IARJSET

ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Vol. 8, Issue 8, August 2021

DOI: 10.17148/IARJSET.2021.8831

Dr. Phish: Phishing Website Detector

Harish Kumar[1] , Anshal Prasad[2] , Ninad Rane[3] , Nilay Tamane[4] , Dr. Anjali Yeole[5]

[1-4]UG-Computer Engineering, Vivekanand Education Society Institute of Technology, Mumbai

[5]Assistant Professor, Vivekanand Education Society Institute of Technology, Mumbai

Abstract : Phishing is an attack on gullible people by making them disclose their personal and unique information. It is a cyber-crime where false sites attract exploited people to give delicate data. This paper describes the various techniques for detecting phishing websites by analyzing different attributes of URLs with the help of ML techniques.This experimentation discusses the techniques used for detecting phishing websites by extracting their features like URL length, port, HTTPS token and many more. We have used data mining techniques for the extraction of the features of an URL in order to get a clear image of URL's structure that spread phishing. To protect the end users from entering these types of phished websites, we can try to predict whether an URL is phished or not. A challenge in this field is that attackers are constantly making new strategies to tackle our defensive methods. To continuously update our system in this domain, we need ML algorithms that adapt to new instances and features of phishing URLs.

Keywords - phishing , anti-phishing , machine learning , cyber-crime , cyber-attack

1 . INTRODUCTION

Phishing attempts to target personal information or data of users that include their usernames, passwords and credit card details or any other details behind the veil of a trustworthy party in this digital world. It was seen that around 76% of businesses (that includes transactions worth billions of dollars) were subjected to phishing attacks in the past year alone. This is because people opt for an antivirus software which doesn't necessarily address phishing. Hence we have decided to propose a project by building a website using our anti-phishing software (Dr.Phish) which works by scanning any nefarious links or possible malware downloads. These programs warn against phishing urls with a high accuracy.

In our proposed project we are using machine learning since multiple datasets with a range of attributes are trained in order to achieve a high accuracy rate. In the year 2004, A Californian teenager replicated some website termed "America Online" and This is where the First phishing lawsuit was filed against. The case where the teenager was able to access user's sensitive information and even access their account credentials, credit card details to withdraw money from their accounts using the same fake website which he created. Other than Email and Website phishing, There are other kinds of phishing, like Vishing i.e. Voice phishing, Smishing i.e. SMS phishing and many different types of phishing detection techniques that the fraudsters come up with. Phishing attackers target vulnerabilities that exist in the system due to the human factor. A cyber-attack by an attacker costs a small business on average $54000. Spear-phishing is a type of phishing which targets data and is aimed specifically at stealing unique information of the users like username, password etc.Henceforth users should be aware about the environment against such intentional loss and here comes Dr.phish tool in the picture. Dr.Phish aims to create a secure environment against phishing attacks so that a warning is given to the user to make him/her alert whenever a malware is detected, protecting the privacy and information of the user. To create a secure environment against phishing attacks so that a warning is given whenever a malware is detected, protecting the privacy and information of the user.The main motive of our website is to warn the user against any tentative phishing attack via a website. The datasets of URLs are trained by ML algorithms. The website shows a warning message if the URL contains malicious content letting safety to the user from phished websites. To implement the logic we have used various ML algorithms like Random Forest, Logistic Regression and Decision Tree.With the help of our website along with continuing the process of notifying safety measures, this process will always be running by which the user is protected from phishing attacks.

2 . LITERARY REVIEW

Many researchers before have studied this subject of detection of malicious and benign URLs. A few of these works used various ML approaches for detection . The efficiency and performance of their system depends mainly on the attribute set, the dataset and the ML algorithm used.

Andronicus et al.[1] has used a random forest classifier for categorisation of spam emails. They have looked to maximize the accuracy and lower the number of features required for its classification. They made a content-based phishing detection model with high accuracy . Here the authors have proposed a model which was based on the features which were extracted appearing in the header and HTML body of URL, which are then categorized using feed forward neural networks. The results show an accuracy of 98.72%.

IARJSET
ISSN (O) 2393-8021, ISSN (P) 2394-1588
International Advanced Research Journal in Science, Engineering and Technology
Vol. 8, Issue 8, August 2021
DOI: 10.17148/IARJSET.2021.8831

Gilchan Park et al.[2] looked to get robust features to differentiate between legitimate and spam emails. The comparison of syntax of the sentence similarity and the difference between subjects and objects of the target verbs between the two types of mails were done.
Further testing from *Tan et al.*[3] were carried out on many classifiers which included Decision Tree and Random Forest. However, it was found that they implemented their model by working with a huge dataset with less number of features which were around 24. Such training often leads to over-fitting.
Ma et al.[4] compared three classifiers namely SVM, Naive Bayes and Logistic Regression on a very good dataset. The Features used were URL of Anchor, Request URL, bag of words (BOW), IP address etc. However, these classifiers in this work could not be deployed. Also the ratio of malicious URLs to benign URLs is 1:3. Usually, there are more benign URLs than malicious URLs.

3. PROPOSED SYSTEM AND ARCHITECTURE

A] System Architecture :

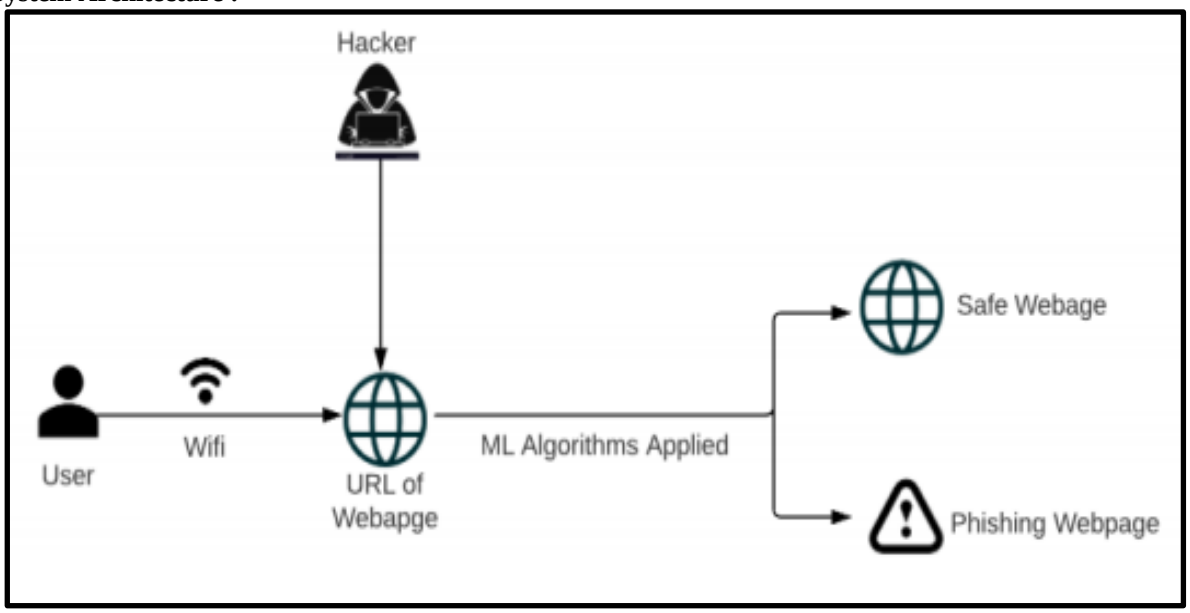

Fig. 1. Basic system architecture

The diagram in Fig. 1 shows the basic system architecture of our proposed system . When a user is connected over a network , the webpage is checked for spam by our website. If the website is identified as spam, a warning message pops up indicating that it is unsafe to use.

B] Dataset :

To evaluate our machine learning techniques, we have used the 'Phishing Websites Datasets' from UCI Machine learning repository.We used 11055 URL entries and then cleaned the data and splitted the dataset into training and testing sets. Out of which are legitimate 4898 URLs and 6157 are phishing URLs. Each instance contains 31 features(Table 1) and 1 target variable to distinguish between legitimate and phishing URLs. Each feature is associated with a rule. If the rule satisfies, it is termed as phishing. If the rule doesn't satisfy then it is termed as benign. The features take three discrete values. '1' if the rule is satisfied, '0' if the rule is partially satisfied and '-1' if the rule is not satisfied.

IARJSET

ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Vol. 8, Issue 8, August 2021

DOI: 10.17148/IARJSET.2021.8831

Sr. No.	Feature	Description
1	id	Number to identify the website
2	having_IP_Address	If an IP address is used instead of the domain name in the URL
3	URL_Length	Phishers can use a long URL to hide the doubtful part in the address bar
4	Shortening_Service	Links to the webpage that has a long URL.
5	having_At_Symbol	Using the @ symbol in the URL leads the browser to ignore everything preceding the @ symbol
6	double_slash_redirecting	The existence of // within the URL which means that the user will be redirected to another website
7	Prefix_Suffix	Phishers tend to add prefixes or suffixes separated by (-) to the domain name
8	having_Sub_Domain	Having subdomain in URL
9	SSLfinal_State	Shows that website use SSL
10	Domain_registration_length	Based on the fact that a phishing website lives for a short period
11	Favicon	If the favicon(icon) is loaded from a domain other than that shown in the address bar , its a phishing URL
12	port	To control intrusions, it is much better to merely open ports that you need
13	HTTPS_token	Having deceiving https token in URL
14	Request_URL	Request URL examines whether the external objects contained within a web page such as images, videos, and sounds are loaded from another domain
15	URL_of_Anchor	An anchor is an element defined by the < a > tag. This feature is treated exactly as a Request URL.
16	Links_in_tags	It is common for legitimate websites to use ¡Meta¿ tags to offer metadata about the HTML document; ¡Script¿ tags to create a client side script; and ¡Link¿ tags to retrieve other web resources.
17	SFH	If the domain name in SFHs(Server Form Handler) is different from the domain name of the webpage
18	Submitting_to_email	A phisher might redirect the users information to his email
19	Abnormal_URL	It is extracted from the WHOIS database. For a legitimate website, identity is typically part of its URL.
20	Redirect	If the redirection is more than four-time , its a phishing URL
21	on_mouseover	Used for hiding link
22	RightClick	It is treated exactly as Using onMouseOver to hide the Link
23	popUpWindow	Showing pop-up windows on the web page
24	Iframe	IFrame is an HTML tag used to display an additional webpage into one that is currently shown
25	age_of_domain	If the age of the domain is less than a month
26	DNSRecord	Having the DNS record
27	web_traffic	This feature measures the popularity of the website by determining the number of visitors
28	Page_Rank	Page rank is a value ranging from 0 to 1. PageRank aims to measure how important a webpage is on the Internet
29	Google_Index	This feature examines whether a website is in Google's index or not
30	Links_pointing_to_page	The number of links pointing to the web page
31	Statistical report	If the IP belongs to top phishing IPs or not

Table 1: Features of URL

C] Classifiers :

This section will tell us in detail about the description of the classifiers that we have used in our project. We used Random Forest, Logistic Regression, Decision Tree for the detection of phishing URLs

● ***Random Forest*** **:** Random Forest[7] is a very popular machine learning formula which is a supervised learning method. This is used for each regression and classification problems in Machine Learning.Random Forest works on the basis of ensemble learning, which is a technique of putting together a number of classifiers to improve the performance of the

IARJSET
ISSN (O) 2393-8021, ISSN (P) 2394-1588
International Advanced Research Journal in Science, Engineering and Technology
Vol. 8, Issue 8, August 2021
DOI: 10.17148/IARJSET.2021.8831

classifiers. Random Forest classifier consists of a wide range of decision trees on a large number of subsets of the input dataset and takes into consideration the typical to improve the accuracy of that dataset.Rather than depending on 1 decision tree, random forest takes count of prediction from every tree and will support the prediction with maximum votes and then predicts the ultimate output.The bigger variety of trees within the forest ends up in more accuracy and prevents the problem of overfitting.

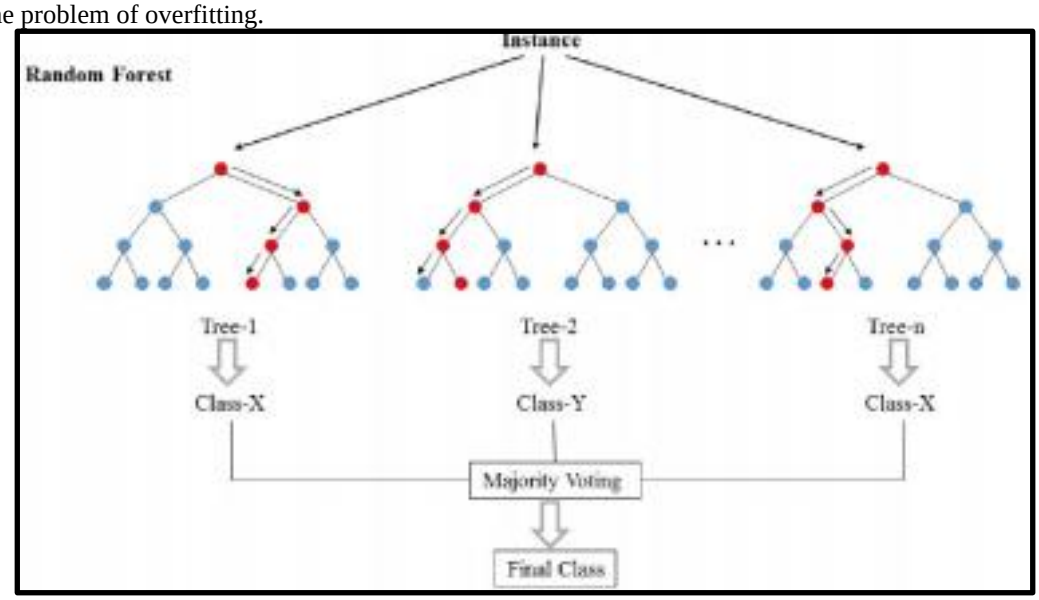

Fig 2. Random Forest Working

Random Forest algorithm works in 2 phases : The 1st phase is to create the random forest by the combination of N decision trees, and 2nd is to form predictions for every tree created within the 1st part. The working of this algorithm can be shown as follows :
1: Select randomly K data points from the set to be trained.
2: The decision trees are built according to these chosen data points
3: Select the amount N for all trees that need to be made.
4: Repeat steps 1 & 2.
5: For the data points which are new,predictions of every decision tree are calculated and the new data points are assigned to the class which has the highest vote.

● ***Logistic Regression*** **:** Logistic Regression[8] is the most suitable regression technique to analyze when the dependent variable is binary or dichotomous. Similar to all other regression analyses, the logistic regression could be a prediction based analysis. Logistic regression explains the relationship between one dependent binary variable with respect to one or additional independent variables where the variables can be nominal, ordinal, interval or ratio-level. Logistic regression is usually known for its core functionality, the sigmoid function also known as logistic function was introduced so that the properties such as increase in ecology, quick rise and maximizing at the carrying capability of the surroundings should be explained. It is a mathematical function that can take any real number and map it to between 0 to 1 in the shape of the letter 'S'.

$$\frac{1}{(1 + e^{-value})}$$

e = base of the logarithms (Euler's number or the EXP() function in the spreadsheet) value = actual numerical value that should be transformed.

● ***Decision Tree*** **:** Decision Tree[9] is a type of Supervised Machine Learning where the input is continuously split according to our requirements. The tree is seen in two entities which are called decision nodes and leaves.
->Leaves- Show the Final decisions and outputs
->Decision nodes - Split the input data.

In the Decision Tree algorithm the biggest challenge is to determine or select the attribute for the root node at every level. Here are 2 measures for the selection of attributes :
1. Information Gain(IG) :When a node is used in a decision tree to separate the input training data into subsets of a smaller size, then it's entropy will change. Information gain is the scale of this difference in the entropy.

IARJSET
ISSN (O) 2393-8021, ISSN (P) 2394-1588
International Advanced Research Journal in Science, Engineering and Technology
Vol. 8, Issue 8, August 2021
DOI: 10.17148/IARJSET.2021.8831

Let S be the group of instances, A be an attribute ,Sv be the subset of subset A=v and Values (A) be a set of all the possible values, then

$$G(S,A) = I(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} I(S_v)$$

2. *Gini Index* : Gini Index measures how many times a randomly chosen element would be identified incorrectly. This means that the attribute which is to be taken into consideration should have a lower gini index.Sklearn encourages "Gini" criteria for Gini Index and takes the "gini" value by default.Gini Index can be calculated as follows

$$\text{Gini} = 1 - \sum_{i=1}^{n} (p_i)^2$$

D] Proposed Approach :

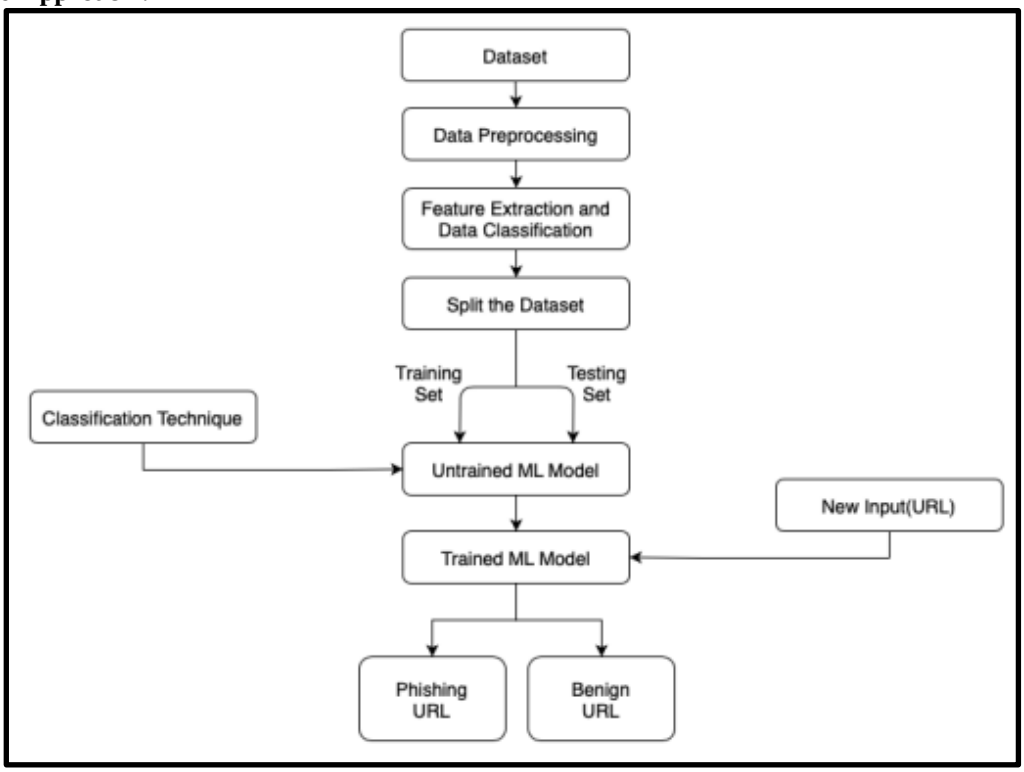

Fig 3. Proposed Approach

The first stage is data gathering and collection at a centralized location after which comes data preprocessing which does the job of removing unnecessary data which makes no sense in running meaningful analysis . Data preprocessing is an important data mining technique that is used to convert the raw data in a useful format. There are three steps in data preprocessing as follows : data cleaning, data transformation and data reduction. After this the features of URLs are extracted and then they are classified as either -1, 0, 1.

Next the dataset is splitted into subsets of two that is a training set and a testing set which are fed to an untrained ML Model which uses any one of the classifiers such as Random Forest , Decision Tree , Logistic Regression to detect phishing URLs. Now a new data input(URL) is fed to the trained model which further predicts whether it is a phishing URL or a benign URL.

4. EXPERIMENTAL RESULTS

This section demonstrates the experimental studies to investigate the predictive accuracy of various Machine Learning Classification Algorithms on the same dataset and also compares it to the existing Machine Learning Techniques.

- The dataset used comprises 11055 URLs out of which 6157 are malicious and 4898 are legitimate websites.
- Each instance had 31 features which were extracted and then fed to the untrained ML classifiers.
- The training set and the testing set consists of 8844 URLs(80%) and 2211(20%) URLs respectively.
- The ML Classification techniques that are used to detect phishing URLs :
  - Random Forest(RF)
  - Logistic Regression(LR)
  - Decision Tree(DT)

IARJSET

ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Vol. 8, Issue 8, August 2021

DOI: 10.17148/IARJSET.2021.8831

The following Table compares the obtained predictive accuracy of our model with the predictive accuracy of the existing models on an average :

Classification Technique	Obtained Accuracy(%)	Existing Technique Accuracy(%)
Random Forest(RF)	95.10	95.50
Logistic Regression(LR)	92.25	94.10
Decision Tree(DT)	89.23	93.90

Table 2 : Accuracy Comparison

From the above results, it is pretty clear to us that Random Forest gives a higher accuracy in terms of classification as compared to other algorithms.
The following Graph compares the predictive accuracies of the above mentioned three ML classifiers :

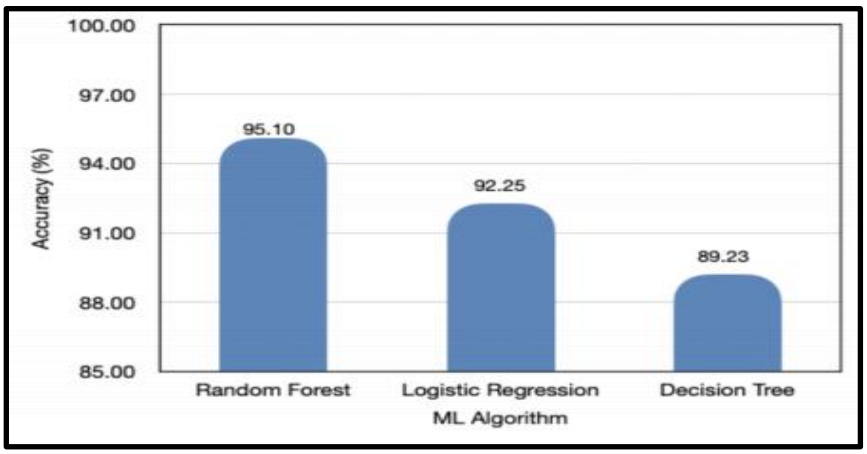

Fig 4. Graph for Accuracy Comparison

5. CONCLUSION

This paper aims to enhance detection methods to detect phishing websites making use of various machine learning algorithms. The attempts made by different researchers to solve this problem through the use of machine learning classifiers was discussed. The features of phishing URLs were extracted using a feature extraction python program. The extracted features were fed into trained models which used algorithms like Random Forest, Logistic regression, and Decision tree to detect phishing URLs. The results of the above classification was uplifting with the highest accuracy achieved was of 95.10% using random forest algorithms for phishing URLs. Also results show that classifiers give better performance when we use more data as training data. Our work has produced motivating results, however , in future this proposed system can be improved by increasing the size of the dataset and creating a browser extension. By including a variety of URLs of both types that are phished and legitimate, our website would be closer to a more accurate model where hackers are upgrading their techniques day by day. Using a larger and a diverse dataset will help us to be ahead of them and protect private information against these criminals.

6. REFERENCES

1. Andronicus A. Akinyelu Aderemi O. Adewumi. Classification of Phishing Email using Random forest Machine Learning Technique 2014.
2. Gilchan Park, Julia M. Taylor, Using Syntactic Features for Phishing Detection 2015, https://arxiv.org/ftp/arxiv/papers/1506/1506.00037.pdf
3. G. Tan, Q. Liu, X. Liu, C. Zhu, and L.Guo,"Malfilter: A lightweight real-time malicious url filtering system in large-scale networks"2018 IEEE ISPA/IUCC/BDCloud /SocialCom/SustainCom.IEEE, 2018,pp. 565-571.
4. J. Ma, L. K. Saul,S. Savage "Beyond blacklists : learning to detect malicious web sites from suspicious urls" in Proceedings of the 15th ACM SIGKDD in international conference on Knowledge discovery and data mining . ACM , 2009 , pp. 1245-1254 .
5. R. Verma "What's in a url: Fast feature extraction and malicious url detection," in Proceedings of the 3rd ACM on International Workshop on Security and Privacy Analytics. ACM, 2017, pp. 55–63.
6. O. K. Sahingoz, E. Buber, O. Demir "Machine learning based phishing detection from urls," Expert Systems with Applications, vol. 117, pp. 345–357, 2019.

IARJSET

ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Vol. 8, Issue 8, August 2021

DOI: 10.17148/IARJSET.2021.8831

7. Pranesh S; Jenita J R; Nisha D; Prabakar D. "Sentimental Data Analysis-To Predict the User Emotions". International Research Journal on Advanced Science Hub, 3, Special Issue ICARD-2021 3S, 2021, 26-29. doi: 10.47392/irjash.2021.056
8. Rani RN; Kumaraswamy HV; Jeyaraj Chellapandi. "Implementation of End to End Automation for BTS commissioning using Python". International Research Journal on Advanced Science Hub, 3, Special Issue 7S, 2021, 37-41. doi: 10.47392/irjash.2021.12118
9. https://www.javatpoint.com/machine-learning-random-forest-algorithm
10. https://machinelearningmastery.com/logistic-regression-for-machine-learning/
11. https://www.xoriant.com/blog/product-engineering/decision-trees-machine-learning-algorithm.html
12. P. Zhang - "Malfilter: A lightweight real-time malicious url filtering system in large-scale networks"2018 IEEE ISPA/IUCC/BDCloud /SocialCom/SustainCom.IEEE, 2018,pp. 565-571
13. Das, "What's in a url: Fast feature extraction and malicious url detection," in Proceedings of the 3rd ACM on International Workshop on Security and Privacy Analytics. ACM, 2017, pp. 55–63.
14. G. M. Voelker, "Beyond blacklists : learning to detect malicious web sites from suspicious urls" in Proceedings of the 15th ACM SIGKDD in international conference on Knowledge discovery and data mining . ACM , 2009 , pp. 1245-1254 .
15. B. Diri, "Machine learning based phishing detection from urls," Expert Systems with Applications, vol. 117, pp. 345–357, 2019.
16. https://williamkoehrsen.medium.com/random-forest-simple-explanation-377895a60d2d
17. https://www.geeksforgeeks.org/decision-tree/

