International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Phishing website analysis and detection using Machine Learning

Ameya Chawla[1,*]

Submitted: 12/08/2021 Accepted : 24/01/2022

Abstract: Cybersecurity has become an essential part of this new digital age with more than 820 million users of internet by the year 2022 there is need of security systems to protect public from phishing scams as it not only effects the wealth but also effects the mental health of public, making people afraid to surf or use the internet services which motivates me to work on this problem to develop efficient solution. Objective of this paper is to analyse some common attributes shown by phishing websites and develop a model to detect these websites. Various models were trained on the dataset like Random Forest Classifier, Decision Tree Classifier, Logistic Regression, K Nearest Neighbours, Artificial Neural Networks and Max Vote Classifier of Random Forest, Artificial Neural Networks and K Nearest Neighbours. Highest accuracy was achieved by Max Vote Classifier of Random Forest (max depth 16), Decision Tree (max depth 18) and Artificial Neural Network of 97.73%. This research can be used in real life by implementing a web application in which user can enter the website link and using the link the application will get values for various factor on which model was trained and it will detect whether a website is phishing website or not.

Keywords: *Cybersecurity, Phishing, K Nearest Neighbour , Support Vector Machine , Artificial Neural Network , Decision Tree , Random Forest , Logsitic Regression , Max Vote Classifier.*

1. Introduction

Phishing is a type of cybercrime in which hackers sends fake websites link to the targets to either gain sensitive information from users like email, passwords, government identification proofs, credit cards, bank details etc. Phishing websites can aslo lead to an event of installation of malicous softwares through which hackers can gain complete access of their system remoltly. Most phishing websites looks exactly similar to original websites and when comparison their domain name with orignal website there is only slight changes between them like a speling mistake or using similar looking characters to fool the target and leading target to enter sensitive information or target giving access by installation of malicous software.[2][5]

2. Dataset

Phishing website dataset provided by uci in their machine learning repository is used for this research. Dataset consisit of information of 11056 websites with 30 parameters about each website, each parameter has value either -1, 0, 1 and on basis of these factors we have to generate an integer value -1, 1 which signifies whether the website tested is a phishing website or not. Dataset is sliced into 9:1 ratio where 90% websites were used for training the model and 10% were used for testing the model which is trained.[10]

2.1. Visualisation of Dataset

The dataset consists of 30 features which makes it a high dimensional dataset and ordinary methods can't be used to plot the dataset, t-SNE is used for dimensional reduction by reducing 30 parameter data into 2 parameters for visualisation and exploratory data analysis of the dataset.

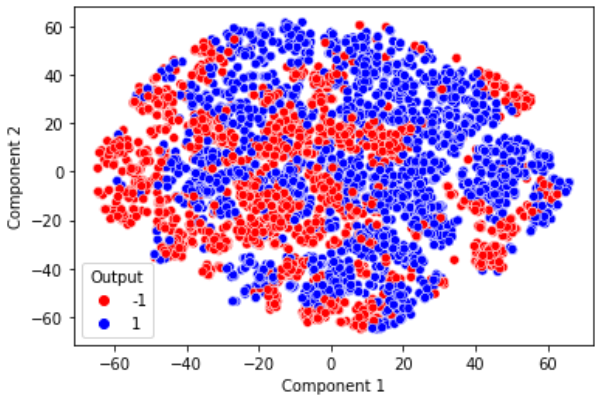

Fig. 1. Scatter plot of Component 1 and Component 2 from result obtained from reducing dimension of the original train dataset. Colour scheme of plot is based on output variable where -1 signifies Phishing website and 1 signifies genuine website.

Scatter plot from Fig. 1. shows that there are many sample points which overlap but majority points can be separated by using a non-linear classification model.

2.2. Data Cleaning

The dataset doesn't contain any missing value or outlier value hence there was no modifications were made on the dataset.

2.3. Feature Selection

Wrapper based feature selection was used to try out all the possible combinations and highest accuracy was achieved when we used all the 30 parameters.

[1] *Guru Gobind Singh Indraprastha University, Delhi, India*
ORCID ID :0000-0002-9917-8807
** Corresponding Author Email: ameya.chawla.ml@gmail.com*

2.4. Data Transformation

Data transformation was not used as data was already defined in 3 integers set of {-1, 0, 1}.

3. Parameter

The dataset has over 30 parameters

3.1. Having IP Address

This parameter signifies whether the website registered domain name or not as non registered domain websites will be shown in form ip address with specified port number in the search bar. This parameter can have only 2 possible values either 0 or 1 where 0 for having ip address.

3.2. URL Length

This parameter signifies about length of the url if the characters length is less than 54 then website is considered legitimate website and value of attribute is 1 if the size is greater than 54 and less than equal to75 is considered suspicious and value of attribute is 0 and more than 75 is considered phishing website and value of attribute is -1.

3.3. URL Shortening

This parameter signifies whether the website link using is shortened which can redirect to Phishing websites. This parameter has 2 values -1 and 1 where -1 for using shortnening service.

3.4. Having '@' symbol

This parameter signifies wheter the website link has "@" symbol in the character set of the link as browsers tend to ignore address before "@" symbol and which could lead to a Phishing site.This parameter can have values -1 and 1 where -1 means website link has "@" symbol.

3.5. Using '//' for redirecting

This parameter signifies whether the website link has "//" in the character set of the link as browser will redirect to page emntioned after "//" so the link has to be checked for last occurance of "//" for pages which have http or https protocol the occurance is 6^{th} position and 7^{th} position respectively if it occurs after 7^{th} position then the site is phishing site and parameter value will be given -1 else 1.

3.6. Addition of prefix opr suffix in url link using '-'

This parameter signifies whether the website link has "-" symbol which can be sued to add prefix , suffix to site leadning user to believe it is a genuine site, this parameter can have 2 values -1 and 1 where -1 for having "-" in url.

3.7. Having Multiple subdomains

This parameters signifies whether the website is Phishing or not on basis of subdomains where removal of top-level domain and second level domain and 'www' subdomain and then checking number of dots left in domain . If their is less than one dot then site is non Phishing if 1 then site is supicous and if greate rthan 1 then phishing site and parameter value is given 1,0,-1 respectively.

3.8. HTTPS

This parameter signifies whether the website is using HTTPS protocol or not and whether the issued certificate is 1 year old and issued by trusted authority. If HTTPS is one year old and issued by trusted authority parameter value is 1 else HTTPS issued by suspicous authority then value 0 else non HTTPS site parameter value -1.

3.9. Domain Registration Length

This parameter signifies whether the website domain is registered for how much more years/months as phishing sites are made for short period of time so if it is registered for elss than 1 year then parameter value -1 else 1.

3.10. Favicon

This parameter signifies whether the website is using favicon from any other website then parameter value is given -1 and website is considered Phishing .

3.11. Using Non- standard port

This parameter signifies whetehr the site is using standard ports like for HTTPS,FTP etc . Website uses nonsstandard ports then site is classified Phishing and parameter value is -1 as any open port not blocked by firewall can give access of target computer to the hacker.

3.12. HTTPS added in domain

This parameter signifies whether the site has added HTTPS in domain name to fool the target if yes then parameter value is -1 else 1.

3.13. Request URL

This parameter signifies whether the images , videos or any graphical content on the website is from any other website or not . In this parameter the percentage of content copied if less than equal to 22% site classified non Phishing and parameter is given value 1 if site has more than 22% and less than 61% then site is classified suspicousand parameter is given value 0 , and if more than 61 % site is classified Phishing and parameter is given value -1.

3.14. URL of Anchor tag

This parameter signifies whether the website anchor tags point towards different domain than this website , if 31% or less anchor tags point to different domains site is classified genuine and parameter value 1 , if more than 31% and less than equal to 67% site is classified suspicious and parameter value is 0 and if more tha 67% website is classified Phishing and parameter value -1.

3.15. Links in <Meta>, <Script> and <Link> tags

This parameter signifies whether the website tags have link to the same domain of the website or different domain, if less than equal to 17% point to different link then site is classified genuine and parameter values is 1, if more than 17% and less than equal to 81% then website is classified suspicious and parameter value is 0 and more than 81% is classified Phishing with parameter value -1.

3.16. Server Form Handler

This parameter signifies what is value of SFH if it contains ''about:blank'' which means it doesnot define where the submitted information will be handled , there can be cases where external domains are mentioned in it. If website has same domain name in SFH then it is classified genuine with parameter value 1, if external domain is mentiones then it is classified as suspicous and parameter value is 0 , and if about:blank is mentioned then site is classified Phishing with parameter value -1.

3.17. Submitting information to email

This parameter signifies whether there is mail() or mailto() function defined as it can lead to direct transmission of data to

Phisher email , if thses functions are present then parameter value is -1 else 1.

3.18. Abnormal URL

This parameter signifies whether the host name is included in the url or not , if not then it is classified Phishing with parameter value -1 else 1.

3.19. Website Forwarding

This parmeter signifies how many times the website have been redirected if more than once then website is classified as Phishing with parameter value -1 else 1.

3.20. Customized Status Bar

This parameter signifies whether the status bar show correct link or not when we hover our mouse over it , it can be checked what happens in ‘’OnMouseOver’’ event if it changes then it is classified Phishing with parameter value -1 else 1.

3.21. Disabled Right Click

This parameter signifies whether right click is disabled or not as Phisher doesn’t wants user to inspect site and check source code, if it is diabled then parameter value is -1 else 1.

3.22. Using Pop-up Window

This parameter signifies whether the site is asking to fill details in pop-up window if yes then it is classified as Phishing with parameter value -1 else 1.

3.23. Redirection using IFRAME

This parameter signifies whether the site is using iframe to display another page without using borders which can fool the target so then the website is classified Phishing with parameter value -1 else 1.

3.24. Age of Domain

This parameter signifies whether the website is at least 6 months old as many Phishing sites are made for short period of time so if site is less than 6 months old then it will be classified Phishing with parameter value -1 else 1.

3.25. DNS Record

This parameter signifies whether DNS exist for the website if yes then genuine site else it is classified Phishing with parameter value -1 else 1.

3.26. Website Traffic

This parameter signifies about the popularity of the website. In Alexa Database if rank is less than 100,000 then site is considered genuine with parameter value 1, if it is more than 100,000 then website is considered suspicious with parameter value 0 and if it is not mentioned then classified Phishing with parameter value -1.

3.27. Page Rank

This parameter signifies how much importance of the webpage is on the internet this value is between 0 and 1 so if page rank less than 0.2 then website is Phishing and parameter value -1 else 1.

3.28. Google Index

This parameter signifies whether the site is indexed by Google or not. If it is not indexed, then classified as Phishing with parameter value -1 else 1.

3.29. Links pointing to the webpage

This parameter signifies number of links pointing to the page if it is less than 1 then considered Phishing with parameter value -1 and if more than 0 and less than 3 then website is suspicious with parameter value 0 and else it is considered genuine with parameter value 1.

3.30. Statistical report-based feature

This parameter signifies whether the host of website belongs to top Phishing IP or Phishing domain then website is classified with Phishing and value is -1 else 1.

4. Machine Learning models

Machine Learning is a field which is subset of Artificial Intelligence which involves creating model based on Machine Learning Algorithms , which is trained on some data and then used to process other data to provide predictions . Models used for this classification problem : Machine Learning is a field which is subset of Artificial Intelligence which involves creating model based on Machine Learning Algorithms , which is trained on some data and then used to process other data to provide predictions .[1, 3, 4, 8] Models used for this classification problem :

4.1. Logistic Regression

It is a supervised machine learning algorithm based on statistical model which gives the probability of certain class as output. It uses its Logistic function to determine the probability. It determines the probability as:

$$\mathbf{P(X)} = \frac{\mathbf{1}}{\mathbf{1 + e^{-(a+bX)}}} \tag{1}$$

X is the input variable
e is the base of natural logarithm
a and b are the weights of Logistic Regression Model

From equation 1 we can figure as the value of X approaches ∞ the value of $P(X)$ approaches to 1 and when X approaches -∞ ,the value of $P(X)$ approaches 0.Output of the Logistic function is in range 0 and 1 including both .

$$\mathbf{Y} = \begin{cases} \mathbf{1}, & \mathbf{if\ P(X) \geq 0.5} \\ \mathbf{0}, & \mathbf{otherwise} \end{cases} \tag{2}$$

To predict class from $P(X)$ we have to apply a threshold which should be minimum value applied to be classified in class 1 and that threshold is 0.5 as shown in equation 2 .

4.2. K Nearest Neighbours

It is a supervised machine learning algorithm which predicts the output on the basis of the similar points to the given data point , K samples are taken which are most similar to given data point and similarity is measured by difference in the values of the features of the points , less the difference more similar are the points . When the points are plotted , the points with less difference in features will be near to each other as their distance between them will be less . When their is a test point given then k points having less distance from that point are selected to predict for that point .
Highest accuracy was achieved with k having value 1.
Distance measure formulas used in KNN :
Euclidean

$$\mathrm{D} = \sqrt{(\mathbf{z_1 - w_1})^2 + (\mathbf{z_2 - w_2})^2 + \ldots + (\mathbf{z_n - w_n})^2} \tag{3}$$

where $z_1, z_2 \ldots z_n$ and $w_1, w_2 \ldots w_n$ are values of n feature of 2

data points z and w.

4.3. Support Vector Machine

It is a supervised machine learning algorithm , in this algorithm n features as input are given and objective of the algorithm is to find a n dimensional plane which separates the two classes data points . Equation of the plane which separates the data $g(x) = w^T x + b$ and here w and b are weights learned by the model.

$$\mathbf{Y} = \begin{cases} \mathbf{1}, & \mathbf{if\ \ g(x) \geq 1} \\ \mathbf{0}, & \mathbf{if\ \ g(x) \leq -1} \end{cases} \quad (4)$$

As the plane acts as the division between them and points above $g(x) = 1$ plane are classified as class 1 and points below $g(x) = -1$ are classified class 0.

4.4. Decision Tree

It is a supervised machine learning algorithm based on the data structure tree , where Decisions are made on each and every node of the tree and based on those Decisions the next step to go to which child node is decided and reaching a leaf node is end point in traversal and the class is predicted at that point.The Decisions are selected by feature in the input data where a feature is selected to split only if the entropy of the system decreases.Highest accuracy was achieved with max depth 18.

$$\mathbf{E(S)} = \sum -\mathbf{p_i log_2 p_i} \quad (5)$$

p_i is probability of class i in our data.

Information Gain is defined as the change in entropy before splitting on basis of a feature and entropy after splitting on the basis of a feature . Feature should be selected to maximize the information gain.

$$\mathbf{IG(Z, W)} = \mathbf{E(W)} - \mathbf{E(W/Z)} \quad (6)$$

$E(Z)$ is the initial entropy, $E(W/Z)$ is the entropy after gaining some information Z

4.5. Random Forest

It is a supervised machine learning algortihm , where it is an ensemble technique which creates multiple Decision Trees while training and when predicting it gives the output as the mode/mean predictions of the individual trees.Highest accuracy was achieved with max depth 16.

4.6. Artificial Neural Network

It is a supervised deep learning algorithm where the model creates structure similar to Biological Brain neurons , model is defined by several layers where each layer has several nodes/neurons.The model structure have each neuron taking input from each neuron of previous layer and giving input to each neuron of next layer. Each input given to the neuron is multiplied by weights defined for each neuron of the model and a bias is added , this sum is then given to input to a function like Sigmoid,Relu . The output from the function applied is then given as input to the next layer neurons. Final layer neurons give the final prediction.Output of each neuron. Highest accuracy was achieved with hidden layer neurons {15,7,3,1}.

$$\mathbf{f(b + \sum x_i w_i)} \quad (7)$$

x_i are the inputs from the previous layer neurons
w_i are the weights for that neuron
b is the bias term

4.7. Voting Classifier

A voting classifier is supervised machine learning algorithm in which the model is trained on the basis of ensemble of many models ,where these models give the predictions and then highest majority voted prediction is taken as the final output.It works on the principle to reduce the wrong predicted outputs or reducing the false negative output and false positive outputs . It also improves the overall performance of the model by decreasing the chance of wrongly predicted outcome by one model being the final outcome Highest accuracy was achieved by combination of Decesion Tree, Random Forest , Artficial Neural Network.
There are mainly two types of voting done in the voting classifier which are :

4.7.1. Hard Voting

In this classifier the class which is voted or predicted maximum times by the models or has the highest probability to be predicted by the model is selected as the final output , it is basically mode of the predictions by the models.

4.7.2. Soft Voting

In this classifier the mean/average of the probabilities of the classes predicted by the model are used to predict the final probability of the model.

5. Evaluation Metrics

Evaluation Metrics are used to measure the quality of the machine learning model.Evaluation Metrics used for this classification problem :

5.1. 5.1 Confusion Matrix

It is a matrix with equal number of rows and columns where n are the number of classes being predicted in the output.

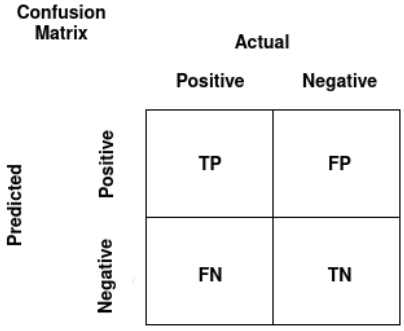

Fig. 2. Confusion Matrix for classification problem having 2 classes

TP are true positives in which actual positive cases were predicted as positives.FP are false positives in which actual negative cases were predicted as positives.TN are true negatives in which actual negative cases were predicted as negatives.FN are false negatives in which actual positive cases were predicted as negatives.
Concise metrics from confusion matrix , these are derived from the the confusion matrix to judge the prediction made by the model.

5.2. Accuracy

$$\frac{TN+TP}{TN+TP+FP+FN} \quad (8)$$

It tells how many data points where classified correctly.

5.3. Precision

$$\frac{TP}{TP + FP} \tag{9}$$

It tell how many predicted positives were actual positives.

5.4. Recall/Sensitivity

$$\frac{TP}{TP + FN} \tag{10}$$

It tells how many positives were predicted correctly out of total positives.

5.5. Specificity

$$\frac{TN}{TN + FP} \tag{11}$$

it tells how many negatives were predicted correctly out of total negatives.

5.6. F_1Score

$$\frac{2 * R * P}{R + P} \tag{12}$$

where R stands for Recall and P stands for Precision F_1 is used in the cases to judge when one model has better score in either Recall and lacks in Precision or vice-versa .

6. Results

The following table shows comparison between confusion matrix for different machine learning models trained.

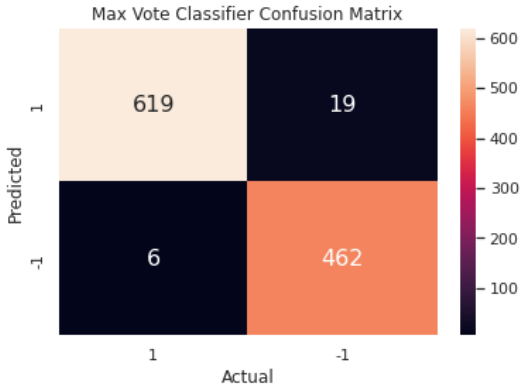

Fig. 3. Confusion Matrix for Max Vote Classifier

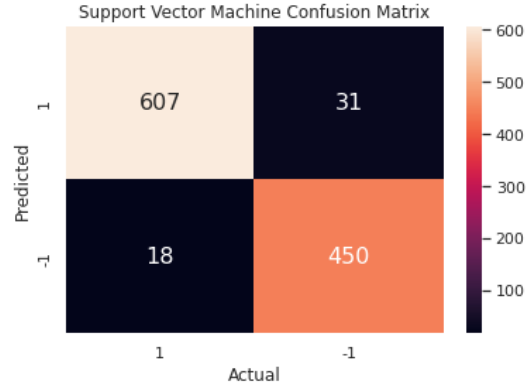

Fig. 4. Confusion Matrix for Support Vector Machine

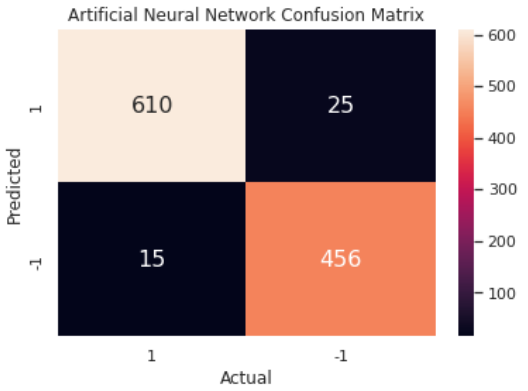

Fig. 5. Confusion Matrix for Artificial Neural Network

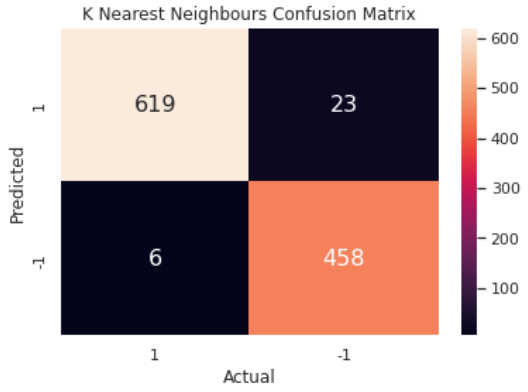

Fig. 6. Confusion Matrix for K Nearest Neighbors

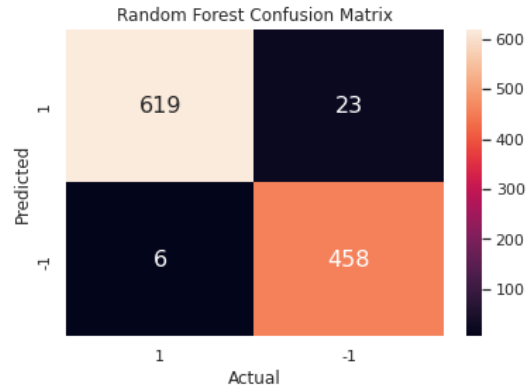

Fig. 7. Confusion Matrix for Random Forest

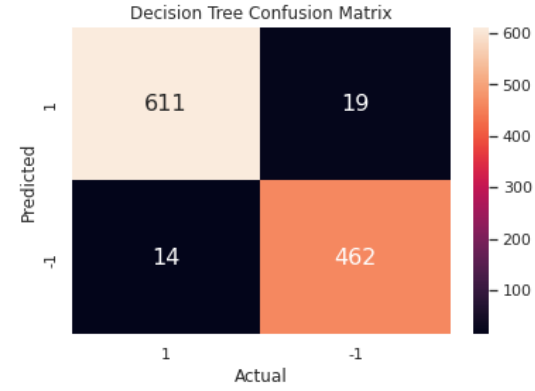

Fig. 8. Confusion Matrix for Decision Tree

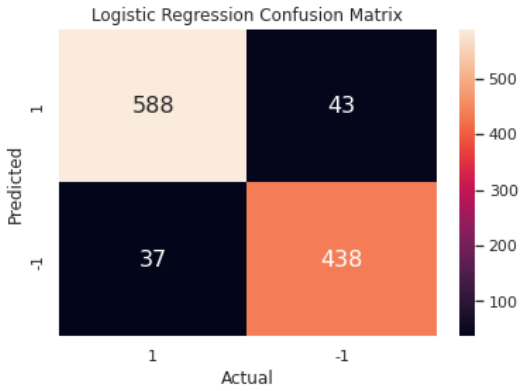

Fig. 9. Confusion Matrix for Logistic Regression

Max Vote Classifier has the highest number of True Negatives and True Positives classified.

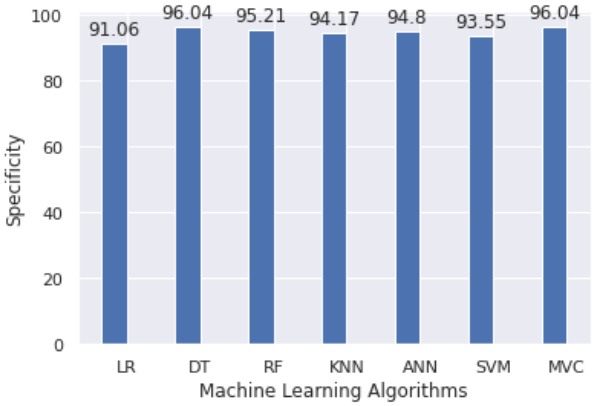

Fig. 10. Specificity comparison of Machine Learning Algorithms

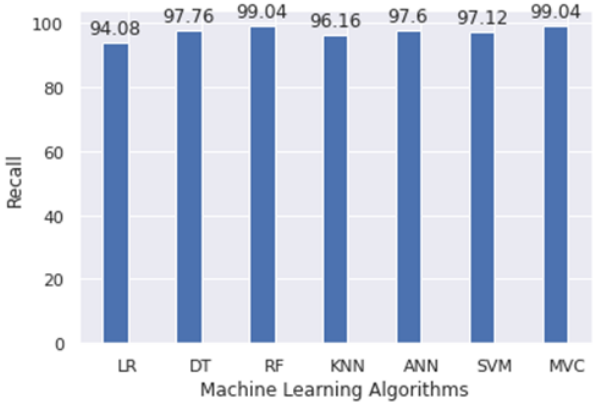

Fig. 11. Recall comparison of Machine Learning Algorithms

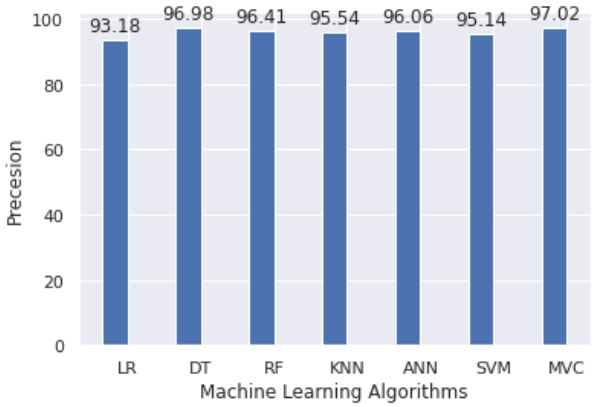

Fig. 12. Precision comparison of Machine Learning Algorithms

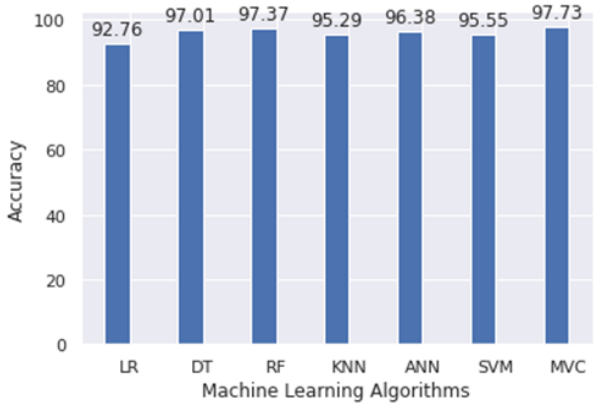

Fig. 13. Accuracy comparison of Machine Learning Algorithms

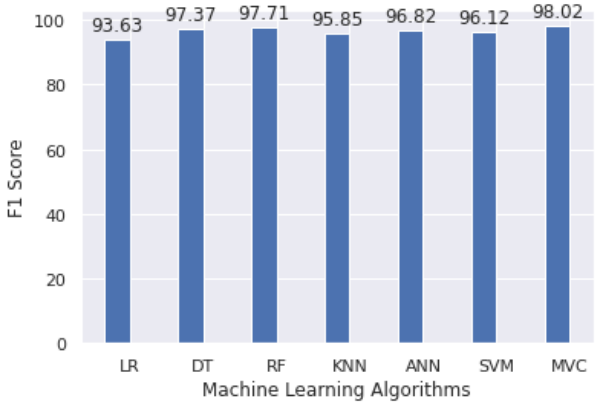

Fig. 14. F_1 Score comparison of Machine Learning Algorithms

Max Vote Classifier scored highest in all the comparison as per the evaluation metrics used hence best choice to solve this problem.

7. Conclusion

Detection of Phishing website at early stage is important as it can save sensitive information and money of the user. Objective of the research is successfully achieved by implementing Machine Learning algorithm which achieves highest score in all evaluation metrics. [9]

8. Future Scope

Max Vote Classifier which achieved highest score in all evaluation metrics can be implemented into a web application where user can enter the URL of the website and detect whether the website is Phishing or not.[6]

References

[1] Patil S, Dhage S. A methodical overview on phishing detection along with an organized way to construct an anti-phishing framework. In: 2019 5th International Conference on Advanced Computing & Communication Systems (ICACCS). IEEE; 2019. p. 588–93.

[2] Geng G-G, Yan Z-W, Zeng Y, Jin X-B. RRPhish: Anti-phishing via mining brand resources request. In: 2018 IEEE International Conference on Consumer Electronics (ICCE). IEEE; 2018. p. 1–2.

[3] Pratiwi ME, Lorosae TA, Wibowo FW. Phishing site detection analysis using artificial neural network. J Phys Conf Ser. 2018; 1140:012048.

[4] Oza Pranali P., Upadhyay D, Gujarat Technological University. Review on phishing sites detection techniques. Int J Eng Res Technol (Ahmedabad) [Internet]. 2020 [cited 2021 Aug 10];V9(04).

Available from: https://www.ijert.org/review-on-phishing-sites-detection-techniques

[5] Alkhalil Z, Hewage C, Nawaf L, Khan I. Phishing attacks: A recent comprehensive study and a new anatomy. Front Comput Sci [Internet]. 2021;3. Available from: http://dx.doi.org/10.3389/fcomp.2021.563060

[6] Jain AK, Gupta BB. A novel approach to protect against phishing attacks at client side using auto-updated white-list. EURASIP J Multimed Inf Secur [Internet]. 2016;2016(1). Available from: http://dx.doi.org/10.1186/s13635-016-0034-3

[7] Patil NM, Dias SP, Dcunha AA, Dodti RJ. Hybrid phishing site detection. Int j adv sci technol. 2020;29(6s):2452–9.

[8] Harinahalli Lokesh G, BoreGowda G. Phishing website detection based on effective machine learning approach. J cyber secur technol. 2021;5(1):1–14.

[9] Jain AK, Gupta BB. Phishing detection: Analysis of visual similarity-based approaches. Secur Commun Netw. 2017; 2017:1–20.

[10] UCI machine learning repository: Phishing websites data set [Internet]. Uci.edu. [cited 2021 Aug 10]. Available from: https://archive.ics.uci.edu/ml/datasets/phishing+websites