DETECTING OF E-BANKING PHISHING WEBSITE -USING MACHINE LEARNING APPROACH-

Smt. Radhikatai Pandav College of Engineering

GROUP MEMBERS :

NIKAHAT QURESHI
DIVYA BANSOD
BHAVANA WAGHMARE
NIKITA SAWANKAR
SUJATA GHODESWAR

GUIDE:
Prof:DRUGA WANJARI

OUTLINE:

. INTRODUCTION
.OBJECTIVE &MOTIVATION
.ADVANTAGE
.SOFTWARE REQUIREMENTS
.RESPONSIBILITIES
.FLOW CHART
. IMPLEMENTATION SCREENSHOTS
.CHALLENGES
.LIMITATIONS & FUTURE ENHANCEMENT
.REFERENCES

INTRODUCTION :

THERE ARE NUMBER OF USERS WHO PURCHASE PRODUCTS ONLINE AND MAKE PAYMENT THROUGH E- BANKING. THERE ARE E- BANKING WEBSITES WHO ASK USER TO PROVIDE SENSITIVE DATA SUCH AS USERNAME, PASSWORD OR CREDIT CARD DETAILS ETC OFTEN FOR MALICIOUS REASONS.

THIS TYPE OF E-BANKING WEBSITES IS KNOWN AS PHISHING WEBSITE. IN ORDER TO DETECT AND PREDICT E-BANKING PHISHING WEBSITE. WE PROPOSED AN INTELLIGENT, FLEXIBLE AND EFFECTIVE SYSTEM THAT IS BASED ON USING CLASSIFICATION DATA MINING ALGORITHM.

WE WILL IMPLEMENTED CLASSIFICATION ALGORITHM AND TECHNIQUES TO EXTRACT THE PHISHING DATA SETS CRITERIA TO CLASSIFY THEIR LEGITIMACY.

MOTIVATION:

THE E-BANKING PHISHING WEBSITE CAN BE DETECTED BASED ON SOME IMPORTANT CHARACTERISTICS LIKE URL AND DOMAIN IDENTITY, AND SECURITY AND ENCRYPTION CRITERIA IN THE FINAL PHISHING DETECTION RATE.

ONCE USER MAKES TRANSACTION THROUGH ONLINE WHEN HE MAKES PAYMENT THROUGH E-BANKING WEBSITE OUR SYSTEM WILL USE DATA MINING ALGORITHM TO DETECT WHETHER THE E-BANKING WEBSITE IS PHISHING WEBSITE OR NOT.

THIS APPLICATION CAN BE USED BY MANY E-COMMERCE ENTERPRISES IN ORDER TO MAKE THE WHOLE TRANSACTION PROCESS SECURE. DATA MINING ALGORITHM USED IN THIS SYSTEM PROVIDES BETTER PERFORMANCE AS COMPARED TO OTHER TRADITIONAL CLASSIFICATIONS ALGORITHMS.

WITH THE HELP OF THIS SYSTEM USER CAN ALSO PURCHASE PRODUCTS ONLINE WITHOUT ANY HESITATION.

ADVANTAGES

1. THIS SYSTEM CAN BE USED BY MANY E-COMMERCE WEBSITES IN ORDER TO HAVE GOOD CUSTOMER RELATIONSHIP.

2. USER CAN MAKE ONLINE PAYMENT SECURELY.

3. DATA MINING ALGORITHM USED IN THIS SYSTEM PROVIDES BETTER PERFORMANCE AS COMPARED TO OTHER TRADITIONAL CLASSIFICATIONS ALGORITHMS.

4. WITH THE HELP OF THIS SYSTEM USER CAN ALSO PURCHASE PRODUCTS ONLINE WITHOUT ANY HESITATION.

Software & Hardware Requirement

Software

Language	Python
Front End	Html, Css
Framework	Flask
Algorithm	SVM, C 4.5
Domain	Machine Learning , Security

Hardware

Processor	i3 or more
RAM	2 GB or more
OS	Windows xp or more

C4.5 Algorithm

C4.5 builds decision trees from a set of training data in the same way as ID3, using the concept of information entropy. The training data is a set {\displaystyle S={s_{1},s_{2},...}} of already classified samples. Each sample {\displaystyle s_{i}} consists of a p-dimensional vector {\displaystyle (x_{1,i},x_{2,i},...,x_{p,i})}, where the {\displaystyle x_{j}} represent attribute values or features of the sample, as well as the class in which {\displaystyle s_{i}} falls.
At each node of the tree, C4.5 chooses the attribute of the data that most effectively splits its set of samples into subsets enriched in one class or the other. The splitting criterion is the normalized information gain (difference in entropy). The attribute with the highest normalized information gain is chosen to make the decision. The C4.5 algorithm then recurses on the partitioned sublists.

This algorithm has a few base cases.

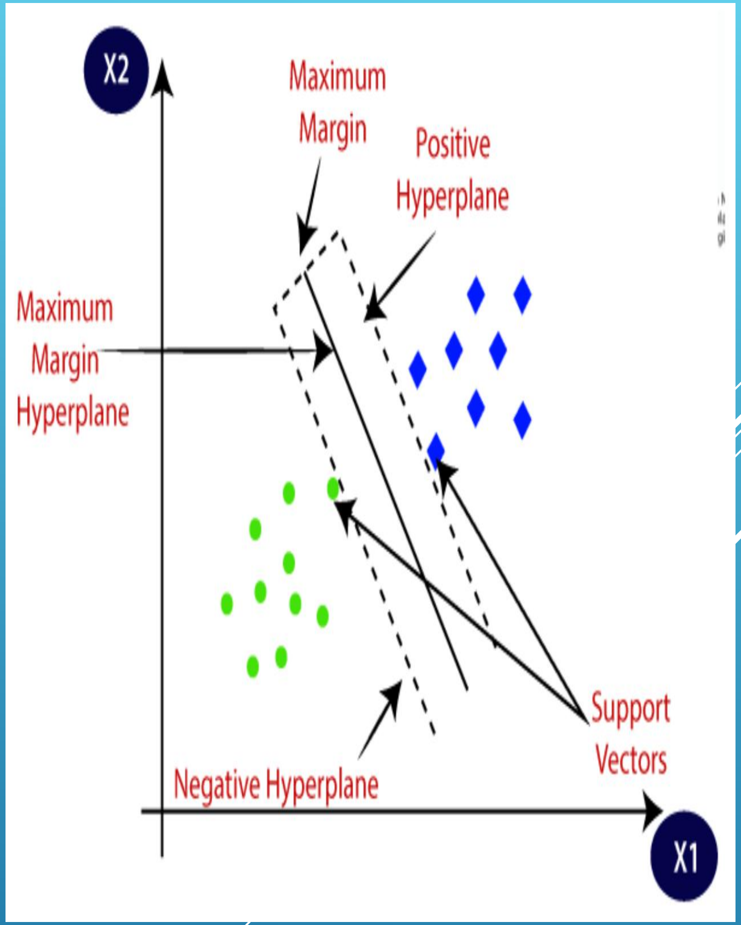
- All the samples in the list belong to the same class. When this happens, it simply creates a leaf node for the decision tree saying to choose that class.
- None of the features provide any information gain. In this case, C4.5 creates a decision node higher up the tree using the expected value of the class.
- Instance of previously-unseen class encountered. Again, C4.5 creates a decision node higher up the tree using the expected value.

SVM Algorithm

Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. However, primarily, it is used for Classification problems in Machine Learning.

The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane.

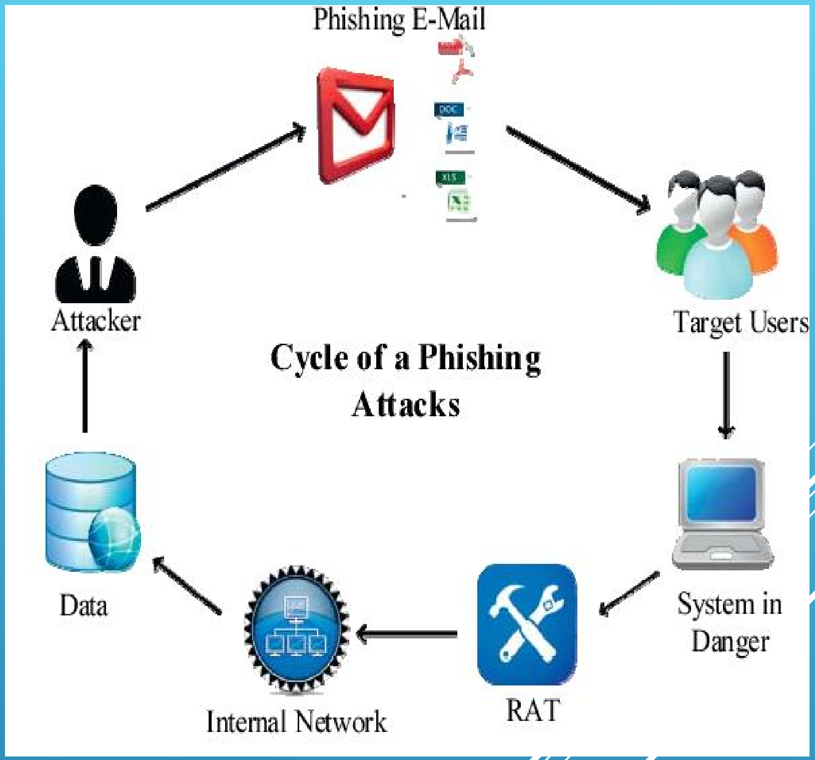
SVM chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called as support vectors, and hence algorithm is termed as Support Vector Machine. Consider the below diagram in which there are two different categories that are classified using a decision boundary or hyperplane:.

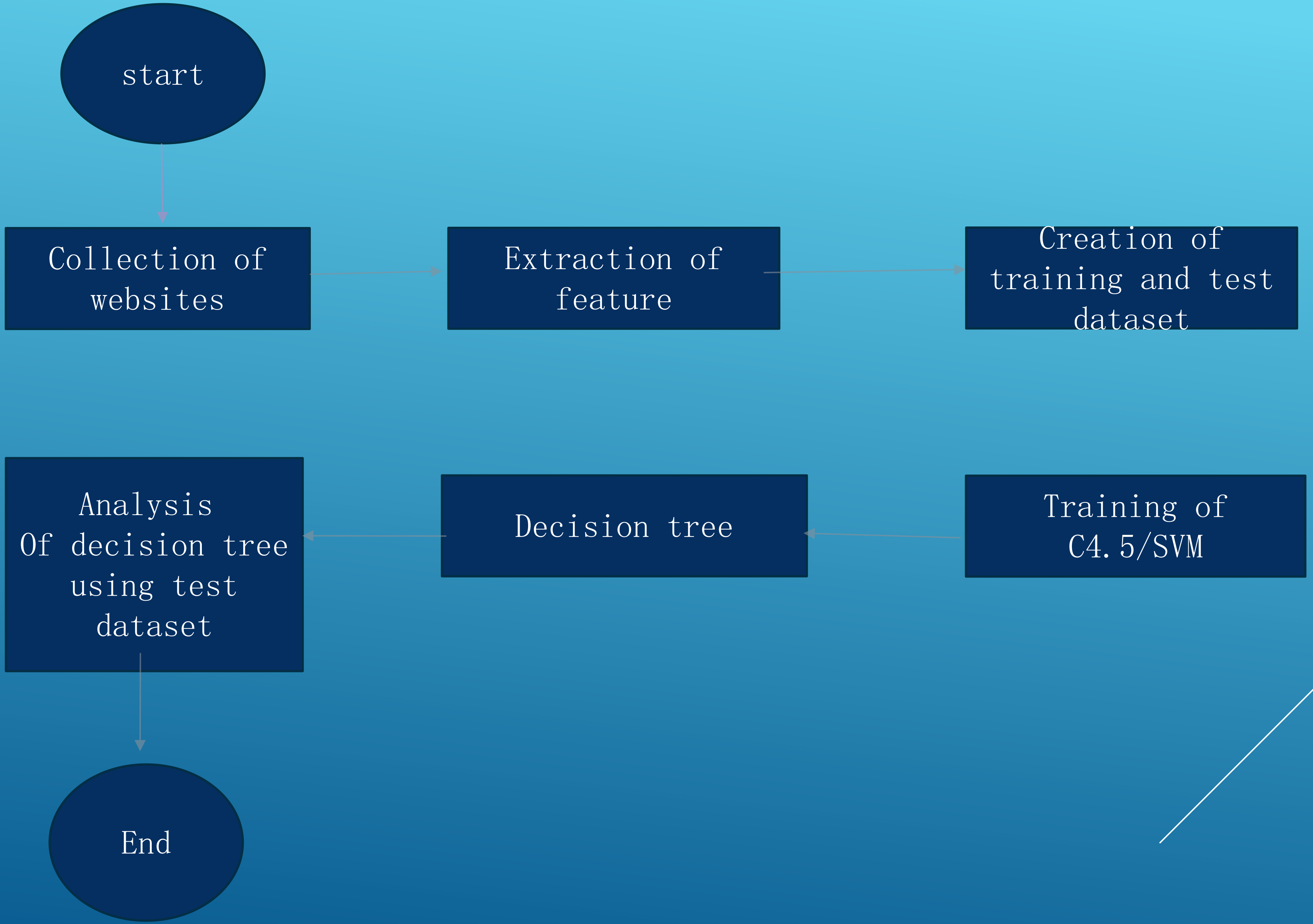


DATASET

In our dataset contains 48 features extracted from 30000 phishing webpages and 10000 legitimate webpages, which were downloaded from January to May 2019 and from May to June 2017. An improved feature extraction technique is employed by leveraging the browser automation framework (i.e., Selenium WebDriver), which is more precise and robust compared to the parsing approach based on regular expressions.

Anti-phishing researchers and experts may find this dataset useful for phishing features analysis, conducting rapid proof of concept experiments or benchmarking phishing classification models..





start

Collection of websites

Extraction of feature

Creation of training and test dataset

Training of C4.5/SVM

Decision tree

Analysis Of decision tree using test dataset

End

We have implemented python program to extract features from URL. Below are the features that we have extracted for detection of phishing URLs.

1) **Presence of IP address in URL**: If IP address present in URL then the feature is set to 1 else set to 0. Most of the benign sites do not use IP address as an URL to download a webpage. Use of IP address in URL indicates that attacker is trying to steal sensitive information.

2) **Presence of @ symbol in URL**: If @ symbol present in URL then the feature is set to 1 else set to 0. Phishers add special symbol @ in the URL leads the browser to ignore everything preceding the "@" symbol and the real address often follows the "@" symbol [4].

3) **Number of dots in Hostname:** Phishing URLs have many dots in URL. For example http://shop.fun.amazon.phishing.com, in this URL phishing.com is an actual domain name, whereas use of "amazon" word is to trick users to click on it. Average number of dots in benign URLs is 3. If the number of dots in URLs is more than 3 then the feature is set to 1 else to 0.

4) **Prefix or Suffix separated by (-) to domain:** If domain name separated by dash (-) symbol then feature is set to 1 else to 0. The dash symbol is rarely used in legitimate URLs. Phishers add dash symbol (-) to the domain name so that users feel that they are dealing with a legitimate webpage. For example Actual site is http://www.onlineamazon.com but phisher can create another fake website like http://www.online-amazon.com to confuse the innocent users

. 5) **URL redirection:** If "//" present in URL path then feature is set to 1 else to 0. The existence of "//" within the URL path means that the user will be redirected to another website

6) **HTTPS token in URL:** If HTTPS token present in URL then the feature is set to 1 else to 0. Phishers may add the "HTTPS" token to the domain part of a URL in order to trick users. For example, http://https-wwwpaypal-it-mpp-home.soft-hair.com

7) **Information submission to Emai**l: Phisher might use “mail()” or “mailto:” functions to redirect the user’ s information to his personal email[4]. If such functions are present in the URL then feature is set to 1 else to 0.

8) **URL Shortening Services “TinyURL”** : TinyURL service allows phisher to hide long phishing URL by making it short. The goal is to redirect user to phishing websites. If the URL is crafted using shortening services (like bit.ly) then feature is set to 1 else 0

9) Length of Host name: Average length of the benign URLs is found to be a 25, If URL’ s length is greater than 25 then the feature is set to 1 else to 0

10) **Presence of sensitive words in URL**: Phishing sites use sensitive words in its URL so that users feel that they are dealing with a legitimate webpage. Below are the words that found in many phishing URLs :- ’confirm’, ’account’, ’banking’, ’secure’, ’ebyisapi’, ’webscr’, ’signin’, ’mail’, ’install’, ’toolbar’, ’backup’, ’paypal’, ’password’, ’username’, etc;

11) **Number of slash in URL:** The number of slashes in benign URLs is found to be a 5; if number of slashes in URL is greater than 5 then the feature is set to 1 else to 0.

12) **Presence of Unicode in URL:** Phishers can make a use of Unicode characters in URL to trick users to click on it. For example the domain “xn--80ak6aa92e.com” is equivalent to ″ａｐｐｌｅ.com″. Visible URL to user is ″ａｐｐｌｅ.com″ but after clicking on this URL, user will visit to “xn--80ak6aa92e.com” which is a phishing site.

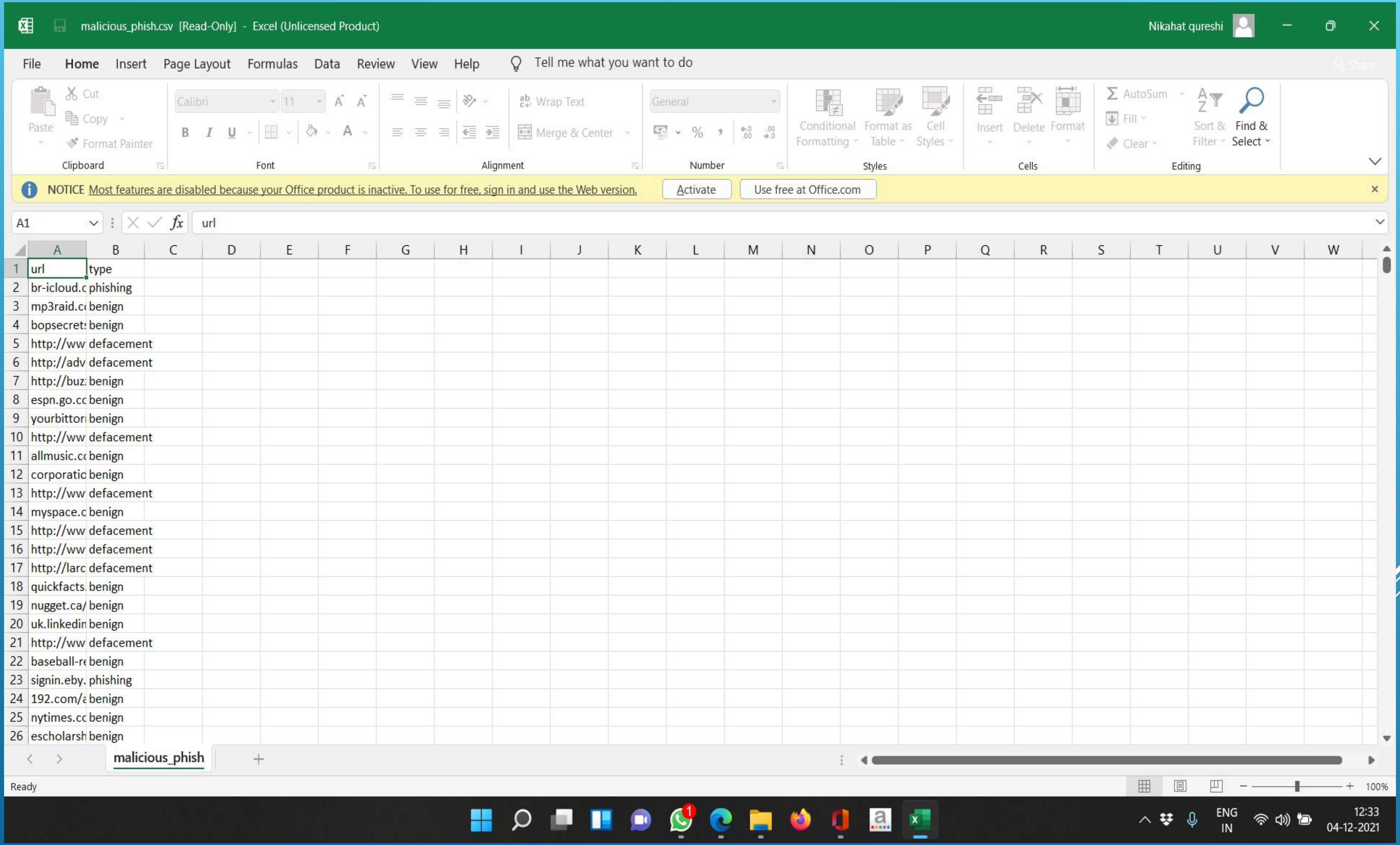
13) **Age of SSL Certificate**: The existence of HTTPS is very important in giving the impression of website legitimacy [4]. But minimum age of the SSL certificate of benign website is between 1 year to 2 year.

14) **URL of Anchor:** We have extracted this feature by crawling the source code oh the URL. URL of the anchor is defined by tag. If the tag has a maximum number of hyperlinks which are from the other domain then the feature is set to 1 else to 0.

15) **IFRAME:** We have extracted this feature by crawling the source code of the URL. This tag is used to add another web page into existing main webpage. Phishers can make use of the “iframe” tag and make it invisible i.e. without frame borders [4]. Since border of inserted webpage is invisible, user seems that the inserted web page is also the part of the main web page and can enter sensitive information. 16) Website Rank: We extracted the rank of websites and compare it with the first One hundred thousand websites of Alexa database. If rank of the website is greater than 10,0000 then feature is

Dataset 1 (consist of 549346 urls data)
https://www.kaggle.com/sid321axn/malicious-urls-dataset

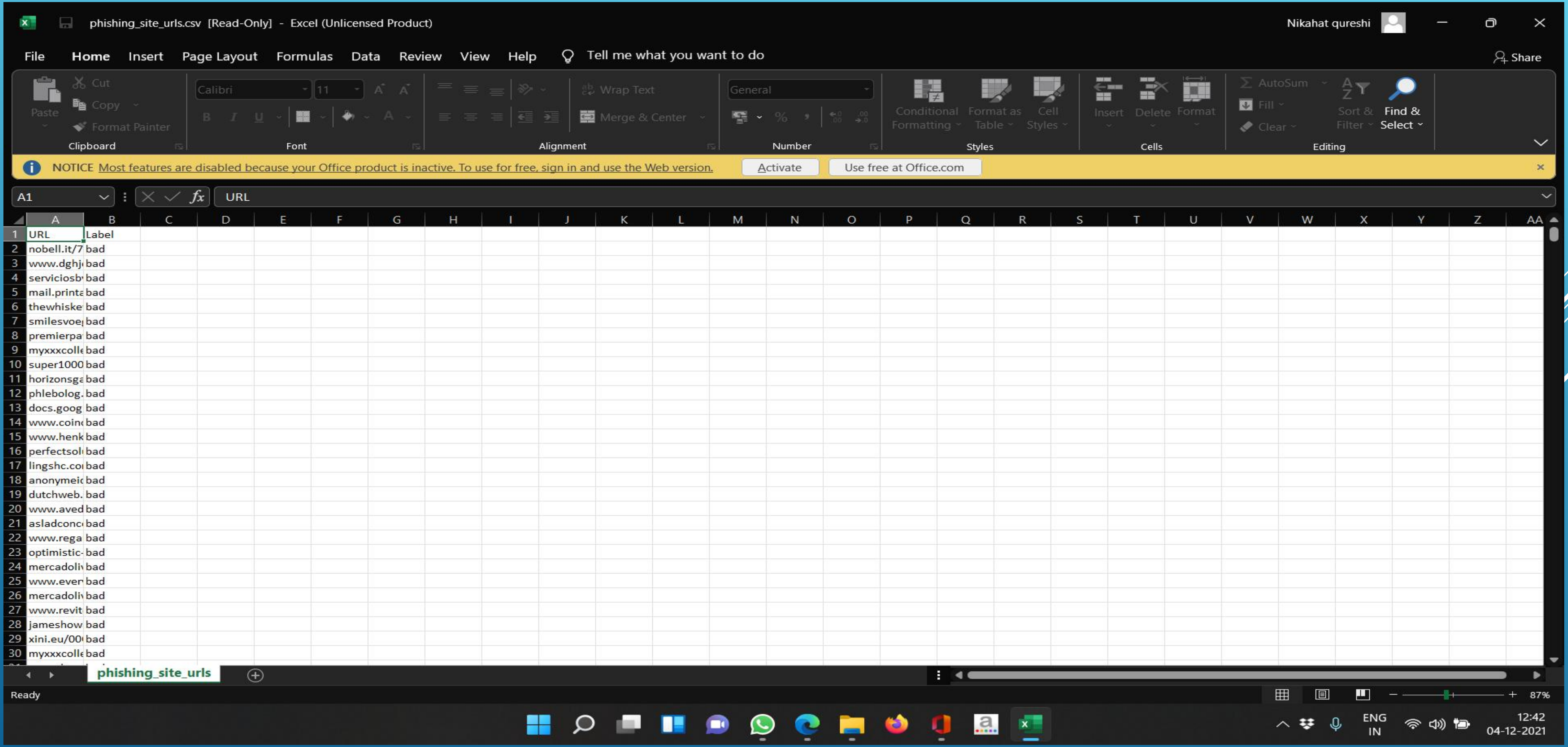
Dataset 2 (consist of 651,191 URLs data)
https://www.kaggle.com/anseldsouza/phishing-url-classification-using-knn-and-lr/data



NOTICE Most features are disabled because your Office product is inactive. To use for free, sign in and use the Web version.

A1 url

	A	B
1	url	type
2	br-icloud.c	phishing
3	mp3raid.co	benign
4	bopsecrets	benign
5	http://ww	defacement
6	http://adv	defacement
7	http://buz	benign
8	espn.go.co	benign
9	yourbittor	benign
10	http://ww	defacement
11	allmusic.co	benign
12	corporatio	benign
13	http://ww	defacement
14	myspace.c	benign
15	http://ww	defacement
16	http://ww	defacement
17	http://larc	defacement
18	quickfacts.	benign
19	nugget.ca/	benign
20	uk.linkedin	benign
21	http://ww	defacement
22	baseball-re	benign
23	signin.eby.	phishing
24	192.com/a	benign
25	nytimes.co	benign
26	escholarsh	benign



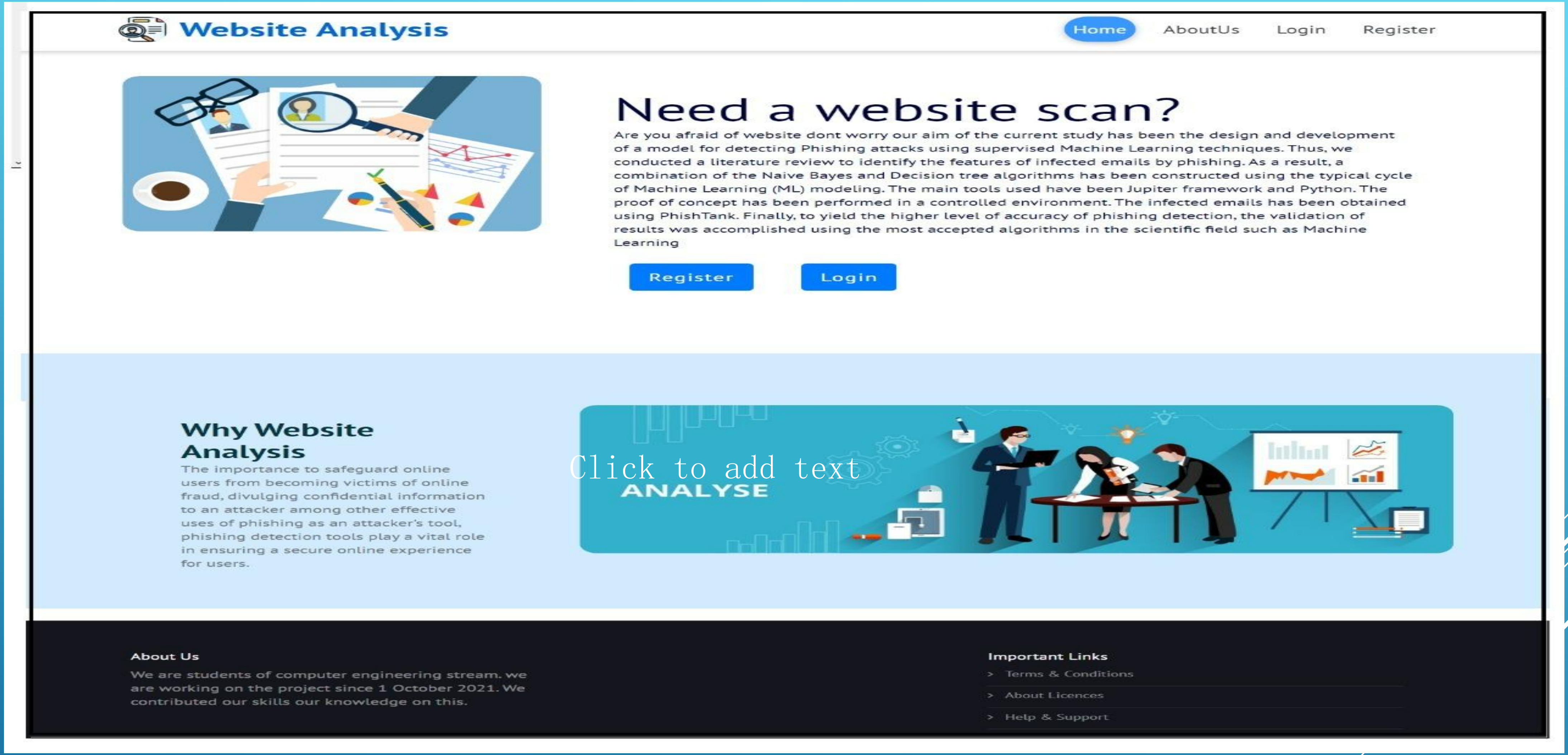
phishing_site_urls.csv [Read-Only] - Excel (Unlicensed Product)

Nikahat qureshi

NOTICE Most features are disabled because your Office product is inactive. To use for free, sign in and use the Web version.

URL

	A	B
1	URL	Label
2	nobell.it/7	bad
3	www.dghj	bad
4	serviciosb	bad
5	mail.printa	bad
6	thewhiske	bad
7	smilesvoe	bad
8	premierpa	bad
9	myxxxcolle	bad
10	super1000	bad
11	horizonsga	bad
12	phlebolog.	bad
13	docs.goog	bad
14	www.coin	bad
15	www.henk	bad
16	perfectsol	bad
17	lingshc.co	bad
18	anonymeic	bad
19	dutchweb.	bad
20	www.aved	bad
21	asladconc	bad
22	www.rega	bad
23	optimistic-	bad
24	mercadoliv	bad
25	www.ever	bad
26	mercadoliv	bad
27	www.revit	bad
28	jameshow	bad
29	xini.eu/00	bad
30	myxxxcolle	bad



Need a website scan?

Are you afraid of website dont worry our aim of the current study has been the design and development of a model for detecting Phishing attacks using supervised Machine Learning techniques. Thus, we conducted a literature review to identify the features of infected emails by phishing. As a result, a combination of the Naive Bayes and Decision tree algorithms has been constructed using the typical cycle of Machine Learning (ML) modeling. The main tools used have been Jupiter framework and Python. The proof of concept has been performed in a controlled environment. The infected emails has been obtained using PhishTank. Finally, to yield the higher level of accuracy of phishing detection, the validation of results was accomplished using the most accepted algorithms in the scientific field such as Machine Learning

Why Website Analysis

The importance to safeguard online users from becoming victims of online fraud, divulging confidential information to an attacker among other effective uses of phishing as an attacker's tool, phishing detection tools play a vital role in ensuring a secure online experience for users.

Click to add text

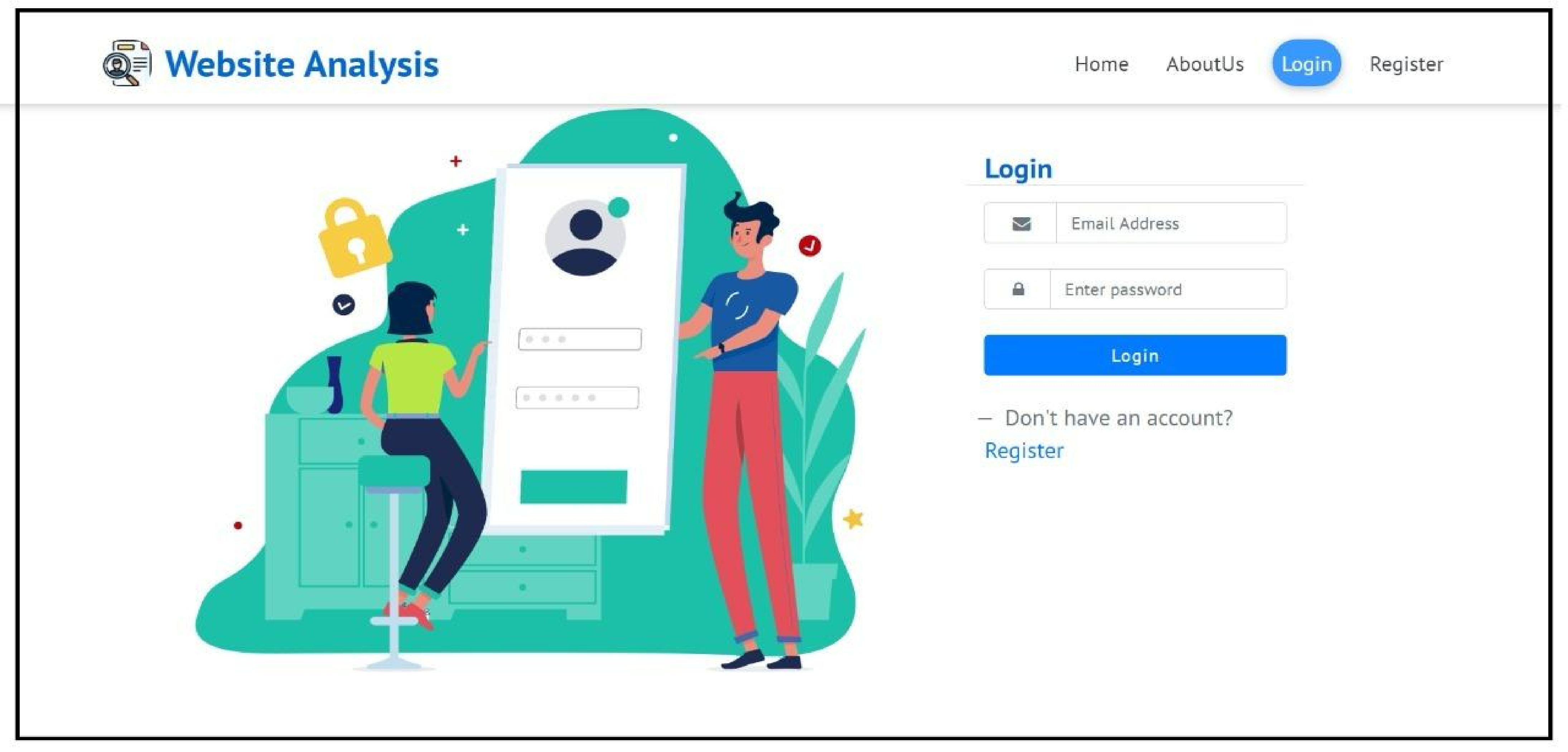
About Us

We are students of computer engineering stream. we are working on the project since 1 October 2021. We contributed our skills our knowledge on this.

Important Links

- Terms & Conditions
- About Licences
- Help & Support

HOME PAGE



Login

Email Address

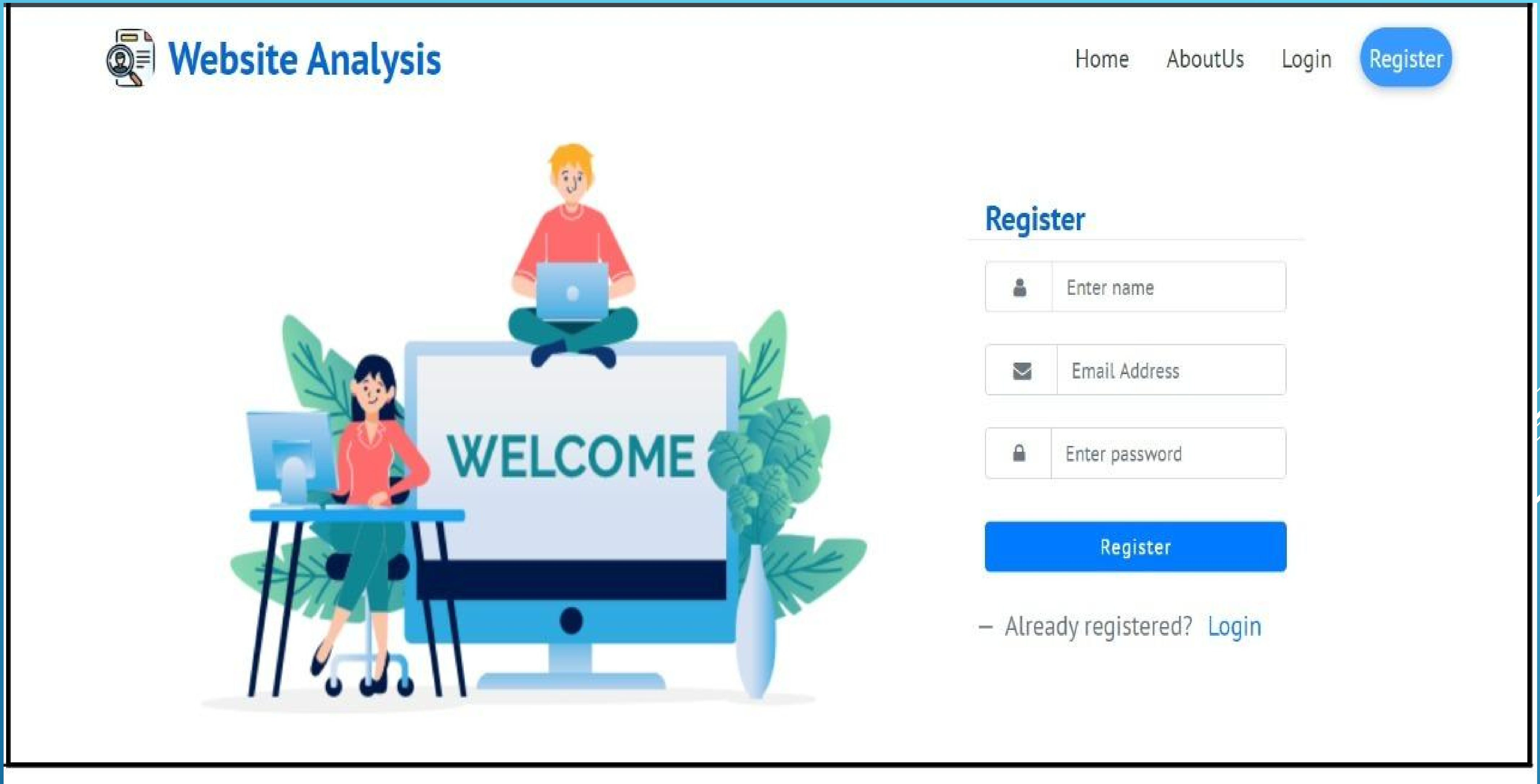
Enter password

Login

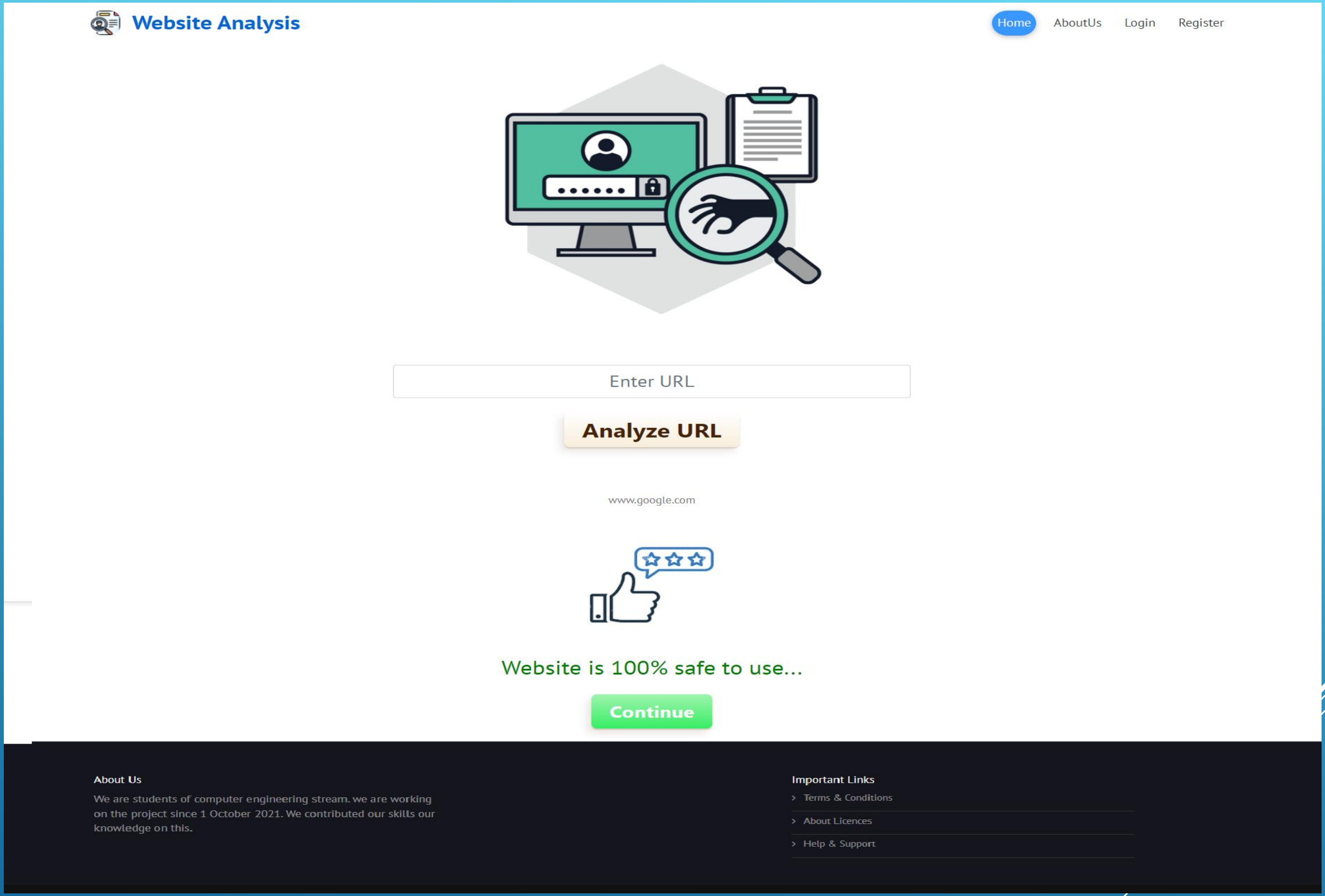
— Don't have an account?

Register

LOGIN PAGE



REGISTER PAGE



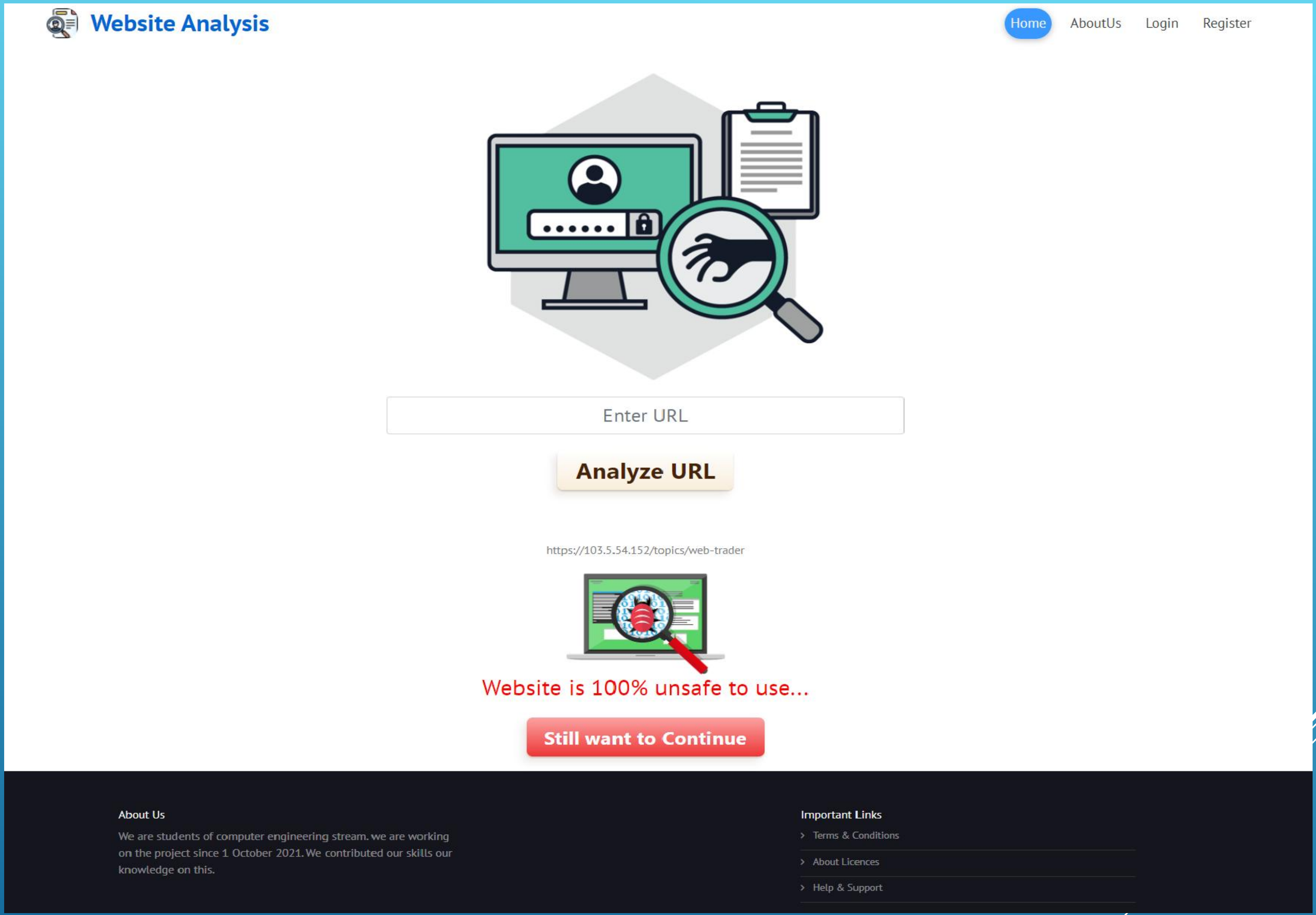
www.google.com

Website is 100% safe to use...

About Us

We are students of computer engineering stream. we are working on the project since 1 October 2021. We contributed our skills our knowledge on this.

LIVE PREDICTION MODUEL



https://103.5.54.152/topics/web-trader

Website is 100% unsafe to use...

About Us

We are students of computer engineering stream. we are working on the project since 1 October 2021. We contributed our skills our knowledge on this.

Important Links

- Terms & Conditions
- About Licences
- Help & Support

LIVE PREDICTION MODUEL