IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Phishing Sites Predictor

[1]K.T. Krishna Kumar, [2]Krishna Prakash Janagam,
[1]Associate Professor and Placement Officer, [2]MCA Final Semester,
[1]Master of Computer Applications
[1]Sanketika Vidya Parishad Engineering College, Visakhapatnam, Andhra Pradesh, India.

Abstract: With the rapid growth of technology, many people are currently working remotely and consume social media more often. Over the last few years, the Web has seen a massive growth in the number and kinds of web services. Web facilities such as online banking, gaming, and social networking have promptly evolved as people perform routine tasks. As a result, a large amount of information is uploaded on a daily basis to the Web. As these web services drive increased opportunities for people to interact, they equally offer new opportunities for criminals. URLs are launch pads for any web attacks such that any malicious intention user can steal the identity of the authorized person by sending the malicious URL. Malicious URLs are a keystone of Internet illegitimate activities. It is a form of cyber-attack, which has an adverse effect on people where the user is directed to fake websites and duped to reveal their sensitive and personal information which includes passwords of accounts, bank details, atm pin-card details etc. Hence protecting sensitive information from malwares or web phishing is difficult Aim of the phishers is to acquire critical information like username, password and bank account details. Cyber security persons are now looking for trustworthy and steady detection techniques for phishing websites detection. Hence protecting sensitive information from malwares or web phishing is difficult. By using Machine learning algorithms, we will be identifying phishing attacks and report their positives and negatives. The proposed approach is that which classifies URLs automatically by using Machine learning algorithm called logistic regression that uses binary classification.

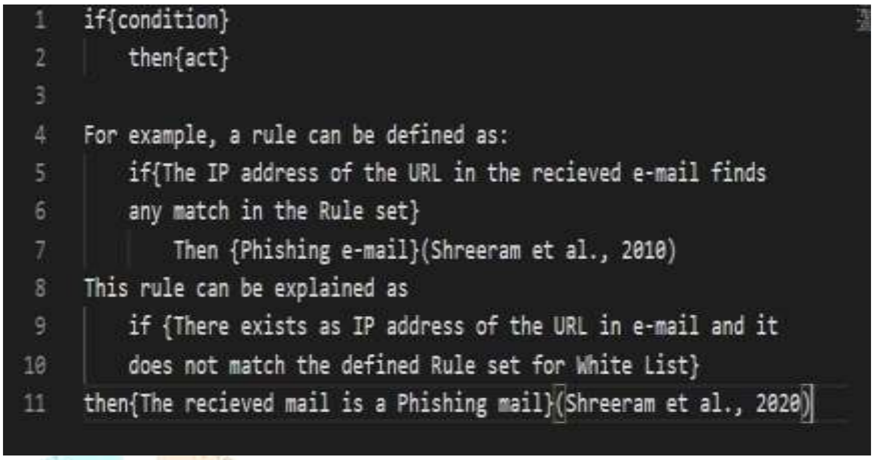
Index Terms -Cyber security, logistic regression, URL prediction, python programming language FAST API, feature extraction, preprocessing.

I. INTRODUCTION

Phishing is one of the most challenging security problems faced by the world today, in part due to the large number of online transactions that take place daily. It refers to the practice of trying to obtain sensitive information, like user names, passwords and credit card details for malicious reasons by mimicking a trustworthy entity, like a well-known and trusted website. It can be carried out by email spoofing, messaging, and generally appears to be from social networking websites, auction sites as well as online pay-mint processing websites. Phishing websites deceive users, and exploit weaknesses of web security technologies.

1.1 Existing System

Phishing attacks [1] have misled a lot of users by impersonating legitimate websites and stealing private information and/or financial data (Afroz and Greenstead, 2011). To protect users against phishing, various anti-phishing techniques have been proposed that following different strategies. In this methodology, content/email is filtered as it enters in the victim's mail box by means of machine learning methods, such as Support Vector Machines (SVM) or Bayesian Additive Regression [2] Trees (BART) (Tout and Hafner, 2009). Blacklist is collection of recognized phishing Websites/addresses published by dependable entities like Google's and Microsoft's blacklist. It involves both a client and a server component. The client component is employed as either an email or browser plug-in that relates with a server component, which in this case is a public website that make available a list of identified phishing sites (Tout and Hafner, 2009). It is an approach that uses genetic algorithm for phishing web pages' detection. Genetic algorithms [3] can be used to develop simple rules for preventing phishing attacks. These rules are used to discern normal web-site from anomalous website. These anomalous websites denote events with probability of phishing attacks. The rules stored in the rule base are typically in the following form (Shreeram et al., 2010 The main advantage is that it provides the feature of malicious status notification before the user reads the mail. It also provides malicious web-link [4] detection in addition of phishing detection.



```
if{condition}
    then{act}

For example, a rule can be defined as:
    if{The IP address of the URL in the recieved e-mail finds
    any match in the Rule set}
        Then {Phishing e-mail}(Shreeram et al., 2010)
This rule can be explained as
    if {There exists as IP address of the URL in e-mail and it
    does not match the defined Rule set for White List}
then{The recieved mail is a Phishing mail}(Shreeram et al., 2020)
```

Figure 1: Existing System

1.1.1 Challenges

- Detection of Zero-Day Attacks
- Sophisticated Social Engineering
- HTTPS and SSL Certificates
- Evasion Techniques

1.2 Proposed System

Here we proposed a new method of anti-phishing technology. The Anti-Phishing Technology using Machine Learning Approach is a mechanism that is proposed in order to ensure high security. In the is mechanism we deal with the URLs (Uniform Resource Locaters) and the URI (Uniform Resource Identifies) check with machine learning technique and predict whether it is phishing website or not. Here we create a web app for browsing imputed URLs. Each time we browse a site the corresponding URL (Uniform Resource Locater) of site will be checked with machine learning technique. Phishing [5] site detection is truly an unpredictable and element issue including numerous components and criteria that are not stable. On account of the last and in addition ambiguities in arranging sites because of the intelligent procedures programmers are utilizing, some keen proactive strategies can be helpful and powerful tools can be utilized, for example, fuzzy, neural system and data mining methods can be a successful mechanism in distinguishing phishing sites. We applied logistic regression algorithms to model our train out model and at the end logistic regression which gave a more accurate prediction was used in our system.

The phishing dataset is collected from Kaggle First [6], the dataset is divided into two set which are then used to train and test our logistic algorithm mode The performance metrics of the reference algorithms based on precision, recall, f1-score and accuracy of our algorithm is analyzed. Implementation of this project is done by creating an app using python programming language and FASTAPI [7].

1.2.1 Advantages

- multi-Faceted Detection Scalability
- User Privacy Protection
- Integration Capabilities

II. LITERARTURE REVIEW

Gather data from diverse sources such as known phishing databases, domain registrars, web crawlers, and user reports. Cleanse and preprocess data to remove noise, standardize formats, and extract relevant features (e.g., domain names, URL structure, SSL certificates). Extract features from URLs such as domain age, length, domain reputation, presence of subdomains, and lexical analysis. Analyze webpage content for phishing indicators like suspicious keywords, deceptive branding, or fake login forms. Provide APIs for integration with other security systems and tools such as web proxies, firewalls, and SIE(SInformationaEvent_Management) solution.

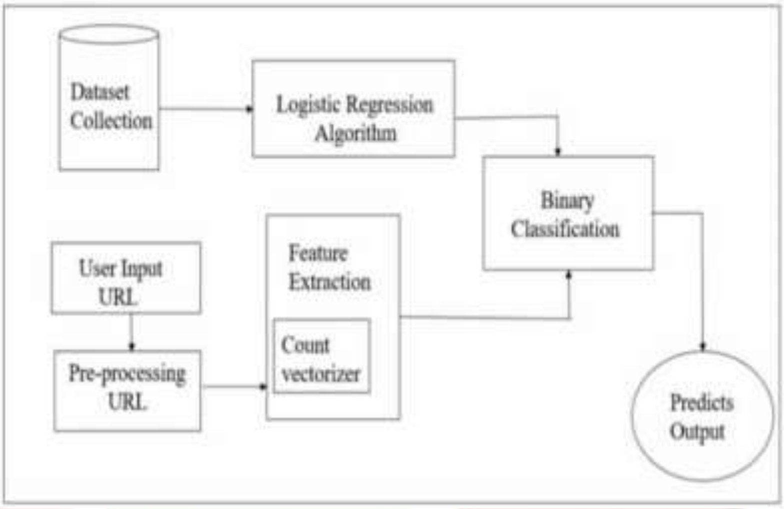


Figure 2: Architecture Diagram

2.1 Algorithm

- Logistic regression is one of the most popular Machine Learning algorithms, which comes under the Supervised Learning technique. It is used for predicting the categorical dependent variable using a given set of independent variables.
- Logistic regression [9] predicts the output of a categorical dependent variable. Therefore, the outcome must be a categorical or discrete value. It can be either Yes or No, 0 or 1, true or False, etc. but instead of giving the exact value as 0 and 1, it gives the probabilistic values which lie between 0 and 1.
- Logistic Regression is much similar to the Linear Regression except that how they are used. Linear Regression is used for solving Regression problems, whereas Logistic regression is used for solving the classification problems.
- In Logistic regression, instead of fitting a regression line, we fit an "S" shaped logistic function, which predicts two maximum values (0 or 1).
- The curve from the logistic function indicates the likelihood of something such as whether the cells are cancerous or not, a mouse is obese or not based on its weight, etc. Logistic Regression is a significant machine learning algorithm because it has the ability to provide probabilities and classify new data using continuous and discrete datasets.
- Logistic Regression can be used to classify the observations using different types of data and can easily determine the most effective variables used for the classification [10].

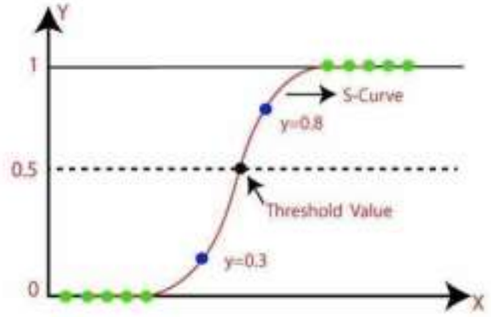


Fig:Logistic Regression

2.2 Techniques

Check the reputation of the domain using known blacklists or reputation databases. Analyse the structure of the URL for suspicious patterns such as Use of IP addresses [11] instead of domain names. Presence of multiple subdomains or unusual characters. Length of the URL (phishing URLs tend to be longer). Extract meaningful tokens from the URL and analyse them for indicators of phishing, such as common phishing keywords ("login," "account," "secure," etc.). Train machine learning models using labelled datasets to classify URLs or websites as phishing or legitimate. Algorithms include:

- Logistic Regression
- Random Forest
- Support Vector Machines (SVM)
- Neural Networks

2.3 Tools

Machine Learning and Data Analysis Tools

- Python Libraries Utilize libraries like scikit-learn, TensorFlow, or Porch for building and training machine learning models for phishing detection.
- Jupiter [12] Notebook Interactive environment for developing and testing machine learning models, visualizing data, and analysing results. Apache Spark Framework for largescale data processing, useful for handling big datasets in phishing detection systems.

Real-Time Monitoring and Analysis Tools

- Splunk: SIEM (Security Information and Event Management) platform for real-time monitoring and analysis of security events, including phishing attempts.
- Elasticsearch and Kibana: Elasticsearch for indexing and searching data, Kibana for visualizing and analysing large volumes of log data related to phishing activities.

Integration And Deployment Tools

- APIs: Use APIs provided by threat intelligence feeds (e.g., Google Safe Browsing API) for integrating real-time phishing site detection capabilities.
- Docker: Containerization tool for packaging phishing site predictor components into standardized units for deployment and scalability.
- Ansible: Automation tool for managing and deploying phishing site predictor updates and configurations across multiple servers or environments.

User interface and Reporting Tools

- Dashboarding Tools: Tools like Grafana or Tableau for creating dashboards to monitor phishing detection metrics, trends, and alerts.
- Custom Web Interfaces: Develop custom web interfaces or applications for security analysts to review phishing detections, manage reported URLs, and analyse phishing trends.

Security and Privacy Tools

- SSL/TLS Certificate Checkers**:** Tools to verify SSL/TLS certificates used by websites to detect potential phishing sites using HTTPS.

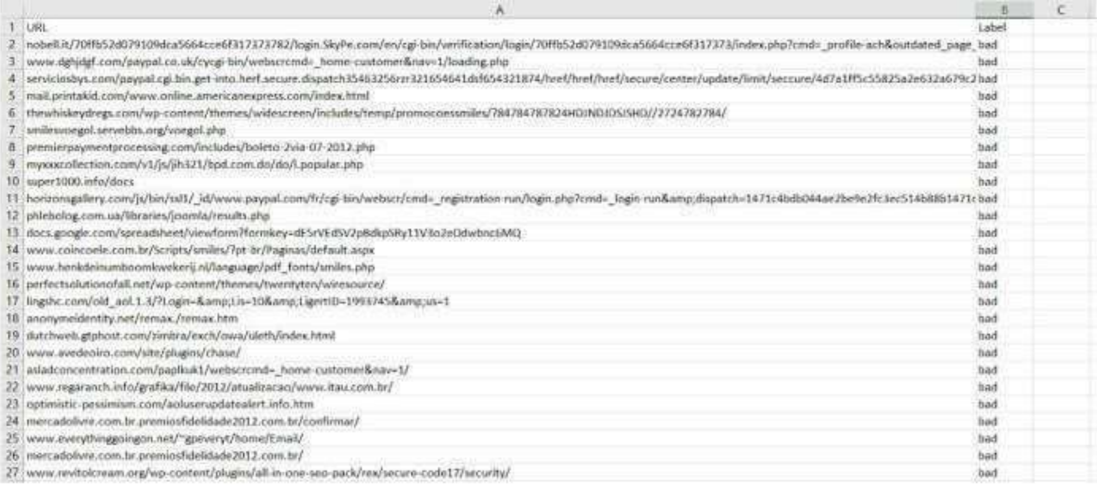
- Privacy-Preserving Techniques: Implement encryption and anonymization techniques to protect user data and ensure compliance with privacy regulations (e.g., GDPRM.)

2.4 Methods

Natural Language Processing (NLP): Analyse textual content on websites for phishing indicators such as deceptive language, urgent requests, etc. Use sentiment analysis to detect emotional manipulation   Blacklists and Whitelists: Incorporate known phishing URLs from databases like Phish Tank, Google Safe Browsing API, etc. Whitelist legitimate domains to reduce false positives. Behavioural Analysis: Monitor real-time behaviour of users on websites [13] to detect suspicious activities (e.g., unexpected redirects, form submissions).
Phishing Kits Detection: Identify common templates and scripts used by phishers to create phishing websites.   Feature Selection and Model Tuning: Use techniques like cross-validation, grid search, and feature importance analysis to optimize model performance.
Methods: Combine multiple models (e.g., bagging, boosting) to improve overall prediction accuracy.   Deep Learning Approaches: Apply neural networks for tasks like image analysis (e.g., detecting fake login pages).   Real-time Detection: Implement systems that continuously monitor and analyse web traffic for suspicious patterns and behaviours.

III. METHODOLOGY

3.1 Input



	A	B	C
1	URL	Label	
2	nobell.it/70ffb52d079109dca5664cce6f317373782/login.SkyPe.com/en/cgi-bin/verification/login/70ffb52d079109dca5664cce6f317373/index.php?cmd=_profile-ach&outdated_page_	bad	
3	www.dghjdgf.com/paypal.co.uk/cycgi-bin/webscrcmd=_home-customer&nav=1/loading.php	bad	
4	serviciosbys.com/paypal.cgi.bin.get-into.herf.secure.dispatch35463256rzr321654641dsf654321874/href/href/href/secure/center/update/limit/seccure/4d7a1ff5c55825a2e632a679c2	bad	
5	mail.printakid.com/www.online.americanexpress.com/index.html	bad	
6	thewhiskeydregs.com/wp-content/themes/widescreen/includes/temp/promocoessmiles/784784787824HDJNDJDSJSHD//2724782784/	bad	
7	smilesvoegol.servebbs.org/voegol.php	bad	
8	premierpaymentprocessing.com/includes/boleto-2via-07-2012.php	bad	
9	myxxxcollection.com/v1/js/jih321/bpd.com.do/do/l.popular.php	bad	
10	super1000.info/docs	bad	
11	horizonsgallery.com/js/bin/ssl1/_id/www.paypal.com/fr/cgi-bin/webscr/cmd=_registration-run/login.php?cmd=_login-run&amp;dispatch=1471c4bdb044ae2be9e2fc3ec514b88b1471c	bad	
12	phlebolog.com.ua/libraries/joomla/results.php	bad	
13	docs.google.com/spreadsheet/viewform?formkey=dFSrVEdSV2p8dkpSRy11V3o2eDdwbnc6MQ	bad	
14	www.coincoele.com.br/Scripts/smiles/?pt-br/Paginas/default.aspx	bad	
15	www.henkdeinumboomkwekerij.nl/language/pdf_fonts/smiles.php	bad	
16	perfectsolutionofall.net/wp-content/themes/twentyten/wiresource/	bad	
17	lingshc.com/old_aol.1.3/?Login=&amp;Lis=10&amp;LigertID=1993745&amp;us=1	bad	
18	anonymeidentity.net/remax./remax.htm	bad	
19	dutchweb.gtphost.com/zimbra/exch/owa/uleth/index.html	bad	
20	www.avedeoiro.com/site/plugins/chase/	bad	
21	asladconcentration.com/paplkuk1/webscrcmd=_home-customer&nav=1/	bad	
22	www.regaranch.info/grafika/file/2012/atualizacao/www.itau.com.br/	bad	
23	optimistic-pessimism.com/aoluserupdatealert.info.htm	bad	
24	mercadolivre.com.br.premiosfidelidade2012.com.br/confirmar/	bad	
25	www.everythinggoingon.net/~gpeveryt/home/Email/	bad	
26	mercadolivre.com.br.premiosfidelidade2012.com.br/	bad	
27	www.revitolcream.org/wp-content/plugins/all-in-one-seo-pack/rex/secure-code17/security/	bad	

Fig:Dataset Structure

3.2 Steps For Execution

Problem Definition and Scope

- Define the Objective: Clearly articulate the goal of the predictor (e.g., to classify websites as phishing or legitimate).
- Scope the Project: Determine the types of phishing attacks and websites you want to detect (e.g., email phishing, fake banking sites).

Data Collection

- Gather Data: Obtain a dataset that includes examples of known phishing and legitimate websites.   • Data Sources: Utilize repositories like Phish Tank, Google Safe Browsing API, or create custom crawlers to collect data.
- Data Labelling: Ensure each website in the dataset is labelled correctly (phishing or legitimate).

Data Preprocessing

- Data Cleaning: Handle missing values, duplicates, and inconsistent formatting.
- Feature Extraction: Extract relevant features from URLs (e.g., length, domain age) and content (e.g., keywords, HTML structure).
- Normalization: Scale numerical features if necessary to improve model performance.

Exploratory Data Analysis (EDA)

- Visualize Data: Understand distributions, correlations [14] and class imbalances.
- Feature Importance: Identify which features are most informative for distinguishing phishing from legitimate sites.

Model Selection and Training

- Choose Algorithms: Select appropriate machine learning models based on the dataset size, complexity, and goals (e.g., Decision Trees, Random Forests, SVMs).
- Split Data: Divide data into training and validation sets (e.g., 70-30 split).
- Train Models: Fit the chosen models on the training data.

Hyperparameter Tuning: Optimize model parameters using techniques like grid search or

randomized search.

Model Evaluation

- Performance Metrics: Evaluate models using metrics such as accuracy, precision, recall, F1-score, and ROC-AUC.
- Cross-validation: Validate model performance across multiple folds to ensure robustness.

Deployment and Monitoring

- Integration: Deploy the model into a production environment (e.g., as a web service or batch processing pipeline).
- Monitoring: Implement monitoring for model performance over time and update the model as needed.
- Feedback Loop: Incorporate feedback to continuously improve the predictor based on new phishing tactics and patterns.

3.3 Output

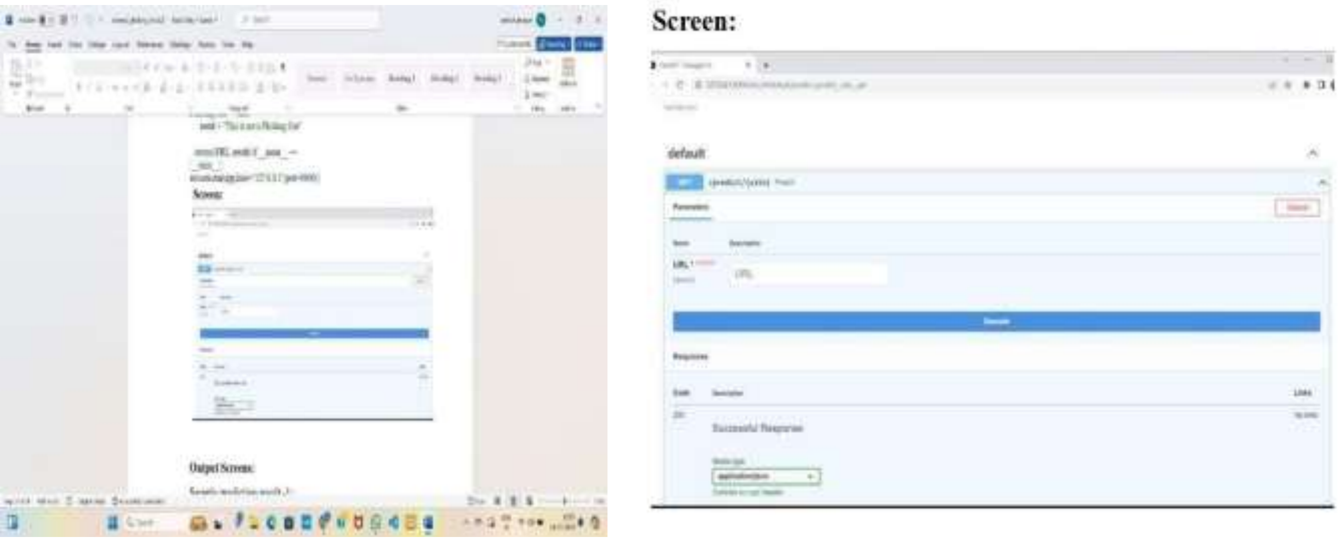
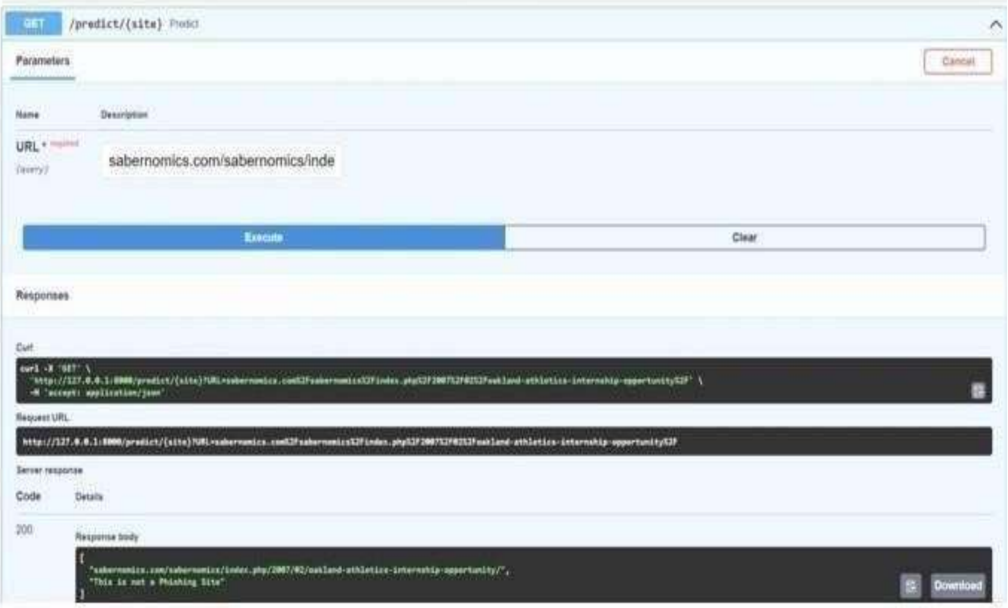


Figure 5,6: Output Scree

IV. RESULT



sabernomics.com/sabernomics/inde

Curl

```
curl -X 'GET' \
  'http://127.0.0.1:8000/predict/{site}?URL=sabernomics.com%2Fsabernomics%2Findex.php%2F2007%2F02%2Foakland-athletics-internship-opportunity%2F' \
  -H 'accept: application/json'
```

Request URL

```
http://127.0.0.1:8000/predict/{site}?URL=sabernomics.com%2Fsabernomics%2Findex.php%2F2007%2F02%2Foakland-athletics-internship-opportunity%2F
```

Server response

Code Details

200 Response body

```
{
  "sabernomics.com/sabernomics/index.php/2007/02/oakland-athletics-internship-opportunity/",
  "This is not a Phishing Site"
}
```

Figure 7: Result Screen

V. DISCUSSION

Developing a phishing site predictor is crucial in today's digital landscape, where online threats continue to evolve in sophistication. This project focuses on leveraging machine learning and data analytics to identify and classify potentially malicious websites accurately. By collecting and analysing data from various sources, such as URLs and website content, the system extracts key features indicative of phishing behaviour. These features include URL structures, domain characteristics, content anomalies, and more. Machine learning models, trained on labelled datasets, play a pivotal role in predicting whether a given website is likely to be a phishing site or not.

The project involves significant challenges, including data preprocessing to ensure data quality and consistency, selecting and fine-tuning appropriate machine learning algorithms for high accuracy and reliability, and implementing real-time scoring mechanisms for prompt detection. User interfaces are designed to provide intuitive access for users to input URLs or content and receive clear, actionable predictions. Security measures are paramount to safeguard against potential threats arising from handling suspicious websites.

VI. CONCLUSION

The malicious web page is created as if a legitimate web page, especially copying the original web page one to one. Therefore, detection of these pages is a very trivial problem to overcome due to its semantic structure which takes the advantage of the human vulnerabilities. However, they are static algorithms and cannot identify the new type of attacks in the system. Therefore, as an efficient solution, we propose the use of logistic regression machine learning system for classifying the incoming URLs. The experimental results show that Logistic regression approach result gives best results of 97 % Accuracy.

VII.FUTURE SCOPE

The phishing site predictor project aims to develop a robust system to detect malicious websites. Key features include data collection from diverse sources, thorough data cleaning, and extraction of relevant features like URL characteristics and content analysis. Machine learning models, such as decision trees or SVMs, will be trained and optimized for classification. Real-time scoring of URLs and intuitive user interfaces for input and result visualization are crucial. Security measures ensure resilience against malicious inputs, scalability, and compliance with privacy regulations.

Continuous feedback mechanisms and performance monitoring will drive iterative improvements. Documentation for users and developers, along with seamless deployment and integration capabilities, will ensure usability and reliability in enhancing web security.

VIII. ACKNOWLEDGEMENT

Kandhati Tulasi Krishna Kumar: Training & Placement Officer with 15 years' experience in training & placing the students into IT, ITES & Core profiles & trained more than 9,500 UG, PG candidates & trained more than 350 faculty through FDPs. Authored 5 books, Guided 40+ papers in international journals for the benefit of the diploma, pharmacy, engineering & pure science graduating students. He is a Certified Campus Recruitment Trainer from JNTUA, did his Master of Technology degree in CSE from VTA and in process of his Doctoral research. He is a professional in Pro-E, CNC certified by CITD He is recognized as an editorial member of IJIT (International Journal for Information Technology & member in IAAC, IEEE, MISTE, IAENG, ISOC, ISQEM, and SDIWC. He published articles in various international journals on Databases, Software Engineering, Human Resource Management and Campus Recruitment & Training.

Mr. Krishna Prakash Janagam is perusing his final semester MCA in Sanketika Vidya Parishad Engineering College, accredited with A grade by NAAC, affiliated by Andhra University and approved by AICTE. With interest in machine learning Mr. Krishna Prakash has taken up his PG project on phishing site predictor and published the paper in connect to the project under the guidance of K. Tulasi Krishna Kumar, associate professor, SVPEC.

IX. REFERENCES

9.1 Book References

[1] A book on Fundamental of Cyber Security by Bhushan linked: http://surl.li/wqulff

[2] A book on Hacking by Harsh Bothra[16] linked: http://surl.li/eyllqh

[3] A book on Information Disclosure on Social Networking Sites: An Exploratory Survey of Factors
Impacting User Behaviour on Facebook by Clare Doherty linked:https://www.igi-global.com/chapter/information-disclosure-on-socialnetworkingsites/228764

[4] A book on Comdex Cyber Security A Complete Solution by Dir. Namrata Agrawal linked: http://surl.li/jtswsg

[5] A book on Cybersecurity for Dummies by Joseph Steinberg linked: http://surl.li/oeyxzj

[6] A book on Cyber Security, w/cd | IM | BS | e by Nina Godbole linked: http://surl.li/fzwouw

[7] A book on Machine Learning Based Prediction Techniques by Manza Prof R R Dr linked: http://surl.li/hvtsgf

[8] A book on Information Security and Cyber Laws by Gupta Sarika linked: http://surl.li/esjfdk

9.2 Web References

[9] A web reference on Phishing web site detection using diverse machine learning algorithms by Ammara Zamir in ISSN linked: https://www.emerald.com/insight/content/doi/10.1108/EL-052019- 0118/full/html

[10] A web reference on A Survey of Machine Learning-Based Solutions for Phishing Website Detection by Lizhen Tang linked: http://surl.li/esjfdk

[11] A web reference on Phishing Website Classification and Detection[17] Using Machine Learning by Jitendra Kumar in IEEE.linked: https://ieeexplore.ieee.org/abstract/document/9104161

[12] A web reference on Detection of Phishing Websites by Using Machine Learning-Based URL Analysis by Mehmet Korkmaz in IEEE. linked: https://ieeexplore.ieee.org/abstract/document/9225561

[13] A web reference on Efficient prediction of phishing websites using supervised learning algorithms by V. Santhana Lakshmi. linked: https://www.sciencedirect.com/science/article/pii/S187770581200940X

[14] A web reference on CANTINA+: A Feature-Rich Machine Learning Framework for Detecting Phishing Web Sites by Guang Xiang in di. linked: https://dl.acm.org/doi/abs/10.1145/2019599.2019606

9.3 Article Reference

[15] A Journal on Predicting Phishing Websites Using Classification Mining Techniques[18] with Experimental Case Studies by M. A. Hossain in IEEE linked: https://ieeexplore.ieee.org/abstract/document/5501434

[16] An article on Phishing web site detection using diverse machine learning algorithms by Ammara Zamir in ISSN linked: https://www.emerald.com/insight/content/doi/10.1108/EL-0520190118/full/html

[17] An article on Phishing website detection based on effective machine learning approach by Gururaj Harina Halli Lokesh linked: https://www.tandfonline.com/doi/abs/10.1080/23742917.2020.1813396

[18] A journal on A machine learning based approach for phishing detection [19] using hyperlinks by Ankit Kumar Jain linked: https://link.springer.com/article/10.1007/s12652-018-0798-z

[19] A Predicting Phishing Websites Using Classification Mining [20] Techniques with Experimental Case Studies by Maher Aburrous in IEEE linked: https://ieeexplore.ieee.org/abstract/document/5501434