PHISHING URL DETECTION USING MACHINE LEARNING

Cybersecurity Internship 2025

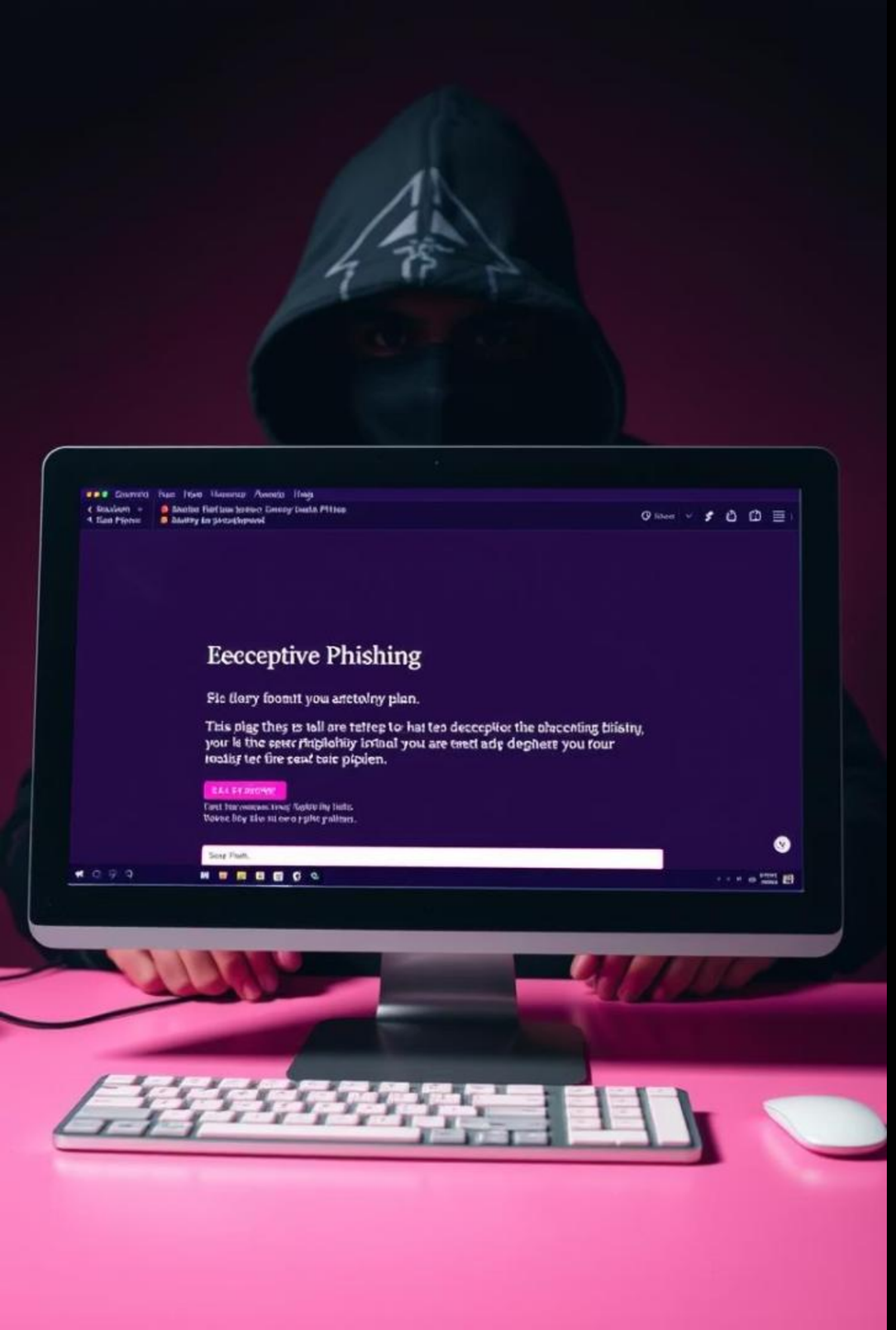
Digisuraksha Parhari Foundation

Powered by Infinisec Technologies Pvt. Ltd.

Presented by:

Saniya Santosh Choughule &

Iqra Mohd Nisar Khan



Understanding Phishing Threats

What is Phishing?

Fake websites trick users to steal sensitive data.

Why Dangerous?

Causes credential theft, financial loss, and data breaches.

Project Objective

Build Detection Tool

Use machine learning to identify phishing URLs.

User Protection

Flag suspicious links before damage occurs.

Technology Stack

Programming Language

Python

Tools & Libraries

- Scikit-learn
- Pandas
- Joblib

Model

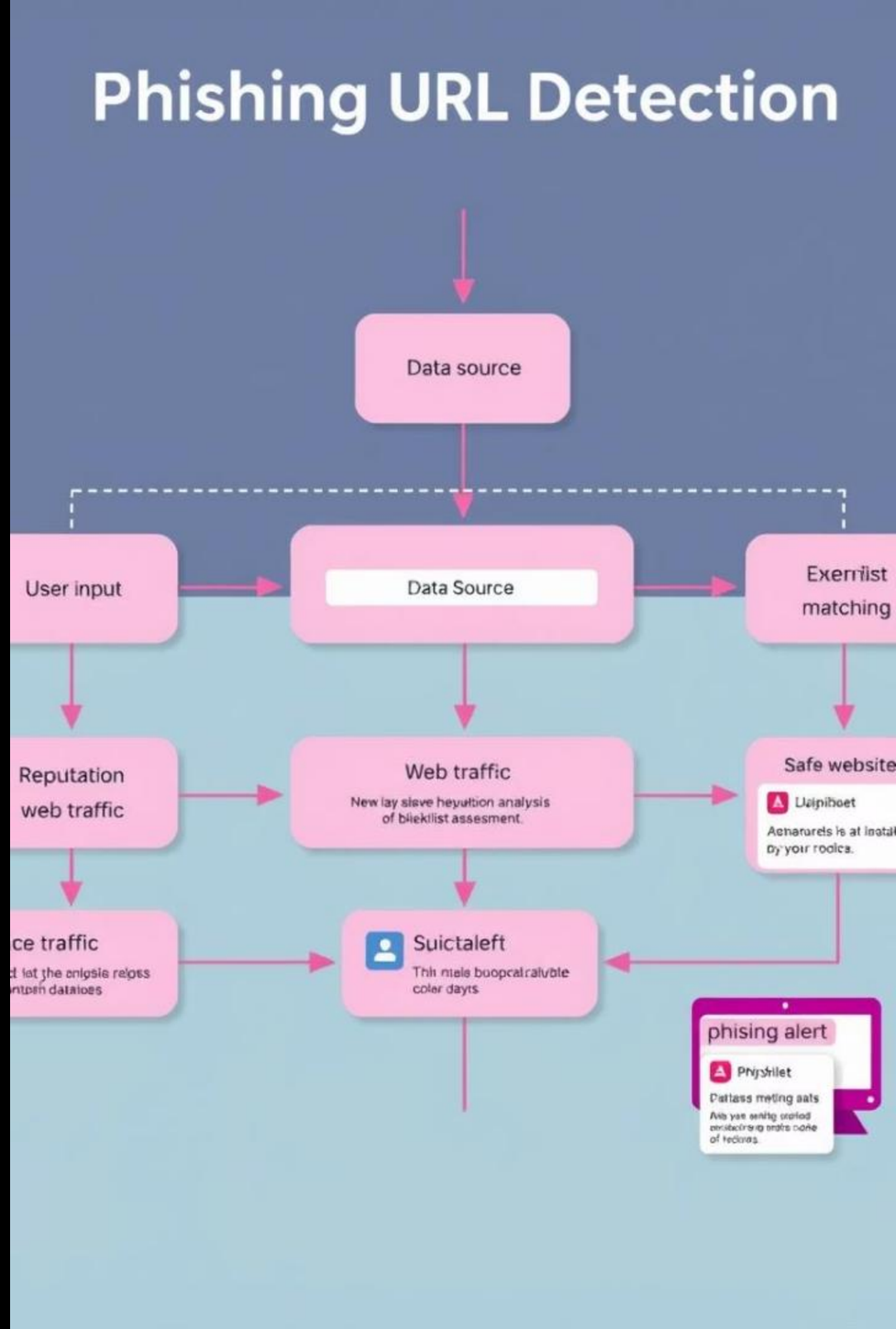
Random Forest Classifier

Environment

Local terminal CLI tool

System Architecture

- **User Input**
  Enter URL for analysis
  - **Feature Extraction**
    Analyze URL properties
    - **ML Model**
      Random Forest predicts legitimacy
      - **Prediction Output**
        Legitimate ( ) or Phishing (❌)



Dataset & Training Details

Data Source

Labeled open-source phishing and legitimate URLs

Features Extracted

- URL length
- HTTPS presence
- Domain structure

Model Accuracy

Achieved ~97% training accuracy

Model Saved

phishing_model.pkl file generated

Demo Overview

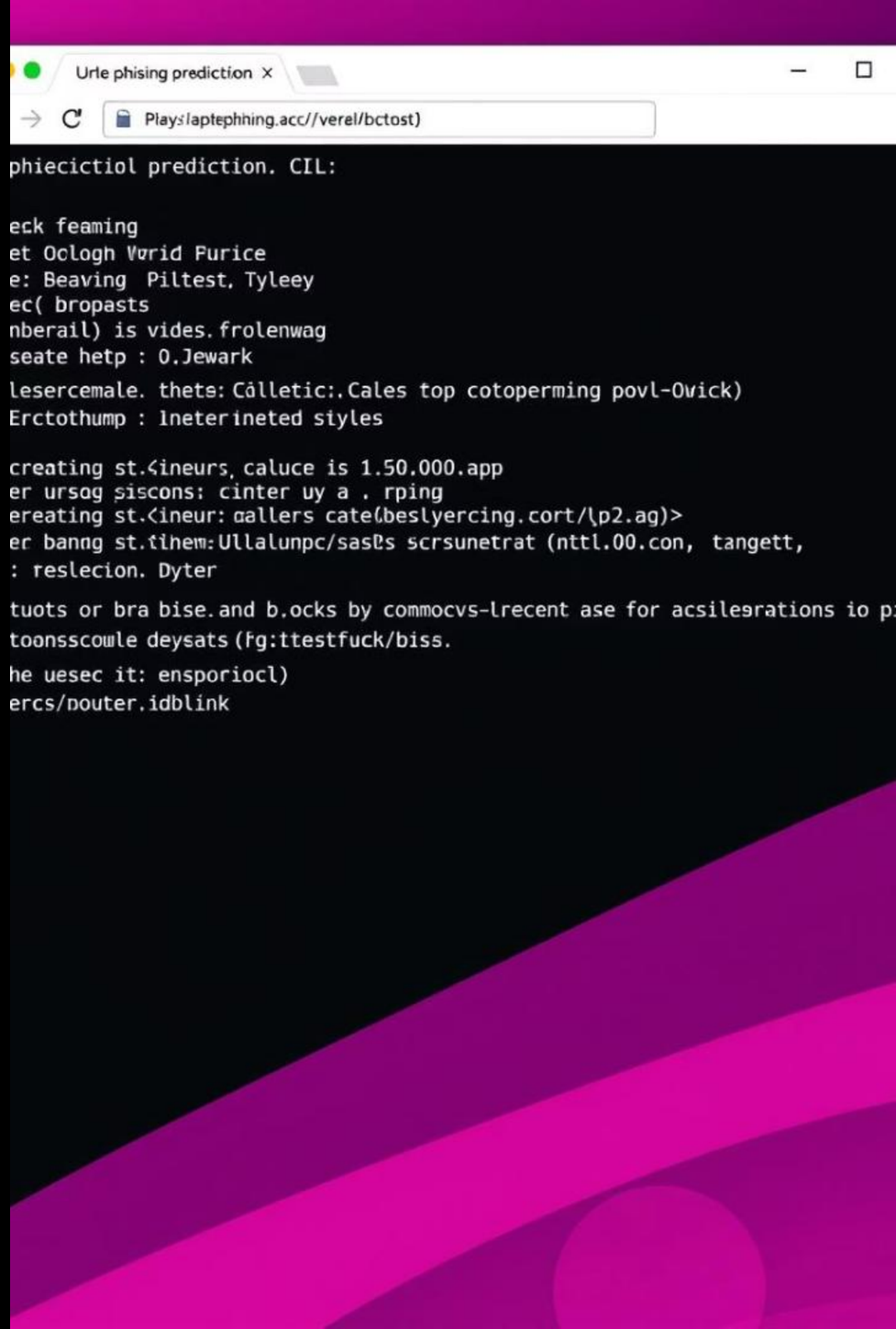
Usage

Enter URL in terminal using predict_cli.py script

Example

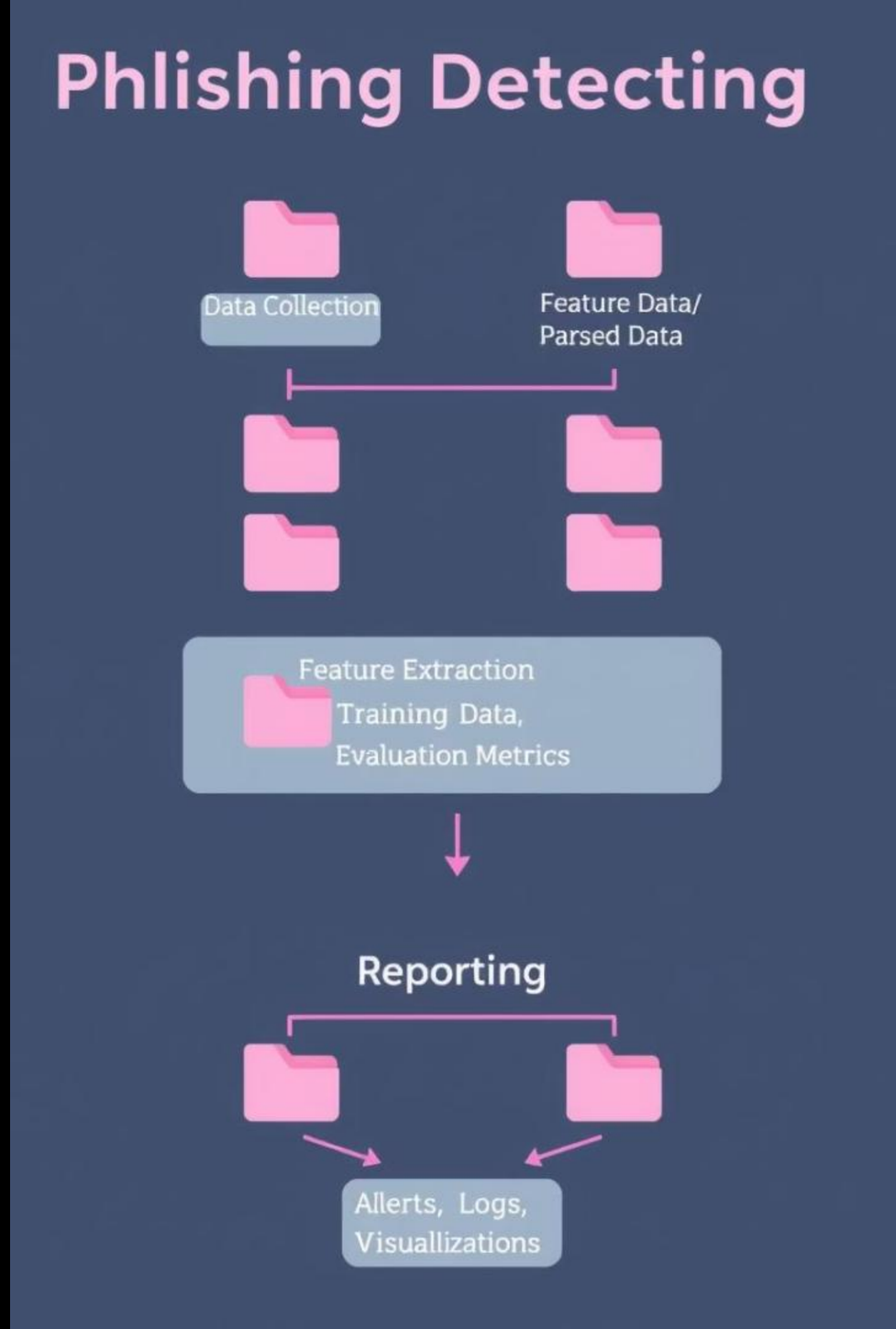
Input: http://www.bank-login.secureverify.com

Output: ❌ Phishing



Project Structure

```
Phishing-Detection/no
  ├── tool/
  ├── source_code/
  │    ├── train_model.py
  │    ├── predict_cli.py
  │    └── phishing_model.pkl
  ├── research-paper/
  ├── presentation/
  ├── demo/
  └── README.md
```



Project Outcome

CLI Tool

Accurately predicts phishing URLs

High Accuracy

Reliable results on test data

Extendable

Easy to integrate into web-based tools

Future Scope & Closing

1

Build Browser Extension

Real-time phishing URL scanning for users

2

Live Data Integration

Enhance model with continuously updated datasets

3

Continuous Improvement

Iterate algorithms for better detection rates

Thank you!

Any questions?