**A Project Work 1 Report Submitted to**

**Jawaharlal Nehru Technological University-A,Anantapur**

***In partial fulfilment of the requirements***

***For the Award of the degree of***

**BACHELOR OF TECHNOLOGY**

IN

**COMPUTER SCIENCE ENGINEERING**

**Submitted by**

VENKATA HANUMA TEJA.D(199Y1A0537)

KASIM HUSSAIN.D(199Y1A0545)

DINESHWAR REDDY.B(199Y1A0516)

MALLIKARJUNA.G(199Y1A0554)

**Under the Esteemed Guidance of**

**Dr. N.RAMANJANEYA REDDY**



**DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING**

**K.S.R.M. COLLEGE OF ENGINEERING**

**(AUTONOMOUS)**

Approved by AICTE, New Delhi & Affiliated to JNTUA, Anantapur.

(Accredited by NAAC, BANGLORE)

**Kadapa, Andhra Pradesh, India– 516 003**

**(2022-2023)**

**K.S.R.M. COLLEGE OF ENGINEERING**

**(AUTONOMOUS)**

Approved by AICTE, New Delhi & Affiliated to JNTUA, Anantapur.

(Accredited by NAAC, BANGLORE)

**Kadapa, Andhra Pradesh, India– 516 003**

# DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING



**CERTIFICATE**

Certified that this is a Bonafede record of the project work 1 report entitled, **“PHISHING WEBSITE DETECTION USING MACHINE LEARNING ”,** done by

**Mr. D.VENKATA HANUMA TEJA (199Y1A0537), Mr.D.KASIM HUSSAIN (199Y1A0545),Mr. DINESHWAR REDDY(199Y1A0516), Mr.G.MALLIKARJUNA (199Y1A0554)**

Submitted to the department of Computer Science and Engineering, in partial fulfilment of the requirements for the Degree of **BACHELOR OF TECHNOLOGY** in **Computer Science and Engineering** from Jawaharlal Nehru Technological University-A, Anantapur during the year of **2022-2023.**

**Guide: HEAD OF THE DEPARTMENT:**

**Dr. N.RAMANJANEYA REDDY Dr. V. LOKESWARA REDDY**

**M.TECH., Ph.D., professor &HOD,**

**Professor, Dept. Of CSE.**

**Dept. Of CSE.**

**DATE: Project Guide Internal Examiner**

# ACKNOWLEDGEMENT

I think this opportunity to my profound gratitude and deep regards to the people who have guided us throughout the project and whose timely guidance has helped us in accomplishment of project.

I express my sincere thanks to my guide, Dr. N.RAMANJANEYA REDDY, professor, Department of CSE, who offered many comments and insights that depended on my understanding of project and its context.

I am highly indebted to Dr. K. CHANDRA OBUL REDDY, Managing Director, Kandula Group of Institutions, prof V.S. S. MURTHY, Principle of K.S.R.M and Dr. V. LOSESWARA REDDY, professor, HOD of CSE department for giving great support all the time that inspired us a lot.

I would like to thank my all-B. Tech classmates and others department friends for their internal support.

Last, but not least, I express my heartfelt thanks to my parents for their support, love, patient sacrifices and providing a favorable environment for the completion of project.

## Table of contents

## Abstract

1.Introduction

2. problem definition and module description

Existing system

Drawbacks of Existing System

Proposed System

Advantages of Proposed System

Modules Description

3.Hardware and Software Requirements

4.References

## ABSTRACT:

Phishing attack is a simplest way to obtain sensitive information from innocent users. Aim of the phishers is to acquire critical information like username, password and bank account details. Cyber security persons are now looking for trustworthy and steady detection techniques for phishing websites detection. This paper deals with machine learning technology for detection of phishing URLs by extracting and analyzing various features of legitimate and phishing URLs.

Decision Tree, random forest and Support vector machine algorithms are used to detect phishing websites. Aim of the paper is to detect phishing URLs as well as narrow down to best machine learning algorithm by comparing accuracy rate, false positive and false negative rate of each algorithm.

**INTRODUCTION**:

“Phishing” is a form of cyber fraud, which means that criiminal use various means to impersonate the URL address and page content of real website or use vulnerabilities in the server program of a real website to insert dangerous HTML code in certain pages of the site. Use this to deceive users bank or credit card account numbers, passwards and other private data. Phishing not only brings economic losses to netizens, but also hinders the deeper environment of the Internet. Preventing phishing is a requirement in the current situation. Although the users care and experience are important, it cannot prevent completely prevent the uers from falling into a phishing scam.

1. Because, in order to increase the success rate of phishing attacks,attackers also consider personality characteristics of end users,especially in deceiving relatively experienced users.
2. Cyber attack targeting end users have resulted in severse loss of sensitive/personal information and even monetary losses to indevisuals whose total loss reached one billion US dollars within one year
3. In Bank of china, the criminals deceive users by changing the “.com” in the URL to “c0m”, which is the most typical method used by phishing websites. BY using this type of fradulent URL, phishing attempt to capture some sensitive and personal information of the victim, such as financial data,personal information, username, passward, etc.
4. In this project we want to collect phishing and non phishing websites and generate a database,using these data and using some algorithms to learn,and detecting phishing website which shows the best effect
5. In this project we want use different algorithms are used for detection among which is best effect
6. The result will be compared with different algorithms and select the one with the highest time efficiency and accuracy as the algorithm we need.

**PROBLEM DEFINATION:**

security defenders struggle to detect phishing domains is because of the unique part of the website domain (the FreeURL). When a domain detected as a fraudulent, it is easy to prevent this domain before an user access to it.

**SUGGESTED SOLUTIONS:**

The components for detection and classification of phishing website are as follows

1. Address bar based feature
2. Abnormal based feature
3. HTML and Java script based feature
4. Domain based feature

**EXISTING SYSTEM:**

An existing system surveys the literature on the detection of phishing attacks. Phishing attacks target vulnerabilities that exist in systems due to the human factor. Many cyber attacks are spread via mechanisms that exploit weaknesses found in end-users, which makes users the weakest element in the security chain. The phishing problem is broad and no single silver-bullet solution exists to mitigate all the vulnerabilities effectively, thus multiple techniques are often implemented to mitigate specific attacks. This paper aims at surveying many of the recently implemented phishing mitigation techniques. A high-level overview of various categories of phishing mitigation techniques is also presented, such as: detection, offensive defense, correction, and prevention, which we belief is critical to present where the phishing detection techniques fit in the overall mitigation process.

**DRAWBACKS OF EXISTING SUSTEM:**

1) The system less effective since it is not implemented for large number of datasets.

2) The system doesn’t implement Data Preprocessing and not compared with number of classifiers.

**PROPOSED SYSTEM:**

1. The Proposed system designs the following concepts which Presence of IP address in URL: If IP address present in URL then the feature is set to 1 else set to 0. Most of the benign sites do not use IP address as an URL to download a webpage. Use of IP address in URL indicates that attacker is trying to steal sensitive information.
2. Presence of @ symbol in URL: If @ symbol present in URL then the feature is set to 1 else set to 0. Phishers add special symbol @ in the URL leads the browser to ignore everything preceding the “@” symbol and the real address often follows the “@” symbol
3. Number of dots in Hostname: Phishing URLs have many dots in URL. For example

http://shop.fun.amazon.phishing.com, in this URL phishing.com is an actual domain name, whereas use of “amazon” word is to trick users to click on it. Average number of dots in benign URLs is 3. If the number of dots in URLs is more than 3 then the feature is set to 1 else to 0.

1. Prefix or Suffix separated by (-) to domain: If domain name separated by dash (-) symbol then feature is set to 1 else to 0. The dash symbol is rarely used in legitimate URLs. Phishers add dash symbol (-) to the domain name so that users feel that they are dealing with a legitimate webpage. For example Actual site is

http://www.onlineamazon.com but phisher can create another fake website like http://www.online-amazon.com to confuse the innocent users.

1. URL redirection: If “//” present in URL path then feature is set to 1 else to 0. The existence of “//” within the URL path means that the user will be redirected to another website
2. **HTTPS token in URL:** If HTTPS token present in URL then the feature is set to 1 else to 0. Phishers may add the “HTTPS” token to the domain part of a URL in order to trick users. For example, <http://https-wwwpaypal-it-mpp-home.soft-hair.com>
3. Information submission to Email: Phisher might use “mail()” or “mailto:” functions to redirect the user’s information to his personal email[4]. If such functions are present in the URL then feature is set to 1 else to 0.
4. URL Shortening Services “TinyURL”: TinyURL service allows phisher to hide long phishing URL by making it short. The goal is to redirect user to phishing websites. If the URL is crafted using shortening services (like bit.ly) then feature is set to 1 else 0.
5. Length of Host name: Average length of the benign URLs is found to be a 25, If URL’s length is greater than 25 then the feature is set to 1 else to 0.
6. **Presence of sensitive words in URL:** Phishing sites use sensitive words in its URL so that users feel that they are dealing with a legitimate webpage. Below are the words that found in many phishing URLs :- 'confirm', 'account', 'banking', 'secure', 'ebyisapi', 'webscr', 'signin', 'mail', 'install', 'toolbar', 'backup', 'paypal', 'password', 'username', etc.

**ADVANTAGES OF PROPOSED SYSTEM:**

1. Proposes a Decision Tree Algorithm which implements for Presence of sensitive words in URL.
2. The proposed system incorporates which Phishes can make a use of Unicode characters in URL to trick users to click on it.

**HARDWARE REQUIREMENTS:**

➢ Processor - Pentium –IV

➢ RAM - 4 GB (min)

➢ Hard Disk - 20 GB

➢ Key Board - Standard Windows Keyboard

➢ Mouse - Two or Three Button Mouse

➢ Monitor - SVGA

**SOFTWARE REQUIREMENTS:**

1. **Operating system :** Windows 7 Ultimate.
2. **Coding Language :** Python.
3. **Front-End :** Python.
4. **Back-End :** Django-ORM
5. **Designing :** Html, css, javascript.
6. **Data Base :** MySQL (WAMP Server).