COGNITIVE COMPUTING PROJECT REPORT

BY

COGNISEC

Navnidhi 102317298

Purvika Bansal 102317121

2Q15(2Q1E)

Submitted to

Mr. Sukhpal Singh

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

THAPAR INSTITUTE OF ENGINEERING AND TECHNOLOGY, (A DEEMED TO BE UNIVERSITY), PATIALA

INDIA

Jan-May, 2025

WEBSITE PHISHING DETECTION REPORT

1. Introduction & Problem Statement

Phishing attacks are a big cybersecurity problem where fake websites pretend to be real ones to steal personal information. Traditional methods that rely on fixed rules often fail because phishing techniques keep changing. This project aims to create a smart system using machine learning to detect phishing websites by analysing their URLs and content.

2. Dataset Overview

To train and evaluate our phishing detection model, we used publicly available datasets from:

- **Kaggle**

The header list (column names) is as follows :

['Using IP', 'Long URL', 'Short URL', 'Symbol @', 'Redirecting //', 'Prefix Suffix-', 'Sub Domains', 'HTTPS', 'Domain Reg Len', 'Favicon', 'Non Std Port', 'HTTPSDomainURL', 'Request URL', 'Anchor URL', 'Links In Script Tags', 'Server Form Handler', 'Info Email', 'Abnormal URL ','Website Forwarding', 'Status Bar Cust', 'Disable Right Click', 'Using Popup Window', 'I frame Redirection', 'Age of Domain', 'DNS Recording', 'Website Traffic', 'Page Rank', 'Google Index', 'Links Pointing To Page', 'Stats Report', 'class']

3. Technology Stack

- **Programming Language:** Python
- **Libraries:** Scikit-learn (data processing & evaluation), TensorFlow (ML model), Pandas (data handling), NumPy (numerical operations), Seaborn & Matplotlib (visualization)
- **Framework:** TensorFlow (MLP model construction & optimization)
- **Tools:** Jupyter Notebook (coding & testing)

4. ML Model Implementation & Evaluation

Feature Engineering

Feature extraction was done based on URL, domain, and content features. Features were normalized and transformed for better model performance.

Model Selection

We experimented with various machine learning algorithms:

- **Logistic Regression-**

  Accuracy score- 0.9172

- **Decision Tree & Random Forest**

  Decision Tree- Accuracy score -0.948

  Random Forest -Accuracy Score-0.9769

- **Neural Networks (Deep Learning)**

  Accuracy Score-0.9611

Performance Evaluation

We used precision, recall, Confusion matrix, and accuracy to measure model performance. The Random Forest model achieved the best accuracy of **97%**, outperforming other models.

5. Results & Insights (Visualizations & Metrics)

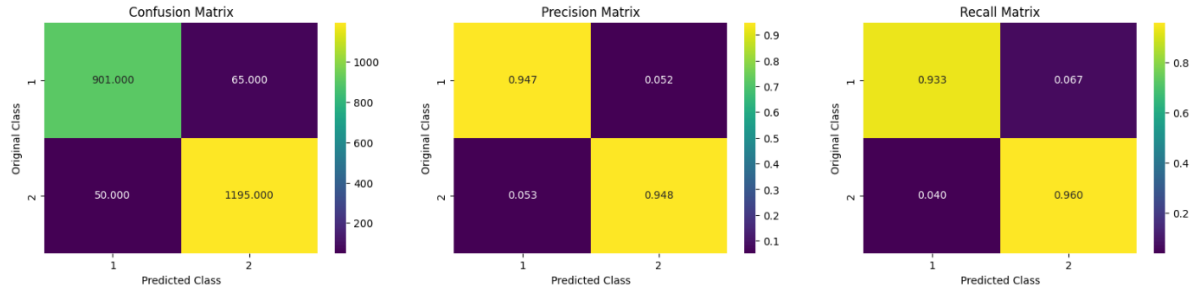
Key observations from our study:

- URL length and presence of special characters were strong indicators of phishing.
- Phishing websites often lack HTTPS security.
- Deep Learning models performed well but required high computational power.

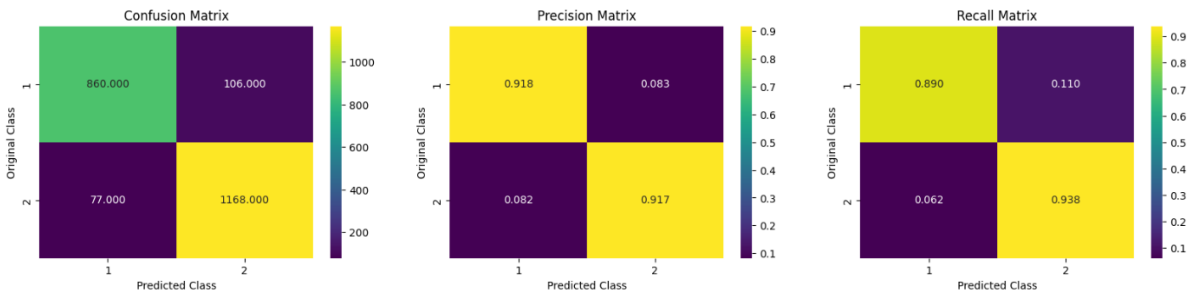
Visualizations included:

- **Confusion Matrix** to analyze model predictions.

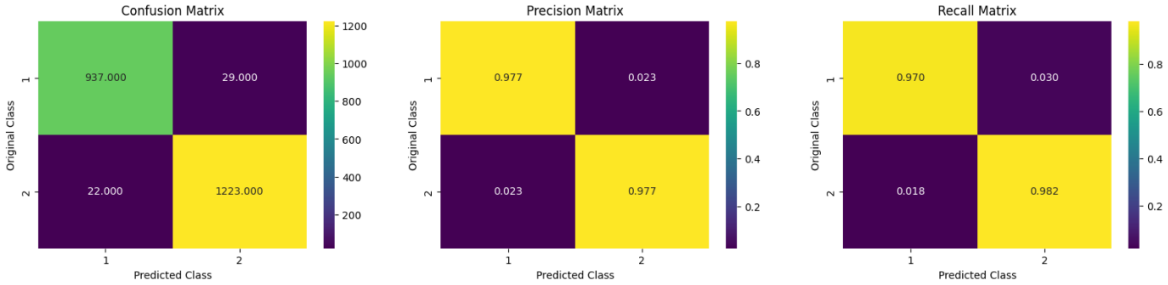
DECISION TREE:



LOGISTIC REGRESSION:



RANDOM FOREST :



6. Challenges & Future Improvements

Challenges:

- Imbalanced Data: More legitimate sites than phishing ones, requiring oversampling.
- Feature Extraction: Some features need real-time checks, making preprocessing harder.
- Deployment Issues: Ensuring smooth integration and fast response in a web-based system.

Future Improvements:

- Real-time Detection: API-based system for dynamic phishing analysis.
- Better Accuracy: Use NLP (BERT embeddings) for improved text analysis.
- Browser Extension: Instant phishing alerts and enhanced feature engineering.

7. Conclusion & Learnings

This project showed how machine learning can detect phishing sites using URL, domain, and content features. We learned the importance of choosing the right features, handling imbalanced data, and making the system work in real-time. Moving forward, we plan to deploy it for real-time detection and create a browser extension for better security.

8. References

UCI Machine Learning Repository: Phishing Websites Dataset
https://archive.ics.uci.edu/ml/datasets/phishing+websites

Life-long phishing attack detection using continual learning
https://www.nature.com/articles/s41598-023-37552-9

A systematic literature review on phishing website detection
https://www.sciencedirect.com/science/article/pii/S1319157823000034