PROJECT REPORT
WEB PHISHING DETECTION

TEAM ID: PNT2022TMID20907

BATCH ID: B10-4A6E

TEAM MEMBERS:

- SHANTHA KUMAR T – TEAM LEADER
- MAHIBALAN J
- SOMASUNDAR M
- BALAMURUGAN S

TABLE OF CONTENTS:

➡ INTRODUCTION:

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.

➡ INSTALLATION:

The Code is written in Python 3.6.10. If you don't have Python installed you can find it [here](https://www.python.org/downloads/). If you are using a lower version of python and you can upgrade using the pip package, ensuring you have the latest version of pip. To install the required packages and libraries, run this command in the project directory after [cloning](https://www.howtogeek.com/451360/how-to-clone-a-github-repository/) the repository:

```
pip install -r requirements.txt
```

➡ <u>DIRECTORY TREE:</u>

- ➢ **PICKLE**
  - ➢ **model.pkl**
- ➢ **STATIC**
  - ➢ **styles.css**
- ➢ **TEMPLATES**
  - ➢ **index.html**
- ➢ **PHISHING URL DETECTION**
- ➢ **PROCFILE**
- ➢ **README.md**
- ➢ **app.py**
- ➢ **feature.py**
- ➢ **phishing.csv**
- ➢ **requirements.txt**

➡ <u>TECHNOLOGIES USED:</u>

- ➧ **NUMPY**
- ➧ **PANDAS**
- ➧ **SCIKIT**
- ➧ **FLASK**
- ➧ **MATPLOTLIB**

➡ RESULT:

Accuracy of various model used for URL detection.

ML Model	Accuracy	F1_SCORE	Recall	Precision
Gradient Boosting Classifier	0.974	0.977	0.994	0.986
Cat Boost Classifier	0.972	0.975	0.994	0.989
XG Boost Classifier	0.969	0.973	0.993	0.984
Multi-layer Perceptron	0.969	0.973	0.995	0.981
Random Forest	0.967	0.971	0.993	0.990
Support Vector Machine	0.964	0.968	0.980	0.965
Decision Tree	0.960	0.964	0.991	0.993
K-Nearest Neighbours	0.956	0.961	0.991	0.989
Logistic Regression	0.934	0.941	0.943	0.927
Naive Bayes Classifier	0.605	0.454	0.292	0.997

➡ CONCLUSION:

➤ The final take away from this project is to explore various machine learning models, perform Exploratory Data Analysis on phishing dataset and understanding their features.

➤ Creating this notebook helped me to learn a lot about the features affecting the models to detect whether URL is safe or not, also I came to know how to tuned model and how they affect the model performance.

➤ The final conclusion on the Phishing dataset is that the some feature like "HTTTPS", "Anchor URL", "Website Traffic" have more importance to classify URL is phishing URL or not.

➤ Gradient Boosting Classifier currently classify URL up to 97.4% respective classes and hence reduces the chance of malicious attachments.