PHISHING URL DETECTION USING MACHINE LEARNING

## A MINI PROJECT II REPORT

***Submitted by***

|  |  |
| --- | --- |
| **ABISHEK PS** | **(1901002)** |
| **ADITYA KUSHWAHA** | **(1901004)** |
| **ARJUN RU** | **(1901016)** |
| **ASHWIN BALAJI PL** | **(1901023)** |

***In partial fulfilment for the award of the degree of***

# BACHELOR OF ENGINEERING

***in***

**COMPUTER SCIENCE AND ENGINEERING**

# SRI RAMAKRISHNA ENGINEERING COLLEGE

[Educational Service: SNR Sons Charitable Trust]

[Autonomous Institution, Reaccredited by NAAC with ‘A+’ Grade]

[Approved by AICTE and Permanently Affiliated to Anna University, Chennai] [ISO 9001:2015 Certified and All Eligible Programmes Accredited by NBA] Vattamalaipalayam, N.G.G.O. Colony Post,

**COIMBATORE - 641 022**

**JUNE 2022**

ANNA UNIVERSITY: CHENNAI - 600 025

# BONAFIDE CERTIFICATE

## 16CS266 - MINI PROJECT – II

This is to certify that Mini Project - II Report, **“PHISHINNG URL DETECTION USING MACHINE LEARNING”** is the bonafide work of **“Abishek PS, ADITYA KUSHWAHA, ARJUN RU, Ashwin Balaji PL”** who carried out the project under my supervision.

|  |  |
| --- | --- |
| **SIGNATURE** | **SIGNATURE** |
| Dr. Grace Selvarani A | Dr. Vijayakumar R |
| **HEAD OF THE DEPARTMENT** | **SUPERVISOR** |
| Professor, | Assistant Professor, |
| Computer Science and Engineering,  Sri Ramakrishna Engineering College, Coimbatore - 641022 | Computer Science and Engineering,  Sri Ramakrishna Engineering College, Coimbatore - 641022 |

## Submitted for the Mini Project Viva-Voice Presentation held on

**INTERNAL EXAMINER EXTERNAL EXAMINER**

## ACKNOWLEDGEMENT

We express our gratitude to **Sri. D LAKSHMINARAYANASWAMY**, Managing Trustee, **Sri. R SUNDAR,** Joint Managing Trustee, SNR Sons Charitable Trust, Coimbatore for providing excellent facilities to carry out our project.

We express our deepest gratitude to our Principal, **Dr. N R ALAMELU, Ph.D.,** for her valuable guidance and blessings.

We are indebted to our Head of the Department, **Dr. A Grace Selvarani, Ph.D.,** Department of Computer Science and Engineering who modeled us both technically and morally for achieving great success in life.

We express our thanks to our Project Coordinator, **Mrs. s PRINCE SAHAYA brighty,** Assistant Professor (Sr. Grade), Department of Computer Science and Engineering for her great inspiration.

Words are inadequate to offer thanks to our respected guide. We wish to express our sincere thanks to **Dr. R Vijayakumar,** Assistant Professor (Sl. Grade), Department of Computer Science and Engineering, who gives constant encouragement and support throughout this project work and who makes this project a successful one.

We also thank all the staff members and technicians of our department for their help in making this project a successful one.

**ABSTRACT**

Most of the financial frauds are caused by phishing attacks. It is one of the most dangerous threats to online accounts and data, because these kinds of exploits hide behind the guise of being from a reputable company or individual and use elements of social engineering to make victims far more likely to fall for the scam.

The important thing is to exercise common sense and a good deal of caution about any message that the user receives which looks faintly suspicious and has tell-tale signs like spelling mistakes or odd phrasing, errors that malware authors often make, urges you to do something ‘right now’, or has a link or attachment which seems even remotely dodgy. A message which comes from a trusted source such as higher officials in the workplace or from a reputed organization or even close friends, where their email address or details could easily have been spoofed.

The model which we have proposed is based on machine learning where the features are extracted from the input URL based on its characteristics and behaviour. The machine learning model used here is extreme gradient boosting or XGBoost, a supervised machine learning algorithm which uses more accurate approximations to find the best tree model using the training dataset to predict the URL is phishing or legitimate as target variable.

**LIST OF FIGURES**

|  |  |  |
| --- | --- | --- |
| **Figure / Table** | **TITLE** | **Page no.** |
| 4.4.1 | Dataset Generation | 14 |
| 4.4.2 | Phishing Detection | 14 |
| 5.1 | Test Accuracy | 16 |
| 5.2 | Train Accuracy | 1 |
| 5.3 | Accuracy Score | 17 |
| 7.2.1 | URL Entry | 30 |
| 7.2.2 | Submitting Legitimate URL | 30 |
| 7.2.3 | Redirection – Not Phishing | 30 |
| 7.2.4 | Submitting fake/phishing URL | 30 |
| 7.2.5 | Warning | 31 |
| 7.2.6 | Prediction – Phishing | 31 |
| 7.2.7 | Continue with RISK | 31 |

**LIST OF ABBREVIATIONS**

|  |  |  |
| --- | --- | --- |
| ML | - | MACHINE LEARNING |
| SVM | - | SUPPORT VECTOR MACHINE |
| XG | - | Extreme Gradient |
| KNN | - | K – Nearest Neighbour |
| URL | - | UNIFORM RESOURCE LOCATOR |
| DNS | - | DOMAIN NAME SYSTEM |

## TABLE OF CONTENTS

|  |  |  |
| --- | --- | --- |
| **CHAPTER** | **TITLE** | **PAGE** |
|  | **ABSTRACT** | i |
|  | **List OF FIGURES** | ii |
| **1** | **INTRODUCTION** |  |
| * 1. Phishing | 1 |
|  | * 1. Phishing Threats | 2 |
|  | * 1. Phishing Attacks | 3 |
| **2** | **LITERATURE REVIEW** |  |
| 2.1 An Enhanced Blacklist Method to Detect  Phishing Websites | 4 |
| 2.2 Malicious URL Detection using Machine  Learning: A Survey | 4 |
| 2.3 Detection of URL based Phishing Attacks using  Machine Learning | 5 |
| 2.4 Limitations | 5 |
| **3** | **SYSTEM DESCRIPTION** |  |
| 3.1 Hardware Requirements | 7 |
| 3.2 Software Requirements | 7 |
| 3.3 Description | 8 |
|  | 3.4 Existing Solutions | 8 |
| **4** | **MODULE DESCRIPTION** |  |
| 4.1 Dataset | 10 |
| 4.2 Feature Extraction | 10 |
| 4.3 Machine Learning Models | 12 |
| 4.4 Flowcharts | 13 |
| 4.5 Merits and Demerits | 15 |
| **5** | **RESULTS** | 16 |
| **6** | **CONCLUSIONS AND FUTURE SCOPE**  6.1Conclusion  6.2 Future Scope | 18  18 |
| **7** | **APPENDICES** |  |
| 7.1 Sample Code | 19 |
| 7.2 Snapshots | 30 |
| **8** | **REFERENCES** | 32 |