e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:05/Issue:11/November-2023 Impact Factor- 7.868 www.irjmets.com

PHISHING WEBSITE DETECTION AND CLASSIFICATION USING DATA MINING

Kislay Kumar[*1], MSS Harshith[*2], M Charan Deepu[*3], Rajvardhan Singh[*4]

[*1,2,3,4]Student, Integrated M.Tech Software Engineering, Vellore Institute Of Technology, Vellore, Tamil Nadu, India.

DOI : https://www.doi.org/10.56726/IRJMETS46448

ABSTRACT

The project entails creating the best-fitting model to handle the challenge of categorizing the website based on criteria such as Has_ URL. Has_ Anchor, URL_ age, web traffic, domain age, and so forth. Based on this information, the chosen model should be able to identify a website as Phishing, Non-Phishing, or Suspicious based on user input. The selected dataset comprises thousands of testcases with 9 distinct criteria that classify phishing websites. Models will be created from these utilizing various methods such as ML-based algorithms and DL-based algorithms. The model with the best fit will be chosen for further evaluation.

Keywords: Phishing Website, Machine Learning, Suspicious.

I. INTRODUCTION

Cyber attackers and spoofers use a variety of online digital platforms to steal the identities of legitimate businesses and trick multiple unaware users into revealing their private information or installing malicious software in their devices, which can range from small-scale spyware and adware to highly resilient ransomwares. Various anti-virus software vendors have designed and implemented numerous successful anti-phishing solutions. These tools' accuracy varies and can occasionally label and warn non-harmful websites as malicious, and vice versa.

Phishing is a frequent technique intended to gain sensitive information by utilising websites that look similar to real websites. Phishing assaults are becoming more common as technology advances. There are various established approaches for distinguishing phishing websites from real websites. The majority of the algorithms used to address this problem are detailed in the study's literature review.

The goal of this project is to tackle the problem of phishing detection using a variety of ML, DL, and evolutionary algorithms that can be upgraded using R language: gives best prototype for respective model or python to implement in the future for real-time purposes. We may use these methods to feature choose a certain fascinating or necessary element that distinguishes a specific page as a phishing website from the rest of the websites.

II. METHODOLOGY

The step to be performed for analyzing the website as phishing, Non-Phishing, Suspicious by using the following steps:

1. Dataset Collection

Sfh, Popup window, SSL_Final state, Request_URL, URL_Anchor, Web traffic, URL Length, Age_of_Domain, Having_IP_Address are among the 10 attribute and 1354 tuples in the data set.

2. Data Preprocessing

This is the initial step in creating any ML model. The data is first visualised, and then it is divided into training and test data sets. The normal default proportion between training and test sets is 70/30. In the case of neural networks, weights and learning rates are also defined.

3. Training

The data set is linked to an algorithm, which uses complex mathematical modelling to learn and produce predictions. We are dealing with categorization models in this specific scenario.

4. Testing

In this step we validate the trained model. We check for the model's accuracy. We change the model until the results are satisfactory.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:05/Issue:11/November-2023 Impact Factor- 7.868 www.irjmets.com

5. Performance metrics

We do other performance metrics besides testing such as F1 score, Precision, Recall and also plot the confusion matrix. The performances of the various models are analyzed using visualization techniques.

6. Analysis

Analyzing the best fit algorithms for the particular case and predicting the result.

III. MODELING AND ANALYSIS

All six methods utilised in this study have been shown in performance metrics, and their accuracy % has also been determined. The entire graphic shows which algorithms will produce the most accurate results.

1. KNN

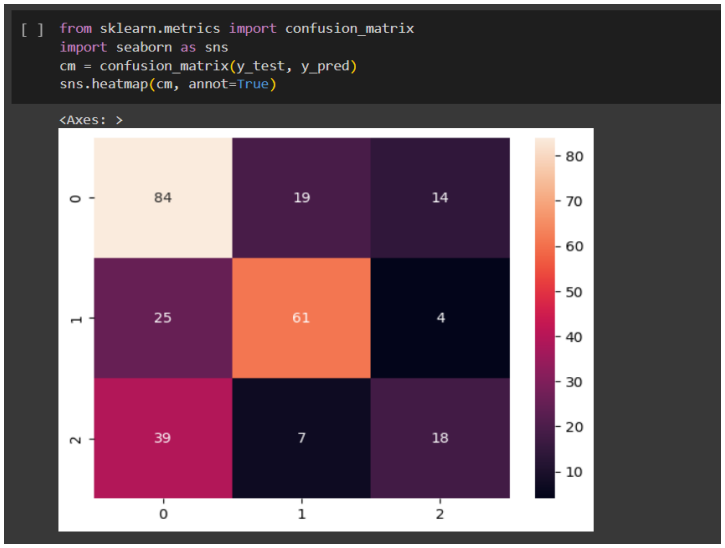

Figure 1: KNN Performance metrics.

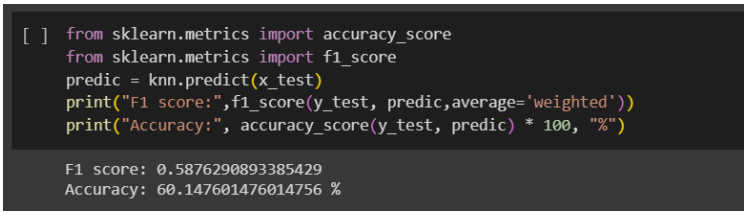

Figure 2: KNN accuracy percentage.

2. Decision Tree

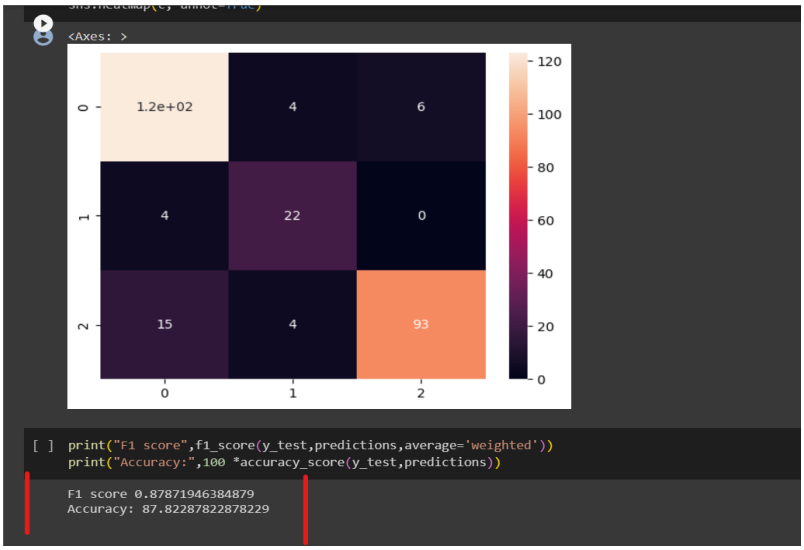

Figure 3: Decision Tree Performance metrics with accuracy percentage.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:05/Issue:11/November-2023 Impact Factor- 7.868 www.irjmets.com

3. Naïve Bayes

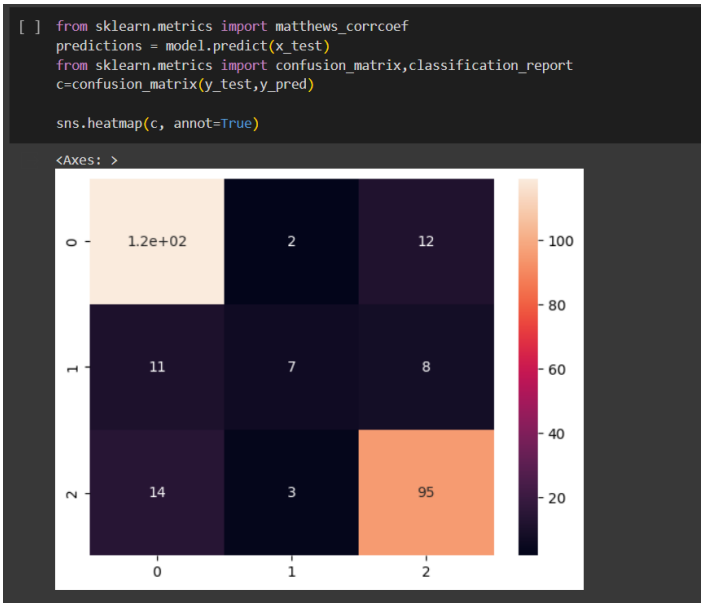

Figure 4: Naïve Bayes Performance metrics.

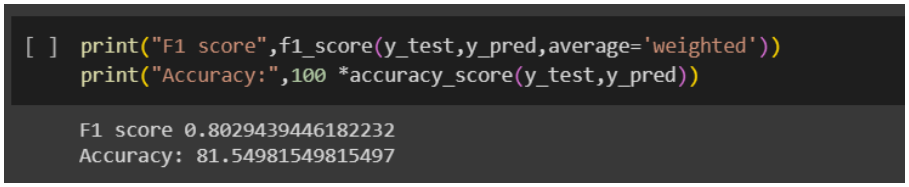

Figure 5: Naïve Bayes accuracy percentage.

4. Single layer perceptron

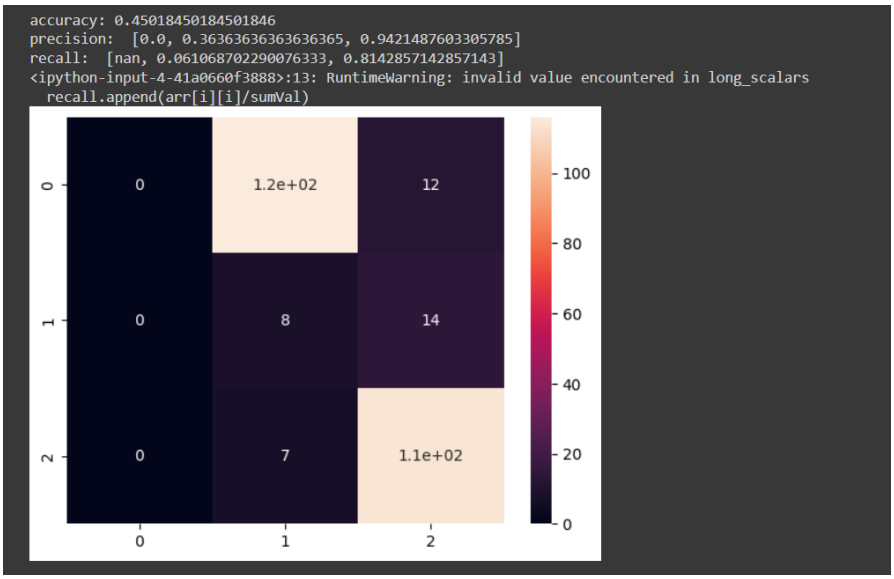

Figure 6: Single layer perceptron Performance metrics with accuracy percentage.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:05/Issue:11/November-2023 Impact Factor- 7.868 www.irjmets.com

5. Multilayer perceptron

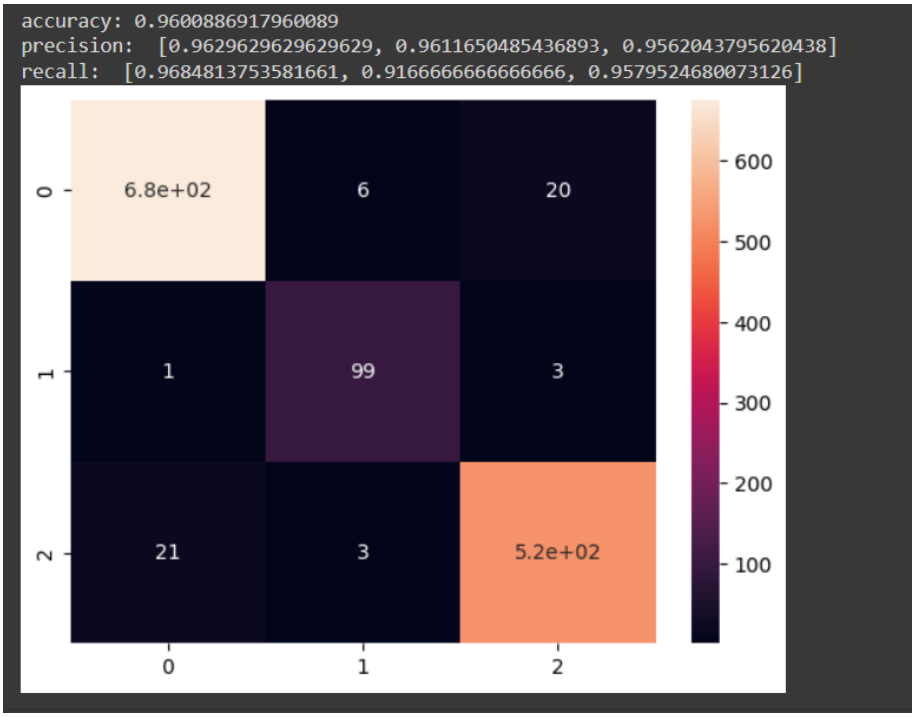

Figure 7: Multilayer perceptron Performance metrics with accuracy percentage.

6. Adaline

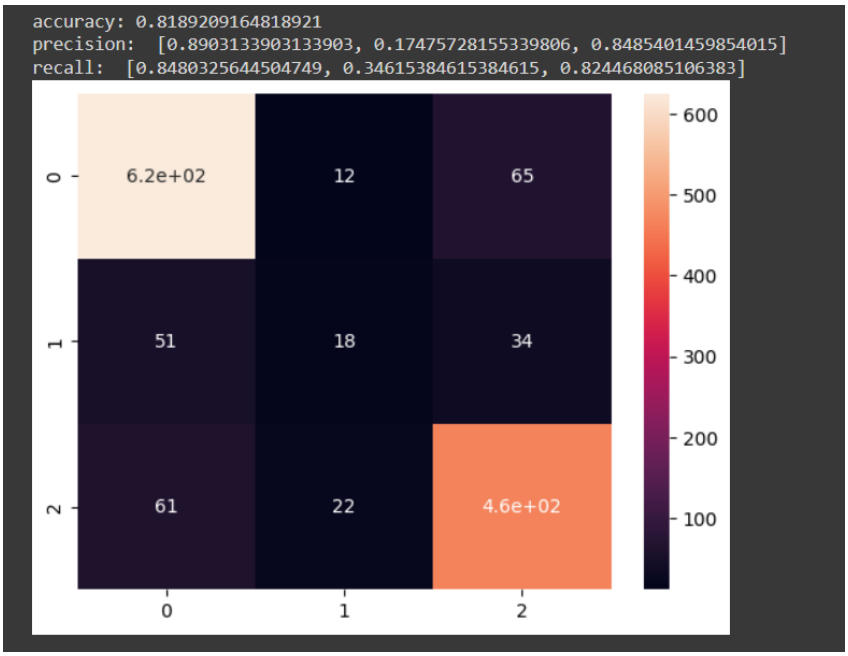

Figure 8: Adaline Performance metrics with accuracy percentage.

IV. RESULTS AND DISCUSSION

In our investigation of numerous machine learning algorithms for the problem of phishing website detection, we examined the performance of K-Nearest Neighbours (KNN), Decision Tree, Single-layer Perceptron, Multilayer Perceptron, Naive Bayes, and Adaline. Accuracy, precision, recall, and F1 score were among the evaluation criteria employed. The Multilayer Perceptron was the most successful of these algorithms, with an accuracy of 96.00%. This result highlights the model's ability to recognise subtle trends in data, resulting in superior categorization of real and phishing websites. While the Multilayer Perceptron demonstrated high accuracy, model selection should also consider problems such as compute economy and interpretability.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:05/Issue:11/November-2023 Impact Factor- 7.868 www.irjmets.com

Furthermore, the Multilayer Perceptron's success illustrates the importance of feature engineering in boosting the model's ability to catch nuanced connections indicative of phishing activity. Despite these hopeful findings, it is crucial to acknowledge the study's limitations and consider future research avenues, such as evaluating additional feature sets and hyperparameter tuning to improve the performance of phishing website detection algorithms. To recap, our findings demonstrate that the Multilayer Perceptron has a lot of potential for this purpose, but choosing the optimum strategy for different applications requires careful consideration of trade-offs.

Table 1. Comparison of algorithms with Accuracy, Precision, Recall

Sl No.	Model	Accuracy	Precision	Recall
1	KNN	0.6885	0.60101	0.6785
2	Decision Tree	0.8753	0.87541	0.8943
3	Naïve Bayes	0.8154	0.8029	0.8098
4	Single layer	0.8523	0.8430	0.7748
5	Multilayer	0.9600	0.9704	0.9405
6	Adaline	0.8240	0.6129	0.6304

V. CONCLUSION

The age of distributed systems, cloud computing, and open-source networks has arrived in information technology. People are encountering increasing security issues in identifying genuine sites from unlawful sites since the introduction of the World Wide Web. Security is a critical issue in online browsing since it has a significant impact on the numerous security methods used to prevent the issue in the first place. Intrusion detection is a critical component of online phishing protection. Thus, by locating the appropriate dataset and applying feature extraction to the model, we may construct an ML, DL, and evolutionary algorithm that best suits the cause of the problem.

Then we may do a comparison study to see how accurate we were in classifying these models. We may employ the Multilayer Perceptron and MLP Regressor now that we know their accuracies are greater than 96. From these, we may create and install an online model that assists users in distinguishing genuine sites from phishing sites, as well as report the sites so that we have a big repository of phishing sites that can be avoided at any time.

VI. REFERENCES

[1] Patil, S., & Dhage, S. (2019). A methodical overview on phishing detection along with an organized way to construct an anti-phishing framework. 2019 5th International Conference on Advanced Computing & Communication Systems (ICACCS). https://doi.org/10.1109/icaccs.2019.8728356

[2] Koreddi, V., Juluru, B. S., Gandipudi, S., Prakash, P. G., Jannegorla, V., & Kalavakuri, C. S. (2023). Phishing detection a predictive model for cyber security. 2023 2nd International Conference on Applied Artificial Intelligence and Computing (ICAAIC). https://doi.org/10.1109/icaaic56838.2023.10141235

[3] Parekh, S., Parikh, D., Kotak, S., & Sankhe, S. (2018). A new method for detection of phishing websites: URL detection. 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT). https://doi.org/10.1109/icicct.2018.8473085

[4] Sindhu, S., Patil, S. P., Sreevalsan, A., Rahman, F., & N., Ms. S. (2020). Phishing detection using random forest, SVM and neural network with backpropagation. 2020 International Conference on Smart Technologies in Computing, Electrical and Electronics (ICSTCEE). https://doi.org/10.1109/icstcee49637.2020.9277256

[5] Liu, G., Qiu, B., & Wenyin, L. (2010). Automatic detection of phishing target from phishing webpage. 2010 20th International Conference on Pattern Recognition. https://doi.org/10.1109/icpr.2010.1010