International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 6, Issue 1, pp 46-53, January-2022

https://doi.org/10.47001/IRJIET/2022.601010

Machine Learning in Policing Counterfeit Websites

[1]Lekha Khobrekar, [2]Qurratulain Munshi, [3]Swapna Naik

[1,2]Student, Information Technology, Shri Bhagubhai Mafatlal Polytechnic, Mumbai, India

[3]Professor, Information Technology, Shri Bhagubhai Mafatlal Polytechnic, Mumbai, India

Mail IDs: [1]lkhobrekar2@gmail.com, [2]munshiqurratulain28@gmail.com, [3]swapna.naik@sbmp.ac.in

***Abstract* - Phishing attacks on websites are a serious cyber-crime whereby mimicking the domain name and appearance of official websites with the aim of stealing confidential information like passwords, credit card information with malicious intent to install malware on the victim's machine or use personalized information to target in multiple types of attacks. When individual victims browse the targeted website, the phishers seize their personal information. Phishing affects assorted fields, such as online business, banking and digital marketing, e-commerce which results in various financial losses and theft of personal and important information. The main purpose of this paper is to present a framework to detect phishing websites using various Machine Learning Algorithms and to focus on several Phishing Attacks along with its Classification Techniques. This detection is done using Uniform Resource Locator (URL). There are many characteristics to distinguish URLs from regular website links, the difference between real and phished website is not visible to the human eye. This paper uses various Machine Learning models and compares their accuracy to profile which model gives best possible analysis. This can benefit website owners to determine best possible mechanism to protect and mitigate phishing attacks.**

Keywords: Phishing, Machine Learning, Supervised Machine Learning Algorithm.

I. INTRODUCTION

Phishing is particularly a vicious form of social engineering which works by gaining the trust of target organization. It does this by pretending to be part of the organization. In recent years, social networks have become a virtual meeting place for the wide-ranging community. Unfortunately, while connecting through social networks, individuals experience phishing attacks. Phishing risks a user's privacy, may execute malware attacks and often steals their confidential information. Phishing is carried out by using various engineering techniques including instant messages, fraudulent emails or mimicking an online bank, auction or payment sites and directing people to fake Web pages that resembles a genuine site. Phishing sites to the tune of 2,145,013were registered by Google as of January 17, 2021.A website phishing attack is carried out by bluffing legal identities, such as an authentic website. A malicious website succeeds at obtaining some user information and can lead a user to additional malevolent links that consequently gain access to even more of the user's fragile or personal information. To achieve this objective, identical websites are created which so closely resemble the actual website that the double-dealing duplication cannot be detected. Phishing attacks cause great economical, intellectual property and damaging to national security.

Phishing destroys industries including e-commerce and digital banking. Several techniques can be applied to safeguard the users from phishing attacks, including the experimental approach, the rule-based approach, and a supervised Machine Learning (ML) approach. Supervised ML Algorithms are extensively used for classification and are more popular among all the techniques used to detect phishing websites.

Figure 1: Phishing

This research study focuses on a supervised ML approach to detect phishing websites. The provisions of this research study are as follows:

- Comparing various Supervised Learning Algorithms including Random Forest (RF), K-Nearest Neighbor,

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 6, Issue 1, pp 46-53, January-2022

https://doi.org/10.47001/IRJIET/2022.601010

Decision Trees, and Logistic Regression on a combination of different feature sets.

- Stacking of the ultimate performing algorithms to enhance the classification accuracy.
- Performance evaluation dimensions, such as accuracy, recall, f-measure, and precision, are used to estimate the performance of all classifiers.

The research paper is further structured as follows: Section 2 reviews the various Phishing Attacks and its Classification Techniques. Section 3 shares the techniques, the dataset and the performance evaluation measures applied in this study. Section 4 displays the experimental results and comparisons of the applied techniques, and Section 5reviews the conclusions of this study.

II. RELATED WORK

This section offers a review of the related research work of phishing attacks in general and concentrates on the classification techniques applied to detect Web Phishing.

A) Phishing Attacks

1. Social Engineering

Social engineering is a falsification technique that exploits human error to acquire confidential information, access, or valuables. In cybercrime, these "human hacking" cons tend to trap unsuspecting users into exposing information, spreading malware viruses, or giving access to restricted systems. It is similar to the motive as that of hacking i.e., to obtain the forbidden access to any system or steel confidential information regarding any organization or any network intrusion. Usually the companies, government agencies or military are the massive targets. There are two degrees of attacks in social engineering: Physical social engineering attack and psychological social engineering attack.

2. Website Phishing

Website phishing is one of the phonological attacks with an aim of targeting a specific individual instead of any organization. Phishing websites are built to trick unsuspecting users into believing they are on a legitimate site. Website attacks can be easily implemented by designing a duplicate copy of any legitimate website. The central objective of creating such illegitimate websites is to carry out fraud transactions with individuals to get their personal and financial data.

3. Hijacking Sessions

The Session hijacking attacks consists of the exploitation of the web session control mechanism, which is normally managed for a session token. Session hijacking is executed either at application level or at network level. At Application level session hijack involves interfering HTTP and at Network level interfering is done at TCP and UDP.

4. Email Phishing

Cybercriminals often target companies and individuals via emails designed to give the impression that they came from a government agency, legitimate bank, or an organization. Email Phishing is the initial step towards the launch of phishing websites. These emails convince recipients to get on a web page where they will confirm personal data or account information to the attacker by phishing server.

5. Key loggers and screen loggers

A key logger, often known as a keystroke logger or keyboard capture, is a type of inspection technology used to observe and track record each keystroke on a specific computer. Key loggers are classified into two: Hardware key logger and software key logger. Hardware key loggers are likely impossible to identify without physical scrutiny. The key logger is a severe threat to the servers or system as users cannot detect their existence. The screen filming software makes the situation harsher because of key logging.

6. Malware-Based Phishing

There is several software designs developed to include malicious contents that gets mounted to the victim's system as he/she installs it. Sometimes users can be deceived into downloading antivirus software, while it is in fact a virus or a malware itself. Malware takes benefit of the weakness in the operating system or in any browser.

B) Classification Techniques

So far, there have been various techniques introduced to get rid of phishing attacks and provide a safe environment to the online users. Detecting Fake URLs and spoofed emails is not straightforward. Classifications of various protective approaches against phishing are as follows:

1. Software-based Defense Approach

Phishing Protection at network level is an approach in which a particular range of IP addresses or few domains are not allowed to enter the network.

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 6, Issue 1, pp 46-53, January-2022

https://doi.org/10.47001/IRJIET/2022.601010

2. User Education Approach

User's education approach is an approach to spread awareness about phishing among all the internet users. This approach spreads information about risks of phishing attacks and their prevention techniques.

3. Authentication-based Approach

In Authentication-based approach, a confirmation message is sent to verify or confirm, it is sent by the valid domain or invalid domain. This method is very helpful in the email communication.

There are many researchers who worked hard to propose approaches to detect such type of fake authentication messages, spoofed emails, or phishing websites. Kumar and Chaudhary[2] proposed an approach that uses hyperlinks in web pages. The features in hyperlinks are used to detect a phishing website. Many Machine Learning approaches are used along with this. Machine Learning techniques are applied on both phishing and non-phishing datasets. This approach is language independent and results up to 99 per cent accuracy. Nagaraj et al.[2] proposed a Machine Learning model to classify phishing sites which was made twofold, applying Random Forest classifier, and integrating the results with a feed forward Neural Network. K-fold cross validation has been used to validate the performance. This resulted an accuracy of 93.41 per cent. Sahingoz et al.[2] executed a real-time anti-phishing system for phishing detection using URLs. The proposed approach uses seven classification algorithms [K-star, AdaBoost, Decision tree, KNN (n = 3), SMO, RF, and Naïve Bayes] and different types of natural language processing-based features. Fette et al.[2]introduced an approach based on ML techniques named PLIFER which requires the age of the domain of URLs. Moreover, ten features are extracted, and Random Forest is applied to detect a phishing website.

The classification in the above papers is achieved by use of complex algorithms which require huge resources to implement. Delving into this problem the paper tries to analyze using simple machine learning algorithms and tries to achieve better accuracy.

III. ANALYTICAL JUSTIFICATION

This section represents a detailed justification of the experiment. It discusses the various Machine Learning Techniques applied in the experiment, the dataset, and the performance evaluation measures considered to profile the best fit model.

A) Applied Machine Learning Techniques

1. Logistic Regression

Logistic Regression is one of the most widely used Supervised Machine Learning Algorithms. It is used for resolving classification problems. This algorithm is used to predict the categorical dependent variable by using a set of given independent variables. Hence, the outcome of this algorithm must be a categorical or a discrete value. But instead of giving the exact values, the algorithm delivers the probabilistic values that lie between 0 and 1. This algorithm is significant Machine Learning because it offers probabilities and classifies new data using continuous and discrete datasets.

2. K-Nearest Neighbor

K-Nearest Neighbor is one of the simplest Supervised Machine Learning Algorithm, it works on analogy-based classification. KNN is also known as Lazy Learner Algorithm and is non-parametric. This algorithm can be used for both Classification as well as Regression problems. On providing test samples to KNN, it picks the number of K of neighbors. Next, it calculates the Euclidean distance to find K-closest neighbor. The K-closest neighbor is used to make decisions of classification. KNN is robust to noisy training data. It works more effectively with Classification problems. If the training data is large, the algorithm can be more efficient.

3. Decision Tree

Decision Tree is a Machine Learning Algorithm which is based on Supervised Learning technique. This algorithm is used for both Classification and Regression problems, but it is mainly preferred for solving Classification problems. This classifier is tree-structured, where internal nodes are represented as the features of a dataset, branches are represented as the decision rules, and each leaf node is represented as the outcome. The tree begins with the root node which contains the complete dataset. It finds the finest attribute in the dataset by means of Attribute Selection Measure (ASM). Then the dataset is divided into subsets that contain likely values for the finest attributes. Next, the decision tree node with the best attribute is generated. It keeps creating new decision trees recursively using the subsets of the dataset which are already created. The process is continued until a stage is reached where further classification of the nodes is not possible and the final node is called as the leaf node.

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 6, Issue 1, pp 46-53, January-2022

https://doi.org/10.47001/IRJIET/2022.601010

4. Random Forest

Random Forest is a renowned Machine Learning Algorithm owing to achieve Classification and Regression successfully with greater performance. RF is based on the idea of ensemble learning, which is used to achieve enhanced accuracy for classification. Random Forest is a classifier that includes several decision trees on the subsets of the given dataset and takes the prediction from each tree and based on the majority votes of predictions, it predicts the final output to improve the predictive accuracy of that dataset.

5. Support Vector Machine

Support Vector Machine is one of the most widely used algorithm based Supervised Machine Learning technique. This algorithm is used for both, Classification and Regression problems. The aim of this algorithm is to create a decision border that can separate n-dimensional space into classes, because of which one can effortlessly classify the new data point in the correct category in the future. This finest decision border is called a hyperplane.SVM selects the extreme vectors that help in forming the hyper plane. These extreme cases are called as support vectors; therefore, the algorithm is called as Support Vector Machine.

B) Data Set

```
 0   UsingIP              11054 non-null  int64
 1   LongURL              11054 non-null  int64
 2   ShortURL             11054 non-null  int64
 3   Symbol@              11054 non-null  int64
 4   Redirection//        11054 non-null  int64
 5   PrefixSuffix-        11054 non-null  int64
 6   SubDomains           11054 non-null  int64
 7   HTTPS                11054 non-null  int64
 8   DomainRegLen         11054 non-null  int64
 9   Favicon              11054 non-null  int64
 10  NonStdPort           11054 non-null  int64
 11  HTTPSDomainURL       11054 non-null  int64
 12  RequestURL           11054 non-null  int64
 13  AnchorURL            11054 non-null  int64
 14  LinksInScriptTags    11054 non-null  int64
 15  ServerFormHandler    11054 non-null  int64
 16  EmailInformation     11054 non-null  int64
 17  AbnormalURL          11054 non-null  int64
 18  WebsiteForwards      11054 non-null  int64
 19  StatusBarCust        11054 non-null  int64
 20  DisableRightClick    11054 non-null  int64
 21  UsingPopupWindow     11054 non-null  int64
 22  IframeRedirection    11054 non-null  int64
 23  AgeofDomain          11054 non-null  int64
 24  DNSRecord            11054 non-null  int64
 25  WebsiteTraffic       11054 non-null  int64
 26  PageRank             11054 non-null  int64
 27  GoogleIndex          11054 non-null  int64
 28  LinksPointingToPage  11054 non-null  int64
 29  StatsReport          11054 non-null  int64
 30  Class                11054 non-null  int64
dtypes: int64(31)
memory usage: 2.6 MB
```

Figure 2: Data Set Features

Data set is simply a collection of data pieces that can be treated by a computer as a single unit for analyzing and prediction purpose. The Data set used in this paper for detecting the phishing websites is freely available on Kaggle (phishing.csv) for research purposes. The dataset contains a collection of website URLs for 11000+ websites. Each sample has 30 attributes and a class label identifying it as a phishing website or not. The data set also provides as an input for project scoping and also specifies the functional and non-functional requirements for it. This data set is further divided into training set and testing set in the ratio of 7:3 respectively to perform predictions.

C) Performance Evaluation Measures

There are various performance evaluation measures when it involves selecting a classification model. The metrics that is selected to evaluate your machine learning model is very crucial. The performance of the machine learning algorithm is influenced by the choice of metrics. The measures that are used in this experiment are accuracy, confusion matrix, recall, precision, and F1-Score.

1. Confusion Matrix

The terms that are used in defining a confusion matrix are TP, TN, FP, and FN.

True Positive (TP):
The predicted value matches with the actual value.
The actual value was positive, but the model predicted a positive value.
True Negative (TN):
The predicted value matches with the actual value.
The actual value was negative and negative value was predicted by the model.
False Positive (FP): Type 1 error
The predicted value was wrongly/falsely predicted.
The actual value was negative, but the model predicts a positive value.
False Negative (FN): Type 2 error
The predicted value was wrongly/falsely predicted.
The actual value was positive, but it was predicted negative by the model.

2. Accuracy

This term tells how many correct classifications were made from all the classifications.
Accuracy = (TP + TN) / (TP + FP +TN + FN)

3. Precision

This shows that out of all that were marked as positive, how many of them are actually truly positive.

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 6, Issue 1, pp 46-53, January-2022

https://doi.org/10.47001/IRJIET/2022.601010

Precision = TP / (TP + FP)

4. Recall

This states that out of all the actual positive cases, how many of them were identified as positive.
Recall = TP/ (TN + FN)

5. F1-Score

The F1-Score is defined as the weighted average of the Precision and Recall.
F1 score = 2* (Precision * Recall) / (Precision + Recall)

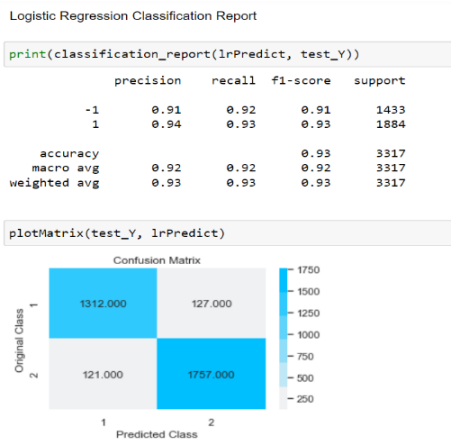

Figure 3: Logistic Regression

The Fig. 3 shown above is an evaluation of Logistic Regression model. This model gives an accuracy of 0.93, with a precision score of 0.94, recall score of 0.93 and f1-score of 0.93.

Figure 4: K-Nearest Neighbor

The Fig.4 shown above is an evaluation of K-Nearest Neighbor model. This model gives an accuracy of 0.94, with a precision score of 0.95, recall score of 0.94 and f1-score of 0.95.

Figure 5: Decision Tree

The Fig. 5 shown above is an evaluation of Decision Tree model. This model gives an accuracy of 0.96, with a precision score of 0.96, recall score of 0.97 and f1-score of 0.97.

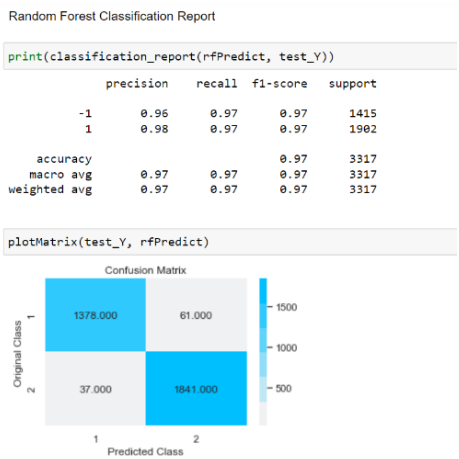

Figure 6: Random Forest

The Fig. 6 shown above is an evaluation of Random Forest model. This model gives an accuracy of 0.97, with a precision score of 0.98, recall score of 0.97 and f1-score of 0.97.

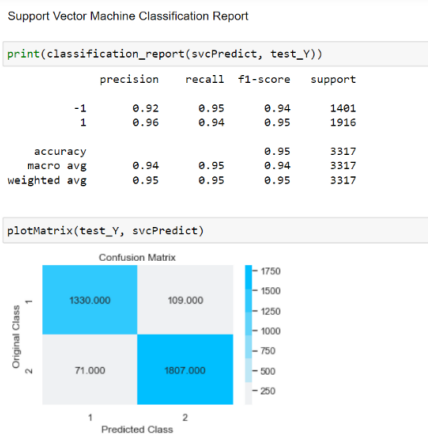

Figure 7: Support Vector Machine

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 6, Issue 1, pp 46-53, January-2022

https://doi.org/10.47001/IRJIET/2022.601010

The Fig. 7 shown above is an evaluation of Support Vector Machine model. This model gives an accuracy of 0.95, with a precision score of 0.96, recall score of 0.94 and f1-score of 0.95.

IV. RESULTS

This section reviews data visualization and experimental results carried out on the selected data set of phishing detection.

A) Data Visualization

Data visualization is the graphical representation of data. By means of graphic elements like graphs, charts, and maps. Data visualization tools offer an understandable way to visualize and understand trends, patterns, and outliers in data.

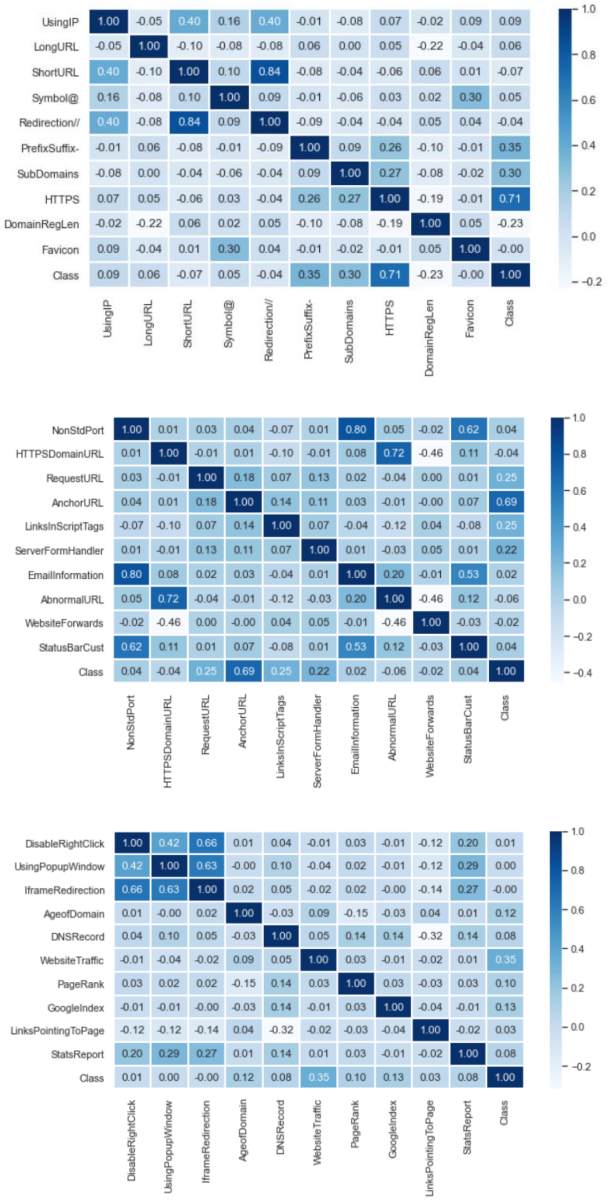

Figure 8: Correlation Heatmap

Correlation heatmap is a graphical representation of correlation matrix. It represents correlation between different attributes and can have values from -1 to 1.The Fig.8.shown above represents Correlation Heatmap which illustrates how the features of the data are linearly correlated for predicting if a site is a phishing site or a legitimate site. The graph plotted into subplots each considering 10 features.

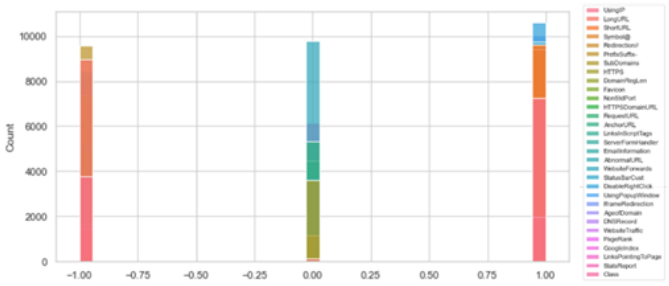

Figure 9: Hist Plot

Histplot is a visualization tool that illustrates the distribution of one or more attributes by counting the observations that lie inside the discrete bins. It shows the fundamental frequency distribution of a set of continuous data. Fig. 9 shown above represents HistPlot which defines the count of the data of each feature in the dataset.

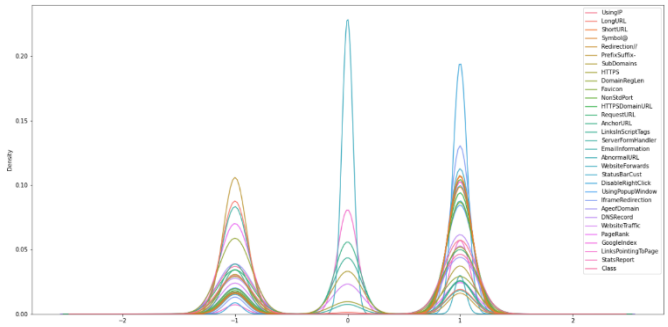

Figure 10: KDE Plot

Kernel Density Estimate (KDE) plot is a tool for illustrating the distribution of observations in a dataset, analogous to a histogram using a continuous probability density curve in one or more dimensions. Fig. 10 shown above represents the KDE Plot which explains the density of the features in the dataset.

B) Experimental Results

After performing a detailed analysis and visualization on the dataset, Machine Learning Algorithms have been applied to validate the accuracy of the classification models. Logistic Regression, K-Nearest Neighbor, Decision Tree Classifier, Random Forest Classifier and Support Vector Machines are the Supervised Machine Learning Algorithms implemented in this experiment. Furthermore, the accuracy of these models is

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 6, Issue 1, pp 46-53, January-2022

https://doi.org/10.47001/IRJIET/2022.601010

compared to find the best fit model. From all the models developed in the experiment, Random Forest Classifier has highest accuracy of 0.97 and followed by Decision Tree Classifier, Support Vector Machine and K-Nearest Neighbor. Logistic Regression with accuracy of 0.93 is the model with the lowest.

	Model	Accuracy
0	LogisticRegression	0.93
1	KNeighborsClassifier	0.94
2	DecisionTreeClassifier	0.96
3	RandomForestClassifier	0.97
4	SupportVectorMachine	0.95

Figure 11: Model Accuracy

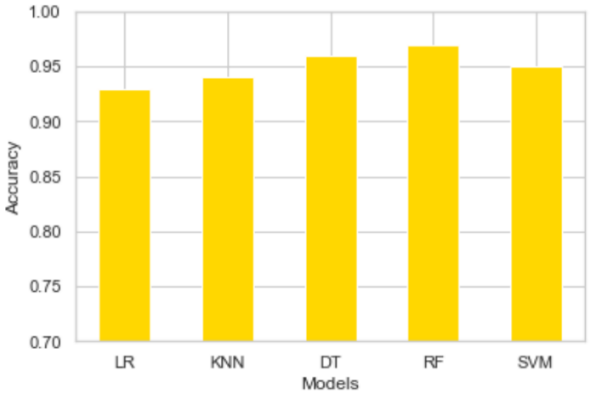

Figure 12: Accuracy Graph

V. CONCLUSION

This research paper presents a detailed review about several Phishing Attacks. It also uses various Classification Techniques and Compares the accuracy of the Supervised Learning models. It is not possible to completely eliminate phishing; however the risk and hazard could be reduced. The main purpose of this research is to compare Machine Learning Algorithms and analyze the best model for certain proposed features to detect a phishing website. The evaluation of the classification models is based on performance evaluation measures such as accuracy, confusion matrix, recall, precision, and F1-Score.The research is performed on the phishing websites data set. It contains data about 11054 URLs with30 attributes and a class label. The research is contingent on the quality and reliability of the features. To carry this work forward, the proposed approaches can be combined with numerous feature extraction models to test its practicality in a real-time scenario.

REFERENCES

[1] P.A. Barraclough, G. Fehringer, J. Woodward, “Intelligent cyber-phishing detection for online”, 2020 Elsevier, 2020.

[2] Ammara Zamir, Hikmat Ullah Khan and Tassawar Iqbal, NazishYousaf, Farah Aslam, Almas Anjum, Maryam Hamdani, “Phishing web site detection using diverse machine learning algorithms”, Emerald Publishing Limited, 2019.

[3] Muhammad Aamir Awan, “PISHING ATTACKS IN NETWORK SECURITY”, LC International Journal of STEM, 2020.

[4] A.A. Orunsolu, A.S. Sodiya, A.T. Akinwale, “A predictive model for phishing detection”, Journal of King Saud University – Computer and Information Sciences, 2019.

[5] Gyan Kamal, Monotosh Manna, “Detection of Phishing Websites Using Naïve Bayes Algorithms”, International Journal of Recent Research and Review, 2018.

[6] Sanjay Kumar, AzfarFaizan, Ari Viinikainen, Timo Hamalainen, “MLSPD - Machine Learning Based Spam and Phishing Detection”, Springer Nature Switzerland AG, 2018.

[7] Alsariera, Yazan Ahmad, Elijah, Adeyemo VictorBalogun, Abdullateef O, “Phishing Website Detection: Forest by Penalizing Attributes Algorithm and Its Enhanced Variations”, Arabian Journal for Science and Engineering, 2020.

[8] Sirisha Derangula, “Identification of phishing websites using ML techniques”, International Journal of Communication and Information Technology, 2020.

[9] Anmar Odeh, Ismail Keshta, Eman Abdelfattah, “Machine LearningTechniquesfor Detection of Website Phishing: A Review for Promises and Challenges”, IEEE, 2021.

[10] https://www.kaggle.com/eswarchandt/phishing-website-detector

[11] Nagaraj K., Bhattacharjee B., Sridhar A., Sharvani G., “Detection of phishing websites using a novel twofold ensemble model”, Journal of Systems and Information Technology, 2018.

[12] Bhattacharjee S.D., Talukder A., Al-Shaer E., Doshi P., “Prioritized active learning for malicious URL detection using weighted text-based features”, IEEE, 2017.

[13] Ibrahim D.R.,HadiA.H., “Phishing websites prediction using classification techniques”, IEEE, 2017.

[14] P.A.Barraclough, Fehringer J, Woodward, "Intelligent cyber-phishing detection for online", Elsevier, 2021.

[15] M. Vijayalakshmi, S. Mercy Shalinie, Ming Hour Yang, Raja Meenakshi U, "Web phishing detection techniques: a survey on the state-of-the-art, taxonomy and future directions", IET Networks, 2020.

ISSN (online): 2581-3048
Volume 6, Issue 1, pp 46-53, January-2022
https://doi.org/10.47001/IRJIET/2022.601010

[16] OzgurKoraySahingoz, Ebubekir Buber, Onder Demir, Banu Diri, "Machine learning based phishing detection from URLs", Elsevier, 2019.
[17] A.A.Orunsolu, A.S.Sodiya, A.T.Akinwale, "A predictive model for phishing detection", Elsevier, 2019.
[18] Said Salloum, Tarek Gaber, Sunil Vadera, Khaled Shaalan, "Phishing Email Detection Using Natural Language Processing Techniques: A Literature Survey", Elsevier, 2021.
[19] Abdullateef O. Balogun, Noah O. Akande, Fatimah E. Usman-Hamza, Victor E. Adeyemo, Modinat A. Mabayoje, Ahmed O. Ameen, "Rotation Forest-Based Logistic Model Tree for Website Phishing Detection", Springer, 2021.
[20] Do Nguyet Quang, Ali Selamat, Ondrej Krejcar, "Recent Research on Phishing Detection through Machine Learning Algorithm", Springer, 2021.
[21] AlMaha Abu Zuraiq, Mouhammd Alkasassbeh, "Review: Phishing Detection Approaches", IEEE, 2019.

Citation of this Article:

Lekha Khobrekar, Qurratulain Munshi, Swapna Naik, "Machine Learning in Policing Counterfeit Websites", Published in *International Research Journal of Innovations in Engineering and Technology - IRJIET*, Volume 6, Issue 1, pp 46-53, January 2022. Article DOI https://doi.org/10.47001/IRJIET/2022.601010

****

