Web Phishing Detection

PROJECT REPORT

SUBMITTED BY

TEAM ID : PNT2022TMID33658

Arivazhagan R (922519104013)
Dinesh K (922519104036)
Dhivakar K (922519104034)
Dhanishsriram K (922519104031)

In partial fulfilment for the award of the degree

Of

BACHELOR OF TECHNOLOGY

IN

DEPARTMENT OF INFORMATION TECHNOLOGY

V.S.B.ENGINEERING COLLEGE ,KARUR

V.S.B ENGINEERING COLLEGE, KARUR

(Approved by AICTE & Affiliated by Anna University, Chennai)

Certified that this mini project report titled **"Web Phishing Detection"** is the bonafide record work by **Arivazhagan R (922519205013),Dinesh K (922519205036), Dhivakar K(922519205034), Dhanishsriram K (922519205031)** for **IBM-NALAIYATHIRAN** in **VII** semester of **B.TECH.,** degree course in **INFORMATION TECHNOLOGY** branch during the academic year of 2022-2023.

Staff-In Charge
Mr.S.Nelson

Evalutor
Dr.M.Parthiban

Head of the Department
Mr.K.Manivannan

ACKNOWLEDGEMENT

First and foremost, we express my thanks to our parents for providing us a very nice environment for doing this mini project. We wish to express our sincere thanks to our founder and Chairman **Shri.V.S.BALSAMY** for his endeavor in educating us in this premier institution.

We wish to express our appreciation and gratefulness to our principal, **Dr.V.NIRMAL KANNAN** and vice principal **Mr.T.S.KIRUBASANKAR** for their encouragement and sincere guidance.

We are grateful to our head of the department **Mr.K.Manivannan** and our Nalaiyathiran project coordinator **Mr.S.Nelson** Department of Information Technology for their valuable support.

We express our indebtedness to the supervisor of our Nalaiyathiran project, **Mr.S.Nelson** Assistant Professor, Department of Computer Science and Engineering, for guidance throughout the course of our project.

Our sincere thanks to all the teaching staff of V.S.B Engineering College and our friends for their help in the successful completion of this IBM Nalaiyathiran project work. Finally, we bow before God, the almighty who always had a better plan for us. We give our praise and glory to Almighty God for successful completion of this IBM Nalaiyathiran.

Web Phishing Detection

Date	19th November 2022
Team ID	PNT2022TMID33658
Project Title	WEB PHISHING DETECTION
Team Members	1.Arivazhagan R 2.Dinesh K 3.Dhivakar K 4.Dhanishsriram K

ABSTRACT:

Use of Web services among people for communication and various other purposes is inevitable in the current era. But also, this always increased the chances of fraudsters to steal data for evil intentions. As usage of internet services has increased drastically over the years, it has also increased the number of various vulnerabilities.

Web phishing has always been one of the most effective way to steal private data. Web phishing means deceiving a web user to somehow reach an certain endpoint of malicious programs that will affect the users personal data or may try to destroy it.

Even though some experienced users will differentiate phishing websites from the malicious ones, the scammers are becoming more creative on ways to cheat the users.

Inexperienced users are the ones who were always fell for the easiest of attacks. So to address this issue there is a need for an solution to prevent this from happening and

protect data that is meant to be private.

Our solution uses machnine learning to achieve this and has a good accuracy and performance. When compared to a blocking service, our approach performs better on generality and content since it uses learning techniques on the provided datasets.
Keywords: Web security, machine learning, phishing.
Git Repo:https://github.com/IBM-EPBL/IBM-Project-41642-1660643612

1. Introduction:

Phishing is a major threat to all Internet users and is difficult to trace or defend against since it does not present itself as obviously malicious in nature. In today's society, everything is put online and the safety of personal credentials is at risk. Phishing can be seen as one of the oldest and easiest ways of stealing information from people and it is used for obtaining a wide range of personal details. It also has a fairly simple approach – send an email, email sends victim to a site, site steals information.

Web phishing aims to steal private information, such as usernames, passwords, and credit card details, by way of impersonating a legitimate entity. It will lead to information disclosure and property damage. Large organizations may get trapped in different kinds of scams. This Guided Project mainly focuses on applying a machine-learning algorithm to detect Phishing websites. In order to detect and predict e-banking phishing websites, we proposed an intelligent, flexible and effective system that is based on using classification algorithms. We implemented classification algorithms and techniques to extract the phishing datasets criteria to classify their legitimacy. 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 a user makes a transaction online when he makes payment through an e-banking website our system will use a data mining algorithm to detect whether the e-banking website is a phishing website or not.

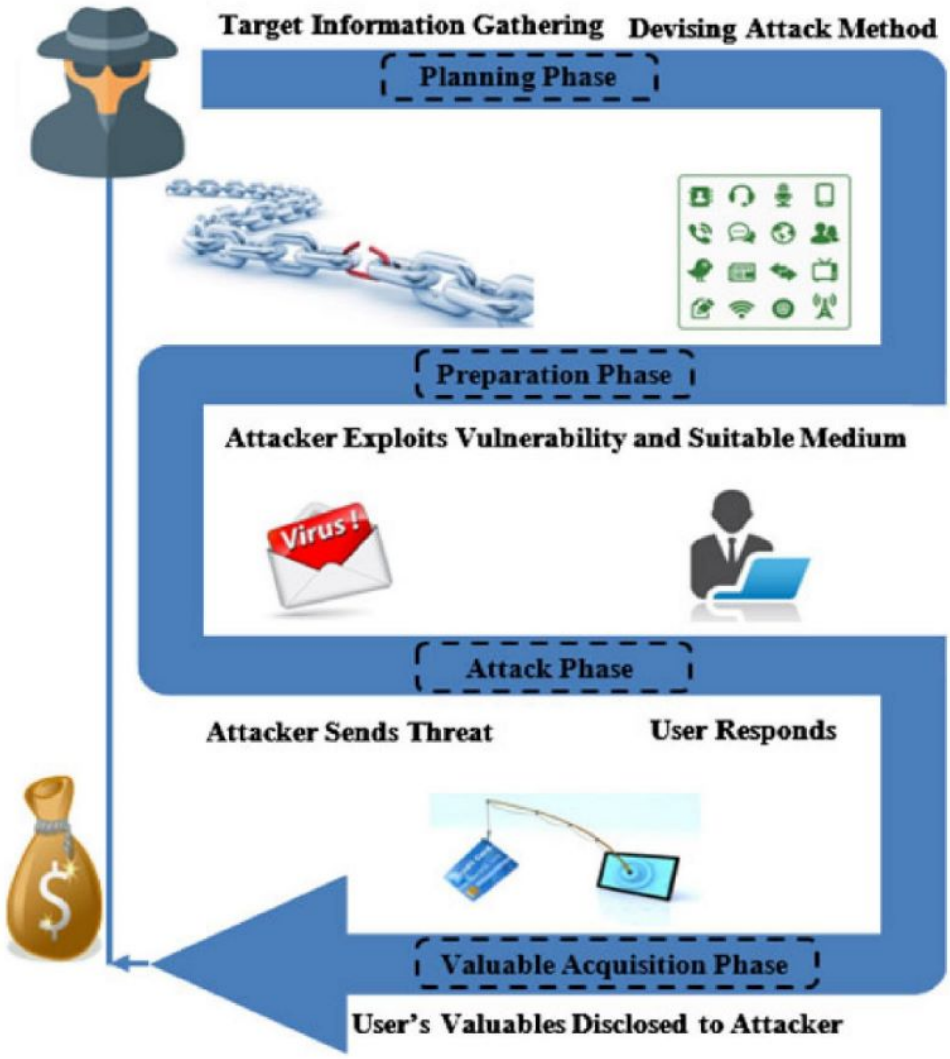
Phishing can come in many different forms, from obvious-to-spot frauds to sophisticated deceptions, but they share some common characteristics. Before you click a link, consider if the message you are reading contains these suspicious attributes:

- Sense of urgency and time constraint
- Fear of losing money or winnings
- Requests to verify accounts or credit card numbers

- Communication from services you do not use
- PDF Attachments from businesses
- Generic email providers
- Poor grammar and spelling
- Confirmations that lack details, such as delivery locations or travel dates
- Any emails from the IRS
- Unexpected, but out of character, emails from people you know
- Files or links that require you to download additional software to view them
- Close, but not quite right, links

Phishing is a field of study that merges social psychology, technical systems, security subjects, and politics. Phishing attacks are more prevalent: a recent study found that nearly 90% of organizations faced targeted phishing attacks in 2019. From which 88% experienced spear-phishing attacks, 83% faced voice phishing (Vishing), 86% dealt with social media attacks, 84% reported SMS/text phishing (SMishing), and 81% reported malicious USB drops. The 2018 Proofpoint annual report has stated that phishing attacks jumped from 76% in 2017 to 83% in 2018, where all phishing types happened more frequently than in 2017. The number of phishing attacks identified in the second quarter of 2019 was notably higher than the number recorded in the previous three quarters.

Phishing Process Flow and Phases



2. LITERATURE SURVEY

1. Oluwatobi Ayodeji Akanbi, ... Elahe Fazeldehkordi, in A Machine-Learning Approach to Phishing Detection and Defense, 2015

This paper presented an intelligent phishing detection and protection scheme by employing a new approach using the integrated features of images, frames and text of phishing websites. An efficient ANFIS algorithm was developed, tested and verified for phishing website detection and protection based on the schemes proposed in Aburrous et al. (2010) and Barraclough and Sexton (2015). A set of experiments was performed using 13,000 available datasets.

Advantage:

The approach showed an accuracy of 98.3%, which so far, is the best-integrated solutions for web-phishing detection and protection.

2. LongfeiWu etal..., "Effective Defense Schemes for Phishing Attacks on Mobile Computing Platforms, "

In this paper, author did a comprehensive study on the security vulnerabilities caused by mobile phishing attacks, including the web page phishing attacks.

Advantage:

Author propose MobiFish, a novel automated lightweight anti- phishing scheme for mobile platforms. MobiFish verifies the validity of web pages, applications, and persistent accounts by comparing thee actual Identity to the claimed identity

3. Surbhi Gupta etal., "A Literature Survey on Social Engineering Attacks: Phishing Attacks," in International Conference on Computing, Communication and Automation(ICCCA2016)

To fool an online user into elicit personal Information. The prime objective of this review is to do literature survey on social engineering attack:

Phishing attacks and techniques to detect attack.

Advantage:

The paper discusses various types of Phishing attacks such as
Tab-napping, spoofing emails, Trojan horse, hacking and how to prevent them.

4. SANS Institute, "Phishing : An Analysis of a Growing Problem",2007.
This paper gives an in depth analysis of phishing : what it is, the technologies and security Weaknesses it takes advantage of the dangers it poses to end users.
Advantage:
In this analysis author explain the concepts and technology behind phishing, show how the threat is much more then just a nuisance or passing trend, and discuss how gangs of criminals are Using these scams to make a great deal of money.

5. Guardian Analytics, "A Practical Guide to Anomaly Detection Implications of meeting new FFIEC minimum expectations for layered security"

Commercial and retail account holders at financial institutions of all sizes are under attacks by sophisticated, Organized, Well-funded cyber criminals
Advantage:Anomaly detection solutions are readily available, are deployed quickly and immediately and automatically protect all account holders against all types of fraud attack with minimal Disruption to legitimate online banking activity.

6. literature survey on Retraction: Phishing website detection using machine learning and deep learning techniques" 1916

In phishing attacks, the intruder puts on an act as if it is a trusted organization with an intention to purloin liable and essential information. The methodology they discovered is a powerful technique to detect the phished websites and can provide more effective defenses for phishing attacks of the future
Advantage:
The association between independent variables as well as dependent variables can be formed without any presumptions about the statistical depiction of the aspect. It contributes positive gains on regression algorithm which includes its competence to act with noisy data.

7. Phishing Website Detection Based on Deep Convolutional Neural Network and

Random Forest Ensemble Learning

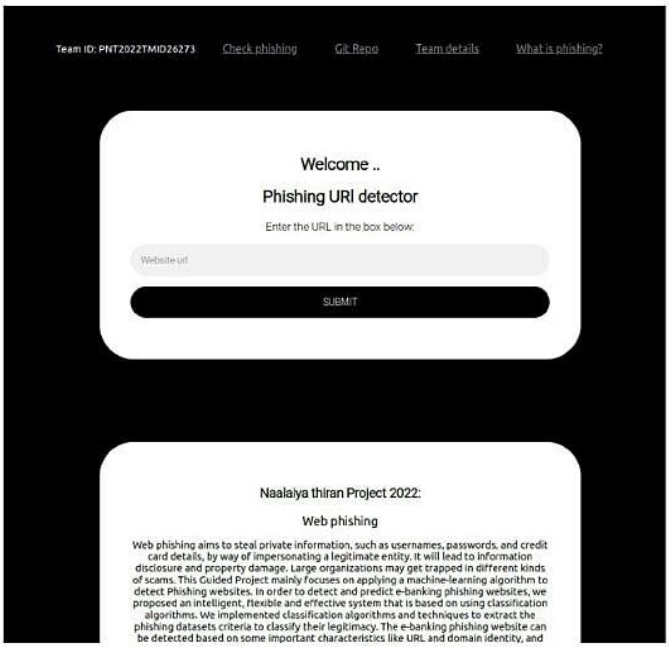
This paper proposes an integrated phishing website detection method based on convolutional neural networks (CNN) and random forest (RF). The method can predict the legitimacy of URLs without accessing the web content or using third-party services. The proposed technique uses character embedding techniques to convert URLs into fixed-size matrices, extract features at different levels Advantage:

A 99.35% correct classification rate of phishing websites was obtained on the dataset. Experiments were conducted on the test set and training set, and the experimental results proved that the proposed method has good generalization ability and is useful in practical applications.

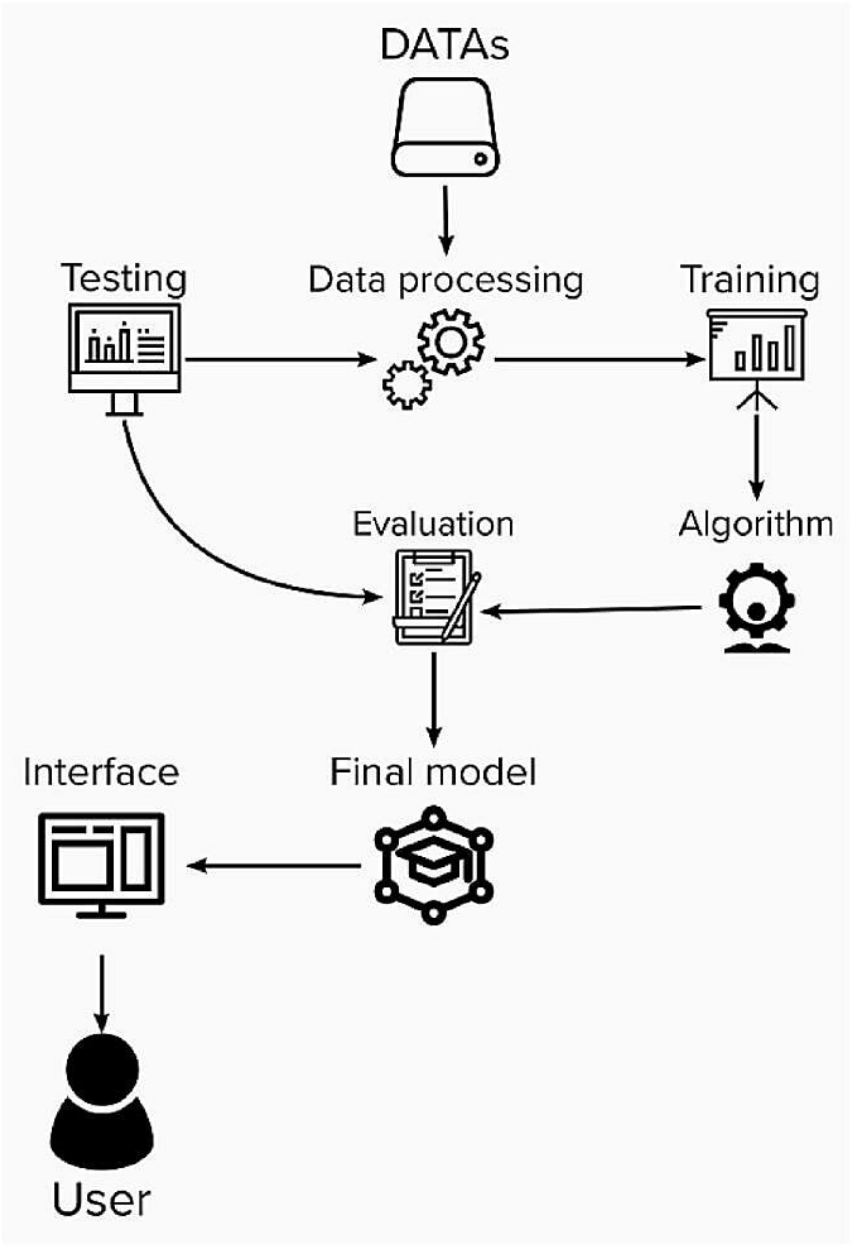
3. PROPOSED SYSTEM

Through Machine learning, user will be able to distinguish a phishing website using its URL(Uniform resource locator).

This web phishing detection project attains the customer satisfaction by discarding various kinds of malicious websites to protect their privacy. This project is not only capable of using by an single individual ,a large social community and a organisation can use this web phishing detection to protect their privacy. This project helps to block various malicious websites simultaneously.


Team ID: PNT2022TMID26273
Check phishing
Git Repo
Team details
What is phishing?
Welcome ..
Phishing URl detector
Enter the URL in the box below:
Website url
SUBMIT
Naalaiya thiran Project 2022:
Web phishing
Web phishing aims to steal private information, such as usernames, passwords, and credit card details, by way of impersonating a legitimate entity. It will lead to information disclosure and property damage. Large organizations may get trapped in different kinds of scams. This Guided Project mainly focuses on applying a machine-learning algorithm to detect Phishing websites. In order to detect and predict e-banking phishing websites, we proposed an intelligent, flexible and effective system that is based on using classification algorithms. We implemented classification algorithms and techniques to extract the phishing datasets criteria to classify their legitimacy. The e-banking phishing website can be detected based on some important characteristics like URL and domain identity, and


Naalaiya thiran Project 2022:
Web phishing
Web phishing aims to steal private information, such as usernames, passwords, and credit card details, by way of impersonating a legitimate entity. It will lead to information disclosure and property damage. Large organizations may get trapped in different kinds of scams. This Guided Project mainly focuses on applying a machine-learning algorithm to detect Phishing websites. In order to detect and predict e-banking phishing websites, we proposed an intelligent, flexible and effective system that is based on using classification algorithms. We implemented classification algorithms and techniques to extract the phishing datasets criteria to classify their legitimacy. 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 a user makes a transaction online when he makes payment through an e-banking website our system will use a data mining algorithm to detect whether the e-banking website is a phishing website or not.
Team ID: PNT2022TMID26273
Team members :
Rohith VS | Pandian S | Santhosh kumar S
Selva viswanath S



Architecture of the proposed system

IV. METHODOLOGY

1. **Datasets collection**

   We have downloaded the dataset (https://www.kaggle.com/datasets/shashwatwork/web-page-phishing-detection-dataset)

2. **Data pre-processing**
   - Checking the number of columns and rows
   - Handling null values.
   - Applying visualization techniques
   - Finding outliers and replacing them
   - Perform encoding
   - Scaling of data for only the important columns

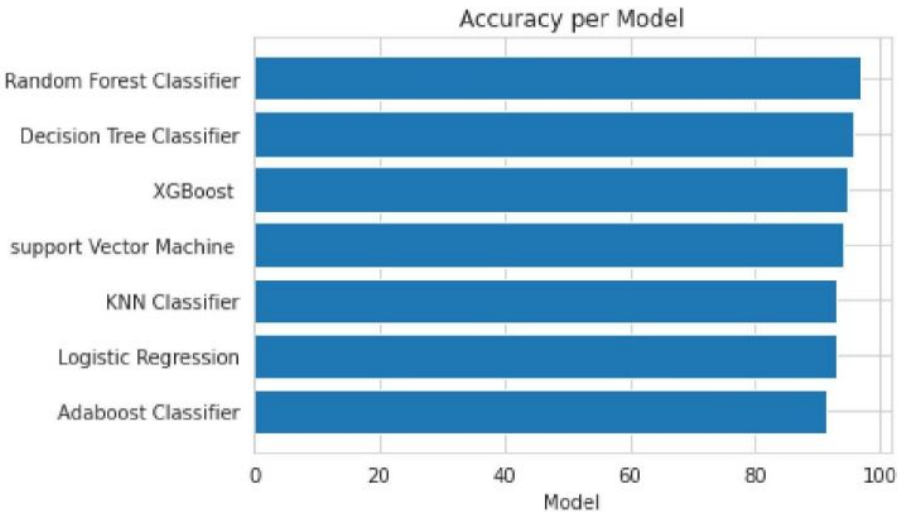
3. **Attribute of URL**

   Extraction of the attributes of the URL provided

4. **Apply various classification algorithms**

   Make predictions from different algorithms
   a. Logistic regression
   b. K-Means
   c. Decision tree
   d. Random forest classifier
   e. Support vector machine classifier
   f. AdaBoost classifier
   g. XGBoost classifier

5. **Compare prediction accuracy of all the classifiers**



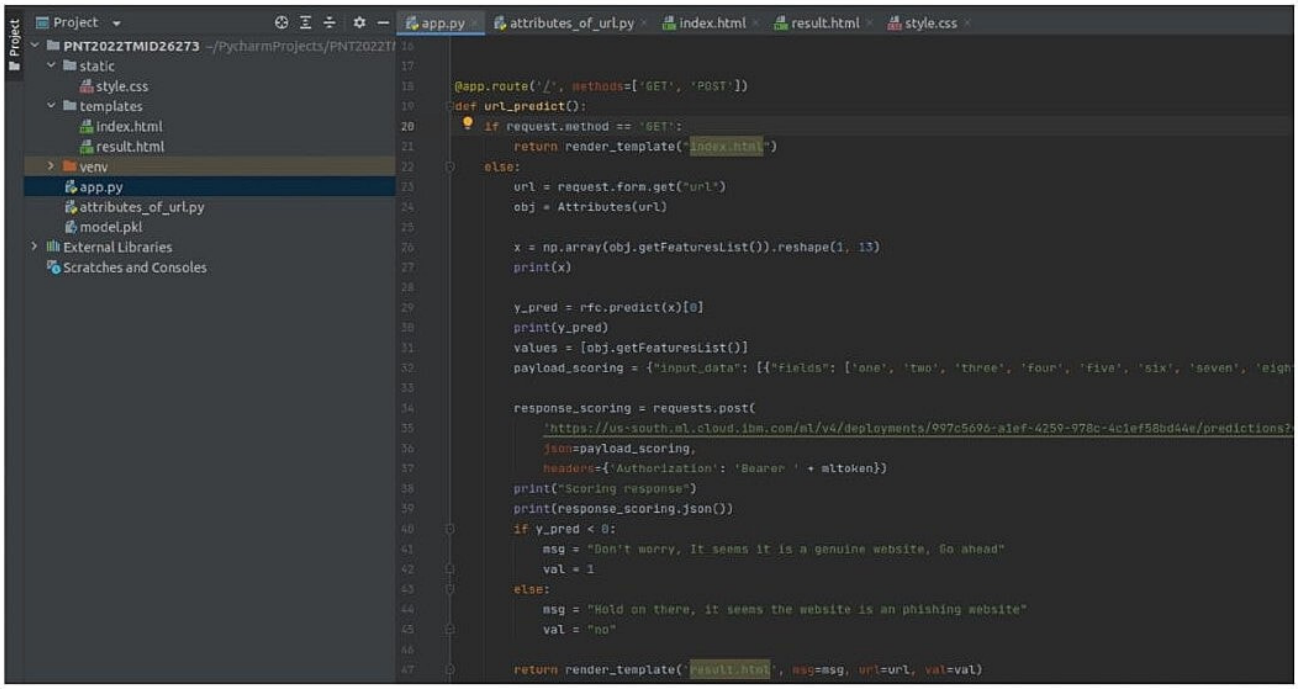

```
Logistic Regression Accuracy: 93.04545454545455
K-Nearest Neighbour Accuracy: 94.8409090909091
Decision Tree Classifier Accuracy: 95.36363636363636
Random Forest Classifier Accuracy: 97.11363636363637
support Vector Machine Accuracy: 94.75
Adaboost Classifier Accuracy: 91.93181818181819
XGBoost Accuracy: 95.45454545454545
```

6. **Choosing Random forest classifier as the model**

7. **Training the Model**

8. **Testing the model**

9. **Developing an Flask application and integrating it with Model**


Project
PNT2022TMID26273 ~/PycharmProjects/PNT2022TM
static
style.css
templates
index.html
result.html
venv
app.py
attributes_of_url.py
model.pkl
External Libraries
Scratches and Consoles
app.py
attributes_of_url.py
index.html
result.html
style.css
@app.route('/', methods=['GET', 'POST'])
def url_predict():
if request.method == 'GET':
return render_template("index.html")
else:
url = request.form.get("url")
obj = Attributes(url)
x = np.array(obj.getFeaturesList()).reshape(1, 13)
print(x)
y_pred = rfc.predict(x)[0]
print(y_pred)
values = [obj.getFeaturesList()]
payload_scoring = {"input_data": [{"fields": ['one', 'two', 'three', 'four', 'five', 'six', 'seven', 'eigh
response_scoring = requests.post(
'https://us-south.ml.cloud.ibm.com/ml/v4/deployments/997c5696-a1ef-4259-978c-4c1ef58bd44e/predictions?
json=payload_scoring,
headers={'Authorization': 'Bearer ' + mltoken})
print("Scoring response")
print(response_scoring.json())
if y_pred < 0:
msg = "Don't worry, It seems it is a genuine website, Go ahead"
val = 1
else:
msg = "Hold on there, it seems the website is an phishing website"
val = "no"
return render_template('result.html', msg=msg, url=url, val=val)

5. Technology Stack used

- Anaconda (Jupyter Notebook) with Python
  For Developing, training and testing the Machine learning model

- Flask (Python web framework):
  For developing Web application

- HTML, CSS:
  - For building the interface

- IBM Watson studio
  - Cloud platform for deploying the application

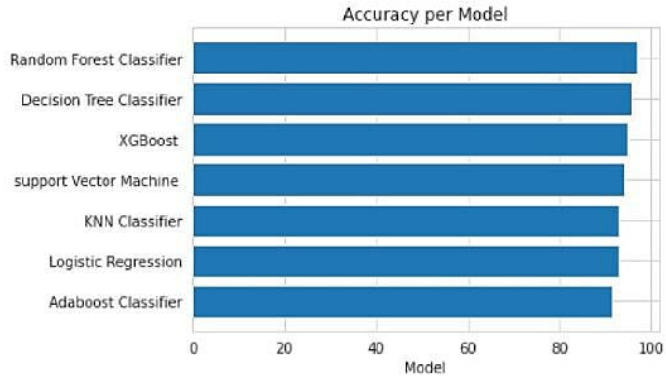
6. Attributes of the URL used

- having_IPhaving_IP_Address
- URLURL_Length
- Shortining_Service
- having_At_Symbol
- double_slash_redirecting
- Prefix_Suffix HTTPS_token
- on_mouseover
- RightClick
- popUpWidnow Iframe
- age_of_domain
- DNSRecord
- Statistical_report

7. Result

Various machine learning algorithms are imported from SciKit-learn library and implemented and Random forest classifier had the highest accuracy.

The accuracy is tabled and the Web application is developed.



References

1. https://towardsdatascience.com/machine-learning-classifiers-a5cc4e1b0623
2. https://www.geeksforgeeks.org/flask-creating-first-simple-application/
3. https://www.kaggle.com/datasets/shashwatwork/web-page-phishing-detection-dataset
4. https://www.geeksforgeeks.org/deploy-machine-learning-model-using-flask/