IJIRSET
International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

e-ISSN: 2319-8753 | p-ISSN: 2347-6710

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 10, Issue 8, August 2021

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 7.569

9940 572 462 | 6381 907 438 | ijirset@gmail.com | www.ijirset.com

International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 7.569|

|| Volume 10, Issue 8, August 2021 ||

| DOI:10.15680/IJIRSET.2021.1008015 |

Web Scrapping and Phishing Detection

Nandana Medhi [1], Kenyuni Keppen [2], Krittika Bhattacharjee [3], Raktim Kakati [4], Purnendu Bikash Acharjee [5]

U.G. Student, Department of Information Technology, Kaziranga University, Jorhat, Assam, India[1]

Associate Professor, Department of Information Technology, Kaziranga University, Jorhat, Assam, India[2]

ABSTRACT: Web scraping is a technique for extracting all of the information from a website. It is important in today's world because, as the internet and the world wide web continue to grow and attract more people, the number of crimes is also that. Sex trafficking, modern slavery, and phishing connections are examples of crimes. The aim of this project is to look at how crimes are still being committed on the internet. The project's goal scheme collects all internal and external links from a webpage, then determines whether or not that webpage contains any phishing links. We're using python and machine learning to accomplish this as python has a huge collection of libraries as well as methods.With the increased usage of mobile devices in recent years, there has been a growing movement to shift nearly all real-world operations to the cyberworld. While this makes our everyday lives easier, it also leads to numerous security breaches due to the Internet's anonymous framework. Many attacks can be avoided with the use of antivirus software and firewall systems. Experienced attackers, on the other hand, attempt to exploit computer users' vulnerabilities by phishing them with fake websites. These pages mimic common banking, social media, e-commerce, and other websites in order to steal confidential data such as user IDs, passwords, bank account, and credit card numbers, among other things. Phishing detection is a difficult issue, and several different solutions, such as a blacklist, rule-based detection, anomaly-based detection, and so on, have been proposed in the market.We proposed a machine learning-based phishing detection program in this paper, which analyses URLs using machine learning algorithms. The experimental results show that the proposed models perform exceptionally well, with a high success rate.

KEYWORDS: Machine learning, phishing link, python libraries.

I. INTRODUCTION

Web scraping is the process of extracting or collecting the data out from the websites. It allows us to acquire any structured data from the websites and convert it into a usable format such as csv file or spreadsheet. We can scrape the data by copying and pasting the website, which we want to scrape. Web scraping is one of the challenging areas of research because web information changes frequently. Moreover, as the internet and world wide continues to expand and amass more users, increase rate of crime occurs. So, this technique is develop to build an understanding of the topics of crime on the internet, and to investigate all the illegal activities undergoing in both surface.

Phishing is the fraudulent attempt to obtain sensitive information or data, such as usernames, passwords, credit card numbers, or other sensitive details, through digital communication by impersonating a trustworthy entity. Phishing frequently directs users to enter personal information on a fake website that appears to be a real or legitimate site. A victim will typically receive a message that appears to have been sent by a known contact or organisation. The message either contains malicious software that targets the user's computer or contains links that direct victims to malicious software websites in order to trick them into disclosing personal and financial information. Phishing is by far the most common attack performed by cyber-criminals, it is popular among attackers, since it is easier to trick someone into clicking a malicious link which seems legitimate than trying to break through a computer's defence systems. So, in this we are using phishing technique to find out whether it is a legit site or phishing site through external links, with the use of logistic regression.

We have used three models :- Logistic Regression, Random Forest, and Deep Learning(KNN). Logistic regression is a linear approach to solving a classification problem, when the dependent variable is categorical. It's a method for analysing a data set with a dependent variable and one or more independent variables in order to predict the outcome of a binary variable, which has only two possibilities. Categorical is the kind of the dependent variable. The goal variable is also known as the dependent variable, and the predictors are the independent variables.

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 7.569|

|| Volume 10, Issue 8, August 2021 ||

| DOI:10.15680/IJIRSET.2021.1008015 |

The Random Forests technique is a supervised learning method that can be used for classification and regression. The algorithm creates a series of decision trees that have been trained on random subsets of features. The mode of the projected classes throughout the decision trees is the output of a random forest model in classification. It outperforms other algorithms in terms of accuracy. It offers a method for guessing missing data that works well and retains accuracy even when a high percentage of the data is missing. On huge datasets, it performs well.
The K-Nearest Neighbors(KNN) algorithm is a simple supervised machine learning technique that may be used to handle classification and regression problems. It's simple to set up and comprehend, but it has the problem of being noticeably slower as the amount of data in use grows

II. LITERATURE REVIEW

"Web Scraping using Python" has been developed by David Methew Thomas, Sandeep Mathur Amity Institute of Information Technology Amity University (AUUP), Sec-125, Noida [1] is created to collect all of the data collected from different sources using the vivid features of web scraping using a script written in Python language, and then analyse it according to the customer's requirements before storing it in the company's database. The web scrappy, which is based on Python, can also assist us in retrieving the desired result, as we analyse the process using unique codes and provide the desired URLs for the iteration to scrape the data from the source URL.
"Web Scraping with Python" has been developed by Ryan Mitchell [2], here it talks about - Part 1 covers web scraping mechanics, including how to use python to request information from a web server, how to handle the server's response, and how to communicate with websites in an automated manner. Part 2 delves further into a number of more specialised methods and applications that can be used in any web scraping scenario. Parse HTML pages that are difficult to understand. Create crawlers using the Scrappy framework and learn how to save the data you scrape. Documents should be read and data extracted. Clean and normalise data that has been poorly formatted. Read and write in natural languages. Forms and logins are crawled into. Crawl via APIs and scrape JavaScript. Use and write programme that converts images to text. Scraping traps and bot blockers should be avoided. Scrapers can be used to test your website.
"Web Scraping with Python" has been developed by Pratiksha Ashiwal, S.R. Tandan, Priyanka Tripathi, Rohit Miri, the proposed project focuses on web page analysis. Created a working model in this project [3]. Web links are transformed into visual blocks using this technique. A visual block is a section of a webpage. The framework uses an automated top-down, tag tree-independent approach to determine the layout of web content. Using beautifulsoup's python script, the block-based page content structure is developed.
"Detecting Phishing using Machine Learning" has been developed by Rishikesh Mahajan MTECH Information Technology K.J. Somaiya College of Engineering,Mumbai, Irfan Siddavatam Prof., Dept. Information Technology K.J. Somaiya College of Engineering,Mumbai [4], using machine learning technology, this paper aims to improve the detection process for phishing websites. Using the random forest algorithm, we were able to achieve 97.14 percent detection accuracy with the lowest false positive rate. In addition, the results show that classifires perform better when more data is used as training data. In the future, hybrid technology will be used to more effectively identify phishing websites, with the random forest algorithm of machine learning technology and the blacklist approach being used.
"Detecting Phishing Websites using Machine Learning" has by developed by Amani Alswailem, Bashayr Alabdullah, Norah Alrumayh, Aram Alsedrani, 2nd International Conference on Computer Applications & Information Security (ICCAIS), 1-6, 2019, Phishing Websites are one of the types of internet security issues that exploit human weaknesses rather than software flaws. It's the process of recruiting web users in order to access personal information including usernames and passwords. It present an intelligent method for detecting phishing websites in this paper. The device works as an extension of an internet browser, notifying the user when it detects a phishing website. The system is based on supervised learning, which is a form of machine learning. [5]
Logistic regression analysis predicts the outcome in a binary variable which has only two possible outcomes. Where the dependent variable is categorical, logistic regression is a linear approach to solving a classification problem. It's a technique for analysing a data set that includes a dependent variable and one or more independent variables in order to predict the outcome of a binary variable with only two options. The type of dependent variable is categorical. The goal variable is also known as the dependent variable, and the predictors are the independent variables[6].
The random forest algorithm can be used to solve both classification and regression problems[7]. It's a supervised classification algorithm that generates a forest with many trees. The Random Forests technique is a supervised learning method that can be used for classification and regression. The algorithm creates a series of decision trees that have been trained on random subsets of features. The mode of the projected classes throughout the decision trees is the output of a

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 7.569|

|| Volume 10, Issue 8, August 2021 ||

| DOI:10.15680/IJIRSET.2021.1008015 |

random forest model in classification. It outperforms other algorithms in terms of accuracy. It offers a method for guessing missing data that works well and retains accuracy even when a high percentage of the data is missing. On huge datasets, it performs well. The more trees in a forest, the more robust the forest seems to be. It could also be argued that the more trees in the area, the higher the precision of the results. Random forest algorithms have many advantages[8].

K-nearest Neighbors is a straightforward algorithm that stores all available cases and categorises new ones using a similarity metric[9]. It is a simple supervised machine learning technique that may be used to handle classification and regression problems. It's simple to set up and comprehend, but it has the problem of being noticeably slower as the amount of data in use grows. It has been used in pattern recognition and statistical estimation. It predicts a new instance (x) by searching the entire training set for the k most related instances and summing the output variable for those k instances. This may be the mean output variable in regression, or the mode class in classification. To decide which of the k instances in the training dataset are most similar to the new input, distance measures such as Euclidean distance, Manhattan distance, and Minkowski distance are used.

III. METHODOLOGY

First, we used Python's BeautifulSoup and request modules for our project. The request module is responsible for sending HTTP requests to the webpage. There are two techniques for making HTTP requests: get() and post(). The get() method is used to retrieve data from a website, whereas the post() function is used to send data to the server for the purpose of producing or updating data. The requests library starts by sending a get() request to the provided url. After that, the server will send back a response with the requested material. To make a get request, we must first import the requests library using the keyword "import." After that, we'll make a get request. Following the execution of the request, we will receive a response. The status code element of the response indicates whether or not the code was successfully downloaded.

Following requests, we utilised the BeautifulSoup library to extract data from the webpage, including all external and internal connections. It uses a parser to parse the data and make it easier to grasp. The term "parse data" refers to the process of giving data a tree-like structure, often known as a "parse tree." If no parser is supplied, the html parser is used by default. Using BeautifulSoup we will extract all the data from the webpage.

A parse tree, also known as a parsing tree, is an ordered, rotatable tree that describes a string's syntactic structure according to a context-free grammar. We don't have to re-parse the data if we make many runs through it. On the tree, we can do transformations. Things can be ranked in any order we wish.

BeautifulSoup extracts the required data, such as all external and internal links, from the requested material using the find all() method after getting it from the requests library. We may extract all occurrences of a certain tag from a page using find all(). It returns a result set object that can be printed using a for loop and provides index-based access to the result of found occurrences. It accepts a list of tags as well as parameters such as ‘id' for finding tags with a unique id and ‘href' for processing tags with ‘href' characteristics.

Up to this we have scrapped the webpage and we have got our required content. After receiving the external and internal links, we put that links in a text file. Our next objective for this project is to find out if there is any phishing links. Phishing is a fraudulent attempt to steal personal data or information, such as usernames, passwords, credit card numbers, or other sensitive information, by impersonating a trustworthy entity in a digital conversation. Phishing is when a user is directed to enter personal information on a false website that appears to be real or legitimate. Typically a victim receives a message that appears to have been sent by a known contact or organization. The mail contains harmful software or links that take users to harmful software websites in order to deceive them into disclosing personal and financial information. Phishing is by far the most prevalent cyber-attack; it is popular among attackers because it is far easier to mislead someone into clicking a malicious link that appears to be legitimate than it is to break past a computer's protection measures. So, utilising phishing techniques and logistic regression, we can determine if a site is legitimate or a phishing site based on external links. The supervised classification algorithm logistic regression uses only discrete values as input. The model is regression-based and predicts if a given data has a chance of 1 or 0. These values can be referred to as any of the categories used to classify data.

Next, we have developed the programme for detection of phishing link. For that we have used python’s library numpy and other features for detecting the phishing url. In phishing URL detection, we import numpy & and import the features we are using. We must now import train and test split from skeleton model selection (It is a Python library that

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 7.569|

|| Volume 10, Issue 8, August 2021 ||

| DOI:10.15680/IJIRSET.2021.1008015 |

offers various features for data processing that can be used for classification, clustering, and model selection. Model_selection is a method for setting a blueprint to analyze data and then using it to measure new data.) to estimate the performance of machine learning algorithms when they are used to make predictions. Now we will import the Logistic Regression algorithm as lr from sklearn.linear_model. It is a linear regression fits a linear model with coefficients w = (w1… wp) to minimize the residual sum of squares between the observed targets in the dataset, and the targets predicted by the linear approximation. Parameters fit_interceptbool, default=True. Then we need to import dataset and separate features and labels. In the next step, we separate the training, testing features and labels. And import the Beatifulsoup library as well as a Selenium web driver. And code to find all external links and internal links and check one by one through domain and sub domains. This code checks all the features like having IP address, URL length, shortening service, having @ symbol, double slash redirection etc which present a malicious URL. If any features are missing from the given URL, it will return -1, indicating that the URL is a phishing URL and will be saved in the Result.txt file, otherwise it will return "1," indicating that the URL is legitimate.

For example:

The URL: http://www/fatebook.com

Our code applied all 30 features on it and declared the final result. After the results are saved in file format, we need to plot for graphical representation. The plotting library used in this program is matplotlib.pyplot.

Method to check the accuracy between algorithms:

We are using google colab for better performance.

First we import pandas library then load the dataset and check the drop index column because its not needed.

Then display if any columns has null value or not.

Then we need to import seaborn to distribution of phishing and non-phishing data.

Then display the data in each columns

Then we import matplotlib.pypot for display how each columns has some relation with each other through heatmap

Then we display the correlation between the result column with other columns

from sklearn.neighbors import KNeighborsClassifier => to import the Kneigbour model

from sklearn.model_selection we import train test split class , for spliting the total data in train and test

Then we create a variable X where train data only is stored and label Result is droped

Y=data['Result'] is variable to store only the labels

Then we split the data in train part 90 % and test part 10 %

From sklearn.metrics we import accuracy score, precision, recall , fscore support

Def calculate_results(y_true, y_pred): is basically use for Calculates model accuracy, precision, recall and f1 score of a binary classification model.

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 7.569|

|| Volume 10, Issue 8, August 2021 ||

| DOI:10.15680/IJIRSET.2021.1008015 |

y_true = true labels in the form of a 1D array

y_pred = predicted labels in the form of a 1D array

Returns a dictionary of accuracy, precision, recall, f1-score.

Then Calculate the model accuracy, precision, recall and f1 score using "weighted" average.

from sklearn.linear_model we import LogisticRegression to check the this Machine learning algorithms accuracy.

For RandomForest algorithm we import RandomForestClassifier from sklearn.ensemble to predict the accuracy.

In Deep learning , we use Dense layers, Functional API and tensor flow framework for calculating the accuracy.

In the end all models results convert in two-dimensional size-mutable, potentially heterogeneous tabular data structure with labeled axes (rows and columns) using pandas and create a plot.

Applications

- It is used to collect the data out from any websites.
- It is used to detect if there is any phishing links or not.
- It is used to check whether there is any illegal links on the website.

IV. RESULTS

1.IN WEB SCRAPPING PART WE HAVE FOUND ALL THE INTERNAL AND EXTERNAL LINKS AS FOLLOWS:

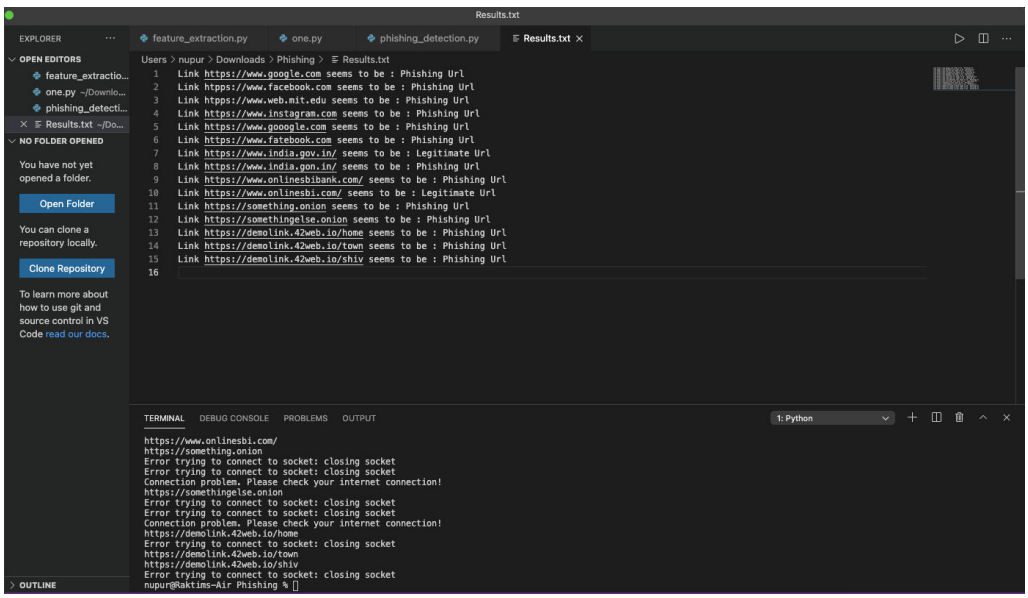

FIG: OUTPUT OF WEB SCRAPPING

2.To detect the phishing url, we have used logistic regression machine learning algorithm. It has 30 features to check whether the link is legit or not. The program will show us all the possible phishing or suspicious links based on the features. All the features are mentioned above in methodology.

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 7.569|

|| Volume 10, Issue 8, August 2021 ||

| DOI:10.15680/IJIRSET.2021.1008015 |

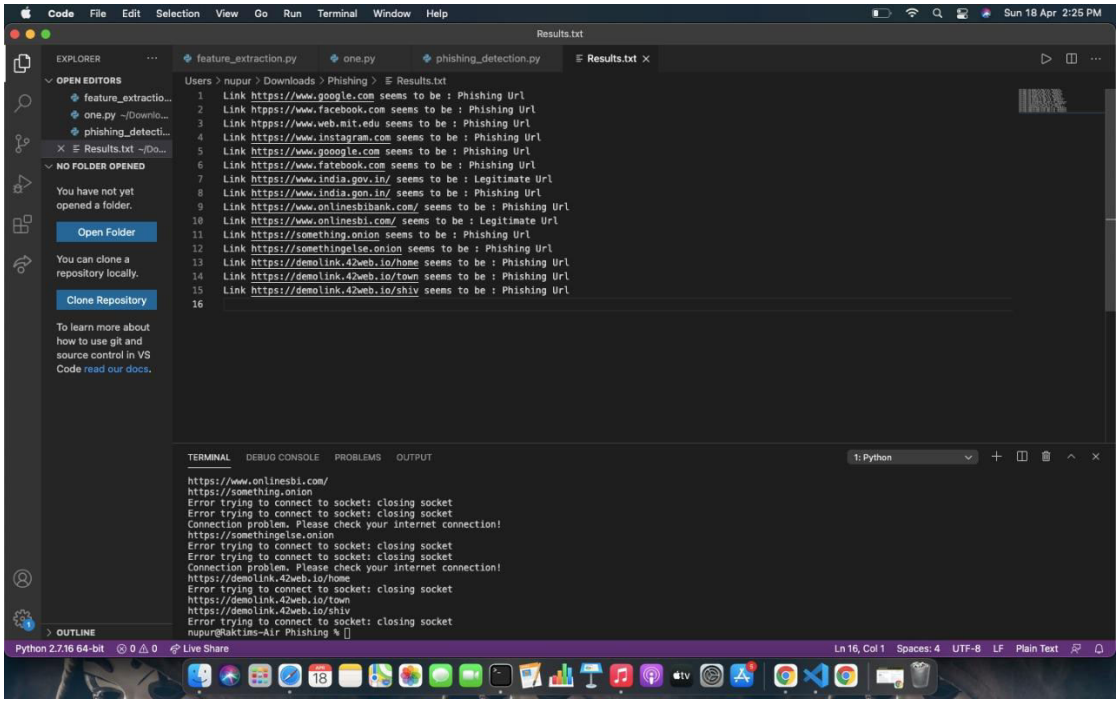

FIG: DETECTING PHISHING URL

3. ACCURACY OF LOGISTICREGRESSION ALGORITHM

FIG: LOGISTIC REGRESSION

4.ACCURACY OF RANDOMFORESTCLASSIFICATION ALGORITHM

FIG: RANDOM FOREST

International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 7.569|

|| Volume 10, Issue 8, August 2021 ||

| DOI:10.15680/IJIRSET.2021.1008015 |

5. ACCURACY OF KNN ALGORITHM

FIG: DEEP LEARNING

6. RESULTS OF THE PROGRAM TO CHECK ACCURACY

	accuracy	precision	recall	f1
Logistic Regression	92.463069	0.925466	0.924631	0.924801
Random Forest	94.995478	0.950599	0.949955	0.950058
Deep Learning	93.518239	0.935416	0.935182	0.935246

FIG: RESULTS

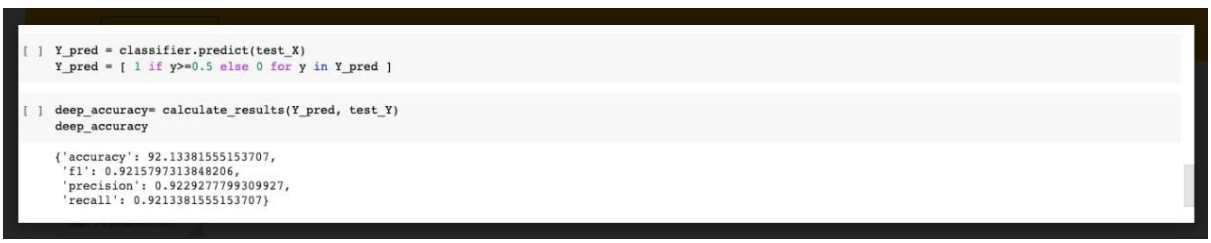

FIG: ACCURACY

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 7.569|

|| Volume 10, Issue 8, August 2021 ||

| DOI:10.15680/IJIRSET.2021.1008015 |

FIG: GRAPH OF RESULT

V. CONCLUSION

A web scraper, even one with legitimate purposes and no intent to bring a website down, can exhibit similar behaviour and, if we are not careful, result in our computer being banned from accessing a website. This is Scrapy's default behaviour, and it should prevent most scraping projects from ever causing problems. Scrapers open up another world of retrieving information without the use of an API, and mostly it is anonymously accessed. Thus, web scraping comes with a number of benefits, but it also faces challenges and it is up to the organisation what they consider best and how they use it for their best interest. Also, phishing is a growing crime and one that we must be aware of. Phishing e-mails are only a small aspect of the overall phishing economy and, until now, the only aspect seen by most people. Phishing is an appalling threat in the web security domain and phishing is becoming an ever-growing threat to users as the attacks evolve and become more difficult to distinguish. And it is a highly profitable activity for cybercriminals. Overall, phishing is real and dangerous. Everyone needs to watch out for it because it happens to everybody and getting scammed can be very costly.

REFERENCES

[1]. David Methew Thomas, Sandeep Mathur Amity Institute of Information Technology Amity University (AUUP), Sec-125, Noida, "Web Scraping using Python".
[2]. Ryan Mitchell, "Web Scraping with Python" "O'Reilly Media, Inc.," 2018
[3]. Pratiksha Ashiwal, S.R. Tandan, Priyanka Tripathi, Rohit Miri, "Web Scraping with Python"
[4]. "Detecting Phishing using Machine Learning", Rishikesh Mahajan MTECH Information Technology K.J. Somaiya College of Engineering,Mumbai, Irfan Siddavatam Prof., Dept. Information Technology K.J. Somaiya College of Engineering,Mumbai
[5]. Amani Alswailem, Bashayr Alabdullah, Norah Alrumayh, Aram Alsedrani, "Detecting Phishing Websites using Machine Learning", 2nd International Conference on Computer Applications & Information Security (ICCAIS), 1-6, 2019,
[6]. https://www.edureka.co/blog/logistic-regression-in-python/
[7]. Jason B, Boinee P."Machine Learning Algorithms"2(3), 138–147. Published on 15, 2016.
[8]. Boser B. E, Guyon I. M.,Vapnik V. N. (1992). "A training algorithm for optimal margin classiers".Proceedings of the 5th Annual Workshop on Computational Learning Theory COLT'92, 152 Pittsburgh, PA, USA. ACM Press, July 1992. On Page(s): 144-152
[9]. Wikipedia.http://en.wikipedia.org/wiki/Support_vector_machine.

INNO SPACE
SJIF Scientific Journal Impact Factor

Impact Factor:
7.569

doi cross ref

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details