Turkish Journal of Computer and Mathematics Education *Vol.11 No.01 (2020), 908-915*
DOI: https://doi.org/10.17762/turcomat.v11i1.13589
Research Article

PyCaret based URL Detection of Phishing Websites

Poonam Rani
Department of Computer Science & Information Technology, Graphic Era Hill University, Dehradun Uttarakhand India 248002

Abstract
The primary objective of the research project is to employ machine learning algorithms to conduct studies and identify instances of phished URLs that might direct people to fraudulent websites. The Kaggle repository, which contains more than 11,000 URLs, is where the authors received a phished dataset for this application. Examples of both genuine and phished URL links can be found in the collection. Also, the dataset contains 31 features that must be obtained using feature engineering stages and methodologies. Nevertheless, this dataset is also available as a csv file and has been further pre-processed to remove redundant and pointless data. This is followed by the feature extraction process, which extracts URL properties including domain-based, content-based, and address-based attributes. The implementation of PyCaret follows, with each line of code being in charge of the entire execution. Nonetheless, the testing at this level consists of three parts. In order to create accuracy, the initial stage of PyCaret's implementation includes running 14 built-in algorithms. The top three accuracy-generating algorithms are combined to build a stacking model in the last stage of the system model's implementation, which is divided into two stages. The second stage of the system model implementation entails taking random forest into account. In the conclusion, the accuracy of each algorithm is assessed together with its performance. After comparison, the technique with the highest generating accuracy is considered to be the optimised model.
Keywords: Malicious, Machine Learning, PyCaret, Phishing, URL

Introduction

One of the biggest threats to cybersecurity is the emergence of phishing attacks [1]. This problem mainly arises in the banking sector and on social networking sites where transactions are performed online. The need for these services has increased significantly as technology has advanced. A survey conducted in 2021 revealed a sharp increase in internet usage, with 7.3 percent more subscribers [2]. Also, 59.5 percent of online users are at risk of assaults and breaches due to the current circumstances [3]. It is difficult for an intruder to gain money online by stealing private and sensitive data because of this vulnerability, which also includes information and transactional loss. The information in the link accidentally places the user in an exposure zone, necessitating a quick response on his part. Due of the user's reaction, he is able to click on fraudulent links that ask him to input his login credentials. Now that he has simple access to the user credentials, the hacker can exploit them to steal personal and financial data [4]. The abstracted data is also used by the attackers to perform crimes and extort the target. The factors that make it possible for a user to click on phishing links are listed below:

- Users are provided URL links that lack technical knowledge. Inability to discern between trustworthy and fake websites results. When someone clicks on one of these links, a hidden website with fake information is visited
- Internet users don't know which websites to believe. As a result, the user can log in and enter his personal information on the website
- Because the complete link and address of the webpage are obscured from him due to redirection, the user tries to obtain legitimate access by providing sensitive information during the transactional process
- People don't take the time to confirm and double-check website URLs. When a "http" is commonly used in place of a "https" and covered by it, certain circumstances occur. This subtle change tempts the user to click on these links, which finally fools him and sends him to a fake page

Because of the Internet's broad use and open accessibility, attackers and intrusion can take control of the infrastructure of online transactions. Private user data may also be stolen. This is the main area that needs improvement and serves the purpose of cyberattacks. Such assaults target both knowledgeable users and inexperienced users. Users continue to fall victim to such threats and fraudulent practises despite adopting several precautions. Many of the aforementioned problems are primarily the result of users' ignorance of the distinctions between trustworthy and fraudulent websites. This situation is exploited by hackers and intruders to eventually gain the needed information. The analogy for phishing assaults comes from the word "fishing" for

particular targets. Attacks including phishing have increased in frequency in recent years. The tactics, strategies, and equipment used by attackers and intrusions have improved throughout time. These methods allow phishers to access phoney websites and produce web pages that closely resemble real ones. As a result, the user's focus is drawn away from a reliable website, leaving him vulnerable to visiting such pages by clicking on them and providing the required information. The comparable GUIs and URLs on the new pages commonly deceive people into abandoning the old page. Even someone with technical expertise who frequently struggles to distinguish between the two may fall prey to such assaults. Due to speed and employment demands, the user frequently overlooks the connected links and URLs, which prompts him to visit the website and activate the fraud page. Such web sites are shared and transmitted via emails by building trustworthy links to them. A typical phishing attack is depicted and explained in Figure 1, along with how it is carried out. The perpetrator of the assault will initially construct a fraudulent and fake website.

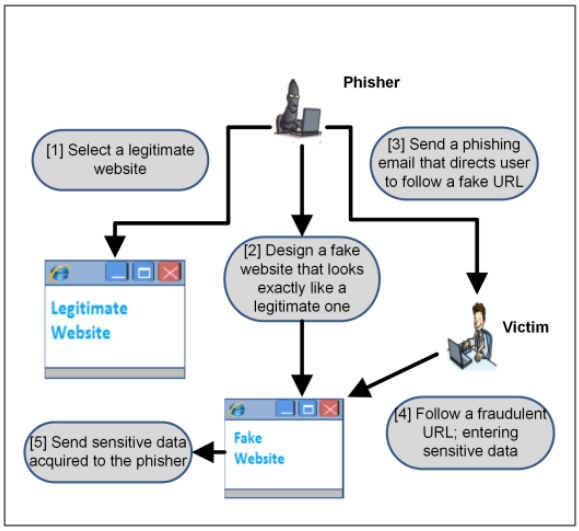


Figure 1: Flow of phishing attacks

This hidden site is the fake site that is used to send links and emails to the real user. The attacker then carefully makes the link or email that is sent look like it came from the original source, making the victim think this is the case. By clicking on the link, the user can show them that the site is real. The user then moves his mouse over the page of the website and tries to log in with his passwords. The attacker has access to the sensitive information once it has been decrypted, and this information can be used to find him.

Related Works

The primary purpose of any phisher is to get user credentials, such as passwords for bank accounts, Social Security numbers, card verification values, and credit card numbers. If someone were to gain access to these extremely sensitive details, it would result in a loss of billions of dollars on a global scale. As a result of the huge increase that has been observed in the frequency of these attacks, they need to be managed in order to keep the consumers' faith during their online transactions. For example, Google bots uncover 9500 new potentially hazardous websites every single day [5]. Phishing scams are difficult to spot, even more challenging to stop, and almost always succeed in doing so without being discovered. Because the process of phishing does not leave any traces behind, the procedure to identify any occurrence of intrusion is further detected in order to ensure that the attacker does not target the user system. This procedure prevents sensitive information from being susceptible while also reducing the risk of its being stolen and misused. The information related to the user needs to be safeguarded against theft so that any potential dangers can be avoided. For this reason, one of the methods that is utilized to ensure the security of such information is the practice of deleting any emails that have phished links inside their bodies.

DOI: https://doi.org/10.17762/turcomat.v11i1.13589

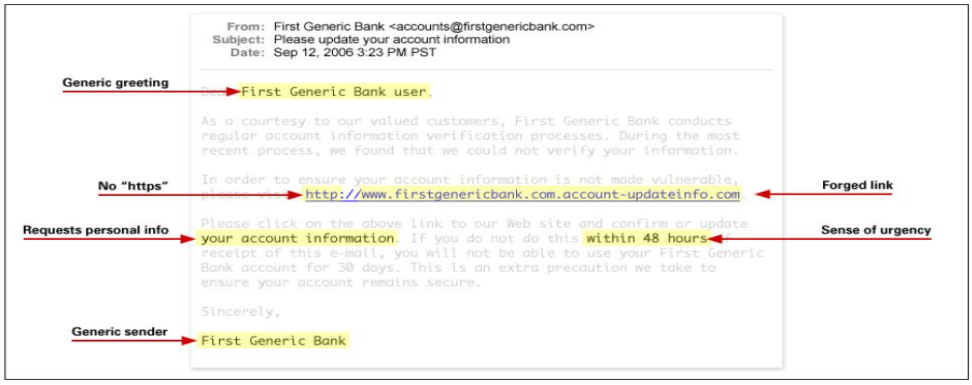


Figure 2: Components of a fraudulent mail with phished URL

The most popular methods for identifying these phishing assaults are machine learning algorithms. One function of machine learning algorithms is to categorize issues into positive or negative groups. These algorithms aid in assessing whether or not phishing attacks are legitimate. The process of developing an automated machine learning system begins with training the samples that were taken from the dataset. The elements in these samples can be found on both trustworthy and phished websites. By utilizing the concepts and procedures of ML algorithms, the URL can be identified and further pre-processed for its complexity. The study of authors in [6] helped to identify phishing attempts made on textual formats of data. They attempted to match the words in the mail with the words that were in the dataset in order to achieve this. In this way, feature extraction methods like the TF-IDF one were employed to extract the keywords. The specific page was checked, and if the Google search phrases matched, it was deemed to be reliable. However, due to the created model's limited responsiveness to English, the study had a number of shortcomings. By expanding on this work, the authors in [7] were able to achieve an ideal accuracy of 92.36 percent. Using diverse datasets, the authors of this work collected keywords from a number of languages. The feature extraction technique makes use of embeddings and TF-IDF. However, the suggested method generated a sizable number of false positives.

The authors of a different research study in [8] recommended using a non-linear approach to find the same. The system model was trained using TF-IDF vectorization and the machine learning techniques SVM and KNN. The text that was extracted from the URL connections in a textual format was vectorized. The reliability of the deployed models was assessed across 15,000 web pages, and the KNN approach produced a precision factor of 92.36 percent. In order to identify phishing attempts on websites, authors in [9] recommended utilizing word vectorization on 209 phrases that were chosen from the URL dataset. When the words were analysed using 17 NLP features, word embeddings were performed on the textual data. Five machine learning techniques were used to enhance the model and achieve optimal levels of accuracy.

The authors have created a machine learning-based approach for identifying such attacks in [10]. This technique may separate 21 components from a URL link and further classify it as a valid or phished website. The dataset the authors obtained from Phish Tank in 2018 was used to categorize 1918 valid web sites and 2141 phished web pages. The developed approach helped to authorize bank accounts and secure financial transactions made through payment gateways. Out of four machine learning techniques, SVM was determined to be the best model because it provided accuracy of 99.39%. The authors in [11] proposed an alternative approach in which they automated the phishing attack process using neural networks. The system model and the Monte Carlo method were used to extract thirty features from the URL. The lengths of the words, word characters, and website URL links were all taken into account throughout the feature extraction process. HTML and Java Script were used as front-end languages. The authors also used two machine learning-based algorithms' basic ideas and performed a final algorithm comparison. The model's overall accuracy was 97.71 percent.

Hence, it is evident that most research scientists employed the method of locating phishing websites by their URLs. The authors in [12] proposed that phishing emails may be recognized by examining the information on the mail packets. For this purpose, they suggested integrating the neural method's guiding principles with the application of reinforcement-based machine learning categorization. The dataset was obtained, and for the mail

packets, a total of 50 features were selected. The relevant web page's body text, links, and headers were additional categories and groups for the URL. Using this method, the authors identified 9118 emails as phished emails and 2506 emails as valid. Nonetheless, the model was able to reach a 98.6% accuracy rate.
Authors in [13] proposed a taxonomical approach that focused on phishing attack defences and its vector outputs. Additionally, the report explained how these attacks highlighted weaker organizational systems. The main objective of this research effort was to provide software developers with guidance for creating anti-phishing strategies that could prevent such attacks. The paper addressed attacks on social media platforms and introduced machine learning techniques to evaluate and detect the malware present in the server system. The study concentrated on the strategies threat actors can employ to trick individuals into providing their personal information on PAN cards and adhar cards.

Methodologies Used

Phishing scams are real, as demonstrated by the results of a number of studies and research projects, and they are not a myth. Attacks using zero-day vulnerabilities, on the other hand, have not yet been identified. But, adopting approaches like blacklists and whitelists still doesn't manage to catch zero-day assaults. Yet, the primary purpose of the phisher is to steal the end user's sensitive data, credentials, and personal information in order to earn monetarily. When a network is the victim of an assault of this kind, it has caused a number of organizations to suffer a loss of integrity and has caused them financial losses of billions of dollars. However, the terminology associated with phishing can vary from one organization to the next, and there are many different kinds of attacks that can be launched against specific applications. Phishing attacks have a lifespan that is strikingly comparable to the process that is followed when a fisherman uses a hook to catch a fish. In a manner not dissimilar to this, phishers utilize network and email attacks as bait. During the process, an individual's personally identifiable information is discarded.

Taxonomy of URL

The URL acronym's primary meaning is Uniform Resource Locator (URL). A URL acts as a user's gateway to a given webpage, website, video, image, etc. It indicates the presence of contextual information and is regarded as the location where an address can be found. The protocol for each URL begins with "https". This is widely utilised and has the ability to route users to a particular piece of material. To detect phishing attempts that occur as a result of improperly positioned URLs, textual information from the URL must be extracted. This extraction makes advantage of the word and character levels. As the URL might be altered by phishers, it is imperative to carefully read every word and character of the link that has been provided. Character groupings that frequently appear together could be a sign of phishing, but character level patterns can spot crucial data in these clusters. A URL is sometimes made up of a number of characters or words, some of which have obscure semantic connotations. Character patterns boost the effectiveness of phishing URL detection and make it simpler to locate this sensitive information. Without a specialist's help, machine learning techniques can be used directly throughout the learning task using the learned character sequence features. The following attributes are shown in a URL:

- The Internet Protocol address of the Uniform Resource Locator (URL): IP addresses are used to locate websites and web pages
- Length of URL: The length of a URL, including any special characters or phrases that are used, can be seen by looking at the URL's length
- Suspicious characters: Suspicious characters include the use of words and characters such as @ and #
- Suffix that was used: Prefixes and suffixes are the most important components of a URL. An authentic uniform resource locator (URL) will always begin with "http" and will lack any unnecessary extensions, such as a hyphen
- The protocol that is being used: The HTTPS protocol can tell you whether a link starts with "http" or "https"

Figure 3 below illustrates the conventional visualization of a URL

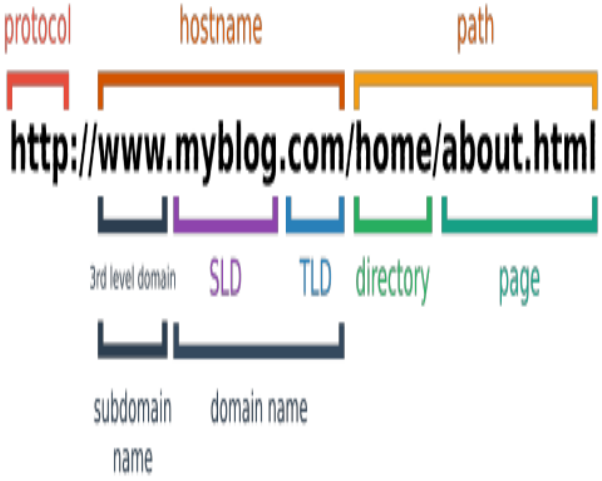


Figure 3: Visual Representation of a URL

Taxonomy of PyCaret

Machine learning algorithms are implemented using the open source library PyCaret. It is a user-friendly low-code library that can be easily implemented with a few lines of code and is mostly used to ease difficult business challenges. PyCaret can be used right from the start of a project's development to help with data preparation in the proper format for model analysis. Instead of implementing and running each block of source code, the full system model deployment takes place in the same format with lines of code and function calls. As a result, PyCaret may be thought of as a Python wrapper module with built-in libraries that may access various techniques for the execution of machine learning tasks. It also includes a number of built-in frameworks, including those from Sci-kit and spaCy. These frameworks have a tendency to make coding execution simpler and the entire process less laborious. It also integrates several pre-processing methods within itself and permits the implementation of codes in single lines of frame. This makes the entire process hassle free and reduces the overall time involved in the implementation process. It helps to fine tune the machine learning models by supporting multiple in-built libraries within itself and process the stage of feature engineering. So, it can be claimed that using PyCaret has a number of benefits over using a single implementation of code executions.

Proposed Methodology

The proposed research study's author employs machine learning methods and approaches to identify such URL-based phishing assaults in order to stop them. However, the innovation in the suggested thesis is achieved by putting it into practice utilizing PyCaret. As was noted in the previous section, PyCaret implements all relevant machine learning algorithms and permits simultaneous calls to internal libraries. This implementation would have been tedious and time-consuming if it had been carried out using individual algorithms.

The main goal of the study is to identify any malicious URL links that are associated to or connected to a particular webpage or website and further identify it as a phishing site. In this case, phishing assaults are carried out by phishers whose goal is to construct phony URL links and send them to the user by email IDs or links that are simple to download. To do this, the phisher employs a variety of strategies for impersonating as a trusted source, increasing the likelihood that a user without a technical background may fall for the scam. But, when a user downloads or otherwise gains access to one of these URL links and clicks on it, he is taken to a fake, disguised page that was created by the attacker. The user is then prompted to enter his name, email address, password(s), and other identifying information. A dataset consisting of malicious URL links and legitimate URL links that would connect to a webpage is retrieved from a repository as part of the process of the proposed methodology. Once data collection is complete, the model is run by invoking the PyCaret function using the "setup" method. By implementing a single function call in PyCaret, 14 algorithms are simultaneously executed as part of this execution. Evaluation criteria including accuracy, precision, and Kappa factors are then used to provide a result. The next step in the research proposal is to choose the algorithm with the best generating

accuracy. Initial deployments revealed that Random Forest produced the highest levels of accuracy, hence it was chosen. The likelihood score of random forest was subsequently determined for the database with respective malicious and genuine URL links. Random forest was used as the specific algorithm to further train and test the entire dataset. The suggested implementation, on the other hand, was further expanded to include creating a hybrid model that combined the top three techniques that produced improved accuracy in the initial phase. Figure below depicts the architectural flow of the proposed methodology.

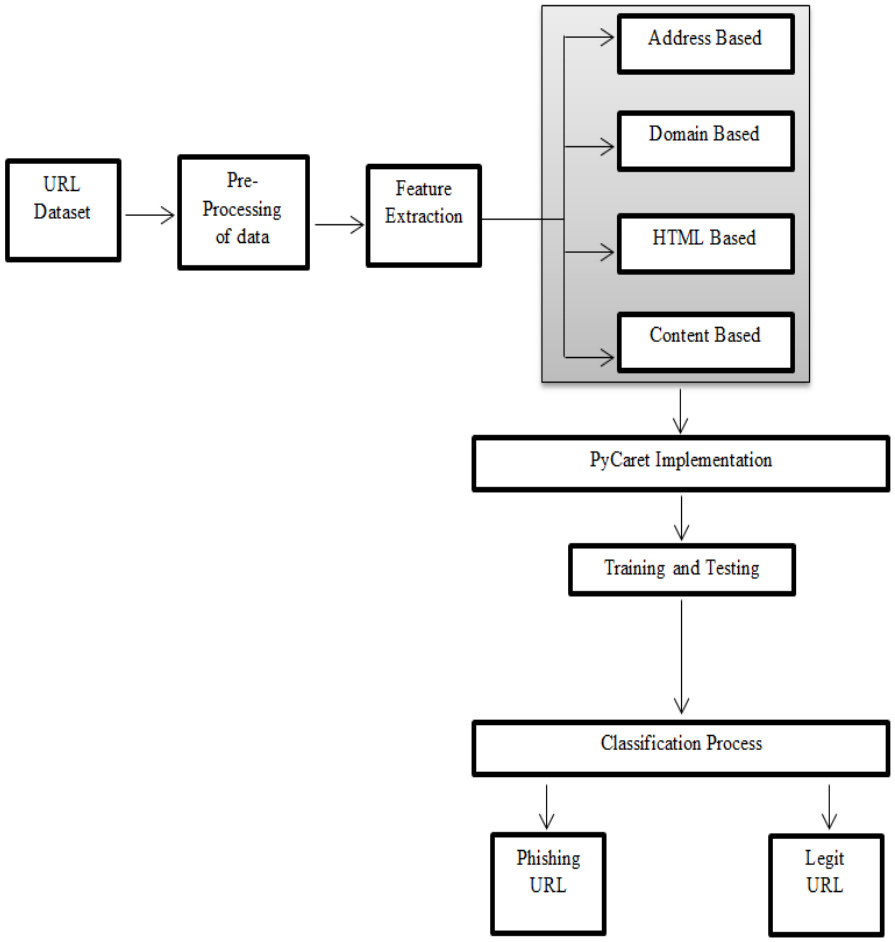


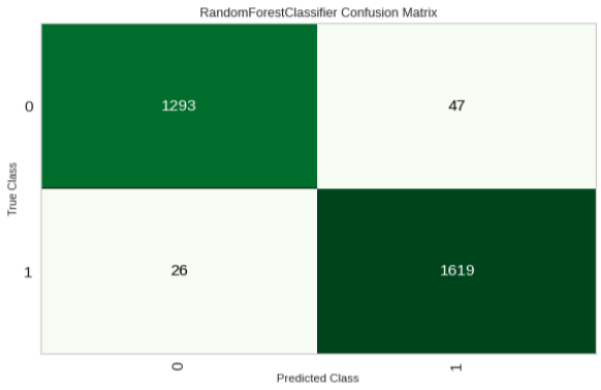
Figure 4: Architecture of the Proposed Methodology

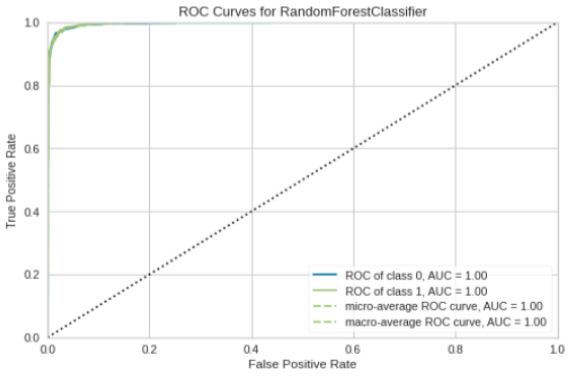
Results

Here, the results obtained by actually doing the proposed work are presented. The findings, however, are presented in two separate case studies. The first case study demonstrates the outcomes that were accomplished through the use of a single classifier, namely random forest. The second case study demonstrates the outcomes that were accomplished through the combination of three algorithms and the formation of a stacking algorithm to accomplish the same thing.

DOI: https://doi.org/10.17762/turcomat.v11i1.13589

Case Study 1: Results generated using random forest

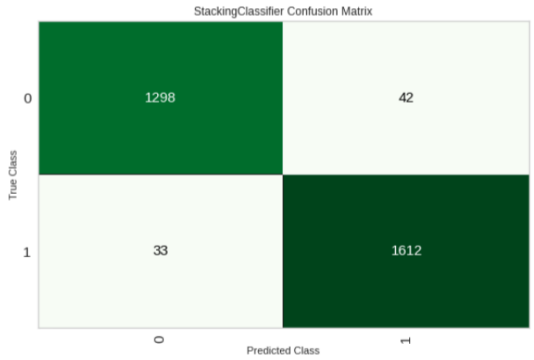


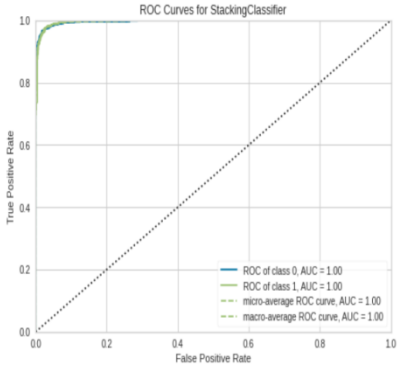


(a) Confusion Matrix **(b) ROC Curve**

The values of 1293, 47, 26, and 1619 respectively denote the TN, FP, FN, and TP categories of information. The TN score indicates that 1293 URLs were expected to be classed as phishing webpages using their URL linkages. These TN values also indicates that 1293 negative predicted values created by the algorithm in the testing phase coincided with the actual negative values of phished webpages acquired from the dataset repository. As a result, and in a manner that is analogous, the TP value indicates that 1619 webpages were predicted to be categorized as legitimate webpages based on the URL links that led to them. These TP values also indicate that 1619 of the positive predicted values produced by the algorithm during the testing phase matched with the actual positive values of legitimate webpages obtained from the dataset repository. This was determined by comparing the actual values to the predicted values.

Case Study 2: Results generated using stacking algorithm





(a) Confusion Matrix **(b) ROC Curve**

The TN score indicates that 1298 URLs were expected to be classed as phishing webpages using their URL linkages. These TN values also indicates that 1298 negative predicted values created by the algorithm in the testing phase coincided with the actual negative values of phished webpages acquired from the dataset repository.

Conclusions

According to the literature review that was conducted, it was discovered that phishing attacks frequently occur and must be prevented from happening. The occurrence of such attacks creates mistrust of regular people in many organisations, which then exposes their personal information. As a result, the primary goal of the research

study is to conduct experiments based on machine learning algorithms and detect the occurrence of phished URLs that might redirect the user to a fake website. These experiments will be conducted in order to fulfil the research study's primary objective. The dataset also comprises 31 features which have to be extracted utilising feature engineering techniques and procedures. However, this dataset can be accessed as a csv file and is further pre-processed for filtering to remove redundant and irrelevant data. Following this step, the feature extraction procedure is carried out, during which URL features such domain-based, content-based, and address-based are extracted. Following this, PyCaret is implemented, with each line of code responsible for the entire execution. But, in this level; the experimentation occurs in three steps. The first phase entails running 14 built-in algorithms as part of PyCaret's core implementation to produce accuracy. The system model is implemented in two steps: the second stage involves considering random forest, and the third and final step involves combining the top three accuracy-generating algorithms to create a stacking model. In the conclusion, both algorithms' performances are evaluated, and the accuracies of each are determined. The method with the highest generating accuracy is declared the optimised model after a comparison.

References

[1] Xiang, G. , Hong, J. , Rose, C. P. , & Cranor, L. (2019). Cantina + : A feature-rich machine learning framework for detecting phishing web sites. ACM Transactions on In- formation and System Security, 14 (2), 1–28 20
[2] Nguyen, H. H. and Nguyen, D. T. (2016). Machine learning based phishing web sites detection. In AETA 2015: Recent Advances in Electrical Engineering and Related Sciences, pages 123–131. Springer
[3] APWG (2015). Phishing activity trends report, 1st ˆaA¸S 3rd quarters 2015. Report, ˇ APWG
[4] Buber, E. , Diri, B. , & Sahingoz, O. K. (2017a). Detecting phishing attacks from URL by using NLP techniques. In 2017 International conference on computer science and Engineering (UBMK) (pp. 337–342)
[5] Ubiquitous Search Engine P Vyas, A Menon, A Ravindran, S Shivadekar International Journal of Computer Applications 117 (20)
[6] Analysis and Study on the Classifier Based Data Mining Methods SN Popat, YP Singh Journal of Advances in Science and Technology| Science & Technology 14 (2)
[7] Jain AK, Gupta BB (2019) A machine learning based approach for phishing detection using hyperlinks information. J Ambient Intell Human Comput 10(5):2015–2028
[8] Izhar, M., Shahid, M., and Singh, V. (2013). Network security issues in context of rsna and firewall. International Journal of Computer Applications, 82(16)
[9] Kang, J. and Lee, D. (2007). Advanced white list approach for preventing access to phishing sites. In Convergence Information Technology, 2007. International Conference on, pages 491–496. IEEE
[10] Openfish (2018). Phishing dataset Accessed 24 July 2018, available at: https://www. openphish.com/
[11] Lastdrager, E. E. (2014). Achieving a consensual definition of phishing based on a systematic review of the literature. Crime Science, 3(1):9
[12] Rao RS, Pais AR (2019) Detection of phishing websites using an efficient feature-based machine learning framework. Neural Comput Appl 31(8):3851–3873
[13] Q. Cui, G. V. Jourdan, G. V. Bochmann, R. Couturier, and I. V. Onut, "Tracking phishing attacks over time," 26th Int. World Wide Web Conf. WWW 2017, pp. 667–676, 2017