IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 3, May 2022
Impact Factor: 6.252

Prediction of Phishing Websites Using Machine Learning

Revathi Pandi[1], Shruthi Suresh[2], Shruthi Siva[3], Dr. Kumaresan G[4]
Students, Department of Computer Science and Engineering[1,2,3]
Assistant Professor, Department of Computer Science and Engineering[4]
SRM Valliammai Engineering College, Chengalpattu, India

Abstract: *The Internet has become an indispensable part of our life, However, It also has provided opportunities to anonymously perform malicious activities like Phishing. Phishers try to deceive their victims by social engineering or creating mock-up websites to steal information such as account ID, username, password from individuals and organizations. Although many methods have been proposed to detect phishing websites, Phishers have evolved their methods to escape from these detection methods. One of the most successful methods for detecting these malicious activities is Machine Learning. This is because most Phishing attacks have some common characteristics which can be identified by machine learning methods. In this paper, we compared the results of multiple machine learning methods for predicting phishing websites.*

Keywords: Phishing, Websites, Machine Learning, Logistic Regression, Random Forest, Decision tree classifier, Naive Bayes Algorithm, SVM, KNN.

I. INTRODUCTION

Phishing is a fraudulent method that makes use of social and technological hints to steal purchaser identity and financial credentials. Social media systems use spoofed e-mails from valid businesses and groups to permit customers to apply fake web sites to disclose info like usernames and passwords [6]. Hackers set up malicious software program on computer systems to scouse borrow credentials, frequently the use of systems to intercept username and passwords of consumers' on-line accounts. Phishers use more than one methods, together with email, Uniform Resource Locators (URL), instant messages, forum postings, cellphone calls, and textual content messages to steal consumer records. The structure of phishing content material is much like the unique content material and trick customers to get admission to the content material that allows you to gain their sensitive data. The number one goal of phishing is to benefit private records for financial benefit or use of identification theft. Phishing assaults are causing intense economic harm across the world. Moreover, Most phishing attacks goal financial establishments and webmail, consistent with the Anti-Phishing Working Group (APWG) today's Phishing sample studies [6].

There had been numerous current research in opposition to phishing primarily based totally at the traits of a domain, including website URLs, internet site content, incorporating each the website URLs and content, the source code of the internet site and the screenshot of the website [7]. However, there may be a loss of beneficial anti phishing tools to locate malicious URL in an organisation to defend its users. In the occasion of malicious code being implanted at the website, hackers can also additionally thieve person facts and install malware, which poses a extreme danger to cybersecurity and person privacy. Malicious URLs at the Internet may be without problems recognised by reading it via Machine Learning (ML) technique [8]. Day by day the cyber security techniques alternate and evolve so one can guard and shield the computer systems from intruders or hackers. On the opposite hand, as the safety techniques evolve, intruders or hackers are evolving and unfolding themselves to increase an increasing number of superior security techniques to evade the respective included systems. With the assist of device gaining knowledge of methods, the cybersecurity structures analyses the numerous information styles and therefore study and find out from the education and evaluation. This respective education and evaluation aids the cyber protection structures to hinder comparable type of net assaults. Another massive thing of machine learning scheme is to react to the changing behaviour of the web attacks in much less amount of time[9]. The motivation at the back of this study is to create a resilient and powerful technique that makes use of Data Mining algorithms and equipment to detect e-banking phishing web sites in an Artificial Intelligent technique. Associative and type algorithms may be very beneficial in predicting Phishing web sites[11].

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 3, May 2022
Impact Factor: 6.252

II. LITERATURE

A literature review is a body of text that aims to review the critical points of current knowledge on and/or methodological approaches to a particular topic. It is secondary sources and discuss published information in a particular subject area and sometimes information in a particular subject area within a certain time period. Its ultimate goal is to bring the reader up to date with current literature on a topic and forms the basis for another goal, such as future research that may be needed in the area and precedes a research proposal and may be just a simple summary of sources. Usually, it has an organisational pattern and combines both summary and synthesis.

A summary is a recap of important information about the source, but a synthesis is a re-organisation, reshuffling of information. It might give a new interpretation of old material or combine new with old interpretations or it might trace the intellectual progression of the field, including major debates. Depending on the situation, the literature review may evaluate the sources and advise the reader on the most pertinent or relevant of them.

Wentao Zhao, Jianping Yin, Jun Long.[1] says that the process of prediction analysis is a process of using some method or technology to explore or stimulate some unknown, undiscovered or complicated intermediate processes based on previous and present states and then speculated the results . In an early warning system, accurate prediction of DoS attacks is the prime aim in the network offence and defense task. Detection based on abnormity is effective to detect DoS attacks.

Jinyu W, Lihua Yin and Yunchuan Guo.[2]says that the prediction results reflect the security situation of the target network in the future, and security administrators can take corresponding measures to enhance network security according to the results. To quantitatively predict the possible attack of the network in the future, attack probability plays a significant role. It can be used to indicate the possibility of invasion by intruders. As an important kind of network security quantitative evaluation measure, attack probability and its computing methods has been studied for a long time.

XiaoyongYuan, Pan He, Qile Zhu, and Xiaolin Li.[3]has reviewed the recent findings of adversarial examples in DNNs. We investigated the existing methods for generating adversarial examples. A taxonomy of adversarial examples was proposed. We also explored the applications and countermeasures for adversarial examples. This paper attempted to cover the state-of-the-art studies for adversarial examples in the DL domain. Compared with recent work on adversarial examples, we analyzed and discussed the current challenges and potential solutions in adversarial examples.

Zhen Yang, YaochuJin, Fellow, and Kuangrong Hao, Member.[4] has the ultimate goal of the Internet of Things (IoT) is to provide ubiquitous services. To achieve this goal, many challenges remain to be addressed. BSCA consists of three layers. The first layer is composed of multiple subpopulations evolving cooperatively to obtain diverse Pareto fronts. Based on the solutions obtained by the first layer, the second layer aims to further increase the diversity of solutions. The third layer refines the solutions found in the second layer by adopting an adaptive gradient refinement search strategy and dynamic optimisation method to cope with changing concurrent multiple service requests, thereby effectively improving the accuracy of solutions. may select one of the extreme solutions or other Pareto optimal solutions on the front according to the service strategy specified by the decision-maker.

Seraj Fayyad, CristophMeinel.[5] says that in this paper we propose a real time prediction methodology for predicting most possible attack steps and attack scenarios. Proposed methodology benefits from attacks history against network and from attack graph source data. it comes without considerable computation overload such as checking of attack plans library. It provides parallel prediction for parallel attack scenarios.

A novel anti-phishing methodology [12] that employs a training intercession for detecting the phishing web sites is proposed and evaluated. The proposed technique aids the users in making the proper choices in differentiating the valid web sites in addition to the phishing web sites.

A straightforward ensemble cataloguing system [13] is proposed to amalgamate the anticipated final results from numerous phishing detection category methods.

III. PROPOSED SYSTEM

The proposed version is to construct a machine learning version for anomaly detection. Anomaly detection is an critical method for spotting fraud activities, suspicious activities, community intrusion, and different unusual activities which can have amazing importance however are hard to detect. The machine learning version is constructed through making use of right data technology strategies like variable identity this is the established and unbiased variables. Then the visualisation of the statistics is accomplished to insights of the statistics. The version is construct base that the preceding dataset wherein

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 3, May 2022
Impact Factor: 6.252

the algorithm learn statistics and get educate done of a kind algorithms are used for higher comparisons. The overall performance metrics are calculated and as compared.

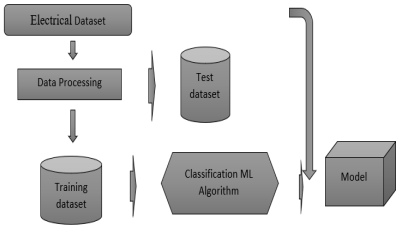


Figure 1: Architecture Of Proposed Model

3.1 Preparing the Data

The phishing trouble is taken into consideration a essential difficulty enterprise specifically e-banking and e-trade taking the quantity of on-line transactions related to payments.

We have diagnose done of a kind capabilities associated with valid and phishy web sites and accrued 1353 one of a kind web sites from distinction sources. Phishing web sites have been accumulated from Phish tank data archive that is a loose network site where in customers can submit, verify, track and proportion phishing data. The valid web sites have been accumulated from Yahoo and start line directories the usage of an internet script evolved in PHP. The PHP script turned into plugged with a browser and we accumulated 548 valid web sites out of 1353 web sites. There is 702 phishing URLs, and 103 suspicious URLs.

When a internet site is taken into consideration SUSPICIOUS which mean sit is able to be both phishy or legitimate, which means the website held a few legit and phish functions.

- Attribute Information:
- URL Anchor
- Request URL
- SFH
- URL Length
- Having
- Prefix/Suffix
- IP
- Sub Domain
- Web traffic
- Domain age
- Class

Accumulated capabilities maintain the explicit values, Legitimate, Suspicious and Phishy, those values were changed with numerical values 1,zero and -1 respectively.

3.2 System Architecture

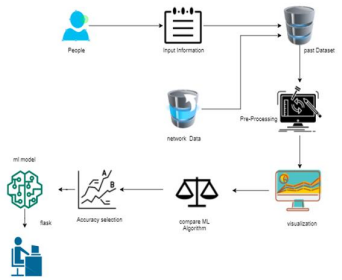


Figure 2: Detailed Description of System Architecture

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 3, May 2022
Impact Factor: 6.252

3.3 Software Requirements

A. Anaconda Navigator

Anaconda Navigator is a computer graphical consumer interface (GUI) blanketed in Anaconda® distribution that allows you to release packages and without problems control conda applications, environments, and channels without the use of command-line instructions. Navigator can appearance for applications on Anaconda.org or in a community Anaconda Repository. Navigator is an easy, point-and- click on on manner to artwork with programs and environments even as now no longer having to kind conda instructions in a terminal window. You can use it to find out the programs you need, installation them in an surroundings, run the programs, and replace them – all inner Navigator.

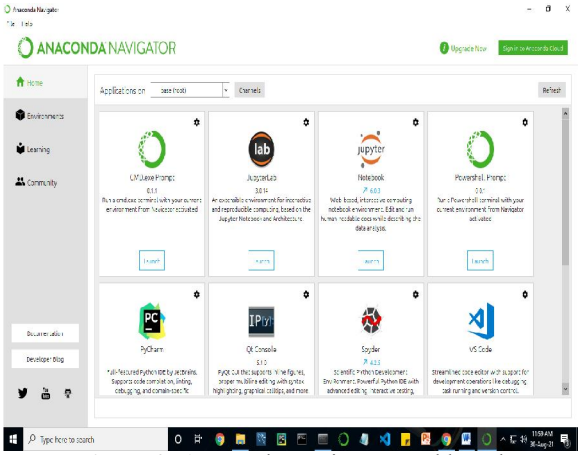


Figure 3: Anaconda Navigator Dashboard

B. Jupyter Notebook

This website acts as “meta” documentation for the Jupyter ecosystem. It has a fixed of assets to navigate the gear and group son this ecosystem, and to help you get started. Notebook files are files produced via way of means of the Jupyter Notebook App, which include each laptop code (e.g. python) and wealthy textual content material elements (paragraph, equations, figures, links, etc…). Notebook files are every human-readable documents containing the assessment description and the results (figures, tables, etc.) similarly to executable documents which can be run to hold out statistics evaluation. Installing the Python anaconda platform.

1. Loading the dataset.
2. Summarizing the dataset.
3. Visualizing the dataset.
4. Evaluating some algorithms.
5. Making some predictions.

C. Python

Python is an interpreted high-diploma general-motive programming language. Its format philosophy emphasises code readability with its use of great indentation. Its language constructs further to its object-orientated approach purpose to help programmers write clear, logical code for small and massive-scale projects. Python is dynamically-typed and garbage-collected. It helps a couple of programming paradigms, which include structured (particularly, procedural), object-oriented and sensible programming. It is regularly defined as a "batteries included" language due to its complete preferred library.

3.4 Module 1 – Data Pre-Processing

A wide variety of various information cleaning tasks the use of Python’s Pandas library and specifically, it has consciousness on in all likelihood the largest information cleansing task, missing values, and it is capable of extra quick clean information. It desires to spend much less time cleansing information, and extra time exploring and modelling.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 3, May 2022
Impact Factor: 6.252

It's vital to apprehend those exclusive sorts of lacking information from a data point of view. The kind of lacking information will impact the way to address filling with inside the lacking values and to locate lacking values, and do a little simple imputation and targeted statistical technique for coping with lacking information.

Figure 4: Data Flow Of Data Preprocessing Module

3.5 Module 2 - Exploration Information Evaluation Of Visualisation

Data visualisation is an essential ability in carried out information and device learning. Statistics does certainly focus on quantitative descriptions and estimations of information. Data visualisation provides an essential suite of tools for gaining a qualitative understanding.

With a bit of area knowledge, information visualisations may be used to explicit and display key relationships in plots and charts which are extra visceral and stakeholders than measures of affiliation or significance.

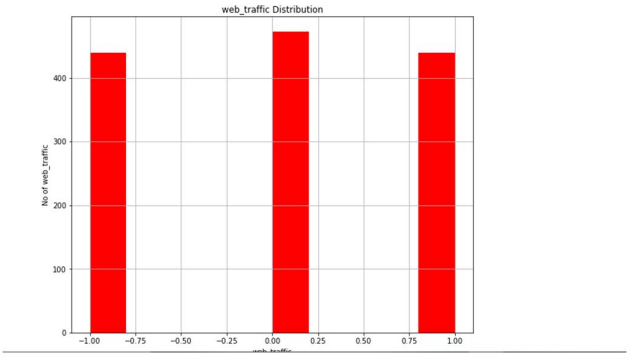


Figure 5: Comparison of Web_Traffic and no of Web_Traffic

3.6 Module 3 - Logistic Regression

- In different words, the logistic regression version predicts P(Y=1) as a characteristic of X. Logistic regression Assumptions:
- Binary logistic regression calls for the established variable to be binary.
- For a binary regression, the factor stage 1 of the established variable have to constitute the favored final results
- Only the significant variables have to be included.
- The independent variables have to be independent of every different. That is, the version have to have little.
- The independent variables are linearly associated with the log odds. Logistic regression calls for pretty massive pattern sizes.

```
Classification report of Logistic Regression Results:

              precision    recall  f1-score   support

          -1       0.82      0.91      0.86       211
           0       0.56      0.16      0.25        31
           1       0.81      0.82      0.81       164

    accuracy                           0.81       406
   macro avg       0.73      0.63      0.64       406
weighted avg       0.80      0.81      0.80       406


Confusion Matrix result of Logistic Regression is:
 [[191   1  19]
 [ 14   5  12]
 [ 27   3 134]]

Sensitivity :  0.9947916666666666

Specificity :  0.2631578947368421

Cross validation test results of accuracy:
[0.79335793 0.82287823 0.8302583  0.84814815 0.84814815]

Accuracy result of Logistic Regression is: 82.85581522481891
```

Figure 6: Confusion Matrix Result of Logistic Regression

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 3, May 2022
Impact Factor: 6.252

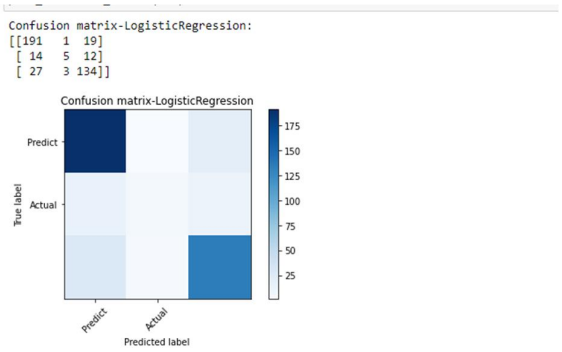


Figure 7: Comparison of Predicted Label and True Label

3.7 Module 4 - Random Forest Classifier

- Pick N random information from the dataset.
- Build a choice tree primarily based totally on those N information.
- Choose the quantity of bushes you need for your set of rules and repeat steps 1 and 2.

In case of a regression problem, for a brand new document, every tree with inside the forest predicts a fee for Y (output). The very last fee may be calculated with the aid of using taking the common of all of the values expected with the aid of using all of the bushes in forest. Or, in case of a class problem, every tree with inside the forest predicts the class to which the brand new document belongs. Finally, the brand new document is assigned to the class that wins the bulk vote.

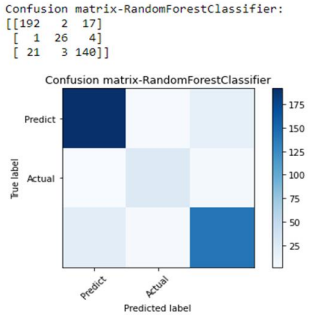


Figure 8: Comparison of Predicted Label And True Label

3.8 Module 5 - Decision Tree Classifier

- It is one of the most powerful and well-known set of rules. Decision-tree set of rules falls under the class of supervised getting to know algorithms. It works for every non-forest all similarly to specific output variables.

Assumptions of Decision tree:

- At the beginning, we remember the whole education set due to the fact the root.
- Attributes are assumed to be specific for statistics advantage, attributes are assumed to be non-forestall.
- On the idea of function values information are allocated recursively.
- We use statistical techniques for ordering attributes as root or internal node.
- Decision tree builds kind or regression models with inside the form of a tree structure.
- This approach is continued on the education set until meeting a termination condition.
- A choice tree can be results easily over-equipped generating too many branches and may mirror anomalies due to noise or outliers.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 3, May 2022
Impact Factor: 6.252

```
Classification report of Decision Tree Classifier Results:

              precision    recall  f1-score   support

          -1       0.86      0.91      0.88       211
           0       0.77      0.77      0.77        31
           1       0.87      0.80      0.83       164

    accuracy                           0.85       406
   macro avg       0.83      0.83      0.83       406
weighted avg       0.86      0.85      0.85       406


Confusion Matrix result of Decision Tree Classifier is:
 [[192   2  17]
 [  4  24   3]
 [ 28   5 131]]

Sensitivity :  0.9896907216494846

Specificity :  0.8571428571428571

Cross validation test results of accuracy:
[0.85239852 0.87453875 0.88191882 0.88518519 0.88888889]

Accuracy result of Decision Tree Classifier is: 87.6586032526992
```

Figure 9: Accuracy Result of Decision Tree Classifier

3.9 Module 6 - Naive Bayes Algorithm

The Naive Bayes algorithm is an intuitive technique that makes use of the chances of every characteristic belonging to every elegance to make a prediction. It is the supervised mastering technique you will give you in case you desired to version a predictive modelling hassle probabilistically.

Naive Bayes simplifies the calculation of chances with the aid of using assuming that the chance of every characteristic belonging to a given elegance cost is impartial of all different attributes. This is a sturdy assumption however outcomes in a quick and powerful technique.

The chance of a category cost given a cost of an characteristic is referred to as the conditional chance. By multiplying the conditional chances together for every characteristic for a given elegance cost, we've a chance of a information example belonging to that elegance. To make a prediction we will calculate chances of the example belonging to every elegance and pick the elegance cost with the very best chance.

Even if those capabilities are interdependent, those capabilities are nevertheless taken into consideration independently. This assumption simplifies computation.

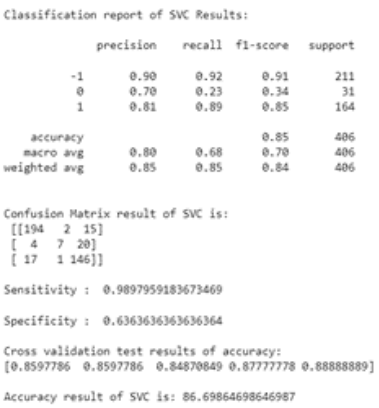
```
Classification report of Naive Bayes Results:

              precision    recall  f1-score   support

          -1       0.83      0.88      0.85       211
           0       0.36      0.16      0.22        31
           1       0.79      0.80      0.80       164

    accuracy                           0.80       406
   macro avg       0.66      0.62      0.62       406
weighted avg       0.78      0.80      0.78       406


Confusion Matrix result of Naive Bayes is:
 [[186   3  22]
 [ 13   5  13]
 [ 26   6 132]]

Sensitivity :  0.9841269841269841

Specificity :  0.2777777777777778

Cross validation test results of accuracy:
[0.79335793 0.80811808 0.83763838 0.82592593 0.82592593]

Accuracy result of Naive Bayes is: 81.81932485991527
```

Figure 10: Accuracy Result Of Naïve Bayes

3.10 Module 7 - Performance Measurements of SVC

- It is powerful in excessive dimensional space.
- It can nevertheless be used whilst the quantity of dimensions exceeds the quantity of samples.
- It is flexible in that one of a kind kernel functions (methods of establishing the selection boundaries) may be distinctive for the selection feature which includes your very own kernel feature

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 3, May 2022
Impact Factor: 6.252



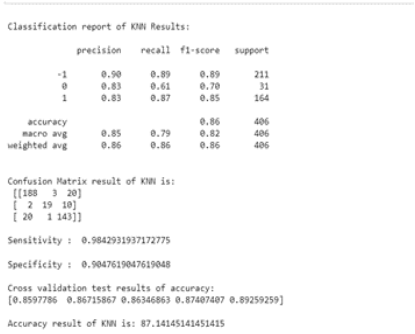
```
Classification report of SVC Results:

              precision    recall  f1-score   support

          -1       0.90      0.92      0.91       211
           0       0.70      0.23      0.34        31
           1       0.81      0.89      0.85       164

    accuracy                           0.85       406
   macro avg       0.80      0.68      0.70       406
weighted avg       0.85      0.85      0.84       406


Confusion Matrix result of SVC is:
 [[194   2  15]
 [  4   7  20]
 [ 17   1 146]]

Sensitivity :  0.9897959183673469

Specificity :  0.6363636363636364

Cross validation test results of accuracy:
[0.8597786  0.8597786  0.84870849 0.87777778 0.88888889]

Accuracy result of SVC is: 86.69864698646987
```

Figure 11: Cross Validation Result Of SVC

3.11 Module 8 - Performance Measurement of KNN

The k-nearest-neighbors algorithm is a supervised classification approach that makes use of proximity as a proxy for 'sameness'. The algorithm takes a group of labelled factors and makes use of them to discover ways to label different factors. To label a brand new point, it appears on the labelled factors closest to that new point. Closeness is commonly expressed in phrases of a dissimilarity function. Once it assessments with 'k' range of nearest buddies, it assigns a label primarily based totally on whichever label the maximum of the buddies have.

Using the geometric distance to determine that is the closes object might not constantly be afford able or maybe possible: the form of the enter may, for example, be text, wherein it isn't always clean how the items are drawn in a geometrical illustration and the way distances ought to be measured. You ought to consequently pick the space metric on a case-by-case basis.



```
Classification report of KNN Results:

              precision    recall  f1-score   support

          -1       0.90      0.89      0.89       211
           0       0.83      0.61      0.70        31
           1       0.83      0.87      0.85       164

    accuracy                           0.86       406
   macro avg       0.85      0.79      0.82       406
weighted avg       0.86      0.86      0.86       406


Confusion Matrix result of KNN is:
 [[188   3  20]
 [  2  19  10]
 [ 20   1 143]]

Sensitivity :  0.9842931937172775

Specificity :  0.9047619047619048

Cross validation test results of accuracy:
[0.8597786  0.86715867 0.86346863 0.87407407 0.89259259]

Accuracy result of KNN is: 87.14145141451415
```

Figure 12: Accuracy Result of KNN

IV. Conclusion

The analytical process began out from facts cleansing and processing, lacking value, exploratory evaluation and eventually version constructing and evaluation. The ML based phishing techniques depend on website functionalities to gather information that can help classify websites for detecting phishing sites. The high-quality accuracy on public take a look at set is better accuracy rating might be locate out. This utility can assist to locate the Prediction of phishing website or not.

References

[1]. Wentao Zhao, Jianping Yin, "A Prediction Model of DoS Attack's Distribution Discrete Probability",2008.
[2]. Jinyu W1, Lihua Yin and Yunchuan Guo, "Cyber Attacks Prediction Model Based on Bayesian Network",2012.
[3]. XiaoyongYuan , Pan He, Qile Zhu, and Xiaolin Li,"Adversarial Examples: Attacks and Defenses for Deep Learning",2019.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 3, May 2022
Impact Factor: 6.252

[4]. Zhen Yang, YaochuJin, Fellow, and Kuangrong Hao , 2018," A Bio-Inspired Self-learning Coevolutionary Dynamic Multi objective",2018.

[5]. Seraj Fayyad, Cristoph Meinel, "New Attack Scenario Prediction Methodology",2013.

[6]. "Anti-Phishing Working Group (APWG) ", https://docs.apwg.org//reports/apwg_trends_report_q4_2019. pdf.

[7]. AlEroud A, Karabatis G, "Bypassing detection of URL-based phishing attacks using generative adversarial deep neural networks. In: Proceedings of the Sixth International Workshop on Security and Privacy Analytics",2020 Mar 16.

[8]. Gupta D, Rani R, "Improving malware detection using big data and ensemble learning", Computer Electronic Engineering, vol. 86, no.106729, 2020.

[9]. E. Sri Vishva and D. Aju, "Phisher Fighter: Website Phishing Detection System Based on URL and Term Frequency-Inverse Document Frequency Values",2021.

[10]. J. Anirudha and P. Tanuja, "Phishing Attack Detection using Feature Selection Techniques ", Proceedings of International Conference on Communication and Information Processing (ICCIP), 2019,

[11]. Maher Aburrous, M. A. Hossain, Keshav Dahal, Fadi Thabtah, "Predicting Phishing Websites using Classification Mining Techniques with Experimental Case Studies",2010.

[12]. Alnajim, A., and Munro, "Ananti-phishing approach that uses training intervention for phishing websites detection",Sixth International Conference on Information Technology: New Generations (pp. 405–410). IEEE,2009.

[13]. Zhuang, W., Jiang, Q., and Xiong, T , "An intelligent anti-phishing strategy model for phishing website detection" ,In 2012 32nd International Conference on Distributed Computing Systems Workshops (pp. 51–56). IEEE,2012.