Journal of Advanced Zoology

ISSN: 0253-7214
*Volume **44** Issue **S-6 Year 2023** Page **930:937***

Analytics of Phishing Attacks Using Machine Learning

Lingampally Shalini[1*], Sunil Kumar S. Manvi[2]

[1]Student, Department of Computing and IT, Reva University, Bangalore-India, [1]shaludd78@gmail.com,
[2]Director, Department of Computing and IT, Reva University, Bangalore-India. [2]ssmanvi@reva.edu.in.

**Corresponding author's E-mail: shaludd78@gmail.com*

Article History

Received: 06 June 2023
Revised: 05 Sept 2023
Accepted: 30 Nov 2023

Abstract

Cyber-attacks which is known as Computer Network Attack (CAN). It is a threat created by cybercriminals by more than one computer against networks. There are various types of cyberattacks among them Phishing Attack is one of them. Phishing is a technique of gathering sensitive information of a target such as username, password, bank details etc. There are various ways to perform phishing attacks, among them URL PHISHING attack is one way to gather user's information. Mainly hackers create a fake website regarding bank details, shopping websites, etc. using social engineering tool which looks like legitimate website. This fake URL will be sent to user via email or through some other resources. In this paper, we will be performing data analysis, data pre-processing, data exploring, training and predicting through machine learning and optimization techniques on dataset which contains two attributes (URL, Label). We will be introducing performance metrics like confusion matrix accuracy, precision, recall and F1 score to know the performance of the model. An optimization technique which is Stochastic gradient descent performed better than logistic regression.

Keywords: *Word Cloud, Tokenizing, Lemmatization, Logistic Regression, Stochastic Gradient Descent, etc.*

1. Introduction

Phishing attack is one of the cyber-attacks. There are various types of phishing attacks like email spoofing, mass phishing, URL phishing, Search engine attack, website spoofing. Phishing is one way of technique to gather the information of user like bank details, passwords, username, email ids etc. At first hacker creates a fake website which looks like a legitimate website using social engineering tools. Sending that fake website URL to the user via email or through other resources. When user clicks on that on particular fake website URL. User will be taken to the fake website and when he/she enters their personal credentials those credentials will be displayed on the terminal screen of the hacker who created fake web site. By this displayed information he/she will get access to the personal details in the absence of the user. These types of phishing attacks are the attacks which performed through URLs to gather personal credentials of the user.

At First, we need to understand or get insights of URL. URL is known as Uniform Resource Locator. URL is made up of various parts like scheme, host address and file path. Here scheme determines the type of the protocol like http, https, ftp etc. Host address which is domain name like IP address or www. File path always starts with forward slash it may consists of directory or folder names. These are the various parts of URL. Legitimate and Phishing URL contains these parts. Legitimate URL is the trustworthy URL in which it starts with scheme like http or https followed by colon and forward slash. Domain name like IP address or www. File path includes directory or name folders and limited number of forward slash and limited number of hyphens, limited length of URL and there are no special characters in the legitimate website. But when we talk about phishing websites URL. It mainly contains scheme, domain name, file path with high number of slashes, more of hyphens and it contains a greater number of special characters.

This paper mainly tells us about how to know whether the URL is phishing or a Legitimate URL and how to get insights of the URL. Objectives of this problem statement are to know all insights of URL by using word cloud for visualization purpose and NLTK tool kit for tokenization, lemmatization and count vectorizer. By implementing machine learning algorithms like Logistic regression and

optimization techniques like stochastic gradient descent, detecting the user input URL whether the URL is benign or malicious. As identified from research papers most of the datasets contain already extracted features. By this model, we can bring awareness.

2. Materials And Methods

I. ADOPTED STUDY

Shahzad Ali, Muhammad Shahbaz, et al. [1], in this paper they described about feature selection methods. They have taken dataset from UCI ML learning repository. They classified phishing websites by using WFS (wrapper feature selection), CFS (Correlation Feature selection), EFS (Entropy feature selection). They have implemented a different machine learning technique like KNN (K Nearest neighbor), DT (Decision tree), RF (Random forest). They have worked on already extracted features. Mahajan Mayuri Vilas, et al. [2], collected the dataset with contains 30 features. They used ELM (extreme learning machine) classification algorithm and machine learning classifiers as ANN, SVM. They worked on already extracted features from URL.

Mohammad Nazmul Alam et al. [3], in this paper they collected the dataset which contains 32 features. They used principal component analysis (PCA). They implemented machine learning algorithms like random forest and decision tree. Abdulhamit Subasi, et al. [4], taken datasets from UCI machine learning repository in which they developed a machine learning model to determine correlations between features. They adopted classifier model that is used for detecting phishing websites of available dataset. Arun Kulkarni and Leonard L. Brown et al. [5], implemented using MATLAB scripts, they taken dataset from University of California, Irvine Machine Learning Repository which contain 9 features. Here they implemented Naïve bayes classifier and other classifiers.

Ishant Tyagi, et al. [6], selected the dataset which contains 30 attributes and implemented through R programming, they used VIF values and then after applied PCA on it so they implemented GLM and Decision Tree on it. Joby James, et al. [7], used detection of phishing websites based on lexical features and host properties. Mainly, there dataset contains 17000 phishing URLs and 20000 benign URLs. They implemented in using matlab where feature set compares host length, path length, number of slashes, number of path tokens. They had implemented using machine learning algorithms such as naïve bayes, SVM, KNN. From all these above research papers, we can observe that the authors used machine learning algorithms like SVM, decision tree, random forest on extracted features of URL and predicted on them. Mostly they concentrated and extracted all special characters of the URL not on words of URL. They used MATLAB programming. So, their motive is to build neural networks like ANN, CNN on already. extracted features to increase the model accuracy.

A. Machine Learning Techniques.

Machine learning techniques are used for training the model and prediction on unknown data As shown in Fig.1, there are different types of machine learning (ML) techniques.

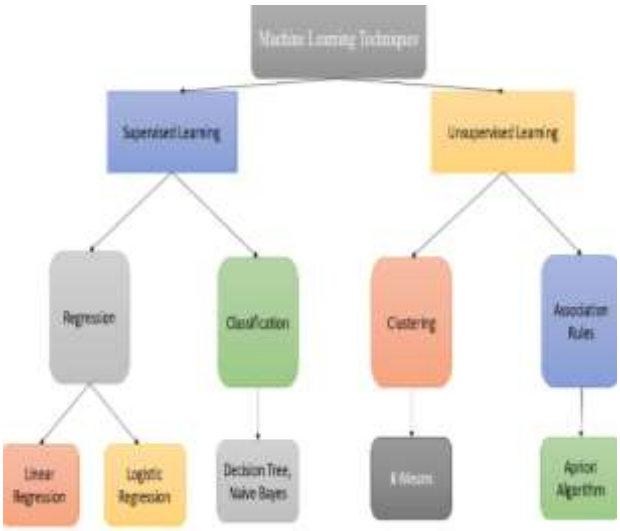

Fig. 1. Various ML Techniques

a) Logistic Regression:

This is another type of machine learning technique which is used when target variable is categorical variable.

In logistic regression sigmoid function is used. Here there is a threshold between 0 and 1. The values above threshold consider into one class and all values below threshold consider as another class. As shown in Fig.2, sigmoid function maps all independent variables to 0 or 1. It forms S-shape.

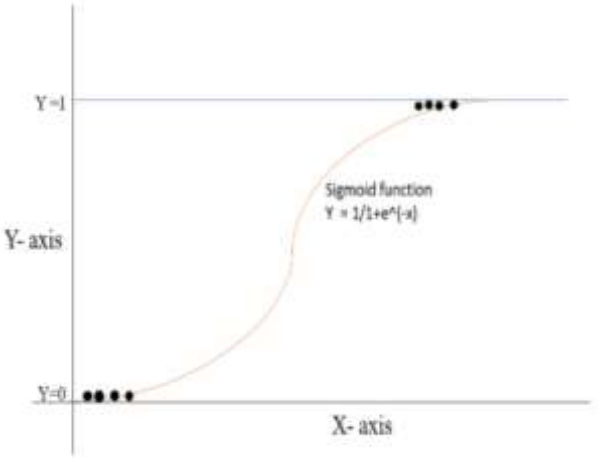

Fig. 2. Logistic Regression

B. Optimization Techniques.

Optimization techniques are mainly used to reduce the error and increase the accuracy of a model As shown in Fig.3. there are different types of optimization techniques.

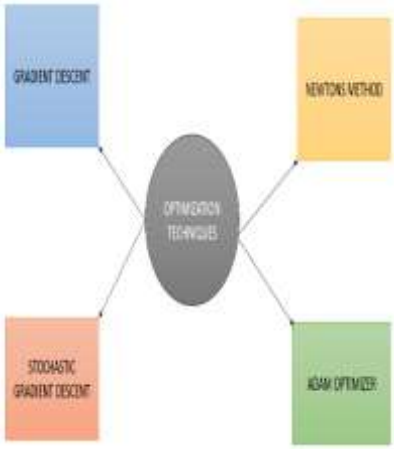

Fig. 3. Various Optimization Techniques

a) Stochastic Gradient Descent:

To calculate the derivatives. Stochastic gradient descent randomly picks a data point from entire dataset. For each and every iteration is performed to reduce the computations.

Hyper parameters: eta (step size)

Pseudocode:

Start

Consider parameters 'w' which is initial vector and step size 'eta'

Repeat till we reach minimum value:

Next shuffling data points in the data set.

for i=1, 2….,n do:

w: = w - etaQ(w) //Q(w) = loss function/old weight

stop.

C. NLPTechniques:

NLP (Natural Language Processing) which is of analyzing and understanding different languages. Tokenization, Lemmatization and Count vectorization.

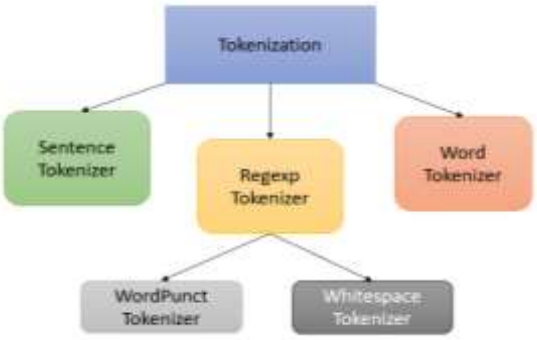

Fig. 4. Tokenization

a) Tokenization:

A process of diving a paragraph into sentences using sentence tokenizer A sentence into words using words tokenizer.

As shown in Fig.4. regex expression is used to extract what we want specifically.

b) Lemmatization:

It is the process of converting all form of words to their base forms. Converted base form is known as "lemma". It gives meaningful words.

c) Count Vectorization:

From text unique words are taken. It creates a sparse matrix. Converts categorical form into numerical form.

D. Proposed work:

In this work, we mainly did the analysis of all the Legitimate URL and Phishing URL. Considering only text from URL not special characters, numbers. Tokenizing the text from URL and performing lemmatization on those words by converting words to their base forms. Performing count vectorization on that text. Training and predicting a model with machine learning techniques one of them is logistic regression. Also applying Stochastic Gradient Descent optimization techniques. As shown in Fig.5, it gives information about how data is analyzed, processed and trained.

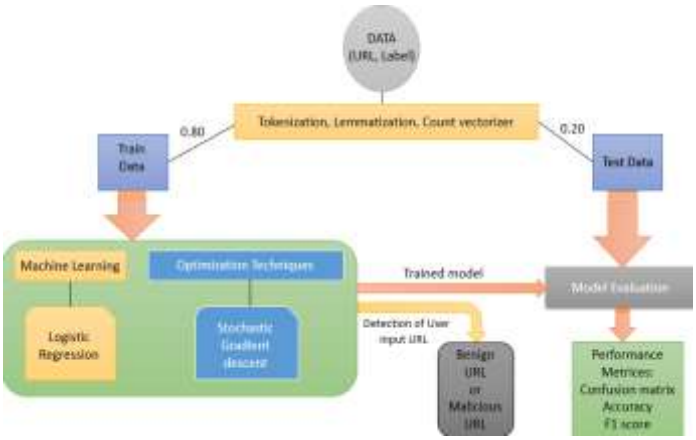

Fig. 5. Phishing Detection System

a) Dataset Description:

Dataset mainly contains 2 attributes with 450176 rows. LABEL is a target variable it is a categorical variable which is malicious and benign. It is the attribute which determines the nature of the URL. Tells us whether the URL is benign or malicious. It is also called as dependent variable. Another attribute which tells us about the URL. URL is made up of various parts like scheme, host address and file path. Here scheme determines the type of the protocol like http, https, ftp etc. Host address which domain name like IP address or www. File path always starts with forward slash it may consists of directory or folder names.

b) Data Analyzing, Pre-processing and Data Exploring:

First understanding and analyzing the data through certain packages in python like matplotlib and seaborn. Through seaborn package we can know the count of label attribute and through word cloud we can visualize the frequency of the words in phishing and legitimate URLs [12-19].

1) Count of Target variable (Label): By using seaborn in python, we can know the count of benign and malicious URL. As shown in Fig.6, indicates the count of benign and malicious URL.

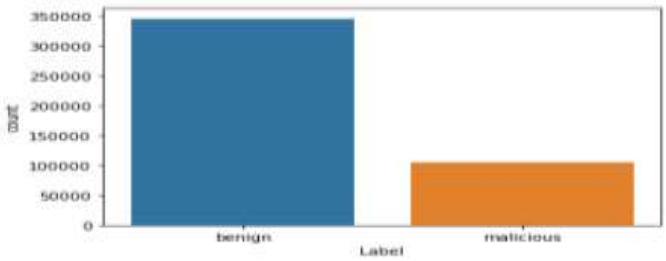

Fig. 6. Count of Label

2) High Frequency words in good sites using word cloud: As URL contains bad sites as well as good sites. So, splitting those sites into storing separately in different variables. Fig.7. gives information about high frequency words in good sites. The word with high bold determines the most occurring word in the URLs. By using word cloud, we can know the words which mostly occurred in the URLs.

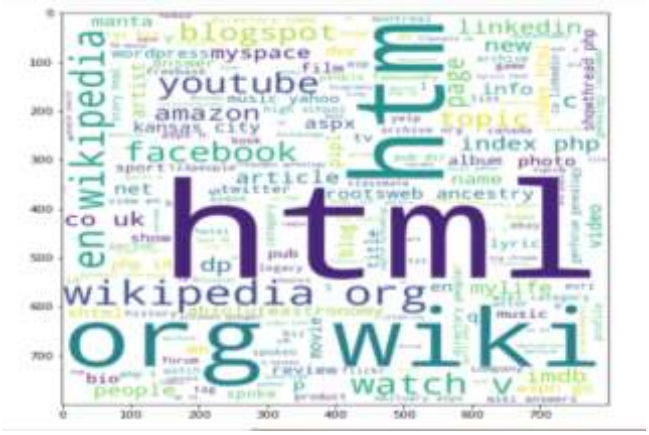

Fig. 7. High Frequency Words in Good Sites.

3) High Frequency words in bad sites using word cloud: As URL contains bad sites as well as good sites. So, splitting those sites into storing separately in different variables. Fig.8. gives information about high frequency words in bad sites. The word with high bold determines the most occurring word in the URLs. By using word cloud, we can know the words which mostly occurred in the URLs.

Fig. 8. High Frequency Words in Bad Sites.

Now performing tokenization by using regex we can select only text from the URL. Undergoing **Tokenization** on that text and performing **Lemmatization** on particular words so that we can convert all the form of words into base form of the word which is meaningful. So, joining those words into text

and applying **count vectorization** on that particular text. That creates a space matrix consists of documents and number of unique features. Now we converted the categorial to numerical form. Machine readable form. By mapping Label column attributes (Benign and Malicious into numeric 0 and 1). So, on that we can train our model and predict.

E. Results:

Python IDE Jupyter is open-source software platform is used for the implementation purpose. Dataset which contains 450176 of two attributes (URL, LABEL). Data is divided into train dataset and test dataset in the ratio of 80%:20%. Performing tokenization, lemmatization and count vectorization on text of the URL. Mapped the target variable (Label) into 0 and 1. Data is now modelled using machine learning algorithms like logistic regression. It gives the accuracy of training data with 99.1% and testing data with accuracy of 98.8%, as shown in Fig.9. now same data is now trained and modelled through optimization techniques using Stochastic Gradient Descent, it gives accuracy of training data with 98.4% and testing data with 98.2%. From these models Stochastic Gradient Descent (SGD) performs. As we can from table Fig.9. training and testing results of SGD is quite better result than Logistic regression. Because how much percent it has trained that much it has tried to predict.

Metrices	Logistic Regression		Stochastic Gradient Descent	
	Training Data results	Test Data results	Training Data results	Test Data results
Performance of model	99.19	98.82	98.4	98.2
Accuracy	0.99		0.99	
Label column (2 classes)	Benign (0)	Malicious (1)	Benign (0)	Malicious (1)
Precision	0.99	0.99	0.99	0.99
Recall	1.00	0.96	1.00	0.96
F1 score	0.99	0.97	0.99	0.97
AUC	0.98		0.98	

Fig. 9. Tabular Form of Results

a) Confusion Matrix: It is a table which tells us about the actual data and how our model is predicted on that data. In Fig.10.represents the actual no, actual yes and predicted no and predicted yes. It contains terms such as True Positive, True Negative, False Positive, False Negative.

```
[[69076   199]
 [  867 19894]]
```

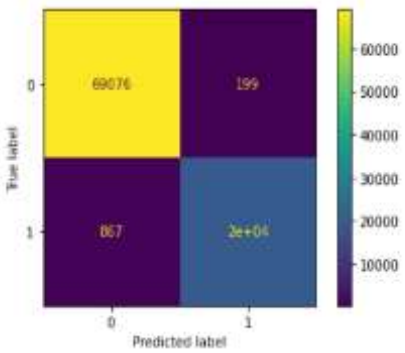

Fig. 10. Stochastic Gradient Descent Confusion Matrix.

b) Performance Evaluation Table: It determines the accuracy, recall, precision, F1- score of the model. As shown in Fig.11. It gives the clear picture of the model performance.

c) Accuracy: It is the ratio of True positive rate and True negative rate to the total. It is represented as (TP+TN)/Total.

d) Precision**:** It determines how our model is positively classified among predicted model. It is the ratio of False Positive rate and True Positive rate to Predicted Yes. It is actually represented as (FP+TP)/Predicted Yes.

e) Recall**:** It determines how our model is positively classified among actual model. It is the ration of False Negative rate and True Positive rate to the Actual Yes. It is represented (FN + TP)/Actual Yes. Which is also known as "Sensitivity".

f) F1-score**:** As recall increases precision increases. It undergoes harmonic mean. It is very important metric which gives overall performance of the model. (2*Precision*Recall)/ (Precision + Recall).

	precision	recall	f1-score	support
0	0.99	1.00	0.99	68948
1	0.99	0.96	0.97	21088
accuracy			0.99	90036
macro avg	0.99	0.98	0.98	90036
weighted avg	0.99	0.99	0.99	90036

Fig. 11. SGD Performance Evaluation Table.

g) ROC curve: It is receiver operating characteristic. It gives the performance of the model. As shown in Fig.12. It creates a graph between False Positive Rate (1-specificity) and True Positive Rate (sensitivity). The line is formed by thresholds. AUC is the area under the curve. When there is more area under the curve. Our model will be the best model since AUC=0.98.

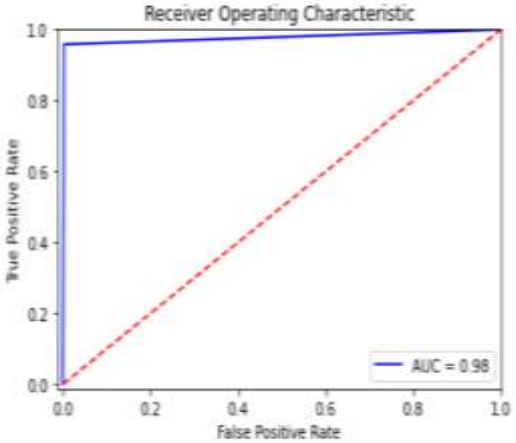

Fig. 12. SGD ROC-AUC Curve

4. Conclusion

Detection of URL plays very important role in real time scenarios. In this paper, we developed a model that actually detects the user input URL. Whether it is Benign URL or Malicious URL. We used python programming language on Jupyter. To understand the insights of data we have performed analyzing, pre-processing and exploring of data by analytical tools and techniques like word cloud, tokenization, lemmatization and count vectorization. By using machine learning techniques i.e., logistic regression and by using optimization techniques i.e., Stochastic Gradient descent. So, 80% of has been trained by these techniques and predicted on 20% of the unknown data that is test data. This model detected the user input URL. From this we can conclude that Stochastic Gradient Descent optimization technique gave appropriate results with high accurate results of 98.2%. In future work, including words all required special characters will be extracted, and another column will be added which is high frequency terms "bad indicator". Will be implementing by using neural networks.

References:

[1] Shahzed Ali, Muhammad Shabaz, Kashif Jamil, "*Entropy-Based Feature Selection Classification Approach for Detecting Phishing Websites*", International Conference on Open-Source Systems and Technologies (ICOSST), 2019 IEEE.

[2] Mahajan Mayuri Vilas, Sawant Purna Jaypralash Kakade Prachi, Ghansham Pawar Shila," *Detection of Phishing Website Using Machine Learning Approach*", International Conference on Electrical, Electronics, Communication, Computer Technologies and Optimization Techniques (ICEECCOT), 2019 IEEE.

[3] A. W. (2018). Impact of IQAC on quality enhancement and sustenance of higher education institutions. International Journal of Research. 5(7):647-651.
[4] Mohammad Nazmul Alam, Dhiman Sarma, Farzana Firoz Lima, Ishita Saha, Rubaiath-E-Ulfath, Sohrab Hossain, "*Phishing Attacks Detection using Machine learning Approach*", International Conference on Smart Systems and Inventive Technology (ICSSIT), 2020 IEEE.
[5] Abdulhamit Subasi, Esraa Molah, Fatin Almkallawi, Touseef J. chaudhery, "*Intelligent Phishing Website Detection using Random Forest Classifier*", International Conference on Electrical and Computing Technologies and Applications (ICECTA), 2017 IEEE.
[6] Arun Kulkarni, Leonard L. Brown," *Phishing Websites Detection using Machine learning*", International Journal of Advanced Computer Science and Applications (IJACSA), 2019.
[7] Ishant Tyagi, Jatin Shad, Shubham Sharma, Siddharth Gaur, Gagandeep Kaur, "*A Novel Machine Learning Approach to Detect Phishing Websites*", International Conference on Signal Processing and Integrated Networks (SPIN), 2018.
[8] Joby James, Sandhya L., Ciza Thomas, "*Detection of Phishing URLs Using Machine Learning Techniques*", International Conference on Control Communication and Computing (ICCC), 2013.
[9] Ebubekir Buber, Onder Demir, Ozgur Koray Sahingoz, "*Feature Selections for Machine Learning based Detection of Phishing Websites*", International Conference on Computing Technologies (ICCT), 2017.
[10] Dr.G. Ravi Kumar, Dr.S.Gunasekaran, Nivetha.R, Sangeetha Prabha.K, Shanthini.G, Vignesh.A.S, "*URL Phishing Data Analysis and Detecting Phishing Attacks using Machine Learning In NLP*", International Journal of Engineering Applied Sciences and Technology (IJEAST), 2019.
[11] Vaibhav Patil, Pritesh Thakkar, Chirang Shah, Tushar Bhat, Prof.S.P. Godse," *Detection and Prevention of Phishing Websites using Machine Learning Approach*", International Conference on Computing Communication Control and Automation (ICCCCA),2018.
[12] Sudhan Murugan Bhagavathi, Anitha Thavasimuthu, Aruna Murugesan, Charlyn Pushpa Latha George Rajendran, A Vijay, Raja Laxmi, Rajendran Thavasimuthu, Weather forecasting and prediction using hybrid C5.0 machine learning algorithm International Journal of Communication Systems, Vol. 34, Issue. 10, Pp. e4805, 2021.
[13] PM Surendra, S Manimurugan, A New Modified Recurrent Extreme Learning with PSO Machine Based on Feature Fusion with CNN Deep Features for Breast Cancer Detection, Journal of Computational Science and Intelligent Technologies, Vol. 1, Issue. 3, Pp. 15-21, 2020.
[14] PK Sadineni, Comparative Study on Query Processing and Indexing Techniques in Big Data, 2020 3rd International Conference on Intelligent Sustainable Systems (ICISS), pp. 933-939, 2020.
[15] AH Omar Baabood, Prajoona Valsalan, Tariq Ahmed Barham Baomar, IoT Based Health Monitoring System, Journal of Critical Reviews , Vol. 7, Issue. 4, pp. 739-743, 2020.
[16] Sajay KR, Suvanam Sasidhar Babu, Vijayalakshmi Yellepeddi, Enhancing The Security Of Cloud Data Using Hybrid Encryption Algorithm, Journal of Ambient Intelligence and Humanized Computing, 2019. https://doi.org/10.1007/s12652-019-01403-1
[17] Bindhia K Francis, Suvanam Sasidhar Babu, Predicting academic performance of students using a hybrid data mining approach, Journal of Medical Systems, 43:162, 2019. https://doi.org/10.1007/s10916-019-1295-4
[18] Vijayalakshmi Y, Manimegalai P, Suvanam Sasidhar Babu, Accurate Approach towards Efficiency of Searching Agents in Digital Libraries using Keywords, Journal of Medical Systems, 43:164, 2019. https://doi.org/10.1007/s10916-019-1294-5
[19] Teena Jose, Dr. Suvanam Sasidhar Babu – Detecting Spammers on Social Network Through Clustering Technique, Journal of Ambient Intelligence and Humanized Computing, pp.1-15, 2019. https://doi.org/10.1007/s12652-019-01541-6