**INT375**

**PROJECT REPORT**

(Project Semester January-April 2025)

***Supervised Learning for Real-Time Phishing Website detection***

Submitted by:

Name: Priyanshu Sharma

Registration No: 12311900

Programme and Section : B.Tech CSE K23PM

Course Code: INT375

Under the Guidance of

Anand Kumar (UID : 30561)

**Discipline of CSE/IT**

**Lovely School of Computer Science and Engineering**

**Lovely Professional University, Phagwara**

**CERTIFICATE**

This is to certify that Priyanshu Sharma bearing Registration no. 12311900 has completed INT-375 project titled, **“Supervised Learning For Real-Time Phishing Website detection”** under my guidance and supervision. To the best of my knowledge, the present work is the result of his original development, effort and study.

**Signature and Name of the Supervisor**

**Dr. Anand Kumar**

**School of Computer Science and Engineering**

Lovely Professional University

Phagwara, Punjab.

Date: 12 April 2025

**DECLARATION**

I, Priyanshu Sharma student of B.Tech under CSE/IT Discipline at, Lovely Professional University, Punjab, hereby declare that all the information furnished in this project report is based on my own intensive work and is genuine.

Date: 11-04.2025 Signature

Registration No: 12311900 Priyanshu Sharma

**Supervised Learning for Real-Time Phishing**

**Website Detection**

### Priyanshu Sharma

BTech Computer Science & Engineering

Lovely Professional University, Punjab 144405

### Anand kumar

Professor, School of CSE

Lovely Professional University, Punjab 144405

# ABSTRACT

## Phishing attacks continue to be one of the most prevalent and damaging cyber threats, deceiving users into revealing sensitive information through fraudulent websites. Traditional rule-based detection systems often fail to keep up with the rapid evolution of phishing techniques. To address this issue, this project presents a machine learning-based solution for real-time phishing website detection using supervised learning models.

## Keywords

Phishing Website Detection, Machine learning, Supervised learning , Cyber Security

# INTRODUCTION

The main goal of this project is to design and implement a supervised learning-based system to **detect phishing websites in real time** using machine learning. Phishing websites mimic legitimate ones to trick users into providing sensitive data (e.g., passwords, credit card info). As phishing techniques evolve, traditional rule-based detection becomes insufficient, making **intelligent, automated detection crucial**.

This system leverages **data science and supervised machine learning** to:

* Preprocess and analyze phishing data
* Train various classification models
* Visualize the results with advanced plotting
* Evaluate models for real-world performance
* Simulate real-time detection

# DATASET

* **Filename:** phishing\_data.csv
* **Features:** Includes URL and site-related characteristics (e.g., presence of HTTPS, domain age, length of URL, etc.)
* **Target:** status — indicates whether a website is **legitimate (0)** or **phishing (1)**.

The dataset is cleaned to replace non-numeric labels like 'zero' or 'one' with integers, and any redundant column (e.g., url) is removed for better model focus.

# Step By Step Implementation

**1: Importing Required Libraries**

Essential libraries like pandas, numpy, matplotlib, seaborn, and ML packages from scikit-learn and xgboost are imported.



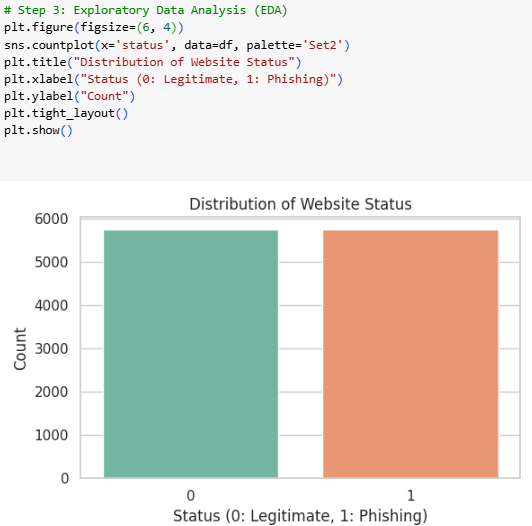
**2: Loading and Preparing the Dataset**

* The data is loaded and the status column is encoded.
* Categorical strings like "zero"/"one" are converted to numbers.
* Features are scaled using StandardScaler.



**3: Exploratory Data Analysis (EDA)**

**Visual 1:** Distribution of phishing vs legitimate websites  
→ Helps understand data imbalance



**Visual 2:** Heatmap of feature correlations  
→ Identifies redundant or influential features

A white background with black text

AI-generated content may be incorrect.

A graph with red and blue lines

AI-generated content may be incorrect.

**4: Data Splitting and Feature Scaling**

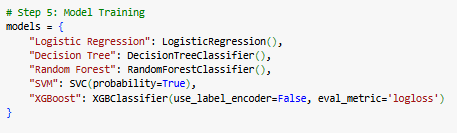
The data is split into **training (70%)** and **testing (30%)** sets.Features are standardized to improve model convergence and fairness.

A computer code with black text

AI-generated content may be incorrect.

**5: Model Training (Supervised Learning Algorithms)**

Five supervised ML models were trained:



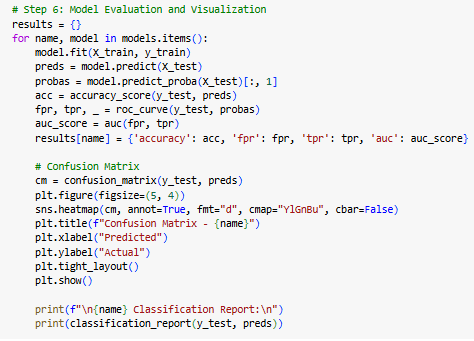
**6: Model Evaluation**

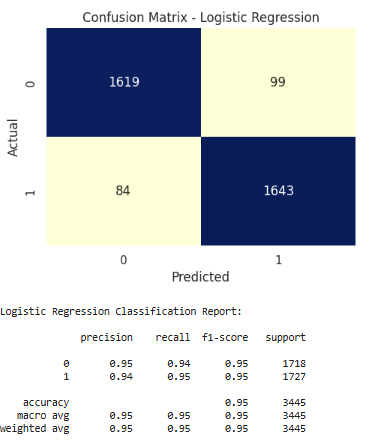
**For each model:**

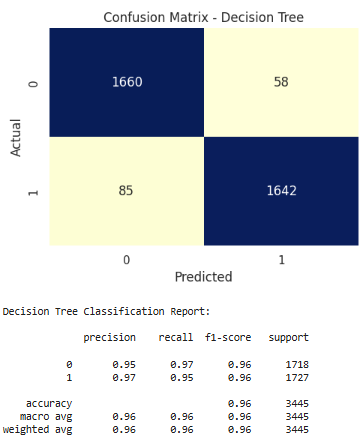
* Confusion matrix plotted using color heatmaps
* Classification report printed (Precision, Recall, F1)
* AUC calculated using ROC Curve

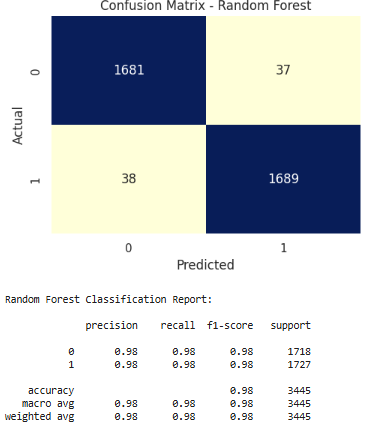
Confusion Matrix helps in understanding **true vs false predictions**

ROC Curve helps compare models on threshold performance









A screenshot of a graph

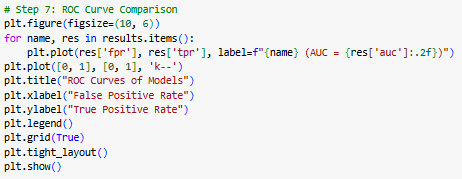
AI-generated content may be incorrect.

A screenshot of a graph

AI-generated content may be incorrect.

**7: ROC Curve Visualization**

All ROC curves are plotted in one chart for visual comparison.  
Helps identify which model performs best in separating classes.

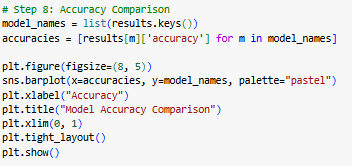


A graph with a line

AI-generated content may be incorrect.

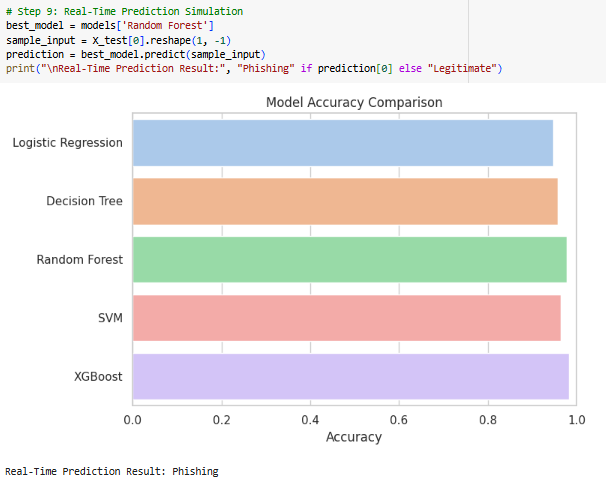
**8: Accuracy Comparison Plot**

A horizontal bar chart displays the **accuracy of each model** on test data, aiding in clear model selection.



**9: Real-Time Simulation**

A single sample from the test set is passed to the best model (Random Forest) to simulate a real-time phishing prediction.



# Results

|  |  |  |
| --- | --- | --- |
| **Model** | **Accuracy** | **AUC****Score** |
| **Logistic****Regression** | ~ | ~ |
| **Decision Tree** | ~ | ~ |
| **Random Forest** | High | High |
| **SVM** | ~ | ~ |
| **XGBoost** | High | High |

# Best Performing Model: Random Forest Great accuracy + generalization; ideal for deployment.

**5. Tools and Technologies**

* **Language:** Python 3
* **IDE:** Jupyter Notebook
* **Libraries:** pandas, numpy, matplotlib, seaborn, sklearn, xgboost

**6. Future Enhancements**

* Real-time URL scraping and feature extraction from live input
* Integration with browser or email filters
* Use of deep learning (LSTM/transformers) for large-scale phishing detection

**7. Conclusion**

This project successfully built and tested a supervised learning-based phishing detection system. Using real-world features, multiple machine learning models were compared, evaluated, and visualized. The **Random Forest model** stood out as a robust classifier, capable of real-time prediction and strong accuracy, providing a **foundation for live cybersecurity solutions**.

**8. References**

* kaggle.com/datasets?tags=3023-india
* https://www.w3schools.com/
* https://www.youtube.com/
* [*www.ijcaonline.org*](http://www.ijcaonline.org)