Identification of Phishing URL’s using Machine Learning.

Nithin Reddy Masireddy[1], Konda Ankitha[2] , Vijaykumar Mantri , M.Tech(Ph.D.)[3]

[1]*Student, Information Technology, B.V Raju Institute of Technology, Telangana, India*
[2] *Student, Information Technology, B.V Raju Institute of Technology, Telangana, India*
[3]*Associate Professor, Information Technology, B.V Raju Institute of Technology, Telangana, India*

ABSTRACT

The growth of the Internet has brought network security to the public's notice. One may argue that the foundation for the Internet's quick and safe development is a secure network environment. Phishing is a crucial subset of cybercrime that involves deceiving consumers into clicking on dangerous links, obtaining their personal information, and then utilising that information to pretend to log into associated accounts in order to steal money. Attack and defence are iterative problems in network security. Both phishing techniques and phishing detection technologies are continually being improved. Blacklists and whitelists are the foundation of conventional techniques for detecting phishing links, however these cannot detect newly created phishing connectionsAs a result, we must figure out how to determine whether a recently discovered link is a phishing website and increase the prediction's precision. Prediction has grown in importance as machine learning technology has matured. This article gives techniques for phishing threat assessments for websites. It begins with a discussion of the phishing life cycle, moves on to a discussion of common anti-phishing techniques, primarily focuses on identifying phishing links, and concludes with a solution that includes data collection and a thorough understanding of machine learning based on feature extraction, modelling, and performance evaluation.. This paper provides a detailed comparison of various solutions for phishing website detection. We. Compare machine learning models like logistic regression, random forest, decision tree, Xgboost, and KNeighbors and identify the efficient model among them. Identifying efficient algorithms can help use highly efficient algorithms instead of testing separately on each algorithm.

Keyword: - *malicious URLs, phishing, neural networks, sensitive data etc.....*

1. Introduction

The Internet has integrated seamlessly into people's daily lives. A world without the Internet is inconceivable. There are 4.66 billion active Internet users globally, or 59.5% of the world's population, according to the January 2021 global digital population survey. Among these, smartphones are used by 92.6% of users to access the Internet. How individuals communicate, shop, talk, and conduct business online has fundamentally altered how people live and work. Several conventional businesses, including catering and retail, have switched from physical to online services due to the pandemic that began at the end of 2019. Several sensitive pieces of information, including usernames, account names, passwords, privacy questions, personal details, and credit card numbers, have been placed on the Internet by users. Cybercriminals utilize various unlawful methods to collect this information, then exploit these individuals as fronts for their illicit online operations. Network security problems have existed since the beginning of the Internet. Network security has faced several difficulties due to the Internet's fast expansion and the quick shift in network attack methods. Cybersecurity challenges are typically grouped into four categories:

- Denial-of-service assaults (DoS)
- Man-in-the-middle attacks (MitM)
- SQL injection attacks
- Zero-day exploits

The riskiest illegal activity on the internet is phishing. As the great majority of customers use the internet to access the services provided by governmental and financial institutions, phishing attacks have increased significantly over the past few years. Phishers started obtaining money, and they are doing this as a successful business. Phishers use a variety of methods to attack the helpless victims, including information, VOIP, mocked connections, and false websites. Making phoney websites that, in terms of structure and content, mimic genuine websites is by no means impossible. In fact, there would be no way to tell these sites from their actual websites in terms of content.

Phishing pretends to be a genuine source while impersonating an email address. This deceives users into clicking on links in phishing emails to visit phished websites. By sending alert messages requesting users to authenticate their accounts, phishers will deceive users into entering their personal information. This is done to make the user believe that it is a task that must be completed on their end. According to the Anti-Phishing Working Group, phishing is the theft of user credentials for bank accounts and identity data. In the APWG 1Q report for the first quarter of 2018, 263538 phishes were recorded. This is 46 percent less than 180577, as of the fourth quarter of 2017 [1].

Using a Universal Resource Locator as a vector, users are caught (URL). Before clicking a URL, the user must do sanity tests, such as paying close attention to how the website's address is spelled and weighing the potential risks of doing so [3]. Recent research has demonstrated that security professionals have created strategies, such blacklisting, to shield consumers from phishing websites. A third party gathers the names of well-known phishing websites when blacklisting is used. The method does not offer full security because no blacklist can be comprehensive and up to date, while having a low query cost. As a consequence, before the link turns up on a blacklist, the user could click on it to go to a phishing website [3].

The term "website phishing," which is a play on the word "fishing," is where the word "phishing" originates. The concept is tossing out bait in the hopes that someone would take it and eat it like the fish. Most often, the bait is an email or an instant messaging service that directs the victim to dangerous phishing websites. Phishing assaults increased in volume and ferocity over time. Users of online banking, payment systems like PayPal, and e-commerce websites are now the focus of phishing assaults. The term "website phishing" is a play on the word "fishing" that gives rise to the word "phishing." In the hopes that someone will pick it up and eat it like the fish, bait is tossed out into the water. The user will often be taken to malicious phishing websites by the bait, which is either an email or an instant messaging service. The frequency and severity of phishing attempts increased with time. Phishing scams now target those who utilise online banking, payment systems like PayPal, and e-commerce websites [4]. This paper discusses various machine learning algorithms and finds an efficient algorithm to detect phishing URLs. We will discuss algorithms as we have mentioned.

2. Existing System

As new phishing strategies are implemented, the old system of phishing detection techniques experiences low detection accuracy and high false alarm rates.

• After that, the most popular solution is the blacklist-based method, which is ineffective at stopping phishing attempts since it has become simpler to register new domains. No exhaustive blacklist can provide a flawless up-to-date database for phishing detection.

Disadvantages of Existing system

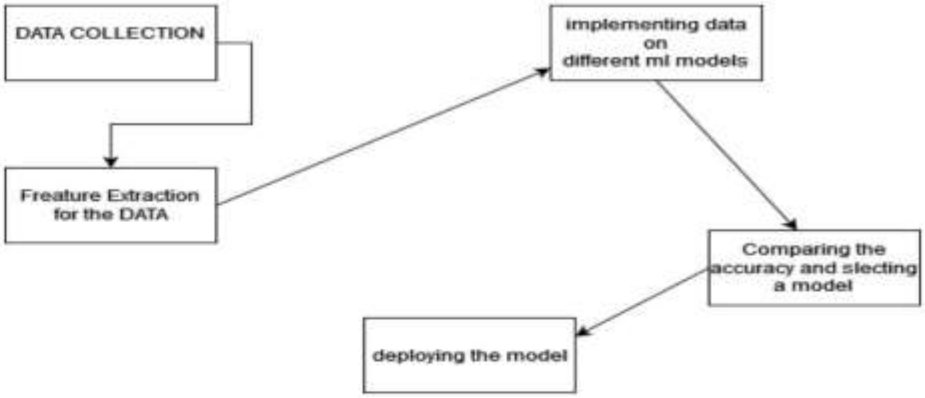
• It can’t predict rare and new phishing URLs.

• Blacklist tools cannot predict all phishing URLs.

3. Proposed System:

To make precise predictions on URLs to classify them as legitimate or phishing

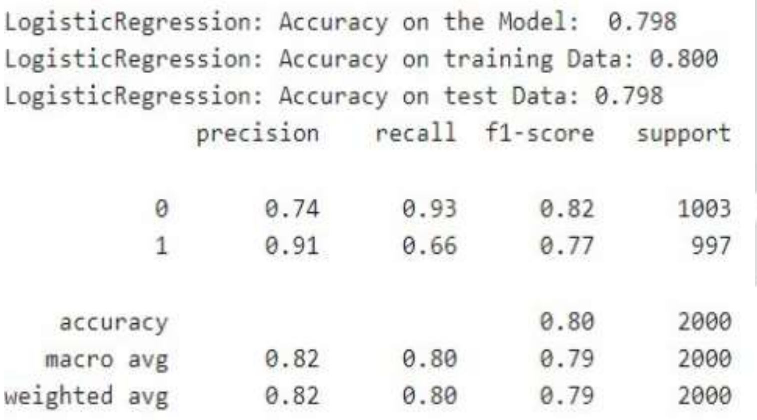
- To implement different models and select the best possible model
- To club the models or use ensemble learning to make better predictions


DATA COLLECTION
Freature Extraction for the DATA
implementing data on different ml models
Comparing the accuracy and slecting a model
deploying the model

4. Performance of Algorithms:

4.1 Logistic Regression Algorithm

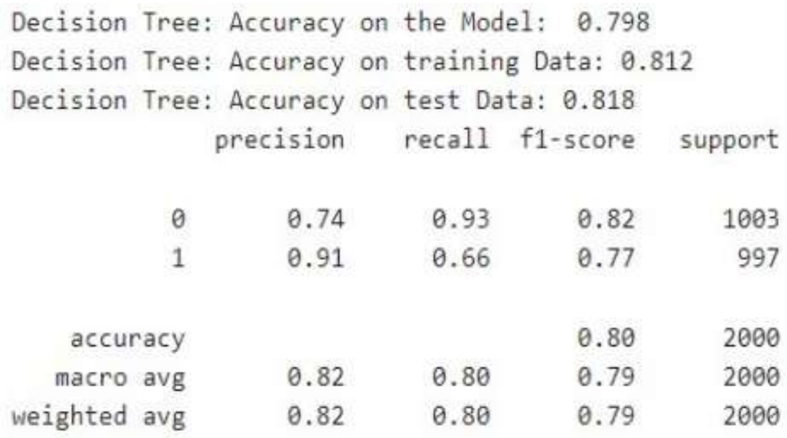
- One of the most often used Machine Learning algorithms, within the category of Supervised Learning, is logistic regression.With a predetermined set of independent factors, it is used to predict the categorical dependent variable.



```
LogisticRegression: Accuracy on the Model:  0.798
LogisticRegression: Accuracy on training Data: 0.800
LogisticRegression: Accuracy on test Data: 0.798
              precision    recall  f1-score   support

           0       0.74      0.93      0.82      1003
           1       0.91      0.66      0.77       997

    accuracy                           0.80      2000
   macro avg       0.82      0.80      0.79      2000
weighted avg       0.82      0.80      0.79      2000
```

4.2 Decision Tree algorithm

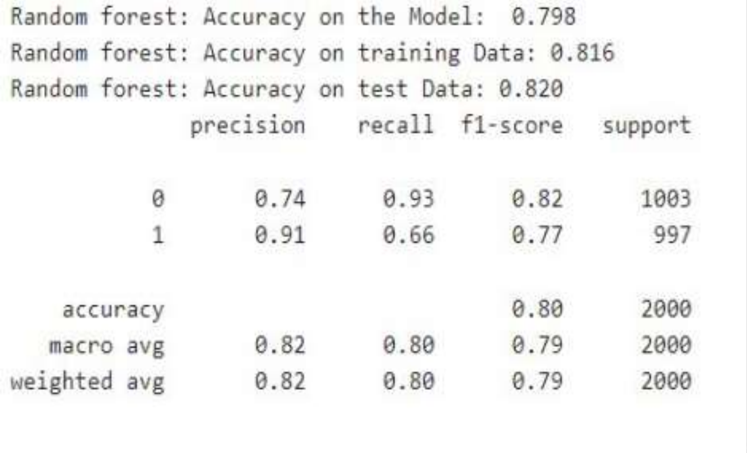
A supervised learning method called a decision tree may be used to solve classification and regression issues, but it is often favored for doing so. It is a tree-structured classifier, where internal nodes stand in for a dataset's characteristics, branches for the decision-making process, and each leaf node for the classification result**.**



```
Decision Tree: Accuracy on the Model:  0.798
Decision Tree: Accuracy on training Data: 0.812
Decision Tree: Accuracy on test Data: 0.818
              precision    recall  f1-score   support

           0       0.74      0.93      0.82      1003
           1       0.91      0.66      0.77       997

    accuracy                           0.80      2000
   macro avg       0.82      0.80      0.79      2000
weighted avg       0.82      0.80      0.79      2000
```

4.3 Random Forest Algorithm

favoured algorithm for machine learning A component of the supervised learning approach is Random Forest. It may be used to solve classification and regression-related ML problems. It is based on the concept of ensemble learning, a technique for combining several classifiers to solve complex problems and improve model performance..

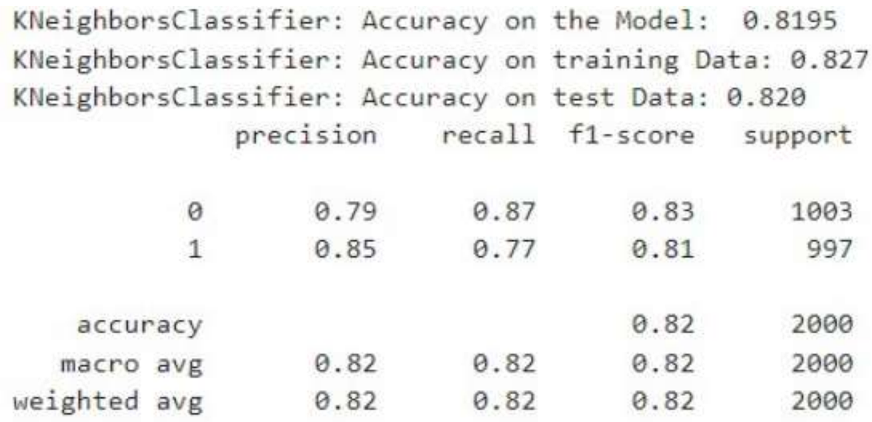


```
Random forest: Accuracy on the Model:  0.798
Random forest: Accuracy on training Data: 0.816
Random forest: Accuracy on test Data: 0.820
              precision    recall  f1-score   support

           0       0.74      0.93      0.82      1003
           1       0.91      0.66      0.77       997

    accuracy                           0.80      2000
   macro avg       0.82      0.80      0.79      2000
weighted avg       0.82      0.80      0.79      2000
```

4.4 KNN Algorithm

The K-NN method makes the assumption that the new case and the existing cases are comparable, and it places the new instance in the category that is most like the existing categories.

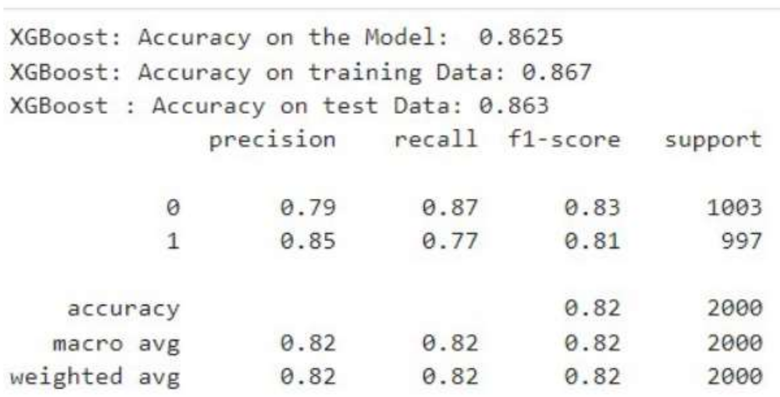
A new data point is classified using the K-NN algorithm based on similarity after all the existing data has been stored. This implies that utilising the K- NN method, fresh data may be quickly and accurately sorted into a suitable category.



```
KNeighborsClassifier: Accuracy on the Model:  0.8195
KNeighborsClassifier: Accuracy on training Data: 0.827
KNeighborsClassifier: Accuracy on test Data: 0.820
              precision    recall  f1-score   support

           0       0.79      0.87      0.83      1003
           1       0.85      0.77      0.81       997

    accuracy                           0.82      2000
   macro avg       0.82      0.82      0.82      2000
weighted avg       0.82      0.82      0.82      2000
```

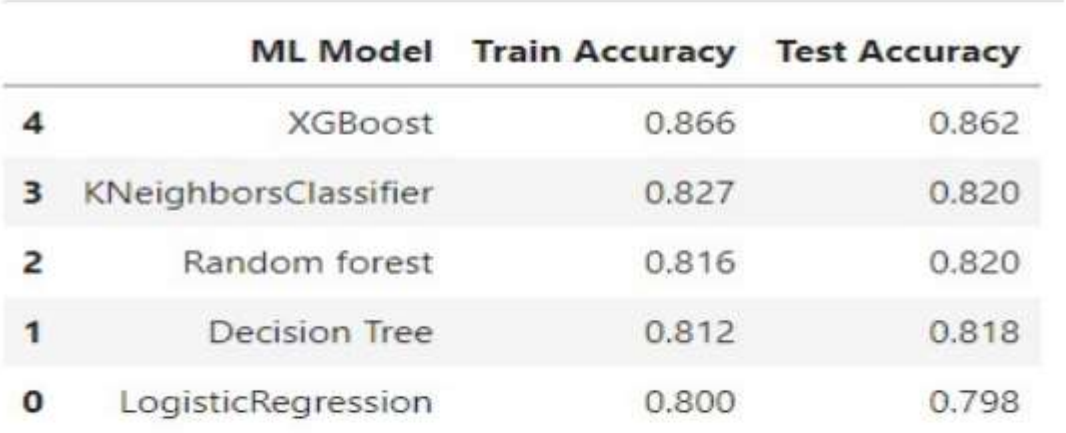
4.4 Xg Boost Algorithm

One of the most widely used machine learning algorithms nowadays is XGBoost. Regardless of the kind of prediction job at hand—classification or regression—XGBoost stands for extreme gradient boosting. A gradient boosted decision tree implementation created for speed and performance is called XGBoost.
Using random bootstrap samples of the data set, Random Forest builds entire decision trees concurrently using a method known as bagging. All of the decision tree forecasts are averaged to provide the final prediction**.**



```
XGBoost: Accuracy on the Model:  0.8625
XGBoost: Accuracy on training Data: 0.867
XGBoost : Accuracy on test Data: 0.863
              precision    recall  f1-score   support

           0       0.79      0.87      0.83      1003
           1       0.85      0.77      0.81       997

    accuracy                           0.82      2000
   macro avg       0.82      0.82      0.82      2000
weighted avg       0.82      0.82      0.82      2000
```

5. Results



	ML Model	Train Accuracy	Test Accuracy
4	XGBoost	0.866	0.862
3	KNeighborsClassifier	0.827	0.820
2	Random forest	0.816	0.820
1	Decision Tree	0.812	0.818
0	LogisticRegression	0.800	0.798

As we observe in the above figure, we can see the performance and efficiency of each algorithm. According to the accuracy, we can conclude that XG boost is the most efficient for performing this task.

6. Conclusion

This study offers a comparison of machine learning methods for predicting URLs. The main objective is to provide security and keep the user from accessing sensitive data. Machine learning algorithms may be used to identify whether a website is trustworthy or not. By comparing XGboost Classifier to other models in the research, we discovered that it has a high accuracy thanks to its 16 characteristics.

7. REFERENCES

[1]. Alkawaz, M. H., Steven, S. J., Hajamydeen, A. I., & Ramli, R. (2021). *A Comprehensive Survey on Identification and Analysis of Phishing Website based on Machine Learning Methods. 2021 IEEE 11th IEEE Symposium on Computer Applications & Industrial Electronics (ISCAIE).* doi:10.1109/iscaie51753.2021.9431794
[2]. Kiruthiga, R., and D. Akila. "Phishing websites detection using machine learning." *International Journal of Recent Technology and Engineering* 8.2 (2019): 111-114.
[3]. Feroz, M. N., & Mengel, S. (2015). *Phishing URL Detection Using URL Ranking. 2015 IEEE International Congress on Big Data.* doi:10.1109/bigdatacongress.2015.97
[4]. Parekh, S., Parikh, D., Kotak, S., & Sankhe, P. S. (2018). *A New Method for Detection of Phishing Websites: URL Detection. 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT).* doi:10.1109/icicct.2018.8473085