Ibn Zohr University
Faculty of Science des Sciences, Agadir
IT Excellence Center
Data Analytics & AI

MODULE : CYBER SECURITY

PHISHING URL DETECTION

Prepared By :
ELQORACHI Hind
JAAFAR Wafa
MISBAH Asmae
BELFAIK Chaymae

Supervised by :
Pr.Boughrous Monsef

Academic year : 2025-2026

PLAN

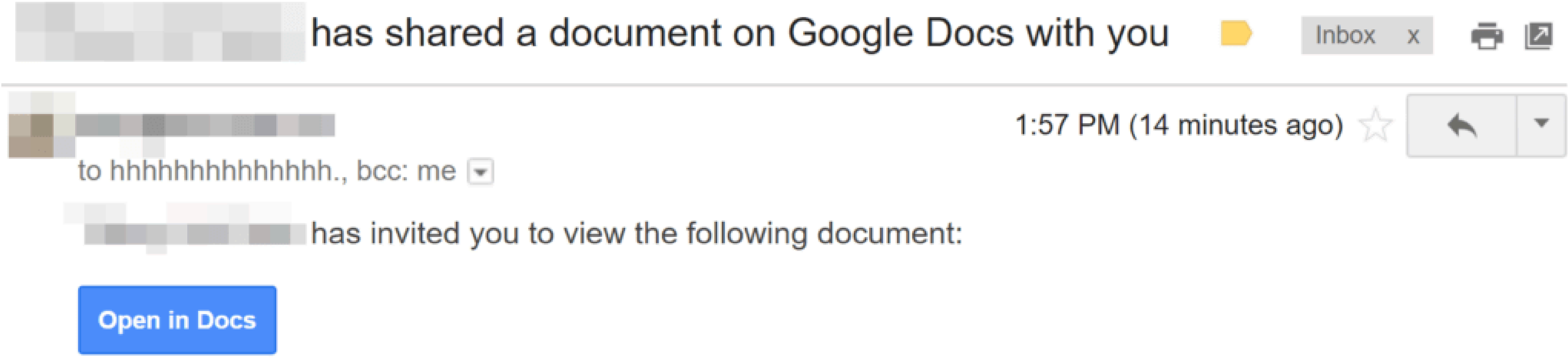
1. Introduction to problematic
2. Pipeline
3. Dataset and Feature Engineering
4. Models
5. Tools
6. Evaluation & Results
7. Demonstration
8. Conclusion & Recommandations

Problematic

‘ Google Study Case’

Google 2017 phishing attack

- **What happened**: Google Docs users were targeted by a highly sophisticated phishing attack.
- **Phishing in context**: Over 90% of cyber-attacks start with phishing. (Source: Huntress, Statistics on Phishing Attacks)
- **Impact**: Financial losses, identity theft, and reputational damage affecting both individuals and organizations.



[redacted] has shared a document on Google Docs with you

[redacted]

1:57 PM (14 minutes ago)

to hhhhhhhhhhhhh., bcc: me

[redacted] has invited you to view the following document:

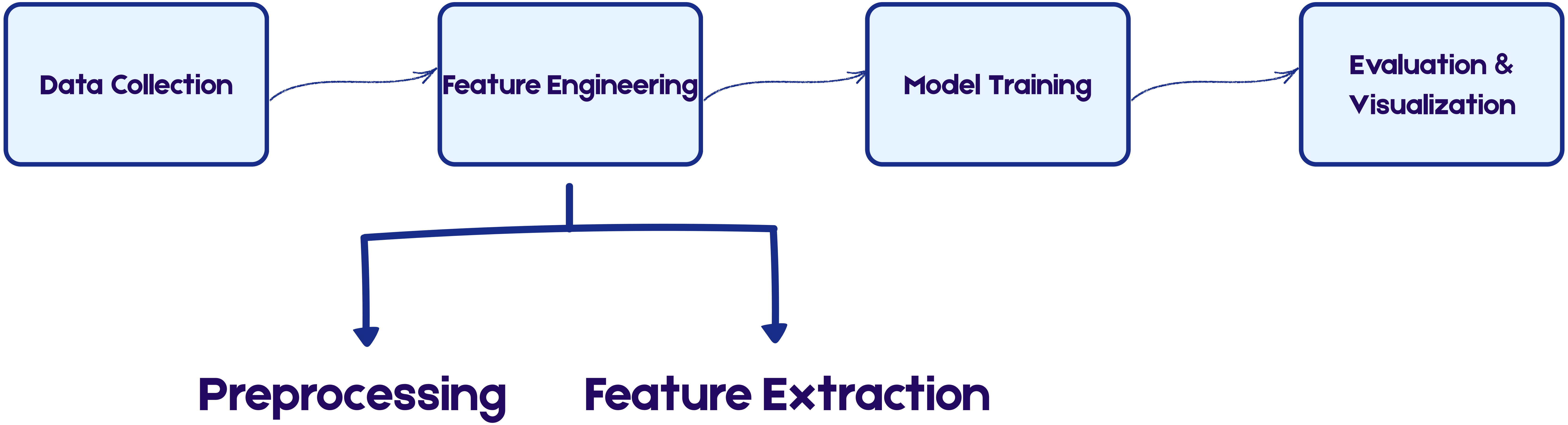
Project Objectives

Project Objectives

1. **Extract key URL features** – lexical patterns, statistical properties, and domain-based attributes.
2. **Train machine learning models** – Random Forest, XGBoost, and SVM for phishing detection.
3. **Evaluate model performance** – accuracy, precision, recall, and F1-score.
4. **Visualize results & build a user interface** – using Streamlit for easy interaction.
5. **Ensure reproducibility** – well-documented code and a reproducible workflow.

ML Pipeline

Work Flow



Data Collection

Feature Engineering

Model Training

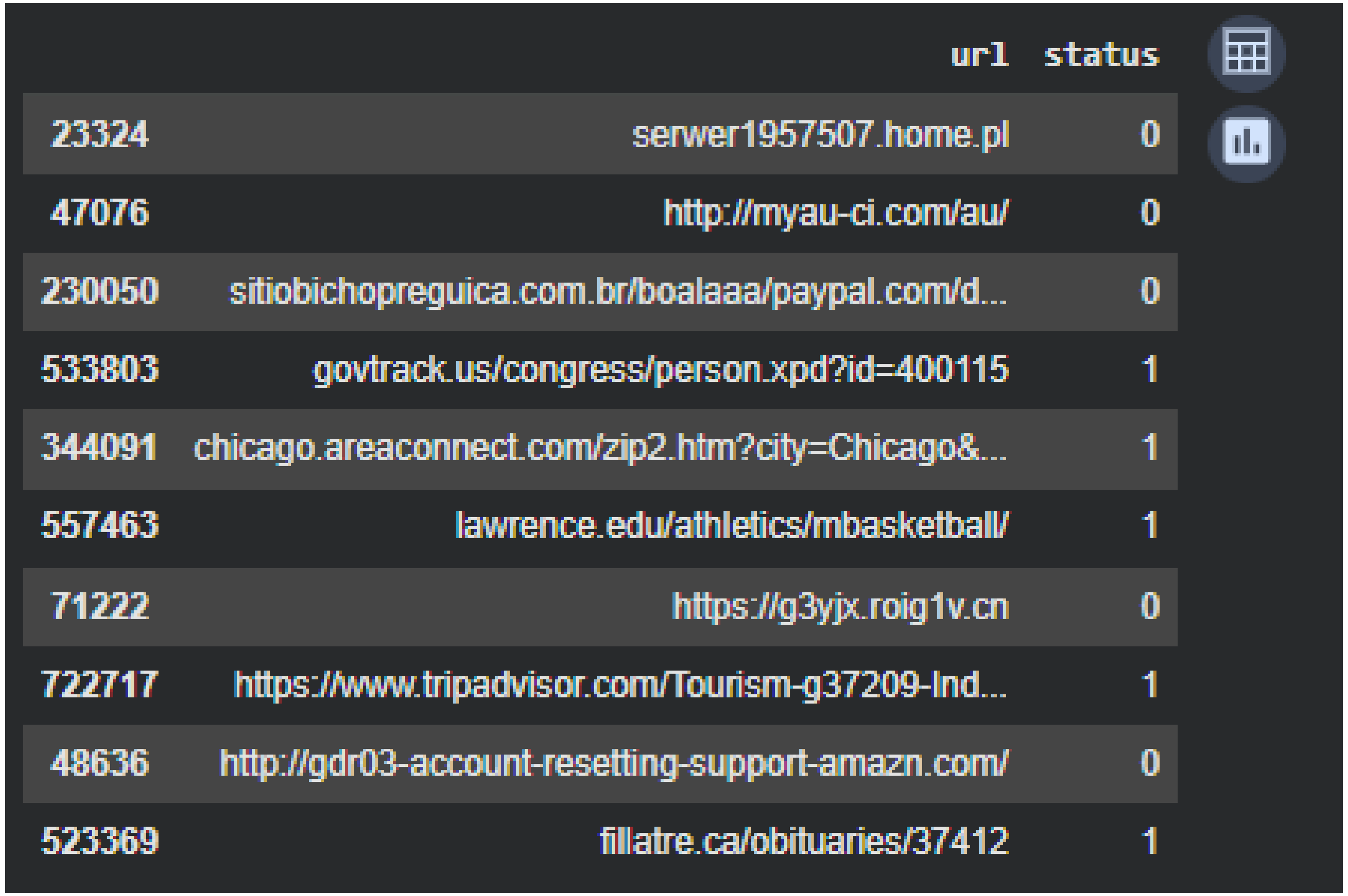
Evaluation & Visualization

Preprocessing

Feature Extraction

Dataset

Dataset Collection



	url	status
23324	serwer1957507.home.pl	0
47076	http://myau-ci.com/au/	0
230050	sitiobichopreguica.com.br/boalaaa/paypal.com/d...	0
533803	govtrack.us/congress/person.xpd?id=400115	1
344091	chicago.areaconnect.com/zip2.htm?city=Chicago&...	1
557463	lawrence.edu/athletics/mbasketball/	1
71222	https://g3yjx.roig1v.cn	0
722717	https://www.tripadvisor.com/Tourism-g37209-Ind...	1
48636	http://gdr03-account-resetting-support-amazn.com/	0
523369	fillatre.ca/obituaries/37412	1

- Kaggle : Phishing and Legitimate URLS : over 800,000 URLs.
- Labels: 1 = Legitimate, 0 = Phishing

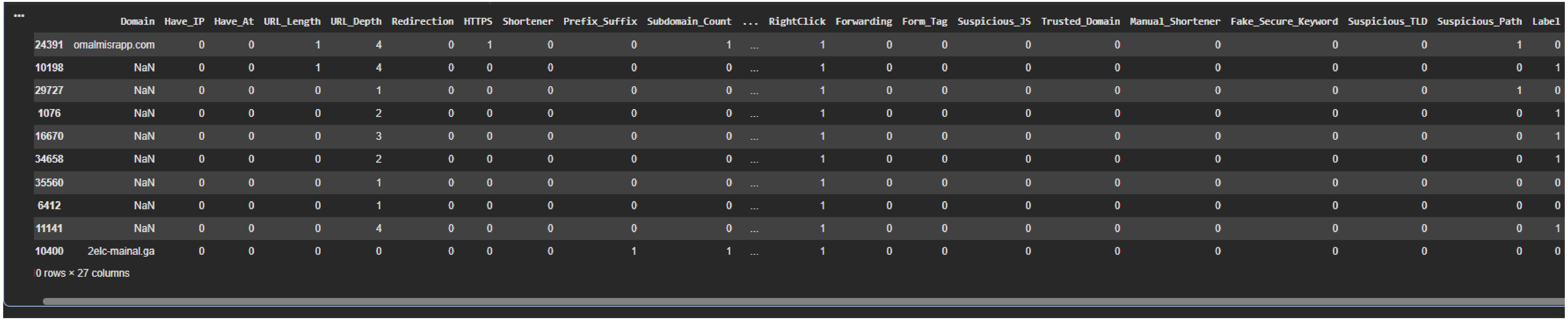
- Extracted 20k Legitimate URLs, and 20k phishing URLs & shuffled them for feature extraction

Feature Engineering

Features Extraction

Category	Features Extracted
Lexical	IP in URL, “@”, length, depth, “//”, HTTPS, shortener, “-”, subdomains, digits, special chars, sensitive keywords
Domain	Domain age (short = suspicious), extension (.com/.org/.net)
HTML/JS	iFrame, mouseover, right-click disabled, forwarding script
Manual Overrides	Tusted Domains, fake secure words,suspicious TLDs

Features Extraction



	Domain	Have_IP	Have_At	URL_Length	URL_Depth	Redirection	HTTPS	Shortener	Prefix_Suffix	Subdomain_Count	...	RightClick	Forwarding	Form_Tag	Suspicious_JS	Trusted_Domain	Manual_Shortener	Fake_Secure_Keyword	Suspicious_TLD	Suspicious_Path	Label
24391	omalmisrapp.com	0	0	1	4	0	1	0	0	1	...	1	0	0	0	0	0	0	0	1	0
10198	NaN	0	0	1	4	0	0	0	0	0	...	1	0	0	0	0	0	0	0	0	1
29727	NaN	0	0	0	1	0	0	0	0	0	...	1	0	0	0	0	0	0	0	1	0
1076	NaN	0	0	0	2	0	0	0	0	0	...	1	0	0	0	0	0	0	0	0	1
16670	NaN	0	0	0	3	0	0	0	0	0	...	1	0	0	0	0	0	0	0	0	1
34658	NaN	0	0	0	2	0	0	0	0	0	...	1	0	0	0	0	0	0	0	0	1
35560	NaN	0	0	0	1	0	0	0	0	0	...	1	0	0	0	0	0	0	0	0	0
6412	NaN	0	0	0	1	0	0	0	0	0	...	1	0	0	0	0	0	0	0	0	0
11141	NaN	0	0	0	4	0	0	0	0	0	...	1	0	0	0	0	0	0	0	0	1
10400	2elc-mainal.ga	0	0	0	0	0	0	0	1	1	...	1	0	0	0	0	0	0	0	0	0

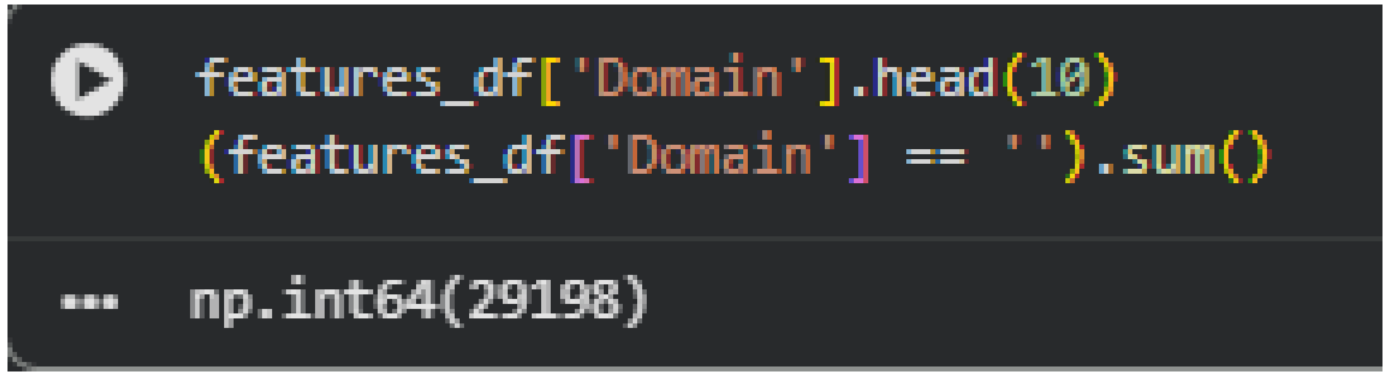
0 rows × 27 columns

<u>Constraints</u> :

Domain name caused too many timeouts in extraction → Dropped before training.

Many Missing values :

Over **70%**



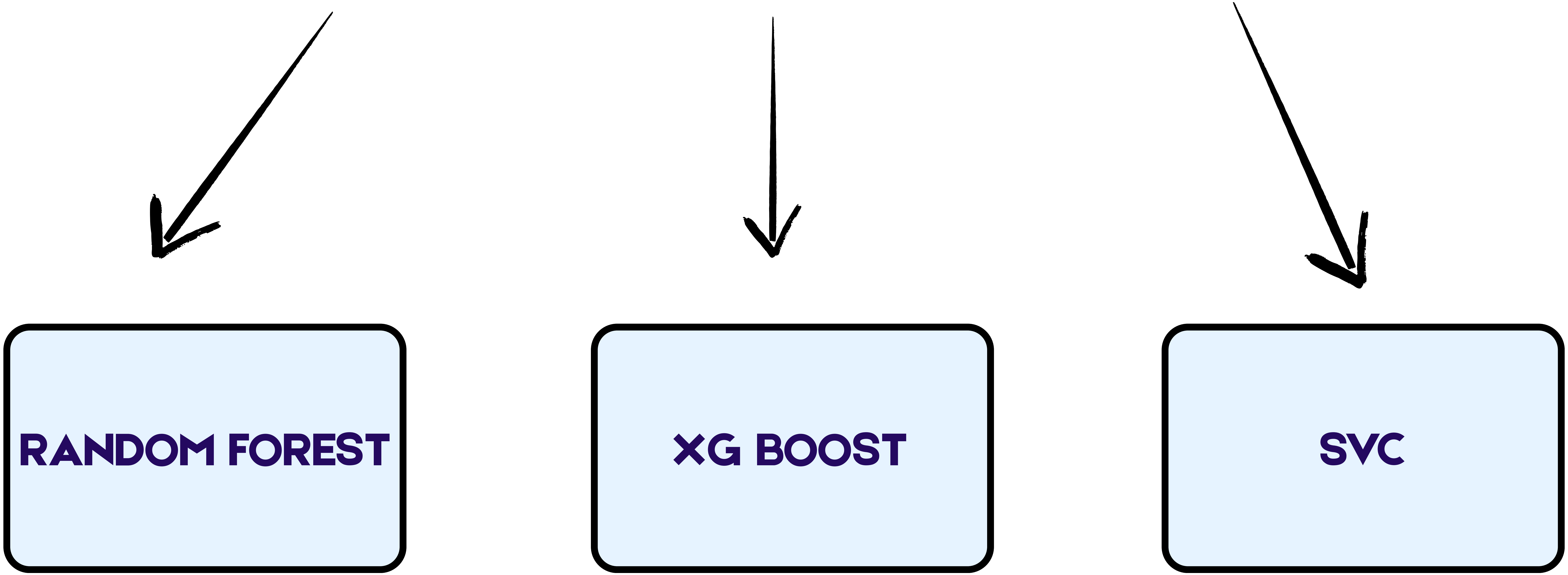
```
features_df['Domain'].head(10)
(features_df['Domain'] == '').sum()
```

```
np.int64(29198)
```

Model Training

Machine Learning Models

THREE MODELS WERE USED FOR TRAINING



RANDOM FOREST

XG BOOST

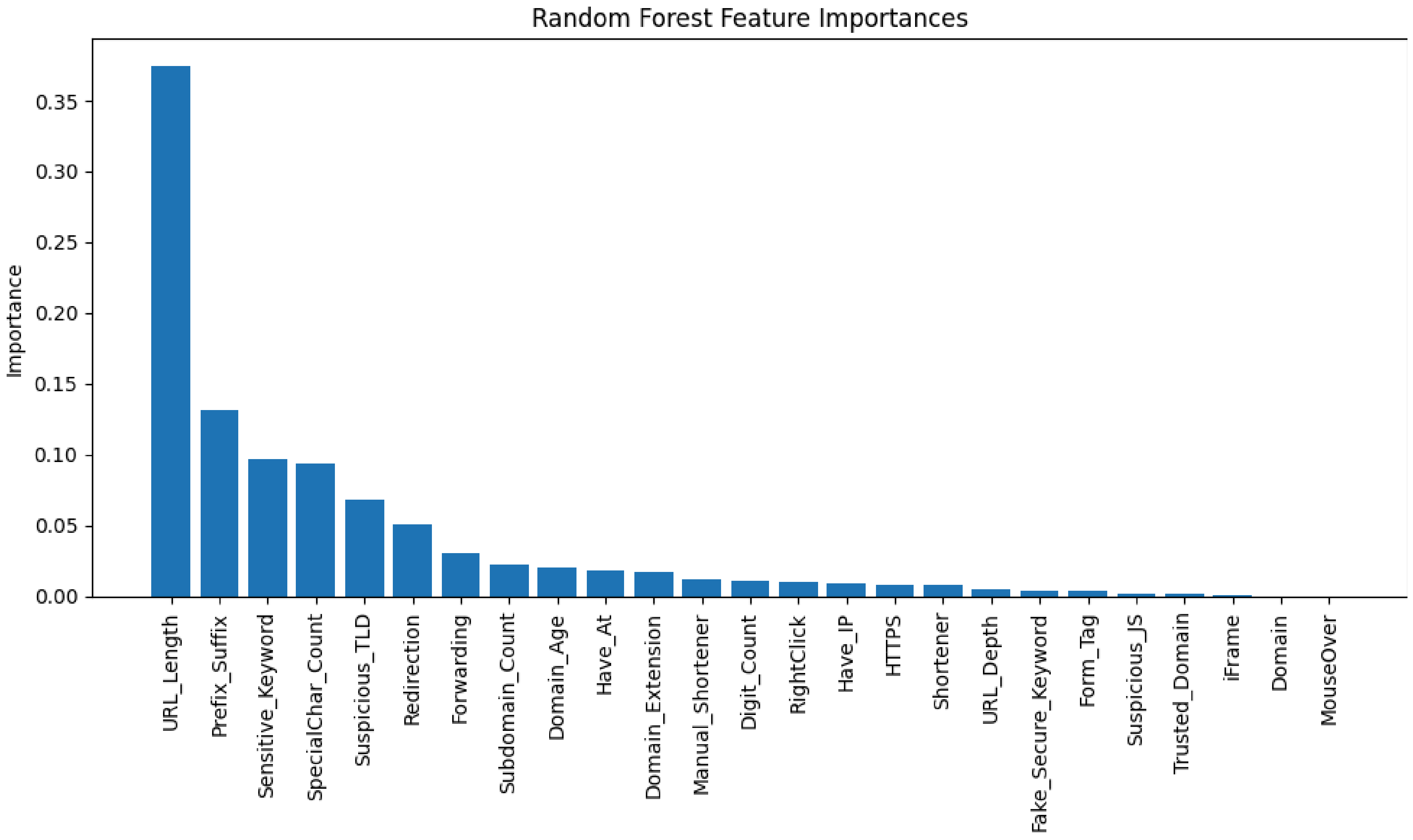
SVC

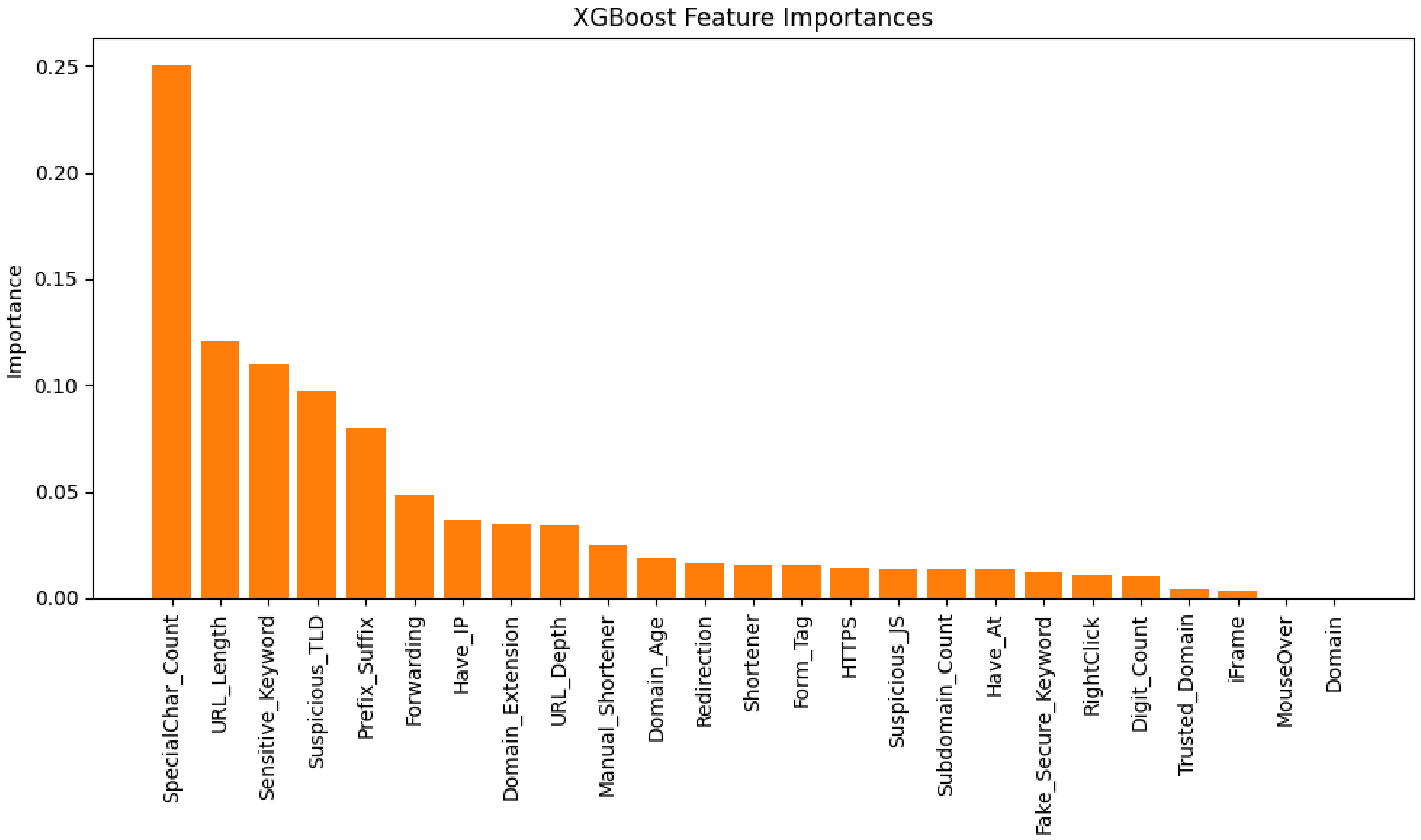
Tools & libraries

Category	Details
Programming language	Python
Libraries	Scikit-learn,sklearn.svm XGBoost,sklearn.ensemble Pandas, NumPy,tqdm ,sklearn.metrics, Matplotlib, Seaborn, Streamlit, pyngrok
Environement	Google Colaboratory

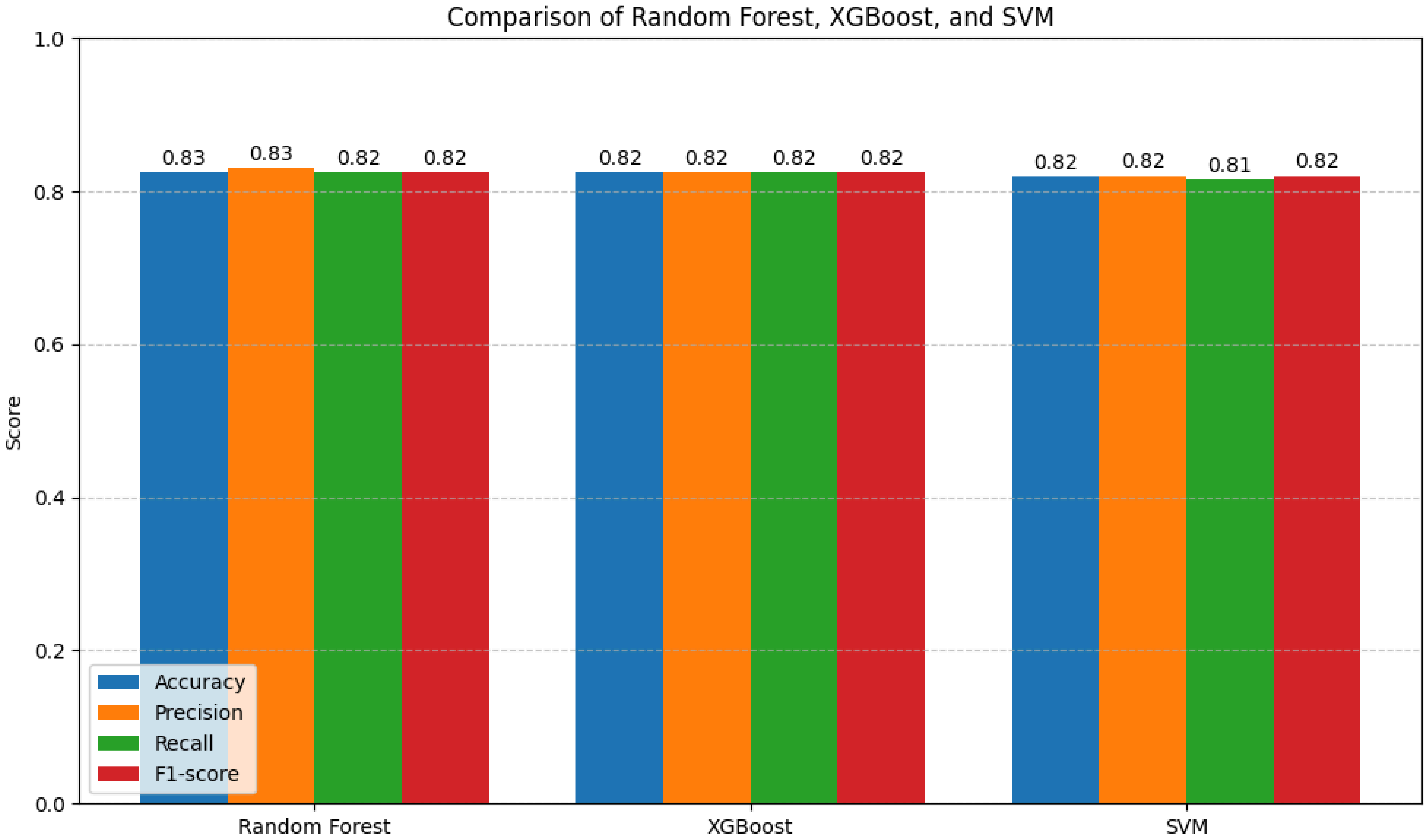
Plots & Results

Feature Importances

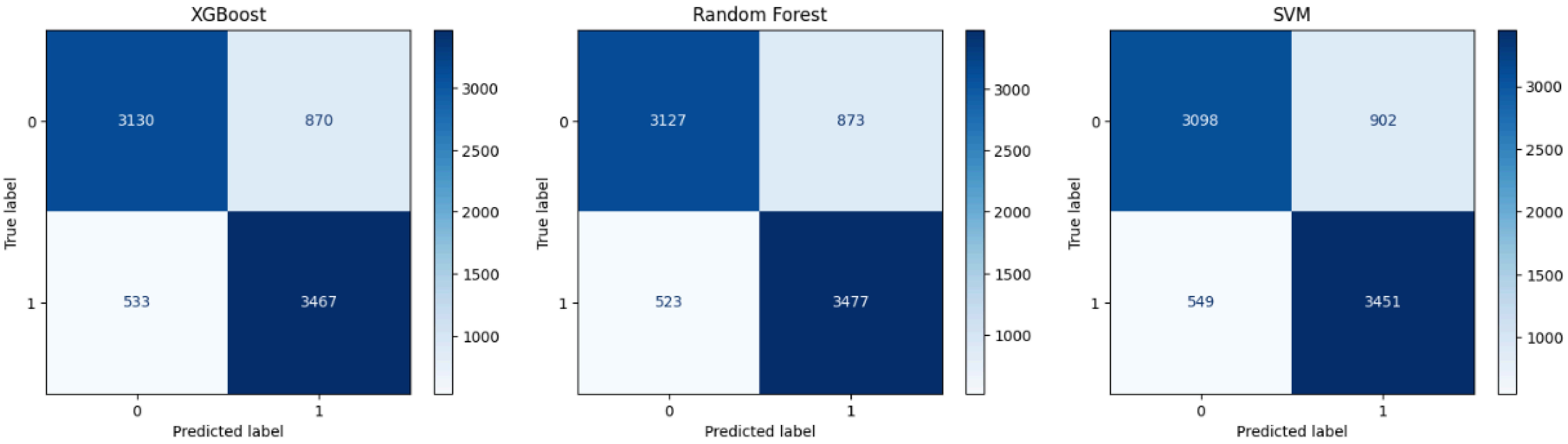

Random Forest Feature Importances


XGBoost Feature Importances

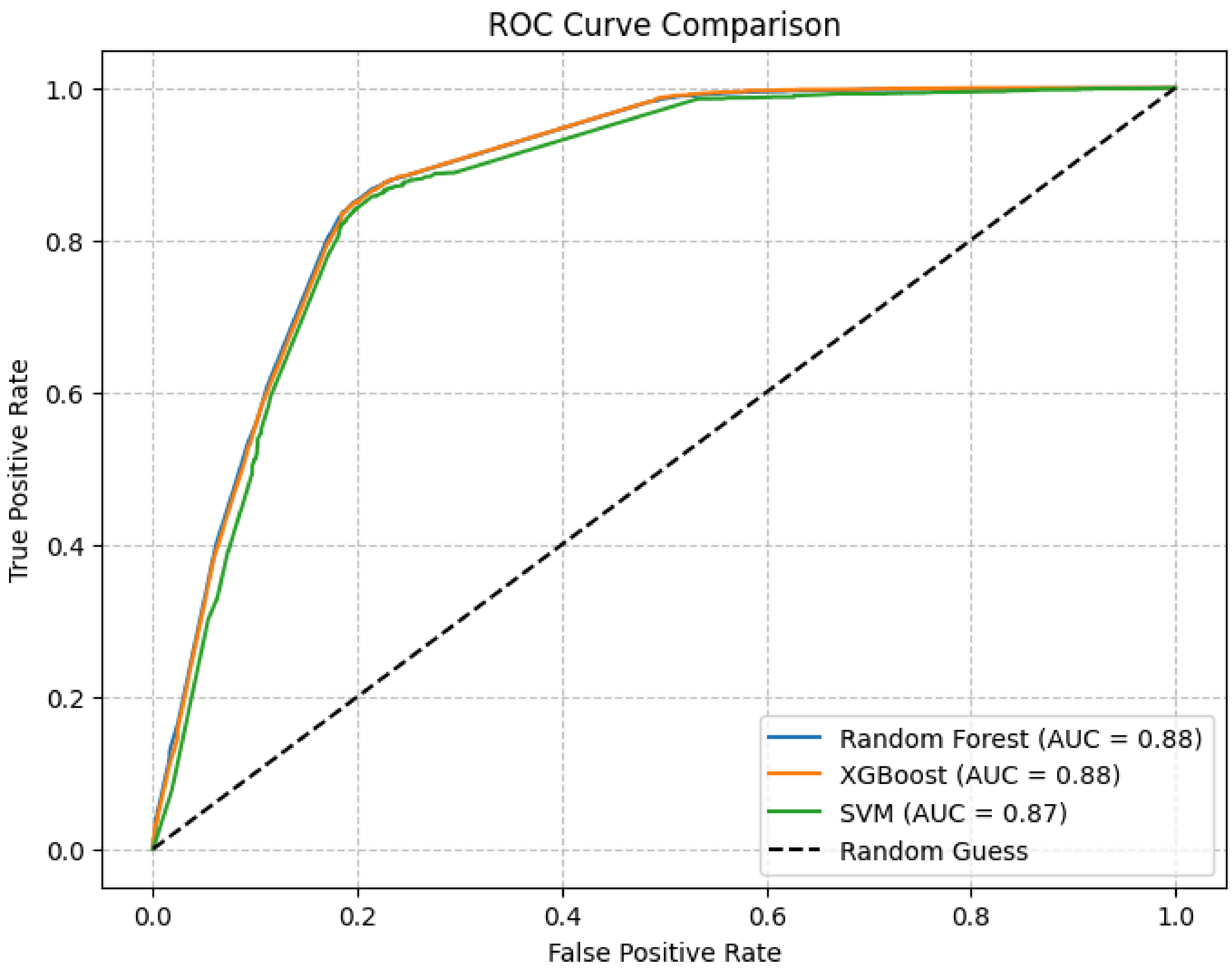
Models Performances comparison



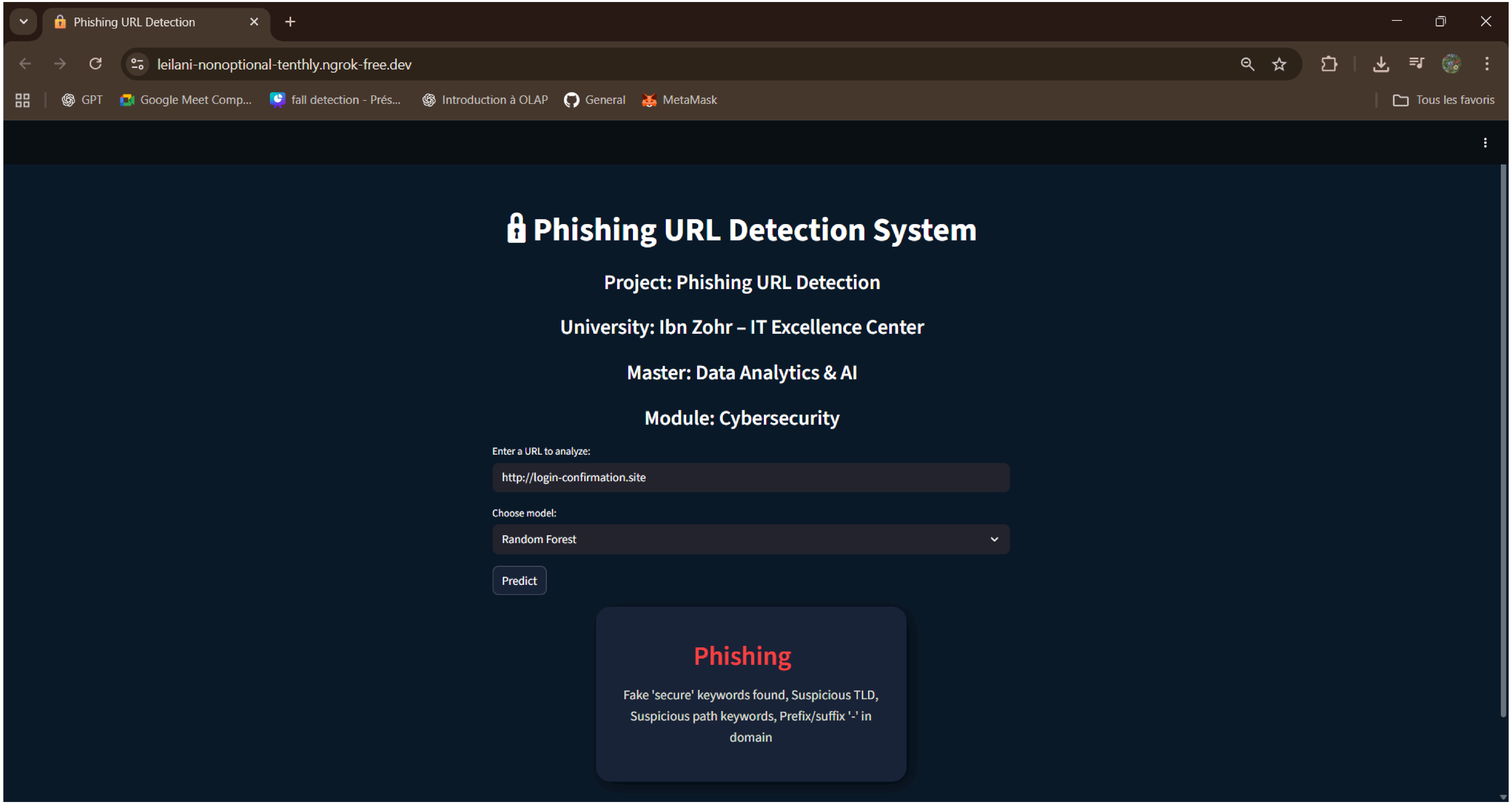
Models Performances comparison



Models Performances comparison - ROC Curves



Demonstration



🔒 Phishing URL Detection System

Project: Phishing URL Detection

University: Ibn Zohr – IT Excellence Center

Master: Data Analytics & AI

Module: Cybersecurity

Enter a URL to analyze:

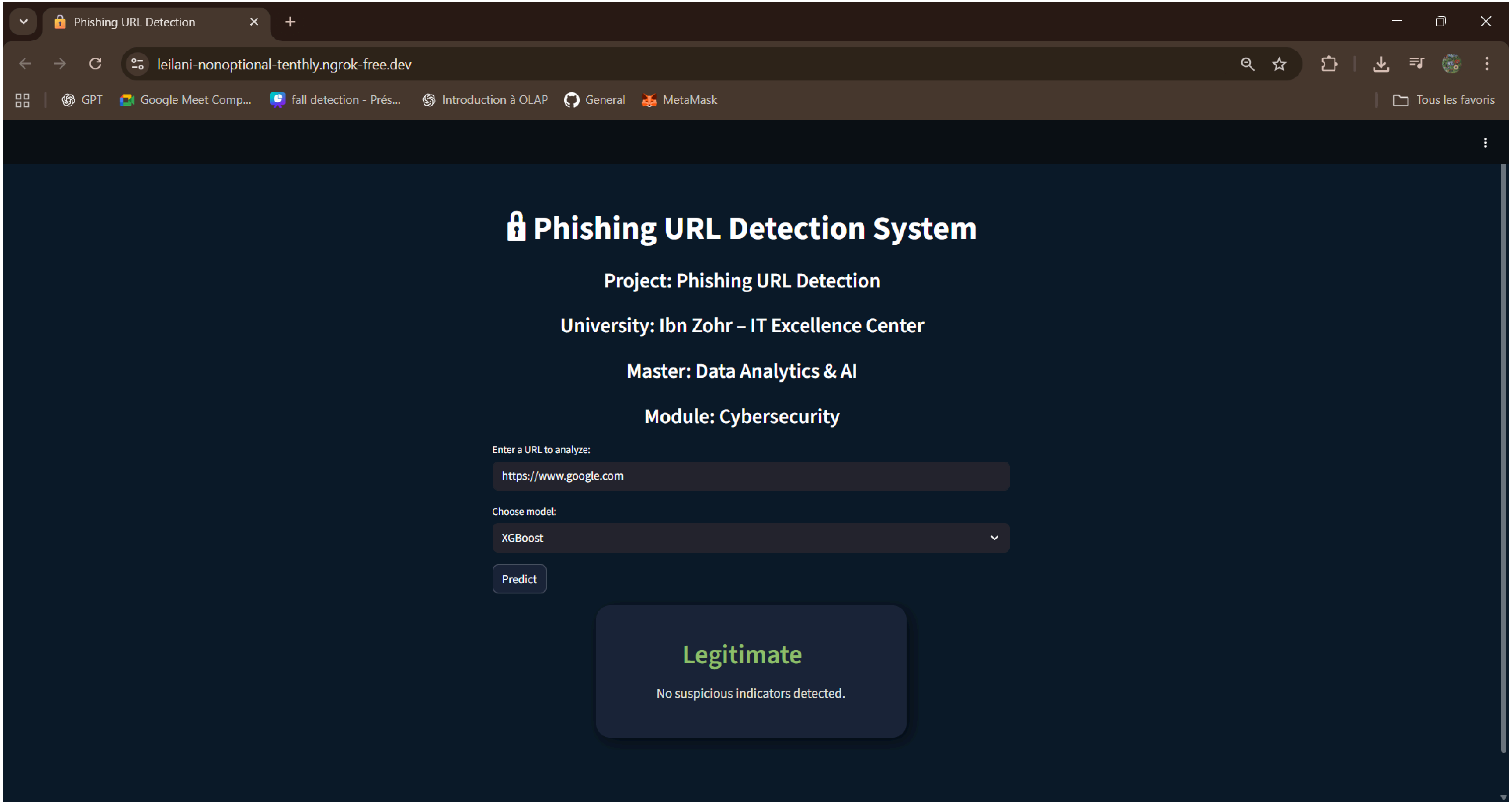
http://login-confirmation.site

Choose model:

Random Forest

Phishing

Fake 'secure' keywords found, Suspicious TLD, Suspicious path keywords, Prefix/suffix '-' in domain



🔒 Phishing URL Detection System

Project: Phishing URL Detection

University: Ibn Zohr – IT Excellence Center

Master: Data Analytics & AI

Module: Cybersecurity

Enter a URL to analyze: https://www.google.com

Choose model: XGBoost

Legitimate

No suspicious indicators detected.

Conclusion

Key Takeaways

1. Automated phishing URL detection is feasible using machine learning.
2. XGBoost performed best, achieving high accuracy.
3. URL characteristics, domain age, and redirection are strong indicators.
4. Balanced datasets improve model reliability and fairness.
5. Combining manual rules with ML enhances detection accuracy.
6. The system can assist users in identifying phishing threats in real time.

Recommendations

Security Measures

1. Deploy real-time phishing detection in email filters or web proxies
2. Educate users on suspicious domains, short URLs, and redirects
3. Keep the dataset updated with new phishing URLs
4. Monitor high-risk features like IPs and short domain age
5. Combine ML detection with traditional cybersecurity tools

References

1. Dataset : **https://www.kaggle.com/datasets/harisudhan411/phishing-and-legitimate-urls?select=new_data_urls.csv**
2. Google Phishing Attack : **https://www.bbc.com/news/business-39798022**
3. Streamlit : **https://streamlit.io/**

Thank you for your attention !

Q&A