

AN EFFECTIVE MODEL OF TERMINATING
PHISHING WEBSITES AND DETECTION BASED
ON LOGISTIC REGRESSION

A PROJECT REPORT

Submitted by

SHAIK RIYAZ [Reg No: RA1611003010231]
R.B.SRINIVAS RAJU [Reg No: RA1611003010184]

Under the guidance of

K.R.JANSI

(Assistant Professor, Department of Computer Sciene & Engineering)

in partial fulfillment for the award of the degree

of

BACHELOR OF TECHNOLOGY

in

COMPUTER SCIENE & ENGINEERING

of

FACULTY OF ENGINEERING AND TECHNOLOGY



SRM
INSTITUTE OF SCIENCE & TECHNOLOGY
Deemed to be University u/s 3 of UGC Act, 1956

S.R.M. Nagar, Kattankulathur, Kancheepuram District

JUNE 2020



Own Work Declaration
Department of Computer Science and Engineering

SRM Institute of Science & Technology

Own Work* Declaration Form

This sheet must be filled in (each box ticked to show that the condition has been met). It must be signed and dated along with your student registration number and included with all assignments you submit – work will not be marked unless this is done.

To be completed by the student for all assessments

Degree/ Course :Btech/ Computer Science and Engineering
Student Name :Rudraraju bhaskar srinivas raju
Registration Number :RA1611003010184
Title of Work :AN EFFECTIVE MODEL OF TERMINATING PHISHING WEBSITES AND
DETECTION BASED ON LOGISTIC REGRESSION

I / We hereby certify that this assessment compiles with the University's Rules and Regulations relating to Academic misconduct and plagiarism**, as listed in the University Website, Regulations, and the Education Committee guidelines.

I / We confirm that all the work contained in this assessment is my / our own except where indicated, and that I / We have met the following conditions:

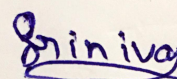
- Clearly references / listed all sources as appropriate
- Referenced and put in inverted commas all quoted text (from books, web, etc)
- Given the sources of all pictures, data etc. that are not my own
- Not made any use of the report(s) or essay(s) of any other student(s) either past or present
- Acknowledged in appropriate places any help that I have received from others (e.g. fellow students, technicians, statisticians, external sources)
- Compiled with any other plagiarism criteria specified in the Course handbook / University website

I understand that any false claim for this work will be penalised in accordance with the University policies and regulations.

DECLARATION:

I am aware of and understand the University's policy on Academic misconduct and plagiarism and I certify that this assessment is my / our own work, except where indicated by referring, and that I have followed the good academic practices noted above.

If you are working in a group, please write your registration numbers and sign with the date for every student in your group.





Own Work Declaration
Department of Computer Science and Engineering

SRM Institute of Science & Technology

Own Work* Declaration Form

This sheet must be filled in (each box ticked to show that the condition has been met). It must be signed and dated along with your student registration number and included with all assignments you submit – work will not be marked unless this is done.

To be completed by the student for all assessments

Degree/ Course :Btech/ Computer Science and Engineering

Student Name :Shaik Riyaz

Registration Number :RA1611003010231

Title of Work :AN EFFECTIVE MODEL OF TERMINATING PHISHING WEBSITES AND
DETECTION BASED ON LOGISTIC REGRESSION

I / We hereby certify that this assessment compiles with the University's Rules and Regulations relating to Academic misconduct and plagiarism**, as listed in the University Website, Regulations, and the Education Committee guidelines.

I / We confirm that all the work contained in this assessment is my / our own except where indicated, and that I / We have met the following conditions:

- Clearly references / listed all sources as appropriate
- Referenced and put in inverted commas all quoted text (from books, web, etc)
- Given the sources of all pictures, data etc. that are not my own
- Not made any use of the report(s) or essay(s) of any other student(s) either past or present
- Acknowledged in appropriate places any help that I have received from others (e.g. fellow students, technicians, statisticians, external sources)
- Compiled with any other plagiarism criteria specified in the Course handbook / University website

I understand that any false claim for this work will be penalised in accordance with the University policies and regulations.

DECLARATION:

I am aware of and understand the University's policy on Academic misconduct and plagiarism and I certify that this assessment is my / our own work, except where indicated by referring, and that I have followed the good academic practices noted above.

If you are working in a group, please write your registration numbers and sign with the date for every student in your group.

SRM INSTITUTE OF SCIENCE AND TECHNOLOGY

(Under Section 3 of UGC Act, 1956)

BONAFIDE CERTIFICATE

Certified that this project report titled “AN EFFECTIVE MODEL OF TERMINATING PHISHING WEBSITES AND DETECTION BASED ON LOGISTIC REGRESSION” is the bonafide work of “ SHAIK RIYAZ [Reg No: RA1611003010231], R.B.SRINIVAS RAJU [Reg No: RA1611003010184]”, who carried out the project work under my supervision. Certified further, that to the best of my knowledge the work reported herein does not form any other project report or disserta-tion on the basis of which a degree or award was conferred on an earlier occasion on this or any other candidate.

SIGNATURE

K.R.JANSI
GUIDE
Assistant Professor
Dept. of Computer Sciene &
Engi-neering

Signature of the Internal Examiner

SIGNATURE

Dr .B.AMUTHA
HEAD OF THE DEPARTMENT
Dept. of Computer Sciene &
Engi-neering

Signature of the External Examiner

ABSTRACT

These days phishing is one of the greatest and the quickest developing risk in light of the fact that phishing assailants will get data which was entered by the client and phishing aggressors will utilize that data without the client information. At present personal information is the most important thing compared to money, So phishing website attackers are focusing on the personal information of the user and they taking advantage when the user enters there personal information in the phishing web-sites. So the main theme of these projects is to avoid misuse of personal information by phishing attackers. To overcome this problem machine learning algorithm (Logistic regression) is used and provided with the massive dataset. From that dataset, it trains the algorithm and helps in detecting the new web links which are fraudulent links. Attackers dis-guise their website as legitimate and try to get data from the user for which they make users visit a website and get the personal information that is needed. Before trying to enter into those websites it is important to check whether the given link is good or a phishing link. By checking the link we can save ourselves from the attackers and can keep our data safe.

ACKNOWLEDGEMENTS

We express our humble gratitude to Dr. Sandeep Sancheti, Vice Chancellor, SRM Institute of Science and Technology, for the facilities extended for the project work and his continued support.

We extend our sincere thanks to Dr. C. Muthamizhchelvan, Director, Faculty of Engineering and Technology, SRM Institute of Science and Technology, for his invaluable support.

We wish to thank Dr. B. Amutha, Professor & Head, Department of Computer Science and Engineering, SRM Institute of Science and Technology, for her valuable suggestions and encouragement throughout the period of the project work.

We are extremely grateful to our Academic Advisor Dr. A.JeyaSekar, Associate Professor, and Dr. R. Annie Uthra, Associate Professor, Department of Computer Science and Engineering, SRM Institute of Science and Technology, for their great support at all the stages of project work.

We would like to convey our thanks to our Panel Head, K.R.JANSI, Assistant Professor, Department of Computer Science and Engineering, SRM Institute of Science and Technology, for his / her inputs during the project reviews.

We register our immeasurable thanks to our Faculty Advisor, K.R.JANSI, Assistant Professor, Department of Computer Science and Engineering, SRM Institute of Science and Technology, for leading and helping us to complete our course.

Our inexpressible respect and thanks to my guide, K.R.JANSI, Assistant Professor, Department of Computer Science and Engineering, SRM Institute of Science and Technology, for providing me an opportunity to pursue my project under his/her mentorship. He / She provided me the freedom and support to explore the research topics of my interest. Her / His passion for solving the real problems and making a difference in the world has always been inspiring.

We sincerely thank staff and students of the Computer Science and Engineering De-partment, SRM Institute of Science and Technology, for their help during my research. Finally, we would like to thank my parents, our family members and our friends for their unconditional love, constant support and encouragement.

Author

SHAIK RIYAZ & R.B.SRINIVAS RAJU

TABLE OF CONTENTS

ABSTRACT	iii
ACKNOWLEDGEMENTS	iv
LIST OF TABLES	ix
LIST OF FIGURES	x
ABBREVIATIONS	xi
1 INTRODUCTION	1
1.1 What is Phishing?	1
1.2 Problem statement	3
1.3 Existing system	3
1.4 Proposed system	4
1.5 Advantages of the proposed system	4
1.6 Summary	5
2 SYSTEM REQUIREMENT SPECIFICATIONS	6
2.1 Python	6
2.1.1 Python Platform	6
2.1.2 Python Library	7
2.2 Machine Learning	7
2.3 HTML&CSS	11
3 LITERATURE SURVEY	13
3.1 A Systematic Review of Software-Based Web Phishing Detection .	13
3.2 Intelligent rule-based phishing websites classification	13
3.3 An fMRI Study of Phishing and Malware Warnings	14

3.4	Phishing-Alarm: Robust and Efficient Phishing Detection via Page Component Similarity	14
3.5	OFS-NN: An Effective Phishing Websites Detection Model Based on Optimal Feature Selection and Neural Network	15
3.6	Protecting Users Against Phishing Attacks with AntiPhish	15
3.7	Detecting Phishing Websites Based on the Study of the Financial Industry Webserver Logs	16
3.8	A Hybrid System to Find & Fight Phishing Attacks Actively	16
3.9	Security Education against Phishing: A Modest Proposal for a Major Rethink	17
3.10	Inference from the survey	17
4	PHISHING DETECTION SYSTEM DESIGN	19
4.1	Architecture	19
4.2	Training and Testing	20
4.2.1	Regression	21
4.2.2	Logistic Regression	22
4.3	Use case diagram	24
4.4	Activity diagram	25
4.5	ER diagram	26
5	PHISHING DETECTION SYSTEM MODULES	28
5.1	List of phishing detection sytem modules	28
5.1.1	Data Collection and Pre-processing	28
5.1.2	Data cleaning and Data transformation	29
5.1.3	Data selection and Data sets	29
5.1.4	Feature extraction	30
5.1.5	Prediction	31
5.1.6	Accuracy	32
6	CONCLUSION & FUTURE ENHANCEMENT	33
A	APPENDIX	35
A.1	Import required fuctions	35

A.2 Import data set from storage	35
A.3 Find type of data set	35
A.4 Show the data set	35
A.5 Preprocessing	35
A.6 Store label column in y variable	36
A.7 Using vectorizer for convert string data set into numerical data ...	37
A.8 Transfrom url list column into numerical and stored in s	37
A.9 Split the training and testing data	37
A.10 Declare logistic regression	37
A.11 Feed the Training data into algorithm	37
A.12 Feed the test data and find and print accuracy	38
A.13 Import tkinter for User Interface	38
A.14 Declare the size of interface size	38
A.15 Give the title of the interface	38
A.16 Create input box	39
A.17 Detecting function	39
 B SCREENSHOTS OF THE PROJECT	 42

LIST OF TABLES

5.1 Feature extraction	30
----------------------------------	----

LIST OF FIGURES

1.1	Phishing attacks	2
4.1	Architecture	20
4.2	Training and Testing	21
4.3	Use case diagram	25
4.4	Activity diagram	26
4.5	ER diagram	27
5.1	The workflow of phishing detection models	28
5.2	Accuracy percentage	32
B.1	Sample output	42
B.2	Result output	43

ABBREVIATIONS

ML	Machine Learning
LR	Logistic Regression
APWG	Anti-Phishing Working Group
URL	Uniform resource locator
APAC	Anti-phishing Alliance of China
GUI	Graphical user interface
ASP	Active server page

CHAPTER 1

INTRODUCTION

1.1 What is Phishing?

The Internet has become an imperative foundation that carries extraordinary comfort to human culture. Notwithstanding, the Internet is additionally described by some unavoidable security issues, for example, phishing, malignant programming, and protection exposure, which have just carried genuine dangers to the economy of clients. The Anti-Phishing Working Group (APWG) characterizes phishing as a criminal instrument using social structure, concentrated trick to take singular character data and budgetary record capabilities of shoppers. Phishing is an electronic assault that utilizes hidden email. Phishing attackers are mainly focus on to make clients fool by sending the email as it is the secure line and they will send that email as it is important and most secure line but it not the secure line, In these way the phishing attackers making fool by sending fake emails address . What really perceives phishing is the structure the message takes: like that, routinely an authentic or possibly certified individual, or an association the loss may work with. It's presumably the most settled sort of cyberattacks, returning to the 1990s, it's so far one of the most no matter how you look at it and toxic, with phishing messages and techniques getting continuously refined. Phishing destinations have such an enormous number of ways to deal with get the client's own one of a kind information, Types of phishing

If there's a mutual factor among phishing ambushes, it's the cover. Normally the phishing attackers sends email address to client by sending it from fake address so the client thinks that the email has received from the secure address. In these way the phishing attackers making fool so clients become fool and use outside. All things considered, there are a combination of systems that fall under the umbrella of phishing. There are a few one of a kind ways to deal with isolated ambushes into characterizations. One is by the inspiration driving the phishing try. By and huge, a phishing exertion endeavors



Figure 1.1: Phishing attacks

These messages intend to trick the customer into revealing noteworthy information —Normally to crack a framework some data is required that is username and public key this is also called the encryption. This can not be done without the public key of the server or the backend server. The exemplary adaptation of this trick includes conveying an email customized to seem as though a message from a significant bank; by spamming out the message to a large number of individuals, the aggressors guarantee that prob-ably a portion of the beneficiaries will be clients of that bank. The casualty taps on a connection in the data and there is a chance of miss using that data in any other private websites some times there is a chance of using that data in the bank site and they try to steal money, so while entering into any websites the client has to take care about the data and the email address. the phishing attackers mainly focus on the email address only because the email was the basis for the phishing attacks if by mistake the phishing attackers get the email address they will send the fake email to the client that client believes that the email was not fake and if he enters the data then the phishing attackers will misuse that data in any other private website and client will suffer so much for the blind mistake.in 2017 it was evaluated that 93% of phishing messages contained ransomware connections.

There are also a couple of particular ways that phishing messages can be centered around. Hoodlums rely upon dubiousness and making a need to continue moving to put forth advance with their phishing attempts. Crises, for instance, the coronavirus pandemic give those gangsters a significant opportunity to bring setbacks into taking their phishing trap There are a huge amount of motivations to utilize antivirus programming. Uncommon engravings that are solidified with antivirus programming guard against known turn of events workarounds and escape from arrangements. Basically

try to keep alert with the latest. New definitions are fused ceaselessly considering the way that new hoodwinks are in addition being conceptualized constantly. Antagonistic to spyware and fire-divider settings should be used to hinder phishing attacks and customers ought to invigorate the ventures reliably. Firewall confirmation thwarts access to malicious records by deterring the ambushes. Antivirus programming checks each record which moves beyond the Internet to your PC. It encourages with forestalling insidiousness to your framework. These are scarcely any strategies to evade the phishing assaults in these techniques fourth is best for keeping away from the phishing assaults so to build up that strategy we need to utilize AI.

1.2 Problem statement

Machine-learning algorithm (Logistic regression) is used and provided with a massive dataset. From that dataset, it trains the algorithm and helps in detecting the new web links which are fraudulent links. Attackers disguise their website as legitimate and try to get data from the user for which they make users visit a website and get the personal information that is needed. The results of using these methods are not 100% success rate but we can get up to 98% of success rate because day by day the phishing attackers are introducing new websites to overcoming the methods. so the success rate is 98%. But using these methods the client can know about the behavior of websites. If the client wants to avoid that website he can do it. So the result of the method increases the security of the client's personal information from the phishing attackers.

Before trying to enter into those websites it is important to check whether the given link is good or a phishing link. By checking the link we can save ourselves from the attackers and can keep our data safe.

1.3 Existing system

Nowadays, To stop the phishing attacks from stealing personal information we have to do manually. The phishing websites to be detected and block that phishing website's these whole process has to be done manually, By doing manually it will consume more

time and there is a chance of doing mistakes while blocking the phishing websites. When Blocking the phishing websites manually there is a chance of security breach or there is a chance of security breach while developing the software or repairing the software this kind of issues are raising in the present software which is blocking phishing websites manually. There is another way to steal the client's personal data, which is sending the spam emails to client emails or they will send to random emails address if the client clicks the email then phishing attackers will steal the data from the client. Nowadays there is some anti-phishing software that is used to inform or block the phishing website not spam emails to the client email address. these type of software can be installed on the client's computer

1.4 Proposed system

This region delineates the proposed model of phishing ambush area. The proposed model bright lights on recognizing the phishing ambush reliant on checking phish-ing destinations features, Blacklist, and WHOIS database. According to relatively few picked features can be used to isolate among genuine and criticized site pages. These picked high-lights are various, for instance, URLs, region character, security and en-cryption, source code, page style and substance, the web address bar, and the social human factor. This study concentrates just on URLs and space name highlights. Features of URLs and space names are checked to use a couple of models, for instance, IP Address, long URL address, including a prefix or expansion, redirecting using the picture.

1.5 Advantages of the proposed system

Improving the performance of the software decreases the risk of stealing personal data from the phishing attackers, In the proposed work we have mentioned the improvement that can be done in this project, By doing the improvements to the software decrease the work that can be done manually. Everything in this software can be done by the software itself but not by humans. But there is some work that can be done by only humans.

So some work has to do manually by humans and these proposed work increase the accuracy rate of the software to find the link from the data sets which are already set by the humans. By using these improvements the speed of the software increase and also catches most of the phishing websites.

1.6 Summary

URLs are sometimes known as “Weblinks” are the primary means by which People use them on computers, to make fetch the data and show some specific asset Our point is to infer order models that identify phishing sites by examination of the lexical and host-based highlights of URLs. To find the spam website different classifying algorithms must be analyzed.

CHAPTER 2

SYSTEM REQUIREMENT SPECIFICATIONS

The software requirements detail is created at the zenith of the examination task. The capacity and execution apportioned to programming as a major aspect of framework engineer-ing are refined by setting up a total data depiction as practical rep-resentation of framework conduct, a sign of execution necessities and plan requirements, proper approval standards.

2.1 Python

Python is a programming language, and it is an object-oriented programming language. The main difference is the Python has its own IDE so no need to search for online IDE like used in JAVA IDE, you can also use the online compiler as well. Python is having functions, classes, methods called a python library, library are special like we can add different features where it will be very helpful and it will be easy to understand the python language, by using these function and all the stuff python makes itself very special. In python, there are not data types like another programming language here you can directly assign the data to the variable without thinking of which type of data you are assigning to the variable this is called dynamic in nature. In python, there are different ways to store the data like List, Tuples, DictionariesActive server page (ASP). which you can add the data it doesn't matter whether you are adding which type of data. But in some cases and in few types we can modify the data present because this python has the ability to use the data whenever it required. this is the main feature in python.

2.1.1 Python Platform

Aside from Windows, Linux and MacOS, CPython usage runs on 21 distinct stages. IronPython is a .NET structure based Python execution and it is cabable of running in the two Windows, Linux and in different situations where .NET system is accessible.

2.1.2 Python Library

Machine Learning (ML) If you want to write any code in python we need python libraries because these are the key elements that will perform actions in the code, like functions, methods by using these functions, and methods we can write the code in python because it is an object-oriented programming language. When you are using functions and methods there are different and a number of inbuilt functions where you can find in Docs sections or you can simply type Help in the IDE so that you can open the documents present in the IDE will open, in the documents you can find every function and inbuilt methods which will help in writing code easily. The built-in Library is helpful because of no need to search for what kind of function required in the code directly you can find in the documents, by using these you decrease the length of the code because we can use functions in place of the code like if you want to generate the random number you will write whole code for that but we can use random module so that it will generate a random number which is in one line code but if you are writing code it will be like more than 10 lines code, like these we can use the Libraries.

2.2 Machine Learning

ML is an interesting issue for some key reasons, and in light of the fact that it gives the capacity to automatically get profound experiences. This significant level comprehension is basic if at any point associated with a dynamic procedure encompassing the utilization of AI, how it can help accomplish business and venture objectives, which AI strategies to utilize, potential entanglements, and how to decipher the outcomes.

Its objective and utilization is to construct new or potentially influence existing calculations to gain from information, so as to fabricate generalizable models that give precise forecasts, or to discover designs, especially with new and inconspicuous comparable information. Imagine a dataset as a table, where the lines are every perception (otherwise known as estimation, information point, and so on), and the segments for every perception speak to the features of that perception and their qualities. At the start of an AI venture, a dataset is normally part into a few subsets. The base subsets are the prepa-

ration and test datasets, and regularly a discretionary third approval dataset is made too. When these information subsets are made from the essential dataset, a prescient model or classifier is prepared utilizing the preparation information, and afterward the model's prescient precision is resolved utilizing the test information. As referenced, AI use calculations to automatically show and discover designs in information, for the most part with the objective of anticipating some objective yield or reaction. These calculations are vigorously founded on measurements and mathematical enhancement. Enhancement is the way toward finding the littlest or biggest worth (minima or maxima) of a capacity, regularly alluded to as a misfortune, or cost work in the minimization case. One of the most famous streamlining calculations utilized in AI is called inclination plunge, and another is known as the ordinary condition. More or less, AI is about naturally learning a profoundly exact predictive or classifier model, or discovering obscure examples in information, by utilizing learning calculations and advancement techniques. cikit-learn, Theano, TensorFlow, Keras, PyTorch, Pandas, Matplotlib

Anaconda Navigator

In the desktop graphical interface Graphical user interface (GUI), Anaconda Navigator is one of the popular interfaces to create the graphics using the python and it is easy to launch the applications it makes the user friendly to install the anaconda default packages, environment, and channels. To install this future, The user need not to worry about the command line argument because the to install this there is no need to use command-line argument. This will search for the packages in all available storage futures like a cloud or local storage. This search is available in all operating systems in world.

NumPy

python contain so many libraries in that libraries numpy is most popular compared to because it is multidimensional cluster and network handling capacities. It is extremely valuable for key logical calculations in Machine Learning. It is standard ticularly valuable for direct polynomial math, Fourier change, and irregular number abilities. In some most popular libraries like tensorflow also uses the numpay in some cases .

SciPy

SciPy contains so many different modules to make optimization like linear algebra, integration, and statistics. So these make the scipy more popular library among the ml libraries. In ml there is one more thing that belongs to scipy is scipy stack. there is some small difference between the stack and the library. the core package from the library makes the scipy stack. In ml, to manipulate the image scipy is used.

Skikit

Skikit is a library that belongs to the ml. The skikit is used for old style ml calculations among the libraries, but this skikit library based on 2 essential python library's that are numpy and scipy. supervised and unsupervised algorithms are supported by the skikit. In data mining and data analysis will use the same skikit, so it makes most important tool who is using the ml.

Theano

Theano as a whole realize that ml is fundamentally arithmetic and measurements. Theano is also one of the most important libraries of python that uses the characterize, assess, and streamline including multi-dimensional clusters in a proficient way. utilization of cpu and gpu can be achieved by it. To diagnose diff types of errors we have to use unit-testing and self-verification extensively. The library which is most powerful in ml is Theano which is used I large scale intensively in big computer projects from a long time in software technology because it very simple to use and it easy to understand and it can be used individuals for there own projects

TensorFlow

ML has some powerful open-source libraries in that TensorFlow one which is devel-oped by the Google team in Google for best performance numerical com-putation. This library is used to work on frameworks that are based on the running computations in-cluding tensor flow. the main future of this library is to prepare and run for neural

networks. so this library can also be used to make the ai applications also. In deep learning research, we can also use TensorFlow widely.

Keras

In the ml libraries, Keras is one of the well-known libraries because it has the capability to running API on top of TensorFlow or Theano which is high-level neural networks and it can easily run on both gpu and cpu. Neural Network can be build and design in ml programming by using the keras libraries. the best thing about the Keras libraries is to allows fast and easy prototyping.

PyTorch

In the desktop graphical interface, PyTorch is one of the popular interfaces to create the graphics using the python and it is easy to launch the applications it makes the user friendly to install the PyTorch default packages, environment, and channels. To install this future, The user needs not worry about the command line argument because to install this there is no need to use command-line argument. This will search for the packages in all available storage futures like a cloud or local storage. This search is available in all operating systems in the world.

Matpoltlib

Whole realize that ML is fundamentally arithmetic and measurements. Matpoltlib is also one of the most important libraries of python that uses the characterize, assess, and streamline including multi-dimensional clusters in a proficient way.utilization of cpu and gpu can be achieved by it. To diagnose diff types of errors we have to use unit-testing and self-verification extensively. The library which is most powerful in ml is Matpoltlib which is used I large scale intensively in big computer projects from the long time in software technology because it very simple to use and it easy to understand and it can be used for individuals for there own projects.

2.3 HTML&CSS

The webpage is a document, commonly written in HTML, that is viewed in an Internet search engine. A site page can be gotten by entering a URL into a program's location bar. A page may contain substance, structures, and hyperlinks to other site pages and reports. webpage plays a major role in the project because we need a platform to enter the link and it has to attract the clients so that they will use our website. To make the website more attractive we use CSS.

CSS is used to portray styles for your site pages, including the structure, for-tangle, and assortments in appear for different devices and screen sizes. properties to HTML elements so that we obtain the most attractive website, In the webpage, Elements plays a major role because of depend-ing on the elements only we can get the different input boxes or the different type of heading that makes the page more. The web page is created with the latest version of html5 because in the html5 has more features compared to old version html4, html5 contain doctype HTML tag in the top of the code which will tell about the version to the compiler, and it contains two major tags there are head and body tags. The <head> tag contains meta, explicit informa-tion about the site page that isn't shown to the client. Metadata furnishes programs and web search tools with spe-cialized data about the site page. The <body> tag characterizes the archive's body. An inside template might be utilized on the off chance that one single HTML page has a one of a kind style. Negative qualities are not permitted.

Content - The substance of the container, where content and pictures show up
Padding - Clears a zone around the substance. The cushioning is straightforward, Border - A fringe that circumvents the cushioning and substance, Margin - Clears a territory outside the outskirts. The edge is a straightforward box model that permits us to include a fringe around components, and to characterize space between components. A layout is a line that is drawn around components, OUTSIDE the fringes, to make the component "stand out". Outline contrasts from outskirts! In contrast to the outskirts, the diagram is drawn outside the component's fringe and may cover with other substance. Like-wise, the blueprint isn't a piece of the component's measurements; the component's all out width and stature are not influenced by the width of the framework. The diagram

style property indicates the style of the framework, and can have one of the accompanying qualities: The layout width property determines the width of the blueprint. The layout balance property includes space between a blueprint and the edge/fringe of a component. The space between a component and its diagram is transparent. The shading property is utilized to set the shade of the content. The shading is determined by If you characterize the shading property, you should likewise characterize the foundation shading. Since the shade of the content must be diff. The content adjust property is utilized to set the even arrangement of a book. A book can be left or right-adjusted, focused, or legitimized. There is one more incentive for content adjust that is defended which makes content arrangement like paper content arrangement There is one more incentive for content adjust that is legitimized which makes content arrangement like paper content arrangement other content enhancement esteems are utilized to enliven content are Underline, Line through, Overline It isn't prescribed to underline message that isn't a connection, as this regularly confounds the peruser.

CHAPTER 3

LITERATURE SURVEY

3.1 A Systematic Review of Software-Based Web Phishing Detection

Belabed et al. (2012) Phishing is a sort of cyberattack that utilizes social structuring technique ologies and other complex systems to assemble singular information from customers of destinations. The ordinary yearly improvement pace of the quantity of amazing phishing locales distinguished by the Anti Phishing Working Group is 36.29% for as far back as six years and 97.36% for as long as two years. In the wake of this ascent, easing phishing assaults has gotten a developing enthusiasm from the cyberse-curity network. Wide inventive work has been directed to distinguish phishing attempts subject to their outstanding substance, framework, and URL characteristics. Existing procedures change in a general sense similar to impulses, in-arrangement assessment methodologies, similarly as appraisal draws near.

3.2 Intelligent rule-based phishing websites classification

Wenyin Liu et al. (2006) Phishing is depicted as the art of reverberating a site of a re-spectable firm proposing to get customer's private information, for instance, user-names, passwords, and normalized investment funds numbers. Phishing locales include a grouping of prompts in-side its substance parts similarly as the program based secu-rity pointers outfitted close by the site. A couple of plans have been proposed to deal with phishing. Coincidentally, there is no single charm shot that can understand this risk significantly.

3.3 An fMRI Study of Phishing and Malware Warnings

Alnajim and Munro (2009) The security of PC frameworks regularly depends upon the choices and activities of end-clients. In this paper, we set out to investigate users' susceptibility to cybercriminal attacks by concentrating on the most fundamental component governing user behavior-the human brain. present a novel neuroscience-based examination technique to illuminate the structure regarding client-focused security frameworks as it identifies with cybercrime. Specifically, report on a useful attractive reverberation imaging study estimating clients' security execution and hidden neural action as for two basic security errands: (1) recognizing an authentic and a phishing site and (2) paying attention to security (malware) alerts

3.4 Phishing-Alarm: Robust and Efficient Phishing Detection via Page Component Similarity

Mao et al. (2017) Informal communities have gotten one of the most famous stages for clients to cooperate with one another. Given the enormous proportion of sensitive data open in relational association stages, customer security confirmation on casual associations has gotten research issues. As a piece of regular data taking methodology, phishing ambushes in spite of everything work in their way to deal with cause a huge amount of assurance encroachment events. In a Web-based phishing attack, an assailant sets up stunt Web pages (professing to be a huge Website, for instance, a casual network access) to attract customers to incorporate their private data, for example, passwords, standardized savings numbers, charge card numbers, etc., such as passwords, social security numbers, credit card numbers, and so on.

3.5 OFS-NN: An Effective Phishing Websites Detection Model Based on Optimal Feature Selection and Neural Network

Yang et al. (2019)

A phishing attack is currently a major danger to individuals' every day life and systems administration environment. By camouflaging illicit URLs as real ones, aggressors can incite clients to approach the phishing website to get private data and different advantages. Viable strategies for recognizing phishing sites are desperately expected to ease the dangers presented by the phishing assaults. As the dynamic takes failure from massive educational collections, the neural framework is extensively used to perceive phishing ambushes. In any case, in the period of getting ready instructive assortments, various trivial and little effect characters will collect the the neural framework model into the issue of the fitting. The issue, generally, infect a readied model that can't enough distinguish phishing locales. So as to diminish the issue, this proposes OFS, a convincing spam site identification reliant the perfect component decision method and neural framework. In the OFS, another record, feature authenticity regard (FVV), is first air conditioning quainted with evaluating the sway of unstable characters on the spam locales acknowledgment. By then, taking into account the new FVV list, a count is planned to pick the perfect features from the phishing locales. This estimation can fitting issue of the key neural framework to colossal degree.

3.6 Protecting Users Against Phishing Attacks with AntiPhish

Prakash et al. (2010)

Normally to crack a framework some data is required that is username and public key this is also called the encryption. This can not be done without the public key of the server or the backend server. The exemplary adaptation of this trick includes conveying

an email customized to seem as though a message from a significant bank; by spamming out the message to a large number of individuals, the aggressors guarantee that prob-ably a portion of the beneficiaries will be clients of that bank. The casualty taps on a connection in the data and there is a chance of miss using that data in any other private websites some times there is a chance of using that data in the bank site and they try to steal money, so while entering into any websites the client has to take care about the data and the email address. the phishing attackers mainly focus on the email address only because the email was the basis for the phishing attacks if by mistake the phishing attackers get the email address they will send the fake email to the client that client believes that the email was not fake and if he enters the data then the phishing attackers will misuse that data in any other private website and client will suffer so much for the blind mistake.in 2017 it was evaluated that 93% of phishing messages contained ransomware connections.

3.7 Detecting Phishing Websites Based on the Study of the Financial Industry Webserver Logs

Zhang et al. (2011) Phishing revelation procedure, which relies upon the examination of certifiable site server log information. The idea relies upon finding the references. By virtue of the references, each time the loss opens the phishing site, the phishing site will insinuate the genuine site by mentioning resources. By then, there is a log, which is recorded by the genuine site server. Through looking at the logs, we find the phishing locales. The idea is unequivocal and centered around. According to our tests, it is convincing and has high accuracy.

3.8 A Hybrid System to Find & Fight Phishing Attacks Actively

Chen and Guo (2006) Customer preparing must focus on testing and changing the mis-interpretations that control current customer direct. Until this point in time, customer

preparing on spam endeavored to persuade them to verify the website links and different different markers, with confined accomplishment. The makers execute the anti-spam mechanical assembly in a sensible fixing—individuals expected to take the data under tension and lose data in case they bring from horrible areas. Though no one individuals bought from areas the instrument evidently perceived as awful, 40 % bet data has goals hailed as perhaps hazardous, yet offering bargains. When lured by a better than average course of action, individuals forgot to concentrate on the cautions

3.9 Security Education against Phishing: A Modest Proposal for a Major Rethink

Kirlappos and Sasse (2012) The standard foe of phishing methodologies and mechanical assemblies reliably focused in a uninvolved direction to get customers' convenience and choose spam website. Conventionally, they are capable enough to discover and chop down spam ambushes. examine spam reports are the Anti-spam Alliance of China. Anti-phishing Alliance of China (APAC) and propose a creamer technique to find spam assaults in a working subject to domain name server request users and known spam websites links. We make and pass on to the structure to report phishing websites link consequently to APA reliably. the structure becomes the guideline redirect in giving spam reports to APA in China and can become better than an average enhancement to the ordinary foe of phishing techniques. can be a decent supplement to the conventional enemy of phishing strategies.

3.10 Inference from the survey

From the survey, we have seen that there are different types of methodologies to find Phishing websites. Most of them are to avoid such type of links by blocking them. Many have a different methodology and have different accuracy in it. The accuracy of finding those is nearly 95% in all these above methodologies. Even then attackers are finding new ways to fool people and escape these securities. By using a different

approach and different methodology we can increase the accuracy in our cases. No one in the previous papers has used this algorithm called logistic regression by using this we can increase the accuracy and even by including the data set of large links.this we can increase the accuracy and even by including the data set of large links.

CHAPTER 4

PHISHING DETECTION SYSTEM DESIGN

4.1 Architecture

In the architecture diagram, there are three main steps to detect the phishing websites, first when the client enters the link in the input field to find whether it is a phishing website or not. The software will check the link in the blacklist which was present in the database maintained by the developer, if the link is present in the blacklist then the software will conclude that the link is a phishing website. If the link was not in the blacklist then the features of the link will be extracted by the software from those features the software will select some features which are used to conclude the behavior of the website like whether the link contains invalid symbols in it. so this whole process is done by the machine learning algorithm (logistic regression algorithm). After the process is complete the logistic regression will conclude the result. If the link is a phishing website then the software will display the result to the client and the website is entered into the blacklist. so that next time when the client enters the same link again we can find that link very easily. Otherwise, if the link is not a phishing website then the software will display the result to the client.

Configuration is a multi-step that centers around information structure programming design, procedural subtleties, technique and so on . . furthermore, interface among modules. The plan methodology likewise interprets the prerequisites into introduction of programming that can be gotten to for exceed expectations hence before coding starts. PC programming configuration changes persistently as novel strategies; improved examination and outskirt understanding developed. Programming proposition is at a moderately essential stage in its insurgency. Along these lines, programming plan procedure comes up short on the profundity, adaptability and quantitative nature that are normally connected with progressively ordinary designing orders. Anyway strategies for programming plans do leave, measures for structure characteristics are existing and structure documentation can be applied.

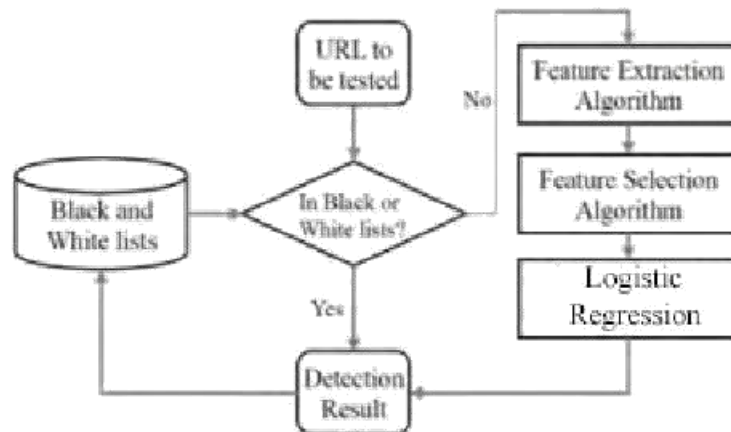


Figure 4.1: Architecture

4.2 Training and Testing

A training dataset is a dataset of models utilized for realizing, that is to fit the parameters. This is utilized to develop our forecast calculation and to change the loads on the neural system. Our calculation attempts to tune itself to the idiosyncrasies of the preparation informational indexes. In this stage, we, for the most part, make various calculations so as to look at their exhibitions during the Cross-Validation Phase. Each kind of calculation has its own parameter choices. A test dataset is a dataset that is free of the preparation dataset, yet that follows similar likelihood dissemination as the preparation dataset. testing information gives a fair assessment. At the point when we feed in the contributions of Testing information, our model will anticipate a few qualities. After the forecast, we assess our model by contrasting it and real yield present in the testing information. This is the way we assess and perceive how much our model has gained from the encounters feed in as preparing information, set at the hour of preparing.

With reference to the 5.2 The preparation informational index is a gathered arrangement of models that are utilized to fit the parameters. The test informational index is the last assessment of the last model fit on the preparation informational collection. So like the last test before it's the genuine article. Testing against one another guarantees the AI model will be increasingly exact.

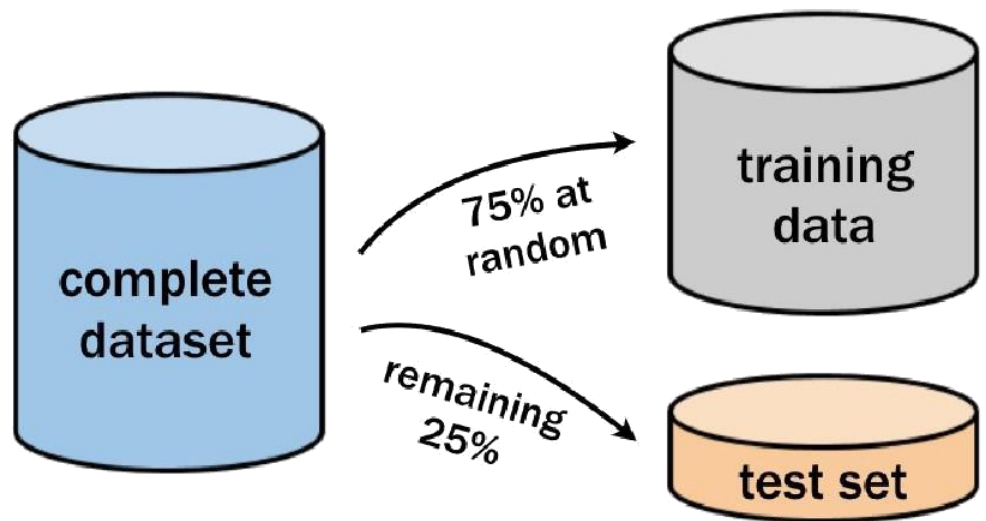


Figure 4.2: Training and Testing

4.2.1 Regression

By applies ml making sense of how to backslide as well. Acknowledge that $y = y_1, y_2, y_3, \dots, y_n$ is the information components and y is the outcome variable. For this circumstance, we can use AI development to convey the yield (x) in light of the data factors (y). You can use a model to convey the association between various parameters as underneath: $x=g(y)$ where g is a limit that depends upon express qualities of the model. In backslide, we can use the standard of AI to overhaul the parameters. To cut the gauge botch and figure the closest possible outcome. We can in like manner use Machine learning for work smoothing out. We can choose to change the commitments to improve model. This gives an as good as can be expected model to work with. This is known as response surface structure.f the models incorporate estimation of lodging value, item value, stock cost and so forth. Relapse

4.2.2 Logistic Regression

In measurements, the calculated model (or logit model) is utilized to demonstrate the probability of a picked class or event existing like pass/miss the mark, win/lose, alive/dead, or strong/cleared out. this might be contacted show a couple of classes of events like choosing if a picture contains a catlike, dog, lion, etc. Each article is recognized inside the image would be given out a probability some place in the scope of 0 and 1 and as such the absolute adding to at any rate 1.

In twofold calculated relapse, the outcome is by and large coded as "0" or "1", as this prompts the most clear interpretation. In case a particular watched result for the dependent variable is the huge possible outcome it is regularly coded as "1" and the contrary outcome (suggested as a "failure" or a "noncase") as "0". Equal determined backslide is used to envision the odds of being a case reliant on the estimations of the self-ruling variables. The odds are portrayed as the probability that a particular outcome is a case segregated by the probability that it is a noncase.

Key loses the faith is besides an honest model in its significant shape utilizes a stop mined capacity to show a consolidated character, however dynamically complex expansions exist. inside the legitimate technique, indispensable lose the faith is as-sessing a limit of a decision model. an equivalent decided model highlights a character that has 2 most important attributes, similar to pass which is tended to by a pointer character, where the 2 qualities name it has "0" and "1". inside these decided model, the log-chances for the value checked "1" was furthermore a straight blend of in any occasion 1 free factor ("pointers"); oneself administering parts can each be a twofold factor or perpetual charecters. The differentiating likelihood of the value named "1" can fluctu-ate b/w 0 (absolutely the value "0") and 1 subse-quently the naming; the breaking point that changes over log-chances to likelihood is that as far as possible, thusly the name. The unit of estimation for the log-chances scale is named a logit, from the figuring unit, hence the decision nms. Like model's with a remarkable sigmoidbreaking point rather than as far as possible besides can be utilized, to a few

degree like the probit model; the describing typical for the key model is that grow-ing one among the independent elements multiplica-tively scales the chances of the

given outcome at a steady rate, with each factor having its own parameter; for an equal variable, this summarizes the chances extent.

In an equal determined backslide model, the variable has two levels (obvious). Yields with exceptionally two characteristics are shown by multinomial determined backslide and, if the various arrangements are mentioned, by ordinal vital backslide (for example the cor-reacting chances ordinal determined model). The determined backslide model itself just models the probability of yield similarly as data and doesn't perform quantifiable gathering ing (is definitely not a classifier), in any case, it is much of the time accustomed make a classifier, for example by picking cutoff regard and request- ing commitments with probability more significant than the cutoff together class, underneath the cutoff considering the way that the other; this will be routinely regular appreciation to making a twofold classifier. The coefficients are all things considered not handled by a shut structure enunciation, as opposed to coordinate least squares; see Model fitting. The key backslide as a general accurate model was at first developed and advanced basically by Joseph Berkson, beginning in Berkson (1944), where he created "logit"; see History.

Advantages of LR

- Logistic Regression (LR) performs good when the dataset is directly distinguish-able.
- LR is inclined to over-fitting however it can over-fit in dimensional dataset. You ought to think about Regularization (L1 and L2) strate-gies to keep away from over-fitting in these situations.
- LR not just gives a proportion of how significant an indicator (coefficient size) is, yet additionally, it's bearing of affiliation (positive or nega-tive).
- LR is simpler to execute, decipher, and extremely effective to prepare.

Disadvantages of LR

- Main limitation of LR is the supposition of linearity b/w the reliant variable and the free factors. In reality, the information is infrequently directly distinguishable. More often than not information would be a cluttered wreckage.
- On the off chance that the quantity of perceptions are lesser than the quantity of highlights, LR ought not be utilized, else it might prompt overfit.

- LR must be utilized to anticipate discrete capacities. Thusly, the reliant variable of LR is confined to the discrete number. This limitation itself is hazardous, as it is restrictive to the forecast of nonstop information.

4.3 Use case diagram

when Clint enters the data, (which is link formate) The futures(__,%,\$,#) from that data are extracted. After extraction, data contain only characters that characters are encrypted into the number formate because security purposes and to give input to the logistic regression algorithm. A data set (data set contain so many numbers of link's which are phishing websites links) which is in the format of the spreadsheet is taken and it will train the data set and future abstraction will take place and after that, it will divide the data set into 80% and 20% of data by using train test split method. 80% percent of the data is given to the logistic regression algorithm.

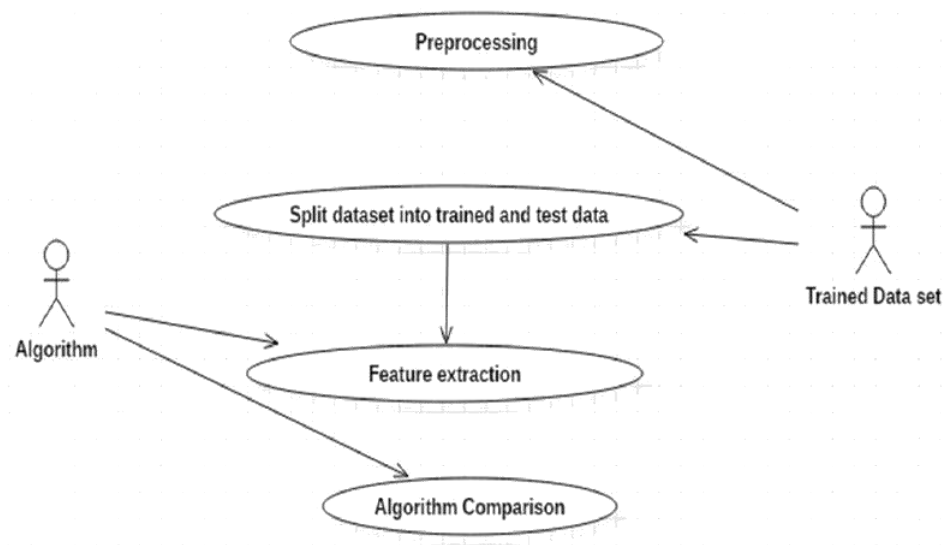


Figure 4.3: Use case diagram

4.4 Activity diagram

An informational collection (informational collection contain such a large number of quantities of connection's which are phishing sites joins) which is in the configuration of the spreadsheet is taken and it will prepare the informational collection and future deliberation will occur and from that point forward, it will part the informational collection into 80% and 20% of information by utilizing train test divide technique. 80% percent of the information is given to the calculated relapse calculation. when Clint enters the data, (which is link formate) The futures from that data are extracted. After extraction, data contain only characters that characters are encrypted into the number formate because security purposes and to give input to the logistic regression algorithm.

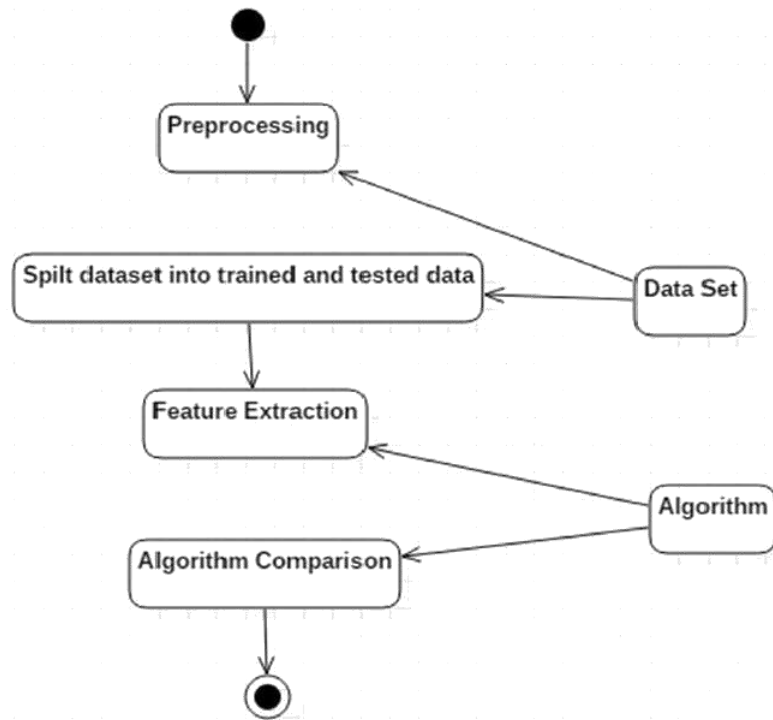


Figure 4.4: Activity diagram

4.5 ER diagram

A data set (data set contain so many numbers of link's which are phishing websites links) which is in the format of the spreadsheet is taken and it will train the data set and future abstraction will take place and after that, it will divide the data set into 80% and 20% of data by using train test split method. 80% percent of the data is given to the logistic regression algorithm. when Clint enters the information, The prospects from that information are removed. After extraction, the information contains just characters that characters are encoded into the number formate in light of the fact that security purposes and to offer a contribution to the strategic relapse calculation.

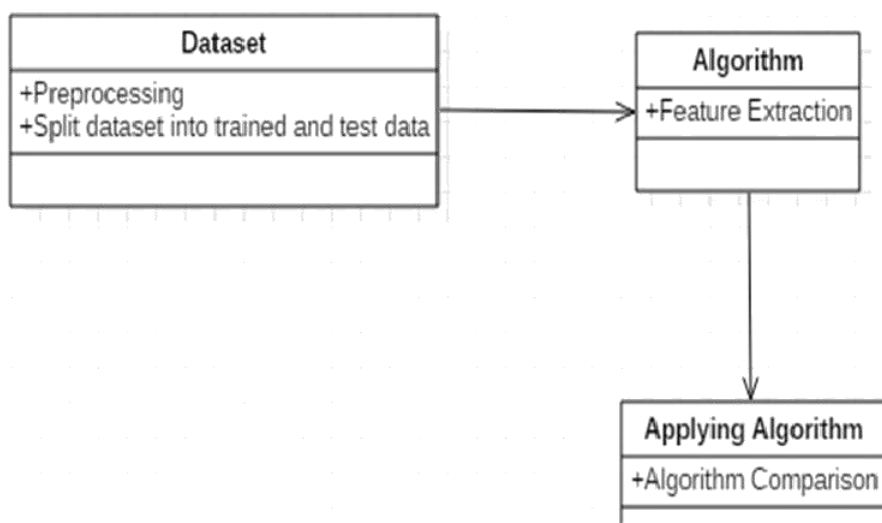


Figure 4.5: ER diagram

CHAPTER 5

PHISHING DETECTION SYSTEM MODULES

5.1 List of phishing detection sytem modules

1. Data Collection and Pre-processing
2. Data cleaning and Data transformation
3. Data selection and Data sets
4. Feature extraction
5. Prediction
6. Accuracy

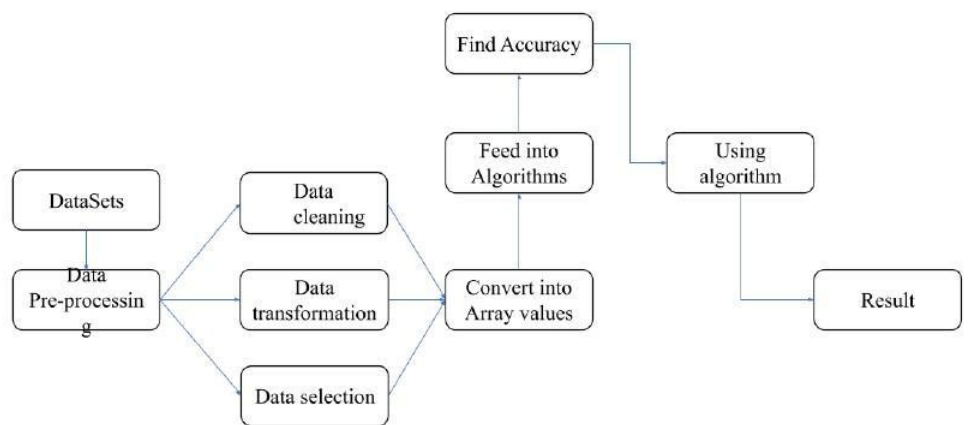


Figure 5.1: The workflow of phishing detection models

5.1.1 Data Collection and Pre-processing

Data Collection is one of the most important tasks in building a machine learning model. It is the gathering of task related information based on some targeted variables to analyse and produce some valuable outcome. However, some of the data may be

noisy, i.e. may contain inaccurate values, incomplete values or incorrect values. Hence, it is must to process the data before analysing it and coming to the results. Data pre-processing can be done by data cleaning, data transformation, data selection. Data cleaning: Fill in missing qualities, smooth uproarious information, recognize or expel exceptions, and resolve

inconsistencies. Data change may incorporate smoothing, accumulation, generalization, transformation which improves the quality of the data. Data selection includes some methods or functions which allow us to select the useful data for our system.

5.1.2 Data cleaning and Data transformation

Data cleaning is the process in which we get noisy data. This noisy data in the sense which is not fully completed and has some missing values that will be written fully. This will be taken and the entire data will be read and will take the possibility of clean data so that the data can go to the next process. Data transformation may include smoothing, aggregation, generalization, transformation which improves the quality of the data

5.1.3 Data selection and Data sets

Data selection includes some methods or functions which allow us to select the useful data for our system. Data input: After finding the best algorithm, that algorithm is used to find the phishing websites. After that input is given to the algorithm and then it gives the output.

A collection of data of a particular instance is stored in a particular sheet which is also known as a dataset and it is used to work on the machine learning that data set can be used for many purposes. Training Data set: A data set that To train our model, it feeds into the logistic regression algorithm. This is the data type used to provide an unbiased evaluation of the final that is completed and fit on the training data set that data is used for the learning purpose

5.1.4 Feature extraction

In order to reach a website on the internet, the URL is used as a medium. Phishing links are generally disguised as legitimate by doing phishing attacks, the attacker can make a user click on their web links which are pretending to be the original ones.

The Table 5.1 displays about feature extraction.

Sr.No	Address bar	Abnormal features	html and Java Script	Domain Features
1	URL contains an ip address	Request URL	Website forwarding	age of domain
2	Length of the input Url	URL of anchor	Status for costumization	DNS record
3	url contains "@" symbol	linksin<meta><script>	Disables right click	page rank
4	URL contains "/" symbol	submits from email	pop up window	google index
5	url contains ":" symbol	Abnormal URL		links pointing to page

Table 5.1: Feature extraction

These phishing links have few identifiable features. As listed in the above 5.1 these features are divide into four classes they are divide into four classes they are

- Address bar
- Abnormal Features
- Domain features

Address bar

In this IP address special characters like (_,%,,\$,#) etc.. and the length of the input URLs are needed to specified to detect phishing attacks accurately. It also has a length of the Uniform resource locator (URL), Short URL existence. We also check the fea-tures like Domain registration length. We check all these features and see whether the given link contains all these or not if the criteria is satisfied then only it will go to the other one.

Abnormal Features

Abnormal features can be checked from the requested URL and the link tags like <META>,<Script>,<Link>, The S F H status information. Checking from the handler. Check Whether the web page having the link that is going to submit the data entered to any mail. Lastly, say it as an Abnormal URL or not.

Domain features

This section of the code generally checks the age of the domain because that phishing websites have a very short time. It checks the DNS record Web Traffic and page rank as the phishing pages have very few visitors it will have the index of 0 or 1 not more than that. It also checks the google index for that and the number of pages pointing to that page all this information of the features is saved in the Statistical report.

Web page

To extract the features of the link first we need a medium to enter the link. This project contains a webpage that is created with HTML and CSS. input, button, h1, a, img, and p are the elements used to make the webpage and By using css style properties like hover, animation, transform, font- style, background and container which makes webpage look premium. After entering the link that webpage directs to python code where it will extract the features using machine learning algorithms like logistic regression

5.1.5 Prediction

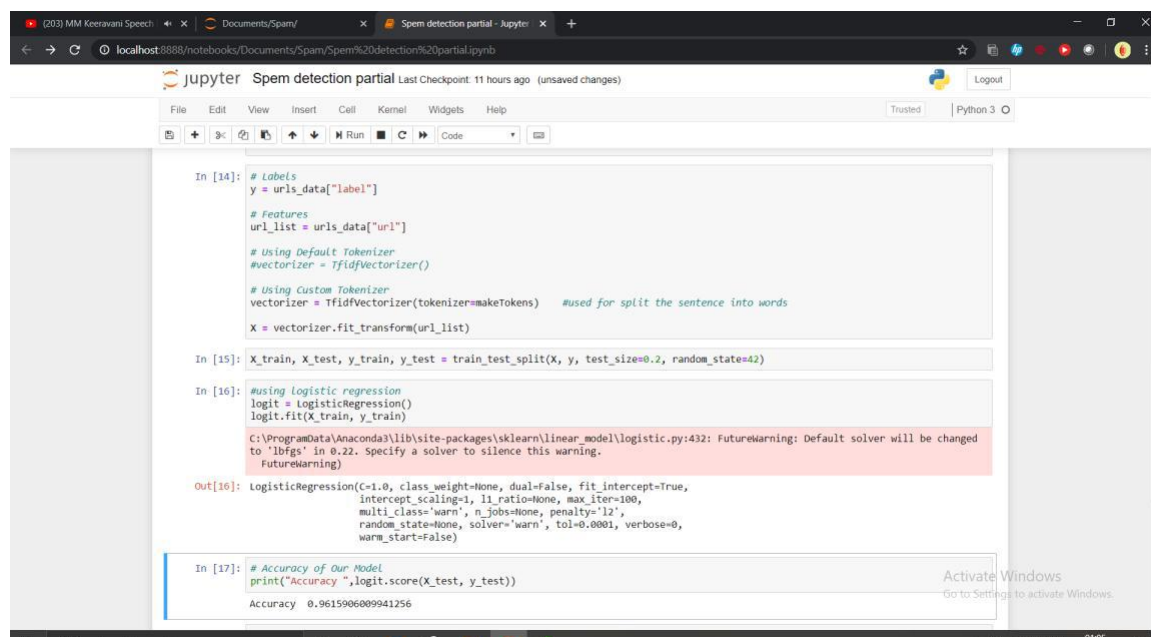
A prediction is what somebody thinks will occur. Pre signifies "previously" and lingual authority has to do with talking. So a forecast is an announcement about what's to come. It's an estimate, once in a while dependent on realities or proof, some future action through the calculation that has been taken in the past and that could have the fortunate to happen for maximum extent. That calculation is taken from the past events and keeping those in the form of data set After training from the data set few features are extracted are required. By the machine learning algorithm, logistic-regression takes

all these features and check those with the link entered and redirects that link weather the given link is good or bad

5.1.6 Accuracy

Exactness is the proportion of the quantity of right forecasts to the all out number of information tests. Exactness functions admirably just if there are an equivalent number of tests having a place with each class. For instance, consider that there are 98% examples of class A and 2% tests of class B in our preparation set.in our case, there are 4,00,000+ info tests and that is taken as contribution to prepare the calculation for highlight extraction this strategic relapse gives the exactness as 96%

Finally we get the result based on our algorithms used. And it will shows the accu-racy and final output.



```
In [14]: # labels
y = url_data["label"]

# Features
url_list = url_data["url"]

# Using Default Tokenizer
#vectorizer = TfidfVectorizer()

# Using Custom Tokenizer
vectorizer = TfidfVectorizer(tokenizer=makeTokens) #used for split the sentence into words
X = vectorizer.fit_transform(url_list)

In [15]: X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

In [16]: #using logistic regression
logit = LogisticRegression()
logit.fit(X_train, y_train)

Out[16]: LogisticRegression(C=1.0, class_weight=None, dual=False, fit_intercept=True,
intercept_scaling=1, l1_ratio=None, max_iter=100,
multi_class='warn', n_jobs=None, penalty='l2',
random_state=None, solver='warn', tol=0.0001, verbose=0,
warm_start=False)

In [17]: # Accuracy of Our Model
print("Accuracy ", logit.score(X_test, y_test))

Accuracy 0.9615906009941256
```

Figure 5.2: Accuracy percentage

CHAPTER 6

CONCLUSION & FUTURE ENHANCEMENT

In the future, The web page is going connect with sign in and sign up page which makes increasing in the security of the web page and clients. If we find any better algorithms to increase the accuracy or for fetching data then replace that algorithm with an old algorithm, so that the speed of the web page will increase

Phishing is unspeakable risk in web field. This episode, the basic man inputs individual data to a bogus site which looks equivalent to a typical site. based on a survey on the phishing methods This provided a good understanding of the attack and many solutions. Various approaches have conversedin the paper for the location of phishing; in any case, the greater part of the strategies despite everything have limits like preci-sion, failing to distinguish objects, and so forth. But In this paper Logistic regression technique finds the accuracy as it is an open challenge in this phishing field the accuracy obtained is 96% percent and it may further increase based on the research.

REFERENCES

1. Alnajim, A. and Munro, M. (2009). "An anti-phishing approach that uses training intervention for phishing websites detection."
2. Belabed, A., Aïmeur, E., and Chikh, A. (2012). "A personalized whitelist approach for phishing webpage detection."
3. Chen, J. and Guo, C. (2006). "Online detection and prevention of phishing attacks." 2006 First International Conference on Communications and Networking in China. 1– 7.
4. Kirlappos, I. and Sasse, M. A. (2012). "Security education against phishing: A modest proposal for a major rethink." *IEEE Security Privacy*, 10(2), 24–32.
5. Mao, J., Tian, W., Li, P., Wei, T., and Liang, Z. (2017). "Phishing-alarm: Robust and efficient phishing detection via page component similarity."
6. Prakash, P., Kumar, M., Kompella, R. R., and Gupta, M. (2010). "Phishnet: Predictive blacklisting to detect phishing attacks."
7. Wenyin Liu, Xiaotie Deng, Guanglin Huang, and Fu, A. Y. (2006). "An antiphishing strategy based on visual similarity assessment." *IEEE Internet Computing*, 10(2), 58– 65.
8. Yang, P., Zhao, G., and Zeng, P. (2019). "Phishing website detection based on multidimensional features driven by deep learning." *IEEE Access*, 7, 15196–15209.
9. Zhang, H., Liu, G., Chow, T. W. S., and Liu, W. (2011). "Textual and visual content-based anti-phishing: A bayesian approach." *IEEE Transactions on Neural Networks*, 22(10), 1532–1546.

APPENDIX A

CODE

A.1 Import required fuctions

```
import pandas as pd
```

```
import numpy as np
```

```
import random
```

```
import train_test_split
```

A.2 Import data set from storage

```
urls_d= pd.read_csv("urldata.csv")
```

A.3 Find type of data set

```
type(urls_d)
```

A.4 Show the data set

```
urls_data.head()
```

A.5 Preprocessing

```
def makeTokens(f):
```



```

tkns_Slash = str(f.encode('utf-8')).split('/')

for i in tkns_Slash:

    tokens = str(i).split('-')

    = []

    for j in range(0,len(tokens)):

        temp = str(tokens[j]).split('.')

    tkns_Dot = tkns_Dot + temp

    total = total + tokens + tkns_Dot

    total = list(set(total))

    //remove redundant tokens

    if 'com' in total:

        total.remove('com')

    #removing .com because it occurs a lot of times return total

```

A.6 Store label column in y variable

```

y = urls_d["label"]

```

A.7 Using vectorizer for convert string data set into numerical data

```
vector = CountVectorizer(tokenizer=makeTokens)
```

A.8 Transfrom url list column into numerical and stored in s

```
s = vector.fit_transform(url_list)
```

A.9 Split the training and testing data

```
s_train, s_test, r_train,
```

```
r_test = train_test_split
```

```
(s, r, test_size=0.2, random_state=42)
```

A.10 Declare logistic regression

```
logit = LogisticRegression()
```

A.11 Feed the Training data into algorithm

```
logit.fit(s_train, r_train)
```

A.12 Feed the test data and find and print accuracy

```
print("Accuracy ",logit.score(s_test, r_test))
```

A.13 Import tkinter for User Interface

```
import tkinter as tk
```

```
root = tk.Tk()
```

A.14 Declare the size of interface size

```
cvas1 = tk.Canvas(root,width = 800, height = 600)
```

```
cvas1.pack()
```

A.15 Give the title of the interface

```
lb1 = tk.Label(root, txt='Phishing Website Prediction')
```

```
lb1.config(font=('italic', 24))
```

```
cvas1.create_window(400, 50, window=lb1)
```

```
lb2 = tk.Label(root, txt='Paste Here:')
```

```
lb2.config(font=('helvetica', 20))
```

```
cvas1.create_window(400, 200, window=lb2)
```

A.16 Create input box

```
etry1 = tk.Entry (root)
```

```
cvas1.create_window(400, 280, window=etry1)
```

A.17 Detecting function

```
def getSquareRoot():
```

```
s1 = etry1.get()
```

```

lb3 = tk.Label(root, text=

'The Result For ' + s1 + ' is:',

font=(italic', 20))

cvas1.create_window(400, 420, window=lb3)


res = vector.transform([x1])

prediction = logit.predict(res)


lb4 = tk.Label

(root, text= prediction,font

=('italic', 20, 'bold'))

cvas1.create_window(400, 460, window=lb4)


bn1 = tk.Button(text='Find',

command=getSquareRoot,

bg='brown', fg='white',

font=('italic', 18, 'bold'))

```

```
cvas1.create_window(400, 350, window=bn1)
```

```
root.mainloop()
```

APPENDIX B

SCREENSHOTS OF THE PROJECT

Based on that data set we can get the result used our decision tree and logistic regression algorithm to predict the result. Here we can also find out the accuracy rate of the algorithm. It will be helpful for finding phishing website whether it is good or bad

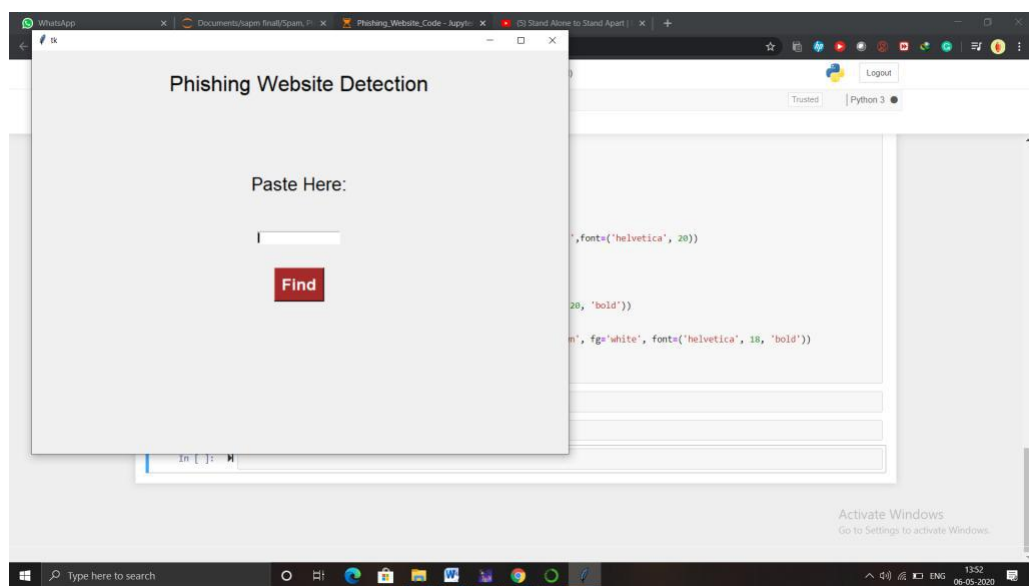


Figure B.1: Sample output

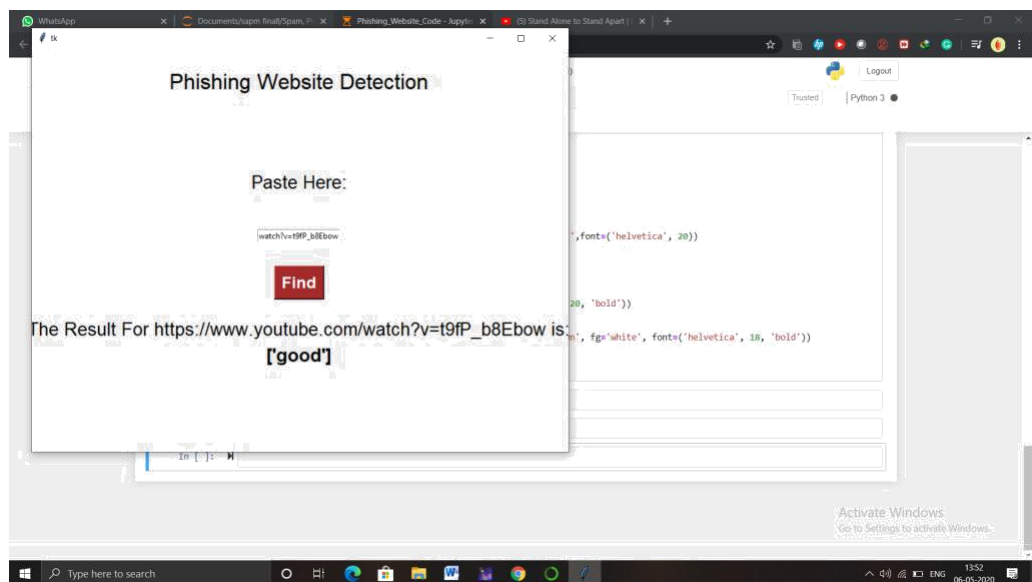


Figure B.2: Result output

Format - I

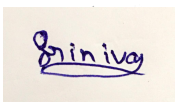
SRM INSTITUTE OF SCIENCE AND TECHNOLOGY

(Deemed to be University u/s 3 of UGC Act, 1956)

Office of Controller of Examinations

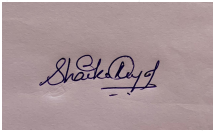
REPORT FOR PLAGIARISM CHECK ON THE DISSERTATION/PROJECT REPORTS FOR UG/PG PROGRAMMES
(To be attached in the dissertation/ project report)

1	Name of the Candidate (IN BLOCK LETTERS)	RUDRARAJU BHASKAR SRINIVAS RAJU
2	Address of the Candidate	Flatno-101, kasthuba nillyam Near yendada supermarket Vivekananda nagar, visakhapatanam Pincode:530045 Mobile Number : 8331055577
3	Registration Number	RA1611003010184
4	Date of Birth	16/05/1999
5	Department	Computer Science and engineering
6	Faculty	Engineering And Technology
7	Title of the Dissertation/Project	An effective model of terminating phishing websites and detection based on logist regression
8	Whether the above project/dissertation is done by	Individual or group : (Strike whichever is not applicable) ----- a) If the project/ dissertation is done in group, then how many students together completed the project : b) Mention the Name & Register number of other candidates : 2 Shaik riyaz RA1611003 010231
9	Name and address of the Supervisor / Guide	K.R.Jansi jansi.k@ktr.srmuniv.ac.in 9600082712 Mail ID : Mobile Number :
10	Name and address of the Co-Supervisor / Co- Guide (if any)	Mail ID : Mobile Number :

11	Software Used	Turnitin		
12	Date of Verification	22/05/2020		
13	Plagiarism Details: (to attach the final report from the software)			
Chapter	Title of the Chapter	Percentage of similarity index (including self citation)	Percentage of similarity index (Excluding self citation)	% of plagiarism after excluding Quotes, Bibliography, etc.,
1	Introduction	3		
2	System Requirement Specifications	2		
3	Literature Survey	2		
4	Phishing Detection System Design	2		
5	Phishing Detection System Modules	0		
6	Conclusion&Future Enhancement	0		
7				
8				
9				
10				
Appendices				
I / We declare that the above information have been verified and found true to the best of my / our knowledge.				
 Signature of the Candidate		Name & Signature of the Staff (Who uses the plagiarism check software)		
Name & Signature of the Supervisor/Guide		Name & Signature of the Co-Supervisor/Co-Guide		
Name & Signature of the HOD				

Format - I

SRM INSTITUTE OF SCIENCE AND TECHNOLOGY <small>(Deemed to be University u/s 3 of UGC Act, 1956)</small>		
Office of Controller of Examinations		
REPORT FOR PLAGIARISM CHECK ON THE DISSERTATION/PROJECT REPORTS FOR UG/PG PROGRAMMES (To be attached in the dissertation/ project report)		
1	Name of the Candidate (IN BLOCK LETTERS)	SHAIK RIYAZ
2	Address of the Candidate	DR-no-1-34,janda chattu,gundalapadu Phirangipuram mandal,guntur dist,pincode-522549,Andra pradesh Sr9087@srmist.edu.in Mobile Number : 9176299235
3	Registration Number	RA1611003010231
4	Date of Birth	20/06/1999
5	Department	Computer Science and Engineering
6	Faculty	Engineering And Technology
7	Title of the Dissertation/Project	An effective model of terminating phishing websites and detection based on logistic regression
8	Whether the above project/dissertation is done by	Individual or group : (Strike whichever is not applicable) a) If the project/ dissertation is done in group, then how many students together completed the project : 2 b) Mention the Name & Register number of other candidates : R.b.s.raju RA1611003010184
9	Name and address of the Supervisor / Guide	K.R.Jansi jansi.k@ktr.srmuniv.ac.in Mail ID : Mobile Number : 9600082712
10	Name and address of the Co-Supervisor / Co- Guide (if any)	Mail ID : Mobile Number :

11	Software Used	Turnitin		
12	Date of Verification	22/05/2020		
13	Plagiarism Details: (to attach the final report from the software)			
Chapter	Title of the Chapter	Percentage of similarity index (including self citation)	Percentage of similarity index (Excluding self citation)	% of plagiarism after excluding Quotes, Bibliography, etc.,
1	Introduction	3		
2	System Requirements Specifications	2		
3	Literature Survey	2		
4	Phishing Detection System Design	2		
5	Phishing Detection System Modules	0		
6	Conclusion & Future Enhancement	0		
7				
8				
9				
10				
Appendices				
I / We declare that the above information have been verified and found true to the best of my / our knowledge.				
 Signature of the Candidate		Name & Signature of the Staff (Who uses the plagiarism check software)		
Name & Signature of the Supervisor/Guide		Name & Signature of the Co-Supervisor/Co-Guide		
Name & Signature of the HOD				

ORIGINALITY REPORT

9%

SIMILARITY INDEX

1%

INTERNET SOURCES

3%

PUBLICATIONS

6%

STUDENT PAPERS

PRIMARY SOURCES

1

Submitted to Oklahoma State University

Student Paper

2%

2

Submitted to University of Adelaide

Student Paper

1%

3

Jun Hu, Xiangzhu Zhang, Yuchun Ji, Hanbing Yan, Li Ding, Jia Li, Huiming Meng. "Detecting Phishing Websites Based on the Study of the Financial Industry Webserver Logs", 2016 3rd International Conference on Information Science and Control Engineering (ICISCE), 2016

Publication

1%

4

Submitted to University of Warwick

Student Paper

1%

5

Submitted to SRM University

Student Paper

<1%

6

ieeexplore.ieee.org

Internet Source

<1%

7

bvucoepune.edu.in

Internet Source

<1%

8	Submitted to Stanmore College Student Paper	<1 %
9	Submitted to Sreenidhi International School Student Paper	<1 %
10	Submitted to Queensland University of Technology Student Paper	<1 %
11	Submitted to Luton Sixth Form College, Bedfordshire Student Paper	<1 %
12	www.ijitee.org Internet Source	<1 %
13	"Root for a Phishing Page using Machine Learning", International Journal of Innovative Technology and Exploring Engineering, 2019 Publication	<1 %
14	Submitted to City and Islington College, London Student Paper	<1 %
15	"Customer Loan Approval Classification by Supervised Learning Model", International Journal of Recent Technology and Engineering, 2019 Publication	<1 %
16	Submitted to Laureate Higher Education Group Student Paper	<1 %

17

Abdulghani Ali Ahmed, Nurul Amirah Abdullah.
"Real time detection of phishing websites", 2016
IEEE 7th Annual Information Technology,
Electronics and Mobile Communication
Conference (IEMCON), 2016

Publication

<1 %

18

docplayer.net
Internet Source

<1 %

Exclude quotes On

Exclude matches

< 10 words

Exclude bibliography On