THAPAR INSTITUTE
OF ENGINEERING & TECHNOLOGY
(Deemed to be University)

Phishing Website Detection

Presented by:
102217005 -Sudansh Rana
102217026 -Saksham Dhiman

Faculty Name:
Dr. Nitin Arora

Content

THAPAR INSTITUTE
OF ENGINEERING & TECHNOLOGY
(Deemed to be University)

Introduction

THAPAR INSTITUTE
OF ENGINEERING & TECHNOLOGY
(Deemed to be University)

Technical background of project:

Phishing attacks are a significant cybersecurity threat where attackers trick users into providing sensitive information, such as passwords or financial details, by masquerading as legitimate entities. Detecting phishing websites is critical to preventing such attacks, and machine learning techniques provide an automated and efficient solution for identifying phishing attempts.

Technical Concepts used :

Phishing Detection as Classification: Binary classification task (phishing = 1, legitimate = 0) using supervised machine learning.
Dataset and Preprocessing: Balanced dataset of 5000 phishing and 5000 legitimate URLs from PhishTank and UNB datasets.
Machine Learning Models: Algorithms: Decision Tree, Random Forest, SVM, XGBoost.
XGBoost performed best in accuracy.
Model Evaluation: Accuracy used as the key metric.Models compared on training and testing performance.
Future Application: Deployment via browser extensions or GUI for real-time phishing detection.

THAPAR INSTITUTE
OF ENGINEERING & TECHNOLOGY
(Deemed to be University)

Motivation:

The project aims to leverage machine learning's predictive capabilities to build an efficient, automated system for phishing detection, ultimately enhancing online security and protecting users from fraud.

Problem Statement:

Phishing attacks are a significant cybersecurity challenge where malicious websites impersonate legitimate ones to steal sensitive user information. Traditional methods, such as blacklists, are often reactive and fail to identify newly created phishing websites effectively

Area of application:

Cybersecurity: Real-time detection of phishing websites.
Web Browsers: Integration for user protection.
Finance: Safeguarding online transactions.

THAPAR INSTITUTE
OF ENGINEERING & TECHNOLOGY
(Deemed to be University)

Dataset and input format:

The dataset for the project consists of phishing and legitimate URLs. Legitimate URLs are sourced from the University of New Brunswick's URL-2016 dataset, while phishing URLs are collected from PhishTank, an open-source repository that provides regularly updated phishing data. A balanced dataset of 5000 URLs from each category is used for training and testing.

Each URL is represented through a set of 17 extracted features, which include **Address Bar Features** (e.g., URL length, presence of IP addresses, and special characters like "@"), **Domain Features** (e.g., domain age, DNS records, and website traffic), and **HTML/JavaScript Features** (e.g., iframe usage and disabling right-click). These features form the input format for machine learning models, enabling them to classify URLs as either phishing or legitimate.

Objective

THAPAR INSTITUTE
OF ENGINEERING & TECHNOLOGY
(Deemed to be University)

Main Objective

The main objective of the project is to develop a machine learning-based system to detect phishing websites by analyzing URLs and their associated features. The system aims to classify websites as either phishing or legitimate by training models on a dataset of 5000 phishing and 5000 legitimate URLs. The goal is to identify key features that differentiate phishing sites from trusted ones, evaluate various machine learning algorithms, and select the most effective model for real-time phishing detection, thereby enhancing online security.

Methodology

THAPAR INSTITUTE
OF ENGINEERING & TECHNOLOGY
(Deemed to be University)

Steps:

Dataset Collection:

- Legitimate URLs are sourced from the University of New Brunswick's URL-2016 dataset.
- Phishing URLs are collected from PhishTank, with 5000 URLs randomly chosen from each category.

Feature Extraction:

- Extract **17 features** from URLs, including Address Bar features (e.g., URL length, IP address presence), Domain-based features (e.g., domain age, DNS records), and HTML/JavaScript features (e.g., iframe usage, right-click disabling).

Data Preprocessing:

- Perform exploratory data analysis (EDA) to clean the data and handle missing values.
- Split the dataset into training and testing sets.

Model Training:

- Train multiple machine learning models, including Decision Tree, Random Forest, SVM, XGBoost, using the training dataset.

Model Evaluation:

- Evaluate model performance using **accuracy** as the primary metric.
- Compare results from different models to select the best-performing one.

Deliverable of each steps or phase:

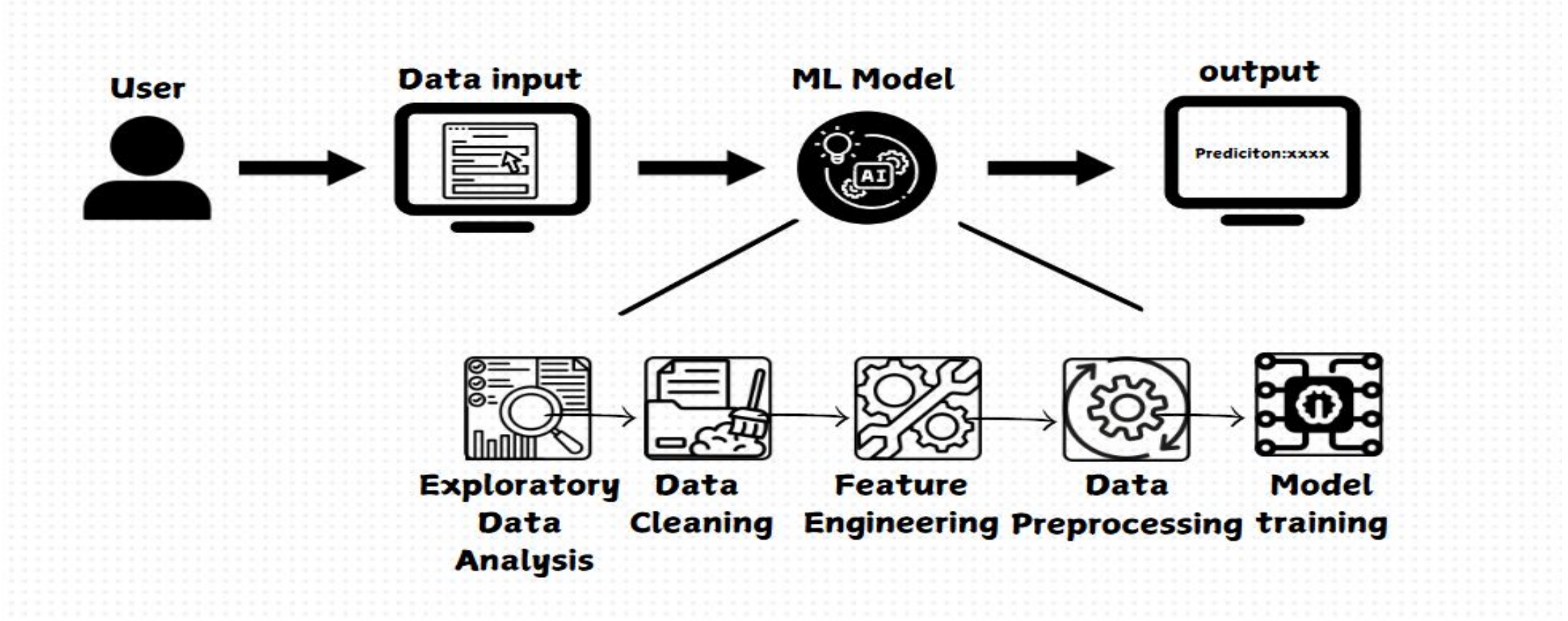
THAPAR INSTITUTE
OF ENGINEERING & TECHNOLOGY
(Deemed to be University)

Dataset Collection: A balanced dataset of 5000 phishing and 5000 legitimate URLs.
Feature Extraction: A dataset with 17 extracted features per URL.
Data Preprocessing: Cleaned dataset, with missing values handled and split into training and testing sets.
Model Training: Trained machine learning models (Decision Tree, Random Forest, SVM,XGboost).
Model Evaluation:Performance evaluation (accuracy) of each model, identifying the best-performing one.

Working Model

Technical Diagram

THAPAR INSTITUTE
OF ENGINEERING & TECHNOLOGY
(Deemed to be University)



THAPAR INSTITUTE
OF ENGINEERING & TECHNOLOGY
(Deemed to be University)

Working Module

The **Data Collection Module** collects phishing URLs from PhishTank and legitimate URLs from the University of New Brunswick's dataset, creating a balanced dataset of 5000 phishing and 5000 legitimate URLs. The **Feature Extraction Module** processes these URLs by extracting 17 key features, including address bar properties, domain information, and HTML/JavaScript features.

The **Preprocessing Module** handles missing values, scales features where necessary, and splits the dataset into training and testing sets, preparing it for model training. In the **Model Training Module**, several machine learning algorithms (Decision Tree, Random Forest, SVM, XGBoost) are trained on the preprocessed data.

The **Model Evaluation Module** assesses each model's performance using accuracy as the primary metric and compares the results to identify the best-performing model.

Attained Deliverable

THAPAR INSTITUTE
OF ENGINEERING & TECHNOLOGY
(Deemed to be University)

Balanced Dataset: A dataset of **5000 phishing URLs** and **5000 legitimate URLs**, collected from PhishTank and the University of New Brunswick.

Feature Set: Extraction of **17 key features** from the URLs, including address bar properties, domain information, and HTML/JavaScript-based characteristics.

Preprocessed Data: A cleaned and processed dataset with missing values handled, features scaled, and split into **training** and **testing sets** for model evaluation.

Trained Models: Multiple machine learning models trained, including **Decision Tree, Random Forest, SVM, XGBoost.**

Model Evaluation: **Performance results** with accuracy metrics for each model, identifying the **best-performing model** (e.g., XGBoost).

Saved Model: The best model is saved and ready for integration into real-time applications such as **browser extensions** or **GUIs** for phishing detection.

Results

Test Cases

a) Input

	Have_IP	Have_At	URL_Length	URL_Depth	Redirection	https_Domain	\
5669	0	0	0	2	0	0	
8800	0	0	0	0	0	0	
3205	0	0	1	3	0	0	
8731	0	0	1	3	0	0	
6412	0	0	0	0	0	0	
...	...	...	...	...	...	...	
794	0	0	1	2	0	0	
142	0	0	1	14	0	0	
275	0	0	1	4	0	0	
8265	0	0	1	1	0	0	
7950	0	0	1	1	0	0	

	TinyURL	Prefix/Suffix	DNS_Record	Web_Traffic	Domain_Age	Domain_End	\
5669	0	0	0	1	0	1	
8800	0	0	0	1	0	1	
3205	1	0	0	1	1	1	
8731	0	0	0	1	0	1	
6412	0	1	1	1	1	1	
...	...	...	...	...	...	...	
794	0	0	0	1	1	1	
142	0	0	0	1	0	1	
275	0	0	0	0	0	1	
8265	0	0	0	1	0	1	
7950	0	0	0	1	1	1	

	iFrame	Mouse_Over	Right_Click	Web_Forwards
5669	0	0	1	0
8800	0	0	1	0
3205	0	0	1	0
8731	0	0	1	0
6412	0	0	1	0
...	...	...	...	...
794	0	0	1	0
142	0	0	1	0
275	0	0	1	0
8265	0	0	1	0
7950	0	0	1	0

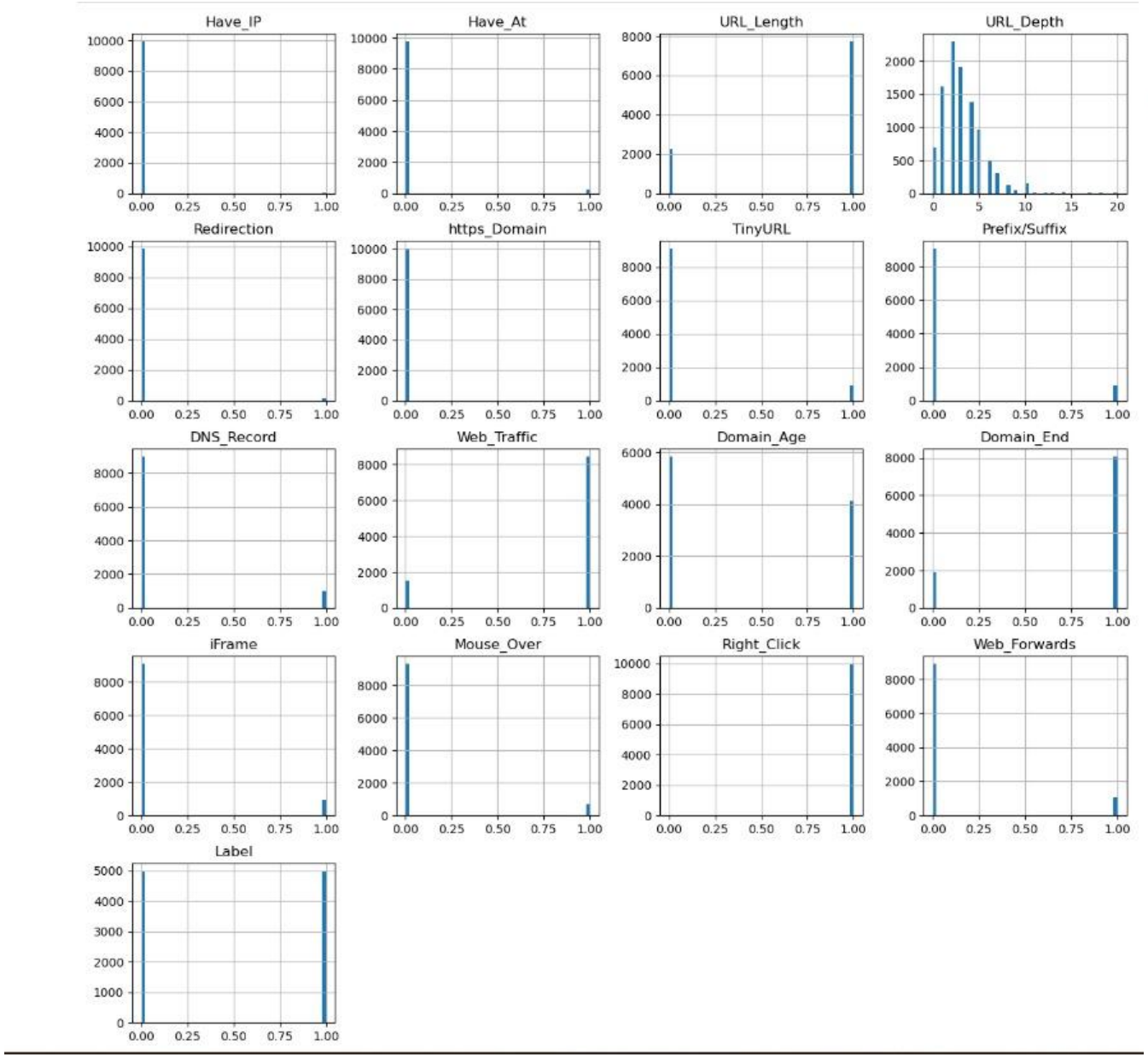
b) output

```
[2000 rows x 16 columns]
[1 1 0 0 1 1 0 0 0 0 1 0 1 0 1 1 0 0 0 1 0 0 1 0 0 1 1 0 1 1 1 1 1 1 1 0 0
 1 1 0 0 1 1 0 1 1 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 0 0 0 0 1 0 1 0 1 1 0 0 0
 0 1 1 0 0 1 0 0 1 0 1 1 1 1 1 1 0 0 0 1 1 0 0 0 0 1 1 1 0 0 1 1 1 0 1 1 0
 0 0 1 0 0 0 0 0 1 0 0 0 1 1 1 0 1 1 1 0 1 0 0 0 0 1 1 1 0 0 1 0 1 1 1 0 0
 0 0 0 0 0 0 1 1 1 1 0 0 0 0 1 0 0 0 0 0 0 1 1 0 0 1 1 1 1 0 0 0 1 0 1 0 1
 0 0 1 0 0 0 0 0 1 1 1 0 0 1 0 1 0 0 0 1 1 1 0 0 1 1 1 1 1 1 1 1 0 0 0 1 0
 0 0 0 1 1 1 1 0 0 1 1 1 1 1 0 0 0 0 0 0 0 1 1 1 0 1 1 0 1 1 0 1 1 0 0 0 0
 1 1 0 0 0 0 0 1 0 0 1 0 0 0 0 0 1 0 0 1 0 1 1 0 1 0 0 1 0 0 1 1 1 1 1 0 1
 1 1 0 0 1 1 1 0 0 1 0 0 0 0 0 0 1 0 0 1 1 0 1 0 0 0 0 0 1 1 1 1 1 1 0 0 1
 0 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 1 0 1 1 0 1 0 0 0 0 0 0 1 0 1 0 0 0 0 1
 0 1 0 1 1 0 1 0 1 0 0 1 1 0 0 0 0 1 1 0 0 0 1 1 0 0 1 0 0 1 0 0 0 1 0 0 0
 1 0 0 0 0 0 0 1 1 1 0 0 1 0 1 1 0 1 1 0 0 0 1 1 0 1 0 1 0 0 0 0 0 0 1 1 0
 1 1 1 0 1 0 0 1 0 0 0 1 0 0 1 0 0 0 1 1 1 1 0 1 0 0 1 1 0 0 0 1 1 1 1 0 0
 0 0 0 1 0 0 1 1 0 1 0 1 1 0 0 1 1 0 0 0 1 0 0 0 0 1 0 1 1 0 1 0 0 1 1 0 0
 0 0 1 0 1 0 1 1 1 0 0 0 0 1 0 0 1 0 1 1 0 0 0 1 1 1 1 0 0 0 0 0 1 1 1 0 0
 1 0 0 0 0 1 0 0 0 1 0 1 1 0 1 0 0 0 0 0 1 0 0 1 1 0 0 1 0 1 1 0 0 1 0 0 1
 0 1 0 0 0 0 0 0 0 1 0 0 1 0 1 0 1 1 0 0 1 0 1 0 1 0 0 1 1 1 1 0 1 1 1 1 0
 0 0 0 0 0 0 0 1 0 1 0 1 0 0 1 0 0 0 0 0 1 0 0 0 1 1 0 0 1 0 1 1 1 1 0 0 1
 1 0 1 0 1 0 0 0 1 1 0 0 0 1 0 0 0 0 0 0 0 1 0 0 1 1 0 1 0 0 0 0 1 0 0 1 0
 0 0 1 1 1 0 0 1 1 0 0 1 0 0 1 1 0 0 0 0 0 0 0 0 1 0 1 0 1 1 1 1 0 0 0 0 0
 0 1 0 0 1 1 0 0 1 0 1 0 0 0 0 1 0 0 1 1 1 0 1 0 1 0 0 0 0 0 0 0 0 0 1 0 0
 1 0 1 1 0 0 0 1 1 0 1 1 1 1 1 1 1 0 0 1 0 1 0 0 0 0 0 1 0 0 0 0 1 1 0 1 1
 0 1 1 0 0 0 0 0 1 0 1 0 1 0 0 0 0 0 1 0 0 1 1 1 0 0 0 1 0 0 1 0 1 0 0 1 0
 1 1 0 0 0 1 0 0 1 0 0 0 0 0 1 0 1 1 0 0 1 0 1 0 0 1 0 0 1 0 0 1 0 0 0 1 0
 1 0 0 0 0 0 1 0 0 1 1 1 0 1 0 0 1 0 0 0 1 0 1 0 1 1 0 0 0 0 1 1 1 0 0 1 0
 1 1 0 0 0 1 1 0 0 1 0 1 0 1 0 0 0 0 0 0 1 1 1 1 1 0 0 1 0 1 1 0 0 1 0 0 0
 0 0 0 0 0 0 0 1 0 1 1 1 0 0 0 1 0 0 0 0 1 1 1 0 1 1 1 0 0 1 0 1 0 1 1 1 0
 0 1 0 0 1 0 0 1 0 1 1 1 0 0 1 0 0 0 0 0 0 0 0 0 1 1 1 1 0 0 0 1 1 0 0 0 0
 0 0 0 1 1 1 1 1 1 0 0 0 0 0 0 1 1 1 0 1 1 0 1 0 0 0 0 0 1 0 0 0 0 0 0 0 1
 0 0 1 1 1 1 0 0 1 1 1 1 0 1 1 1 1 0 1 0 0 1 1 0 0 0 1 0 0 0 0 0 1 0 0 1 0
 0 0 0 0 0 0 1 0 0 1 1 0 1 0 0 0 0 1 0 0 1 1 1 0 0 1 1 0 1 0 1 1 1 0 1 0 1
 1 0 0 0 1 1 1 0 1 0 0 0 1 1 0 1 1 0 0 0 1 1 1 0 1 0 0 0 0 1 1 0 1 0 0 1 1
 0 0 0 1 1 0 1 0 0 0 0 1 1 0 1 1 1 1 1 0 0 0 1 1 0 0 0 1 0 0 1 1 0 0 0 0 1
 1 1 0 0 0 0 0 0 1 0 0 0 0 0 1 0 1 0 0 1 0 1 0 0 0 1 0 0 1 0 1 0 1 1 0 0 0
 0 0 1 0 1 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 1 0 0 1 0 1 1 1 0 0 1 1 1 0 1 0 0
 1 1 0 0 1 1 1 1 1 0 0 0 0 0 0 1 0 1 1 1 1 1 1 1 0 0 1 1 1 1 1 1 1 0 0 0 1
 0 0 0 1 1 1 1 0 0 0 0 0 0 0 0 1 0 1 0 1 1 1 0 0 0 0 0 1 0 1 0 1 0 1 1 0 0
 0 1 0 1 0 0 0 1 1 1 0 1 1 1 0 0 1 0 0 1 1 0 1 0 1 0 0 0 1 0 1 1 0 0 1 1 0
 0 0 0 1 0 1 1 1 1 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 1 0 1 1 0 0 0 1 1 0
 0 0 0 0 0 0 1 1 0 1 0 0 1 0 1 1 0 0 1 1 1 0 1 0 1 0 1 1 1 0 1 0 1 1 1 0 0
 0 1 0 0 0 1 0 1 0 0 0 1 1 0 0 1 0 0 0 1 0 1 0 1 0 0 0 0 1 0 1 0 1 0 0 0 1
 0 1 0 0 1 1 1 0 1 0 1 0 1 1 0 0 0 0 0 0 1 0 0 1 0 0 1 0 1 0 1 1 1 0 1 1 1
 0 1 0 1 0 1 0 0 0 1 1 0 1 0 0 0 0 1 0 0 1 1 0 1 1 0 0 0 0 0 0 0 1 0 0 1 1
 1 1 1 1 1 1 0 1 0 1 1 1 0 1 1 0 0 0 0 0 1 0 0 0 0 1 0 1 0 0 0 0 1 0 0 0 1
 0 0 0 1 1 0 1 1 0 0 0 0 0 1 1 0 0 0 0 0 1 0 0 1 1 1 0 1 0 1 0 1 1 0 1 0 1
 1 0 1 1 0 0 1 1 0 0 0 0 1 1 1 0 0 0 0 0 0 0 1 1 1 0 1 1 0 1 0 1 1 1 0 0 1
 1 0 0 1 0 0 0 1 1 1 1 1 0 0 0 0 0 0 1 1 1 1 1 0 1 1 0 1 0 0 0 0 0 0 0 1 0
 0 0 0 1 1 0 0 1 1 0 0 0 0 1 1 0 0 0 0 0 0 0 1 0 0 0 1 0 0 1 1 0 0 1 1 1 0
 0 1 1 1 1 1 0 1 1 1 1 0 1 1 0 0 0 1 1 1 1 1 1 1 1 1 1 0 0 0 0 1 1 1 0 0 0
 1 1 0 1 0 1 0 1 1 1 1 0 1 0 0 1 0 0 0 0 1 0 0 0 1 0 1 0 0 1 1 0 1 1 0 1 0
 0 1 1 1 0 0 1 0 0 1 1 1 0 0 0 0 1 0 0 1 1 0 1 1 1 1 1 1 0 0 0 1 1 0 0 0 0
 1 0 1 1 1 1 0 1 0 0 0 1 0 1 0 0 0 0 1 0 0 0 0 0 1 0 1 0 1 0 1 0 0 0 0 0 0
 1 1 0 0 0 0 0 0 0 1 0 0 1 1 1 1 1 0 1 0 0 1 1 0 0 0 0 0 1 0 1 1 1 1 1 1 1
 0 0 1 0 1 0 1 1 0 1 0 0 0 0 1 1 0 1 0 1 0 0 0 0 0 1 1 1 1 1 1 1 0 0 0 0 0
 0 0]
```

Outcome Graphs

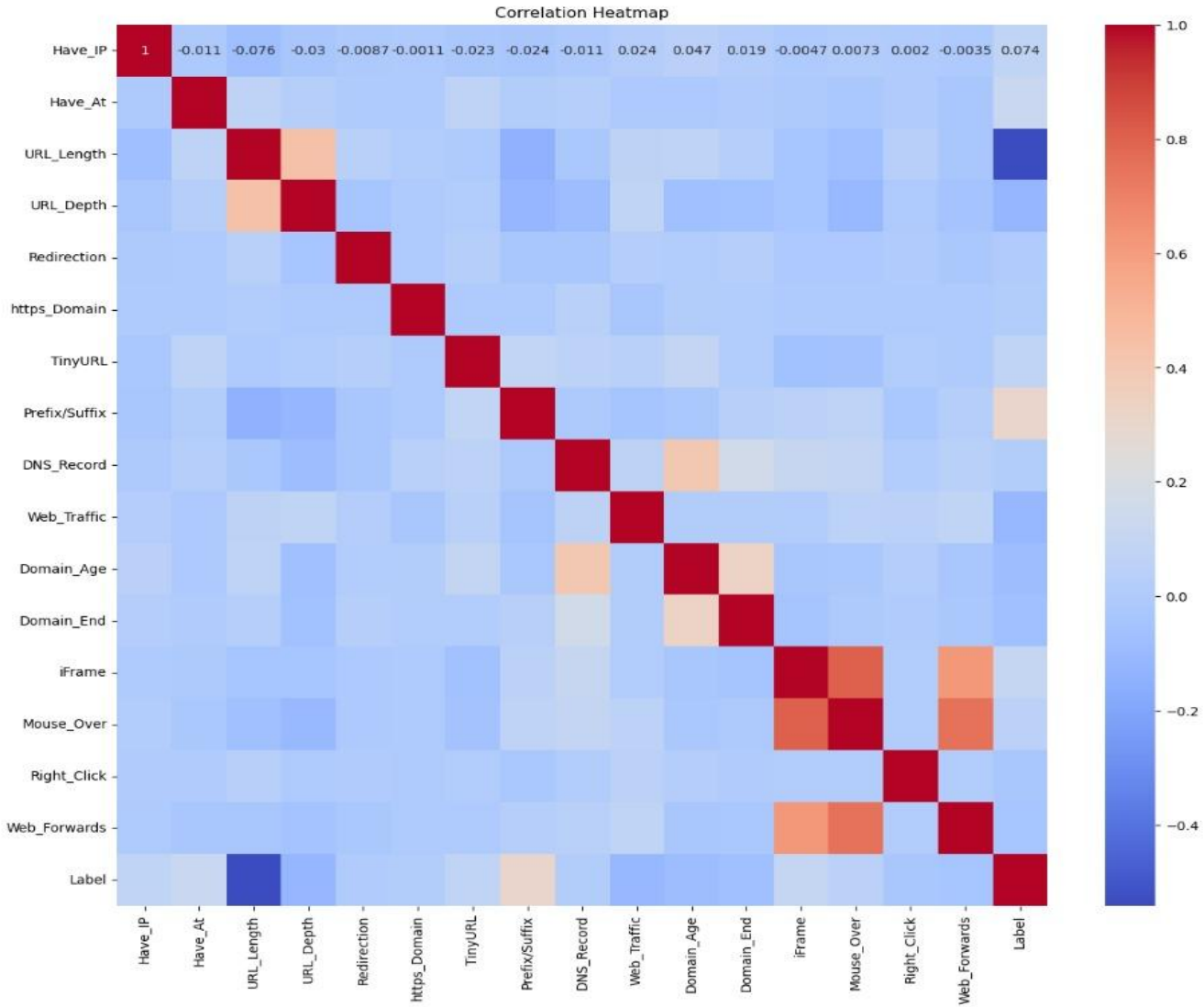
Visualization of data

Few plots and graphs are displayed to find how the data is distributed and the how features are related to each other.



THAPAR INSTITUTE
OF ENGINEERING & TECHNOLOGY
(Deemed to be University)

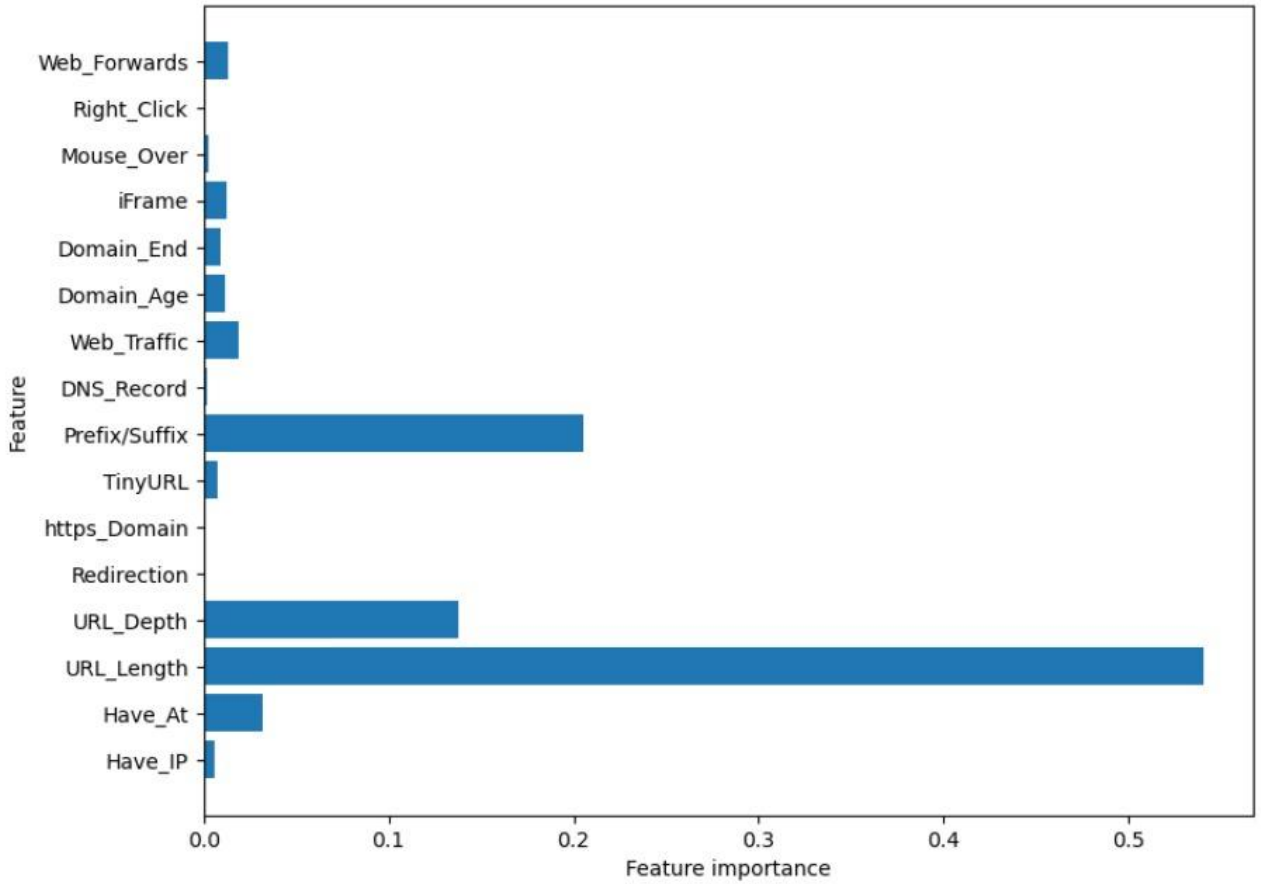
Correlation Analysis of Key Features



THAPAR INSTITUTE
OF ENGINEERING & TECHNOLOGY
(Deemed to be University)

Feature Importance using Random forest

THAPAR INSTITUTE
OF ENGINEERING & TECHNOLOGY
(Deemed to be University)

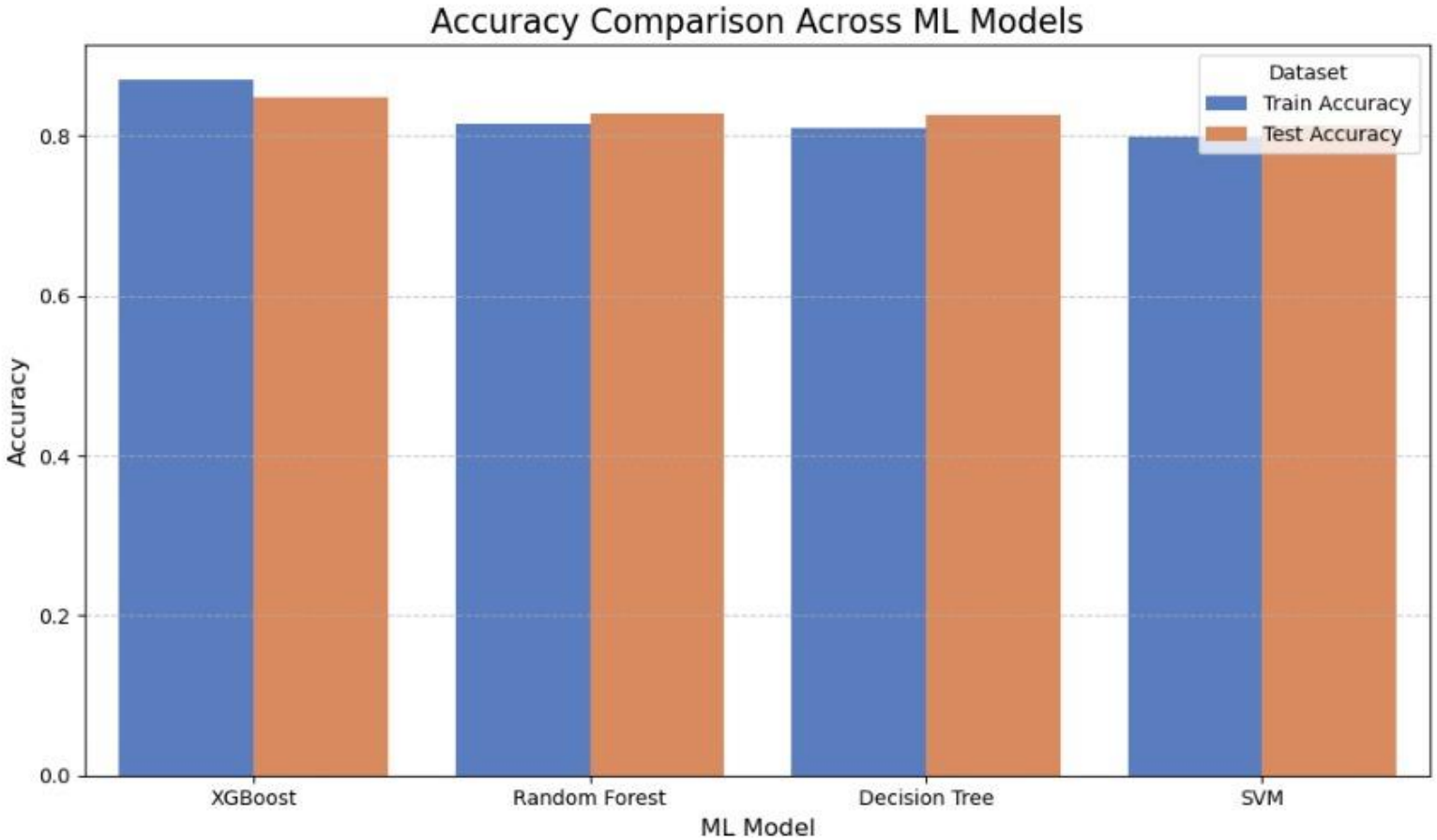


Evaluation of Different Algorithms Across Metrics

THAPAR INSTITUTE
OF ENGINEERING & TECHNOLOGY
(Deemed to be University)

ML Model	Train Accuracy	Test Accuracy	Train F1 Score	Test F1 Score
Decision Tree	0.810	0.826	0.805	0.820
Random Forest	0.810	0.824	0.804	0.818
XGBoost	0.870	0.848	0.870	0.848
SVM	0.799	0.814	0.793	0.808

THAPAR INSTITUTE
OF ENGINEERING & TECHNOLOGY
(Deemed to be University)



THAPAR INSTITUTE
OF ENGINEERING & TECHNOLOGY
(Deemed to be University)

Conclusion

A. Justification of Objectives:

The project successfully developed a machine learning-based system to detect phishing websites by analyzing key URL features. Among the models tested, **XGBoost** delivered the highest accuracy, proving the effectiveness of machine learning over traditional methods like blacklists. This approach justifies the use of machine learning for phishing detection, as it offers a proactive, scalable solution that can identify new phishing sites efficiently. The system is ready for real-time deployment, such as in browser extensions, to enhance online security and protect users from phishing attacks.

B. Future Scope:

THAPAR INSTITUTE
OF ENGINEERING & TECHNOLOGY
(Deemed to be University)

Real-Time Detection: Integrating the model into browser extensions or security tools for real-time phishing website detection and alerts.

Model Enhancement: Incorporating additional features like website behavior analysis (e.g., user interaction or page loading patterns) to improve detection accuracy.

Deep Learning: Exploring deep learning models (e.g., Convolutional Neural Networks) for better pattern recognition and handling more complex data.

Cross-Platform Deployment: Extending the model's use to mobile applications, email filters, and other platforms vulnerable to phishing attacks.

Continuous Learning:Implementing a continuous learning system where the model is updated with new phishing data to adapt to evolving threats.

References

Dataset: https://archive.ics.uci.edu/ml/datasets/Phishing+Websites

Research paper: https://eprints.hud.ac.uk/id/eprint/24330/6/MohammadPhishing14July2015.pdf

Kaggle: https://www.kaggle.com/

Scikit-learn Documentation: https://scikit-learn.org/

Python Documentation for Data Analysis Libraries: Pandas, NumPy, Matplotlib, and Seaborn, https://docs.python.org/

Youtube Channel: https://www.youtube.com/@campusx-official

THAPAR INSTITUTE
OF ENGINEERING & TECHNOLOGY
(Deemed to be University)

Thank You