messages that square measure created to seem substantial and instructed to start out from true blue sources like money connected institutions, online business goals, etc, to draw in customers to go to phony destinations through joins gave within the phishing website.

Literature Survey

Authors Name	Project name	Algorithms	Year	Relevance with the present work
Rishikesh Mahajan, Irfan Siddavatam,	Phishing Website Detection using Machine Learning Algorithms	Decision Tree, random forest, Support vector machine algorithm	2018	97.14% detection Using 3 algorithms
Aarthi Krishna V, Anusree A, Blessy Jose, Karthika Anilkumar, Ojus Thomas Lee	Phishing Detection using Machine Learning based URL Analysis: A Survey	Naive Bayes, Decision Tree, Random Forest, Support Vector Machine, Logistic Regression, K-Nearest-Neighbor	2021	93% detection using 6 algorithms

Problem Definition

We are creating a fake URL using WhPhiser How attackers send links to Victim's we are Creating this URL for awareness

We have developed our project using a **Scikit-learn** Algorithm in This We used Logistic Regression, Train Test Split,TfidfVectorizer.

Logistic Regression:

This is a Scikit-learn algorithm that we will use to train our model. This algorithm will enable our model to understand patterns and relationships in our dataset. The model will gain useful knowledge and insight, which it will use to make predictions

train_test_split:

This is the function in Sklearn model selection for splitting data arrays into two subsets: for training data and for testing data.

TfidfVectorizer:

This package will enable the model to understand and manipulate text data. Text is a big problem for machines, machines cannot consume text in its raw form. We need to convert text into vectors of numbers that machines can read and understand.

*We can use Convert the text data into vectors of numbers, we convert the text using TfidfVectorizer and pass make Tokens as a parameter. make Tokens is the function used to clean our text

*After converting text, we will save our vectors of numbers into a new variable

* We will use train_test_split to split our dataset into two sets. One set will be used for model training and the other one will be used for model testing

* We will initialize the Logistic Regression algorithm

* After initializing the algorithm, we will fit the algorithm onto our training dataset. The model will learn from this dataset.

* To make predictions, we will use several URLs and see if the model can classify if the URL is bad or good

* We use the vectorizer. Transform method to convert the text to vectors of numbers. Then we apply the logit. Predict method to make the actual predictions

Address Bar based Features

Sr. No	Feature name	Description
1	IP address	Whether domain is in the form of an IP address
2	Length of URL	Length of URL
3	Suspicious character	Whether URL has @, //
4	Prefix and suffix	Whether URL has –
5	Length of subdomain	Length of subdomain
6	Number of /	Number of / in URL
7	HTTPS protocol	Whether URL use https.
8	Phishing words in URL	Whether URL has phishing terms
9	Number of.	Number of dots. in URL

PROJECT REQUIREMENTS

Hardware Requirements: -

- 2GB RAM (minimum)
- 100GB HDD (minimum)
- Intel 1.66 GHz Processor Pentium 4 (minimum)
- Internet

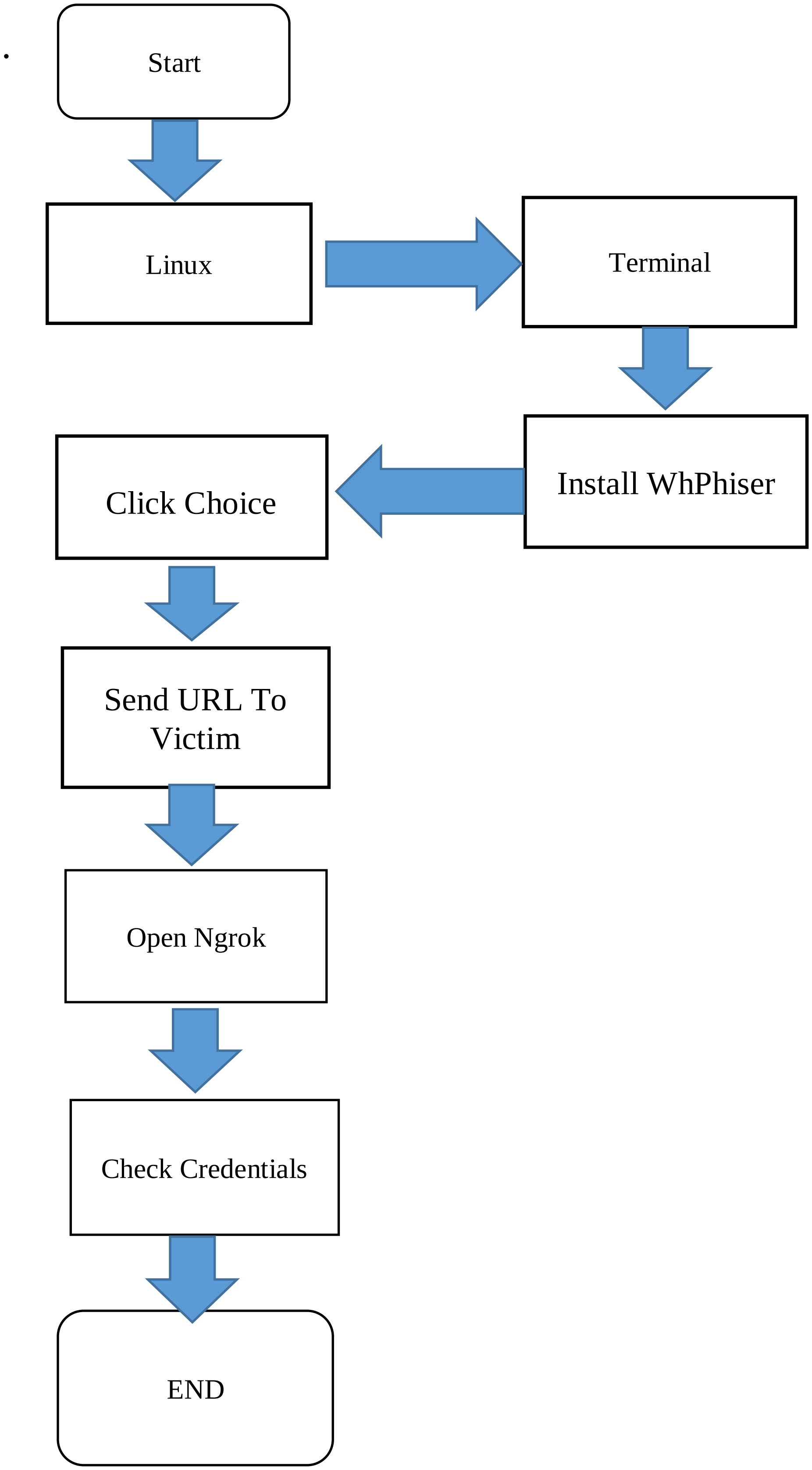
Software Requirements: -

- WINDOWS 7 or higher
- Visual Studio Code or Jupiter

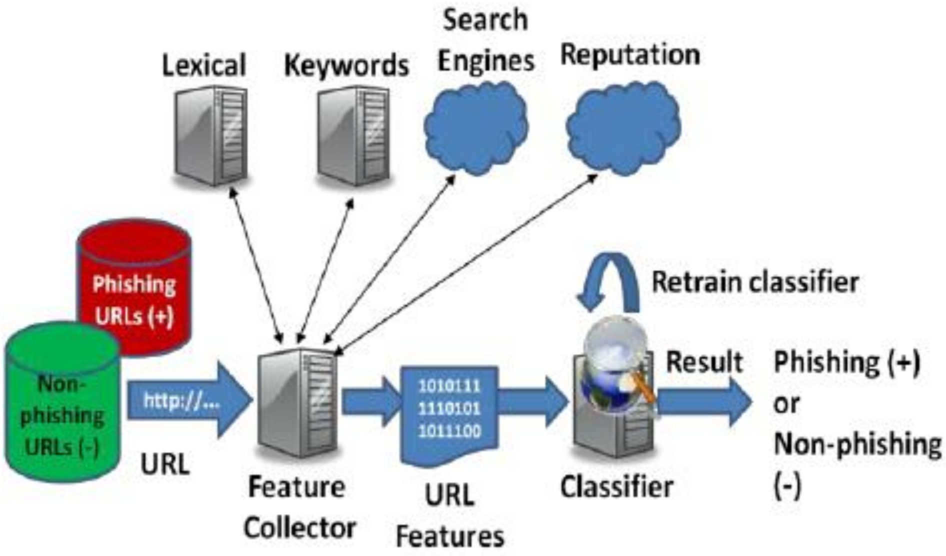
. Linux

Diagram

Creating Fake Website



SK-Learn ALGORITHM



Reference :-

1) (PDF) Phishing Website Detection using Machine Learning Algorithms (researchgate.net)

2) What Is a Malicious URL? (And How You Can Avoid Them) (cheapsslsecurity.com)

3) Getting Started with Logistic Regression in Python | Engineering Education (EngEd) Program | Section

4) scikit-learn: machine learning in Python — scikit-learn 1.1.1 documentatio