

# WEB PHISHING DETECTOR

BDC\_Group 3  
Tp Hồ Chí Minh, tháng 9 năm 2022



# NỘI DUNG

- ◆ Giới thiệu đề tài
- ◆ Phương pháp phát hiện
- ◆ Các thuật toán sử dụng
- ◆ Kết quả demo
- ◆ Giá trị khoa học
- ◆ Kết luận

A photograph of a diverse group of people working together at a wooden table. There are laptops, tablets, and smartphones open, displaying various screens. The scene is overlaid with several large, semi-transparent geometric shapes in blue, purple, and white.

# Giới thiệu thành viên



Phạm Đại Hoàng An



Trần Hoàng Công Toại



Nguyễn Trọng Anh



Đỗng Hoàng Sơn

# **GIỚI THIỆU ĐỀ TÀI**

---

4.66 tỷ người dùng  
Internet

92.6% người sử dụng  
điện thoại



# DIGITAL 2021

## GLOBAL OVERVIEW REPORT

THE LATEST INSIGHTS INTO HOW PEOPLE AROUND THE WORLD USE  
THE INTERNET, SOCIAL MEDIA, MOBILE DEVICES, AND ECOMMERCE

JAN  
2021

# GLOBAL DIGITAL GROWTH

THE YEAR-ON-YEAR CHANGE IN DIGITAL ADOPTION

INTERNET USER NUMBERS NO LONGER INCLUDE DATA SOURCED FROM SOCIAL MEDIA PLATFORMS, SO VALUES ARE NOT COMPARABLE WITH PREVIOUS REPORTS



TOTAL  
POPULATION



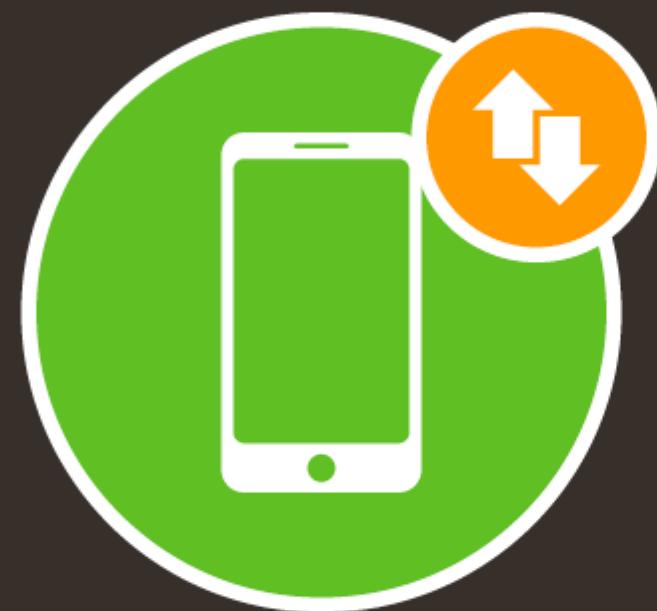
we  
are.  
social

+1.0%

JAN 2021 vs. JAN 2020

+81 MILLION

UNIQUE MOBILE  
PHONE USERS



+1.8%

JAN 2021 vs. JAN 2020

+93 MILLION

INTERNET  
USERS\*



K  
KEPIOS

+7.3%

JAN 2021 vs. JAN 2020

+316 MILLION

ACTIVE SOCIAL  
MEDIA USERS\*

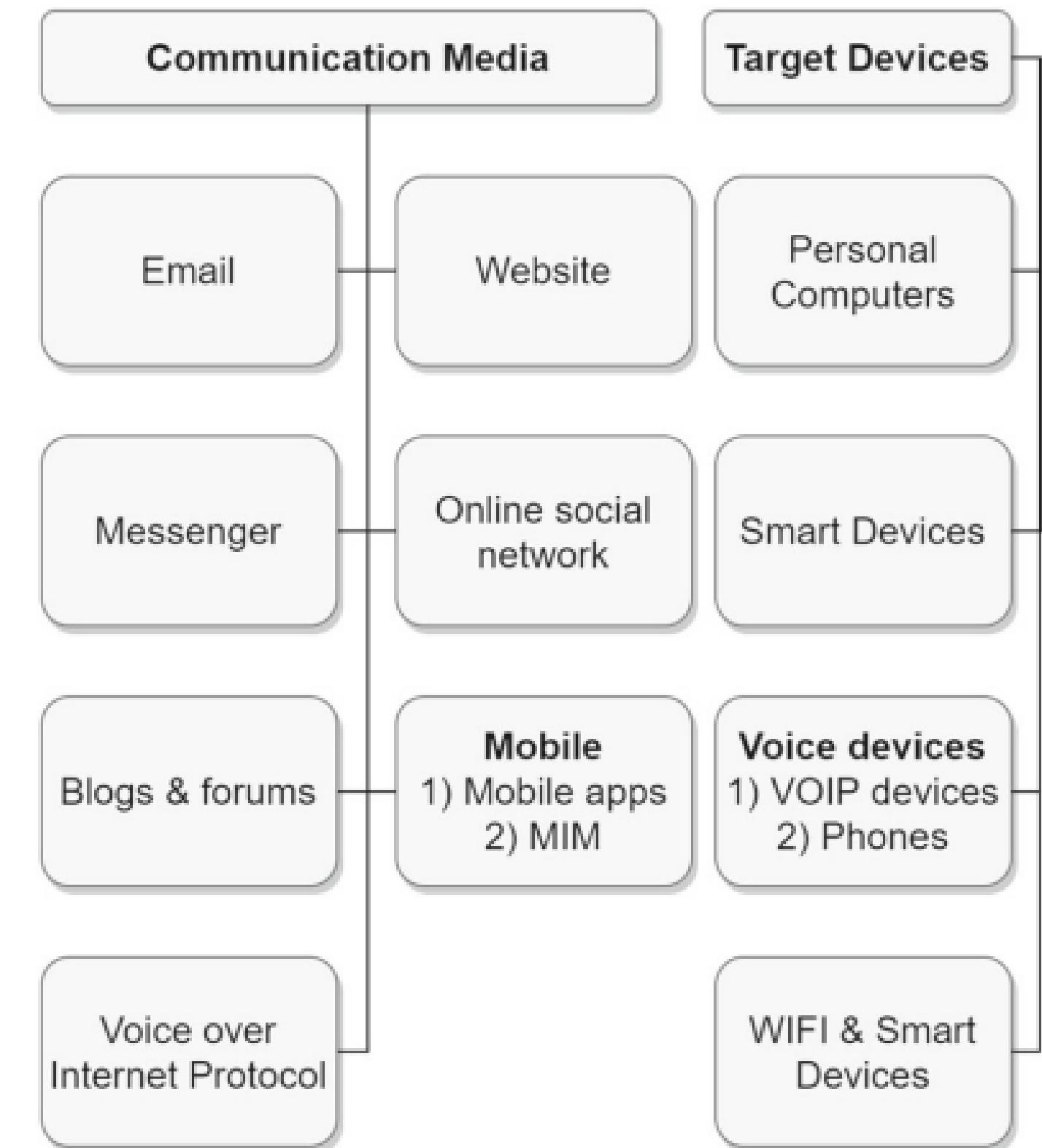


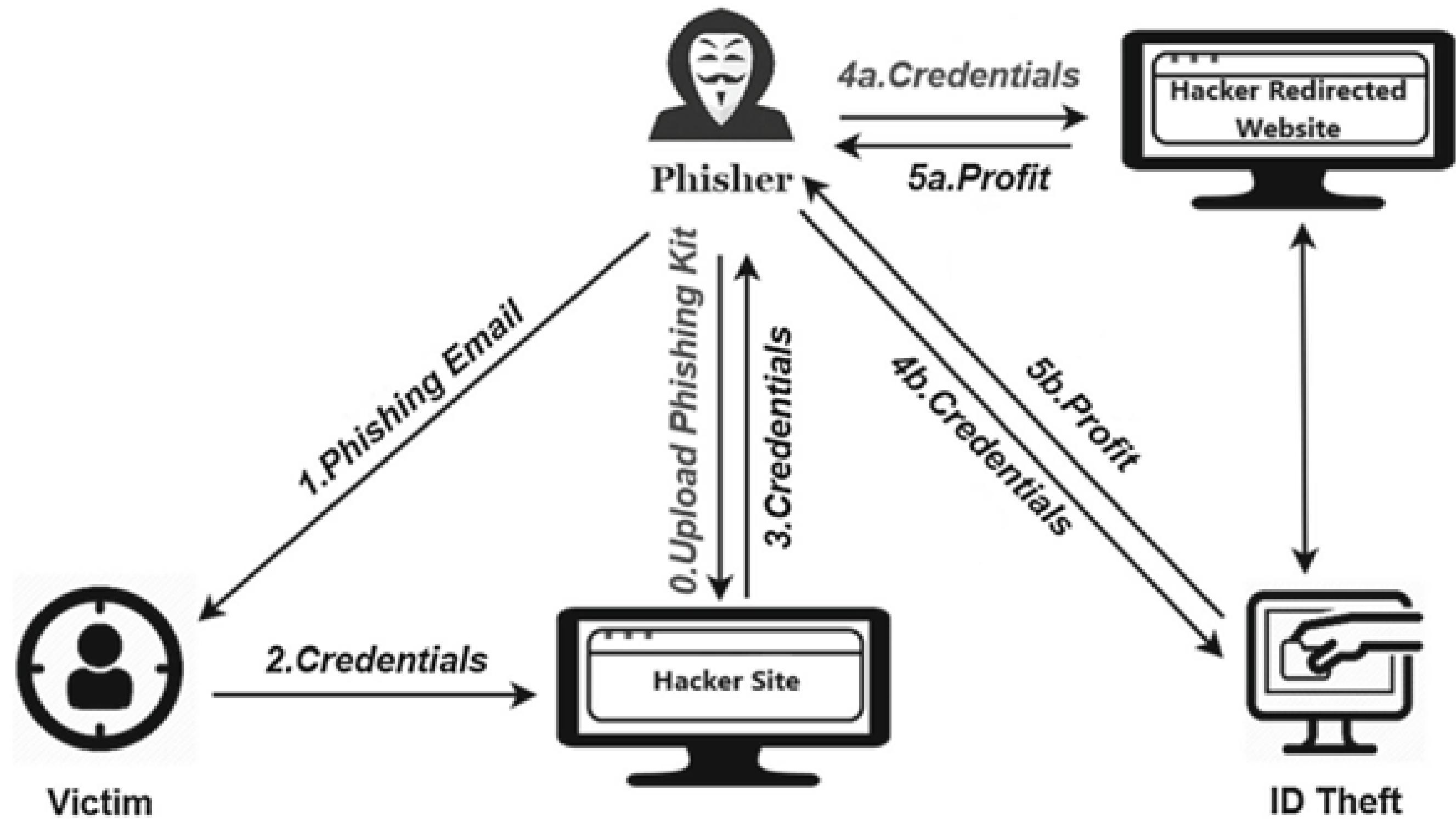
+13.2%

JAN 2021 vs. JAN 2020

+490 MILLION

SOURCE: NATIONAL GOVERNMENT BODIES; GSMA INTELLIGENCE; ITU; GWI; EUROSTAT; CNNIC; APJII; SOCIAL MEDIA PLATFORMS' SELF-SERVICE ADVERTISING TOOLS; COMPANY WEBSITE; HOOPTIC; MASCOP. \*ADVISORIES: INTERNET USER NUMBERS NO LONGER INCLUDE DATA SOURCED FROM SOCIAL MEDIA PLATFORMS, SO VALUES ARE NOT COMPARABLE WITH PREVIOUS REPORTS. SOCIAL MEDIA USER NUMBERS MAY NOT REPRESENT UNIQUE INDIVIDUALS. ♦ COMPARABILITY ADVISORY: SOURCE AND BASE CHANGES.







# Mục tiêu đề tài

Làm một ứng dụng phát hiện trang web lừa đảo



# PHƯƠNG PHÁP PHÁT HIỆN

---

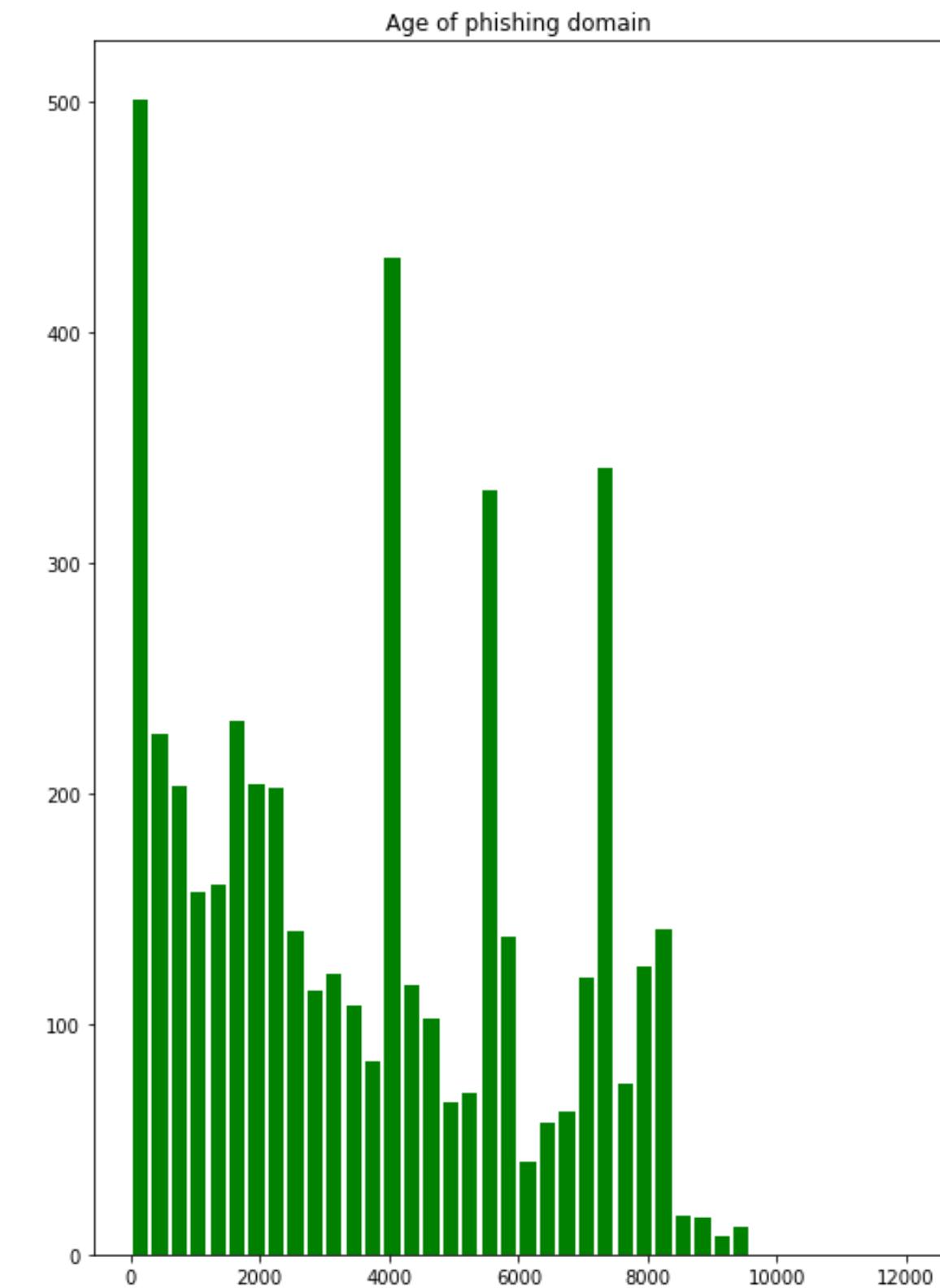
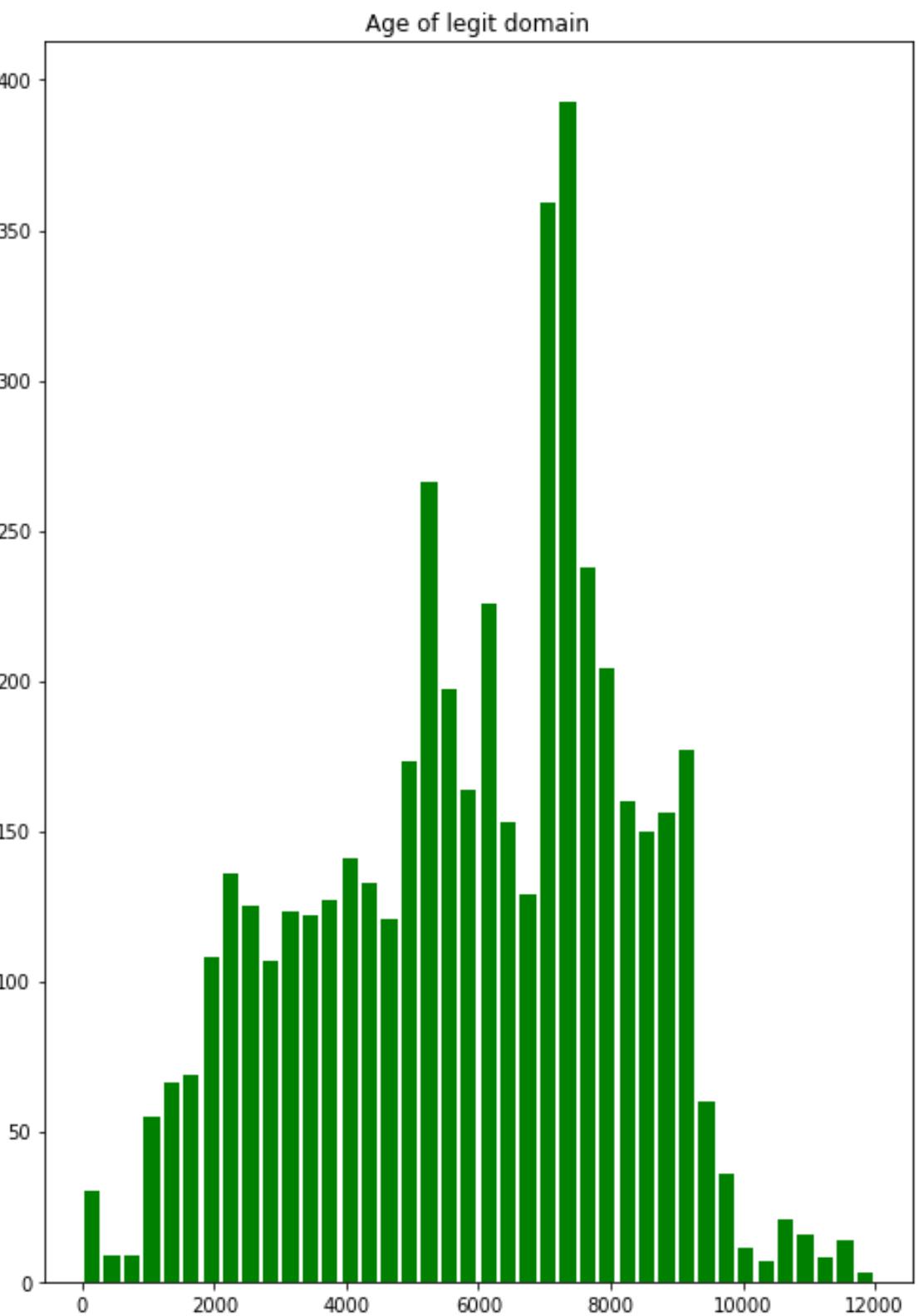


# Tiền xử lý dữ liệu

Ở giai đoạn tiền xử lý ta cần chuẩn bị tập dữ liệu là các địa chỉ URL, Các địa chỉ trang Web trong tập dữ liệu phải được đánh dấu là giả mạo hay không để dễ dàng xử lý.

	url	status
0	<a href="http://www.crestonwood.com/router.php">http://www.crestonwood.com/router.php</a>	legitimate
1	<a href="http://shadetreetechnology.com/V4/validation/a...">http://shadetreetechnology.com/V4/validation/a...</a>	phishing
2	<a href="https://support-appleld.com.secureupdate.duila...">https://support-appleld.com.secureupdate.duila...</a>	phishing
3	<a href="http://rgipt.ac.in">http://rgipt.ac.in</a>	legitimate
4	<a href="http://www.iracing.com/tracks/gateway-motorspo...">http://www.iracing.com/tracks/gateway-motorspo...</a>	legitimate
...	...	...
11425	<a href="http://www.fontspace.com/category/blackletter">http://www.fontspace.com/category/blackletter</a>	legitimate
11426	<a href="http://www.budgetbots.com/server.php?Server%20...">http://www.budgetbots.com/server.php?Server%20...</a>	phishing
11427	<a href="https://www.facebook.com/Interactive-Televisio...">https://www.facebook.com/Interactive-Televisio...</a>	legitimate
11428	<a href="http://www.mypublicdomainpictures.com/">http://www.mypublicdomainpictures.com/</a>	legitimate
11429	<a href="http://174.139.46.123/ap/signin?openid.pape.ma...">http://174.139.46.123/ap/signin?openid.pape.ma...</a>	phishing

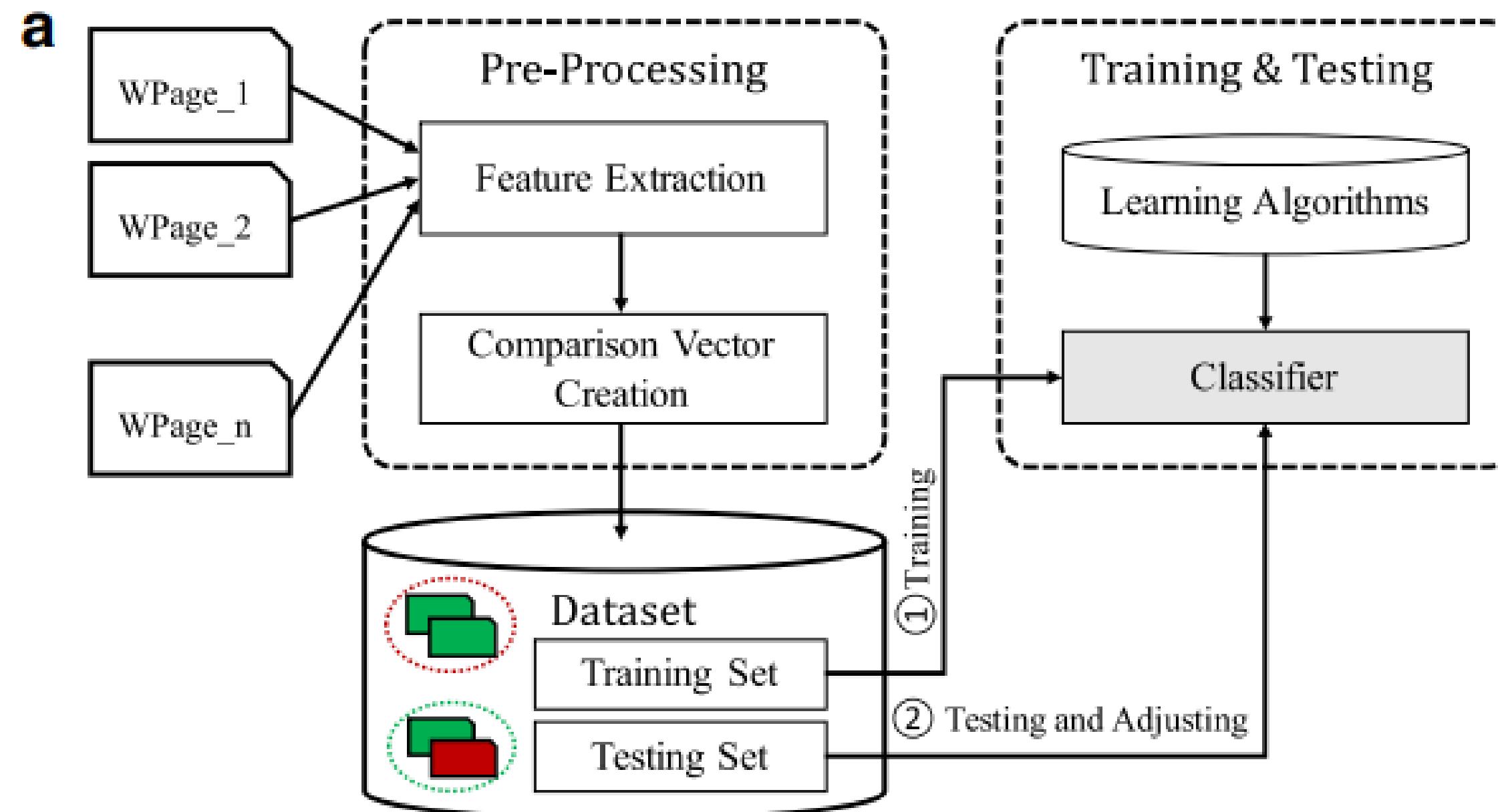
Từ các địa chỉ URL đó, ta trực quan hóa dữ liệu để phân biệt sự khác biệt từng đặc trưng giữa một trang web giả mạo và hợp lệ.



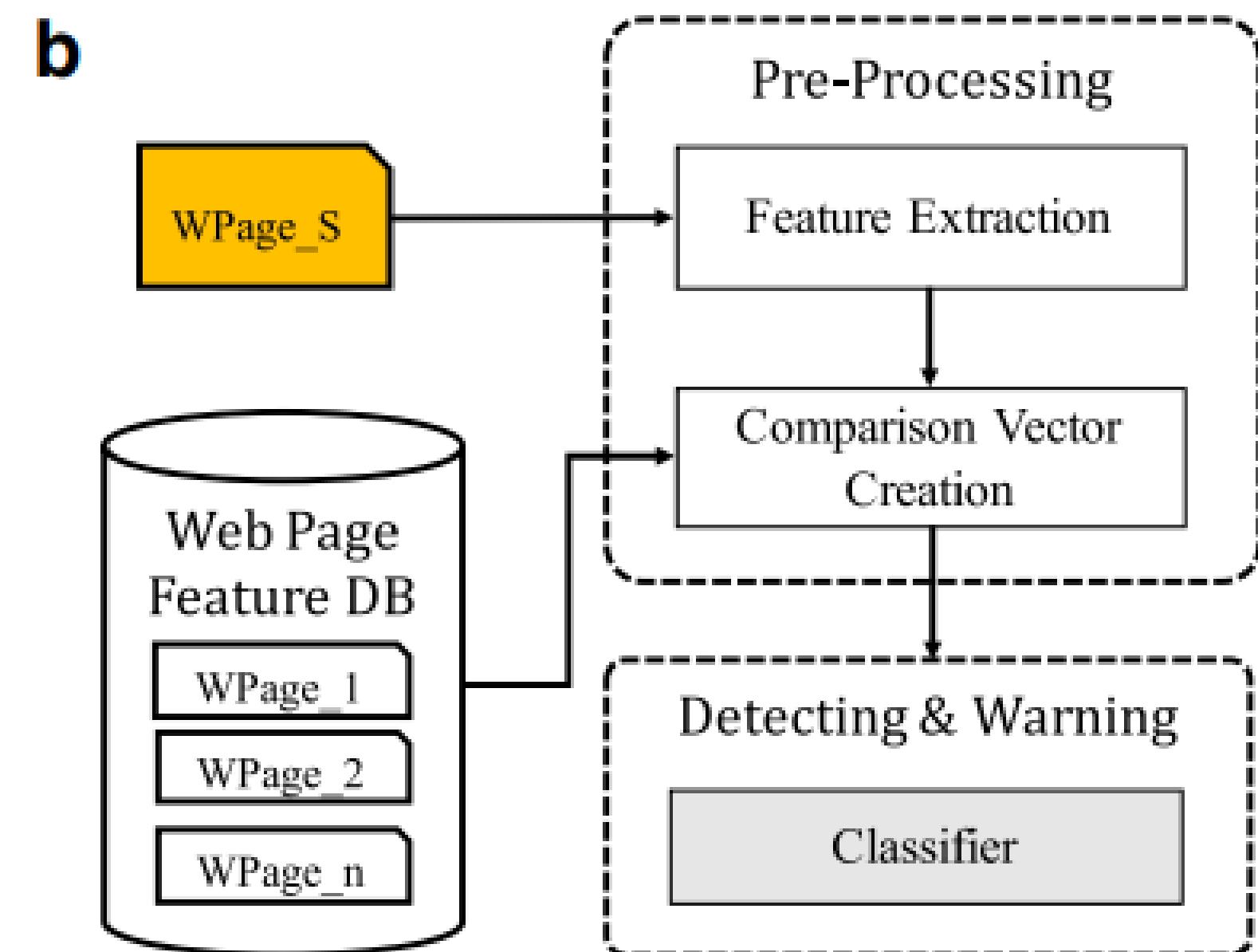
Từ các đặc trưng đó ta sử dụng kỹ thuật filter để lấy các thông tin cần thiết từ dữ liệu tổng như length\_url, prefix\_suffix,... Sau đó số hóa dữ liệu sang số liệu mình sẽ dùng

	UsingIP	LongURL	ShortURL	Symbol@	Redirecting//	PrefixSuffix-
0	1	1	1	1	1	-1
1	1	0	1	1	1	-1
2	1	0	1	1	1	-1
3	1	0	-1	1	1	-1
4	-1	0	-1	1	-1	-1
...	...	...	...	...	...	...
11049	1	-1	1	-1	1	1
11050	-1	1	1	-1	-1	-1
11051	1	-1	1	1	1	-1
11052	-1	-1	1	1	1	-1
11053	-1	-1	1	1	1	-1

# Trainning mô hình



# Phát hiện trang Web giả mạo



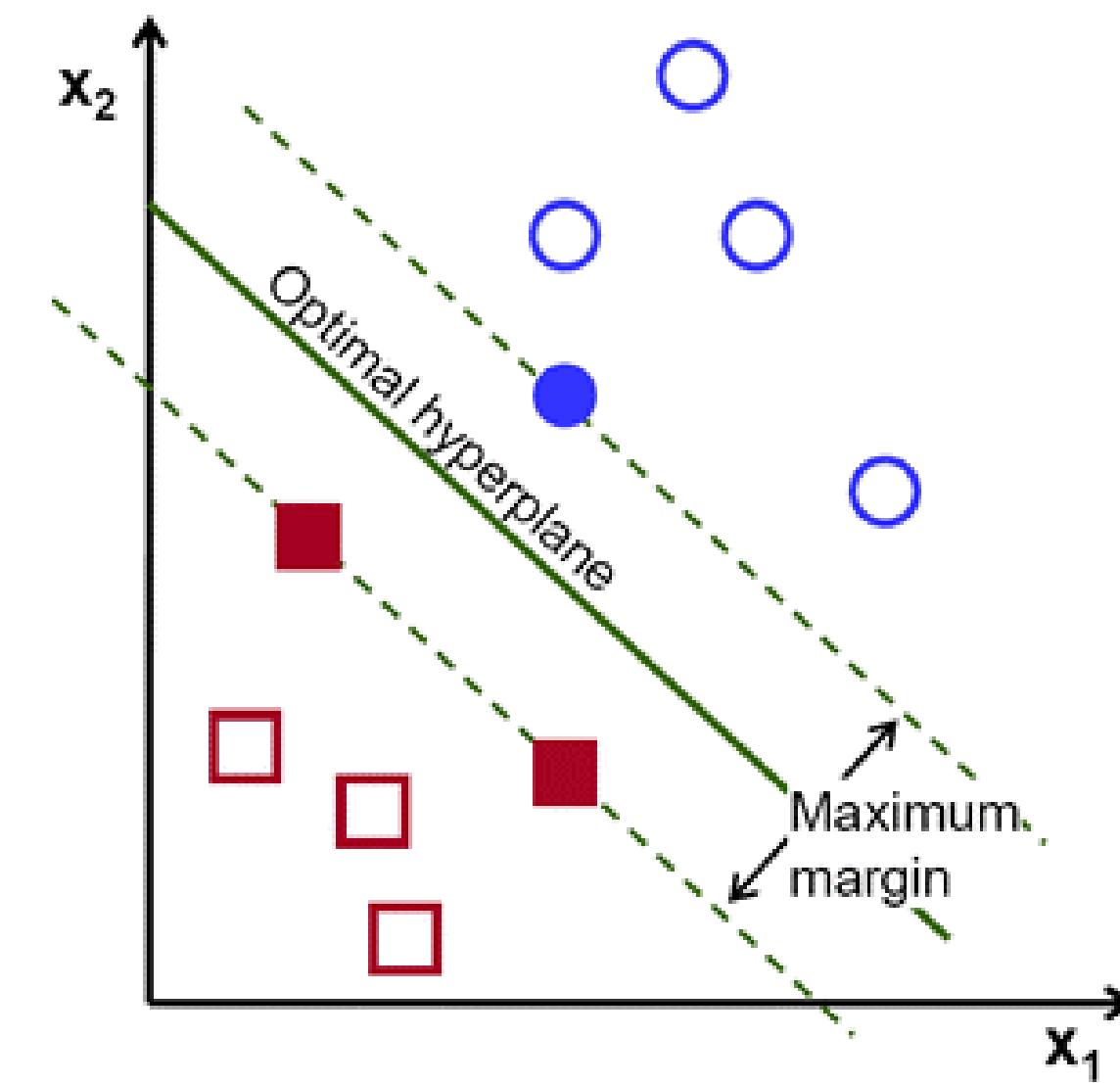
# CÁC THUẬT TOÁN SỬ DỤNG



# Support Vector Machine (SVM)

SVM là thuật toán phổ biến trong bài toán phân lớp.

Thuật toán này tìm siêu phẳng (hyperplane) tốt nhất phân chia hai lớp.

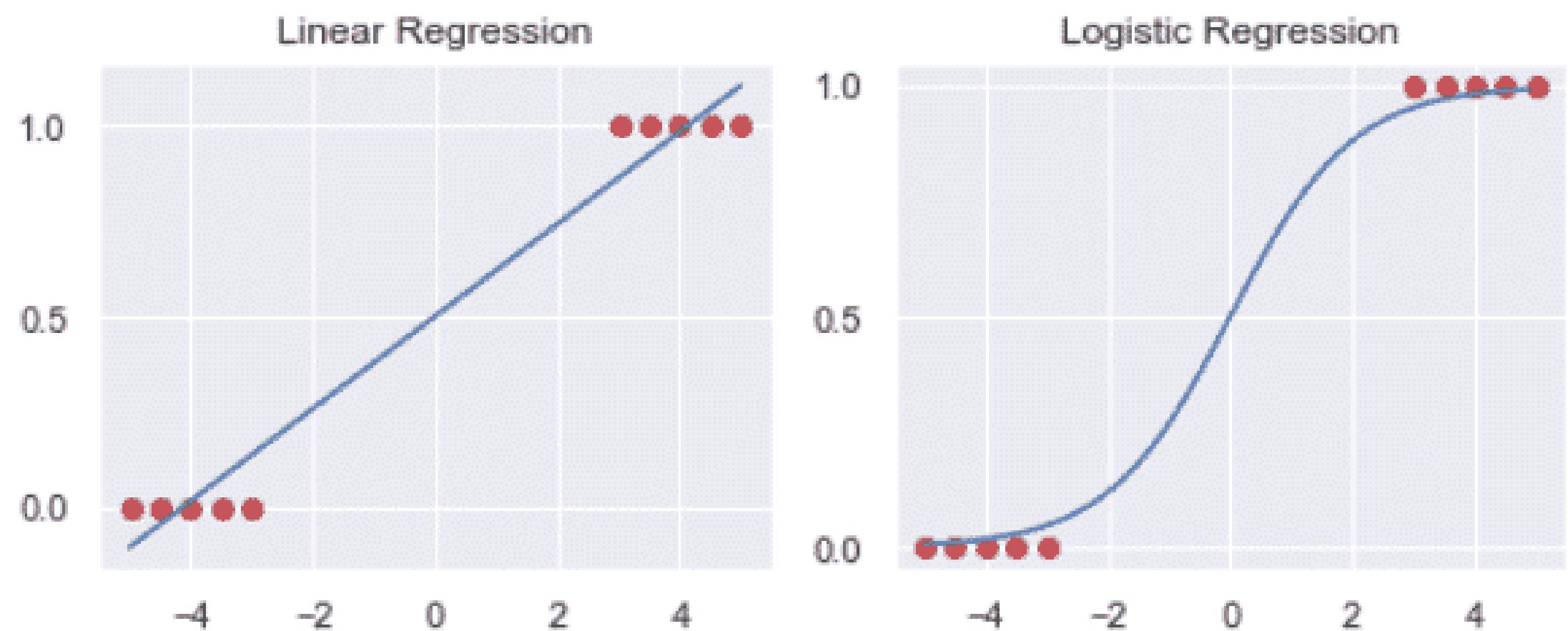


Nguồn: viblo.asia

# Logistic Regression

Hồi quy Logistic (Logistic Regression) là thuật toán phân lớp đơn giản phát triển từ Linear Regression.

Thuật toán được áp dụng để phân lớp tập dữ liệu linear separable.

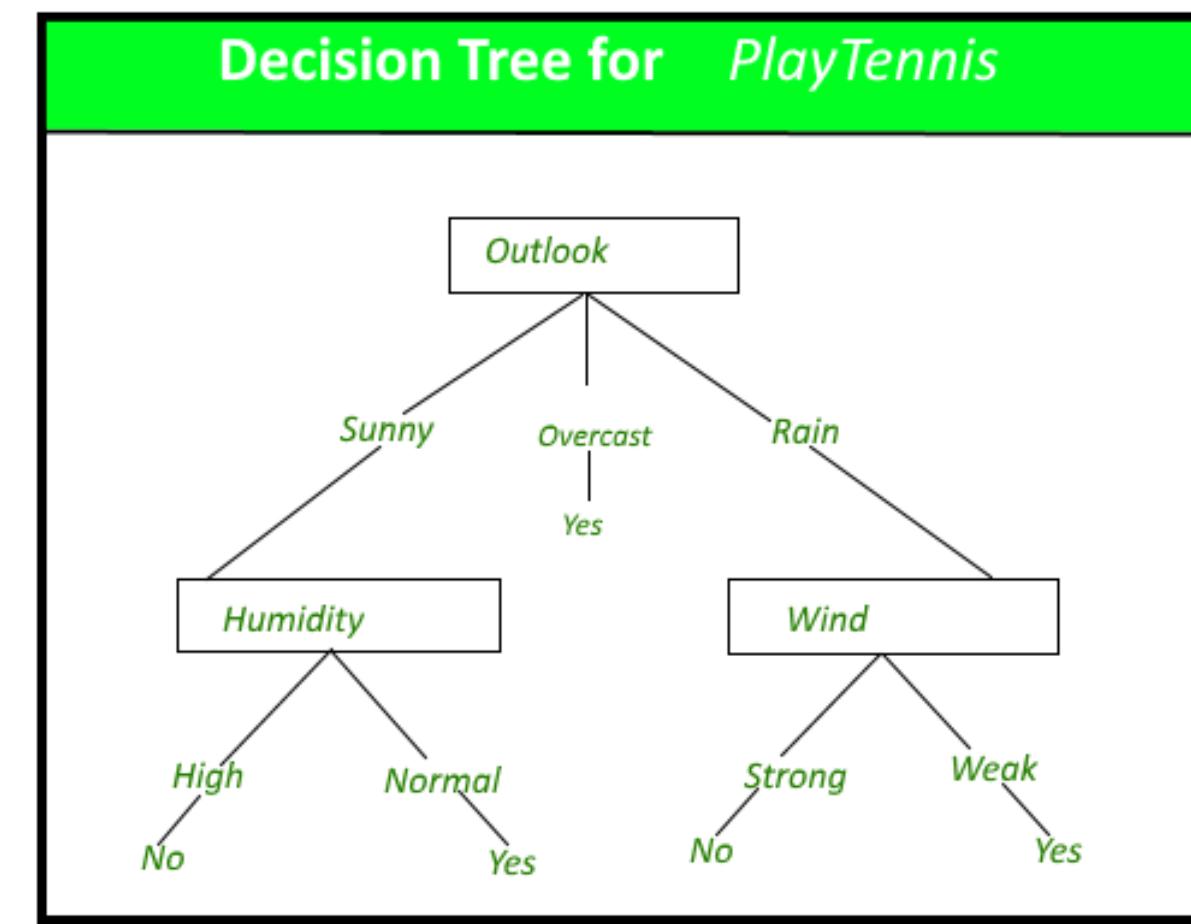


Nguồn: jcchouinard.com

# Decision Tree

Decision Tree là thuật toán phân loại dựa trên cây phân cấp, có thể biểu diễn dưới dạng các luật.

Mô hình dễ hiểu và dễ giải thích.



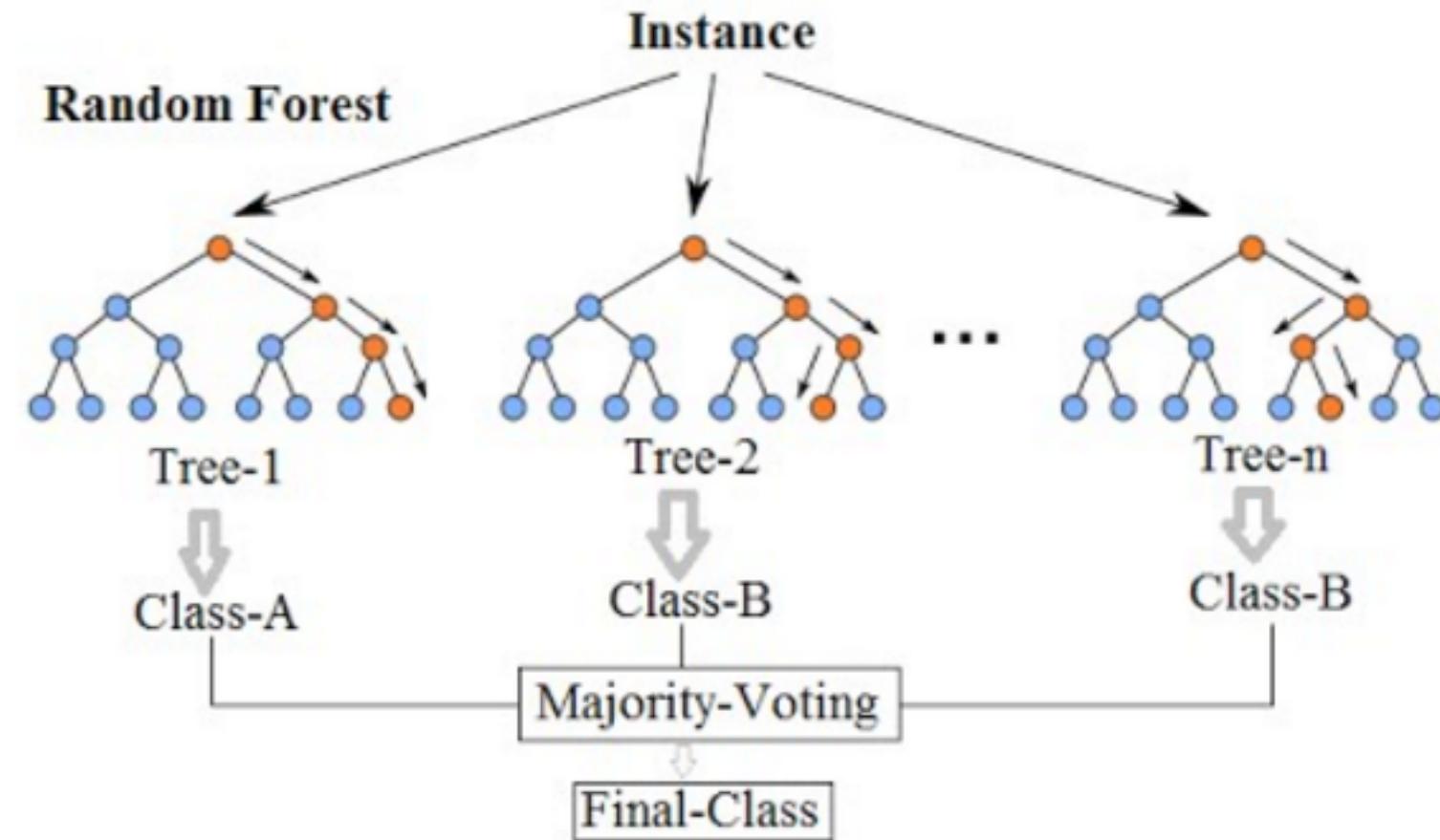
Nguồn: [geeksforgeeks.org](http://geeksforgeeks.org)

# Random Forest

Random Forest là thuật toán được xây dựng từ Decision Tree, nhưng ở đây không phải là một mà là nhiều cây.

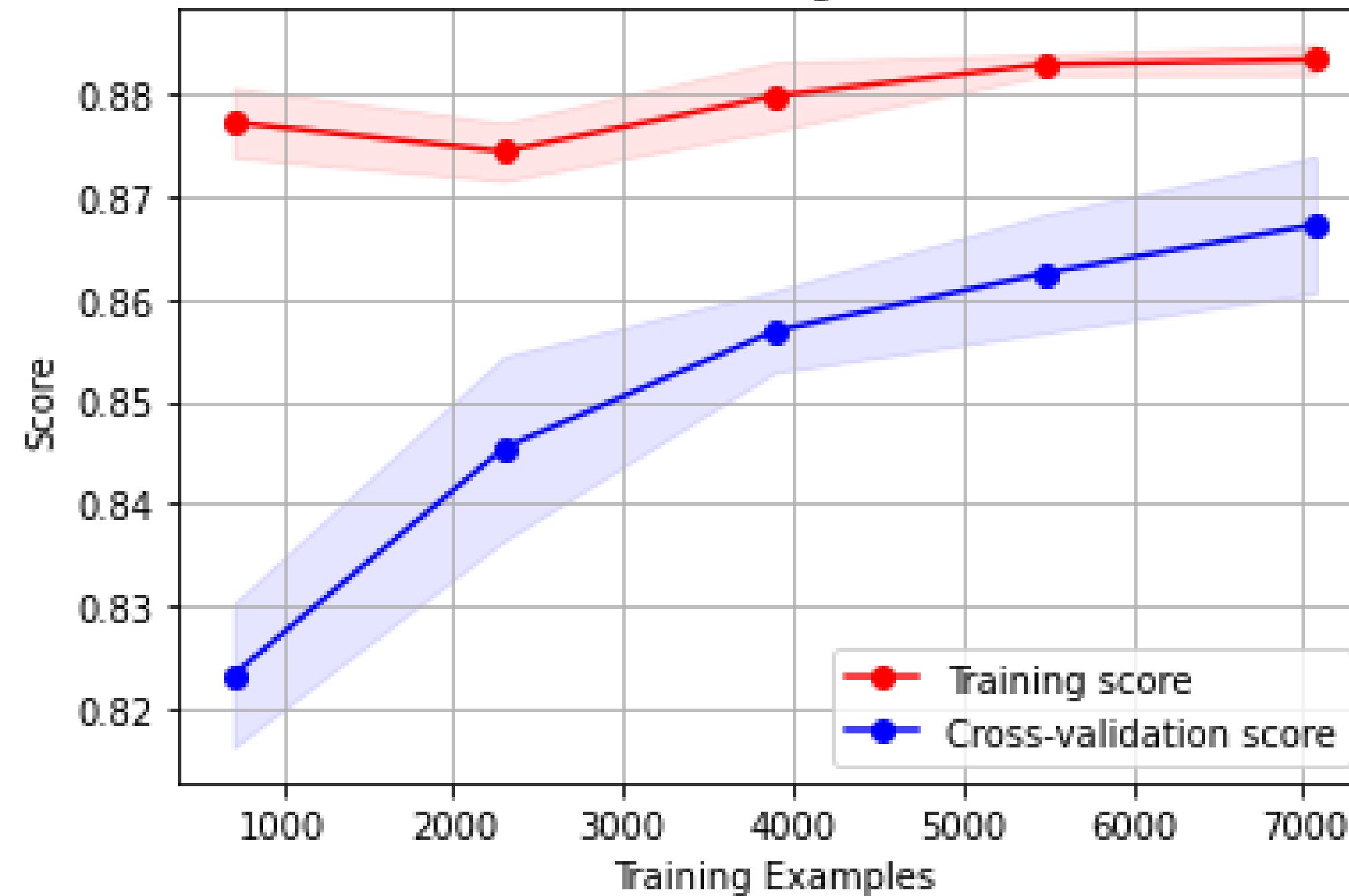
Kết quả dựa trên bỏ phiếu.

**Random Forest Simplified**

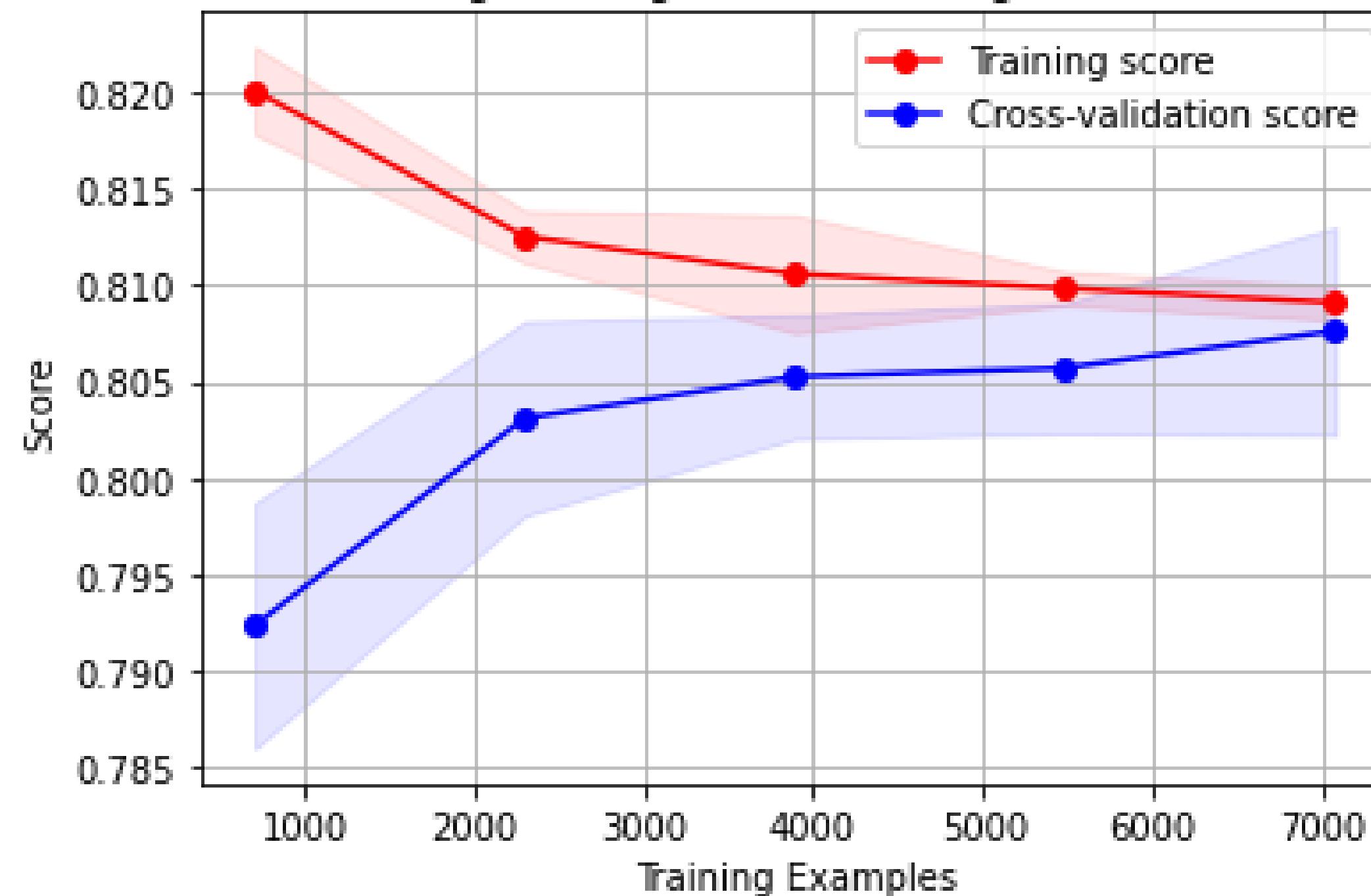


Nguồn: Wikipedia

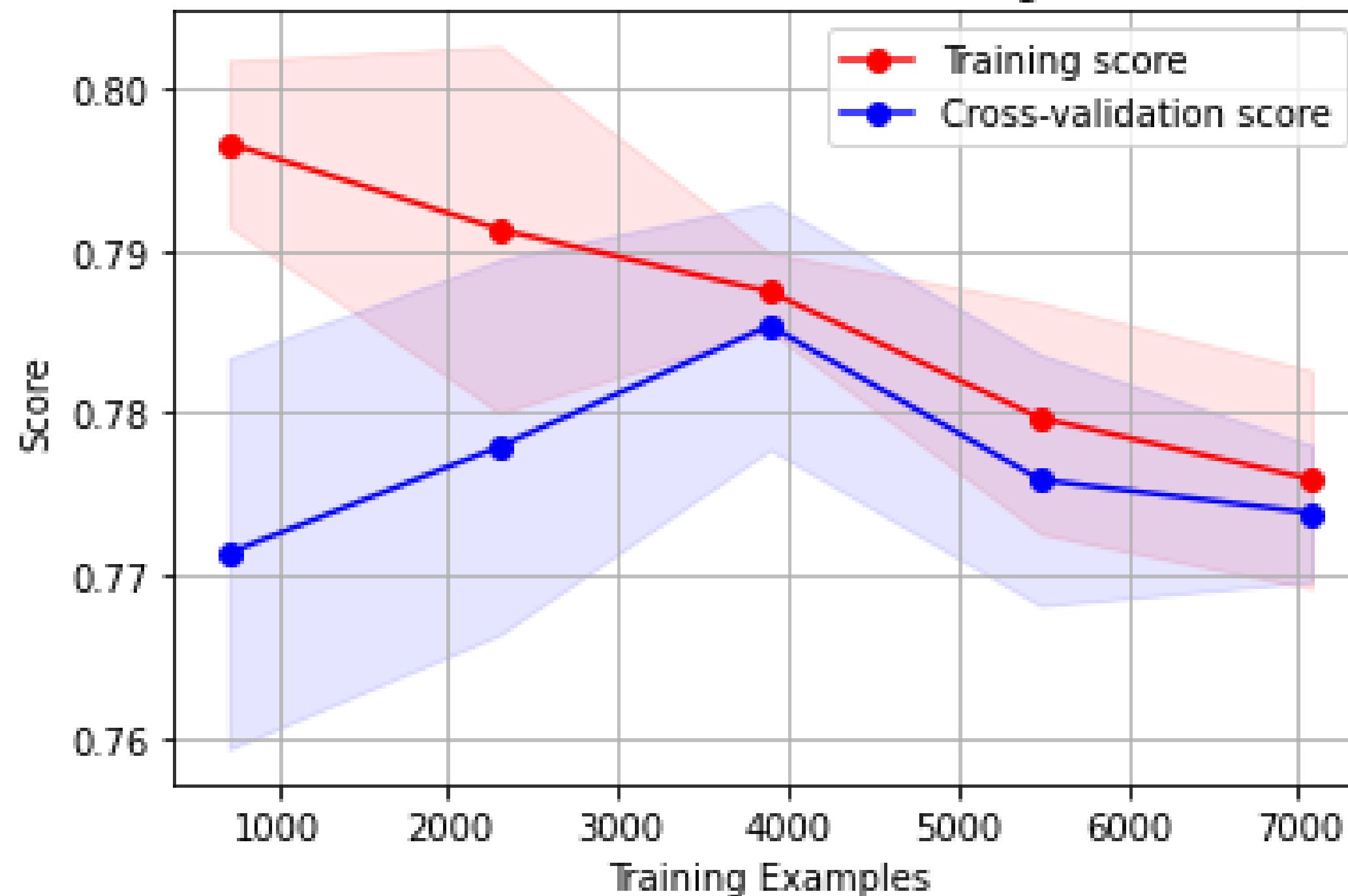
### SVM learning curves



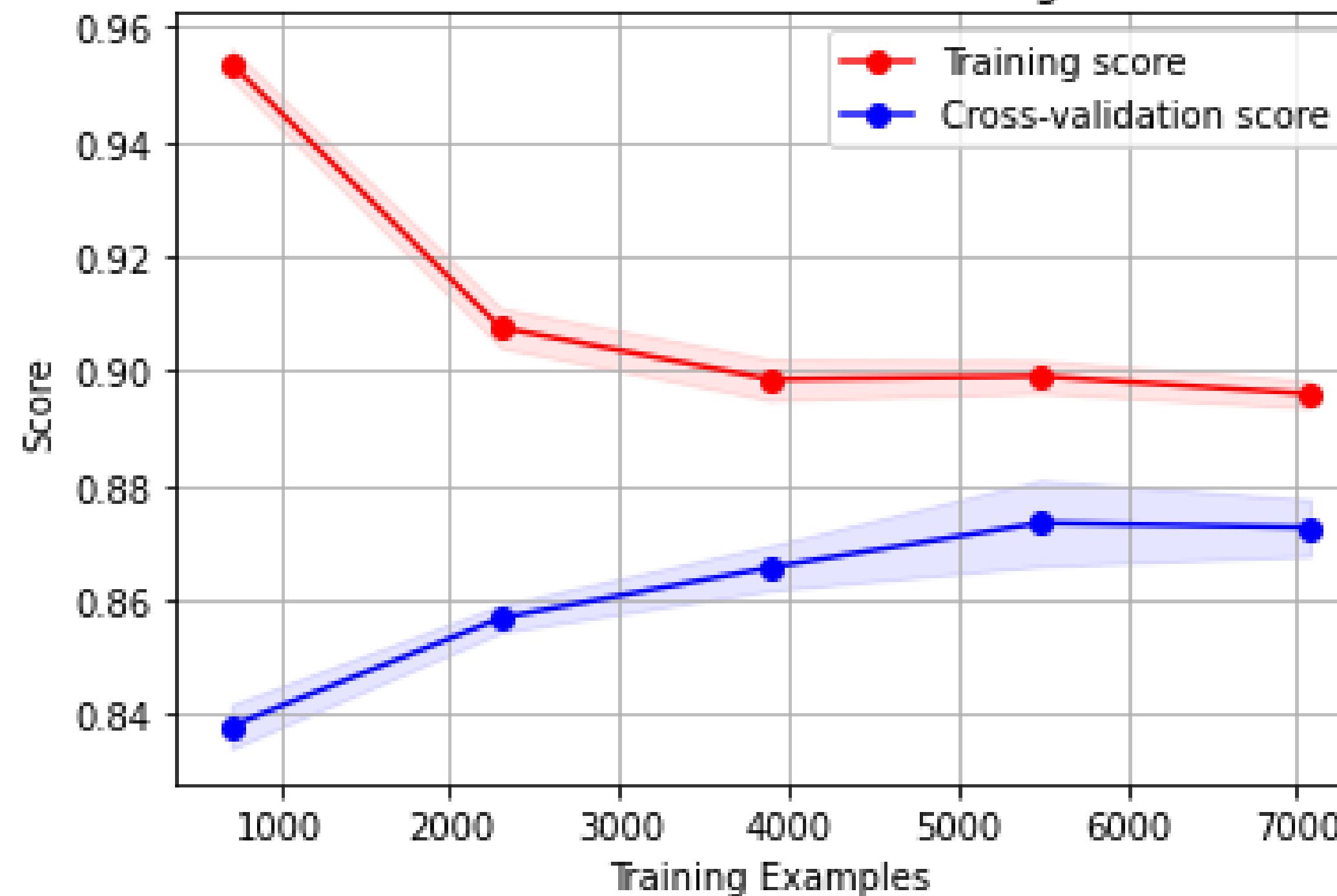
## Logistic Regression learning curves



## Decision Tree Classifier learning curves



## Random Forest Classifier learning curves



# Đánh giá mô hình

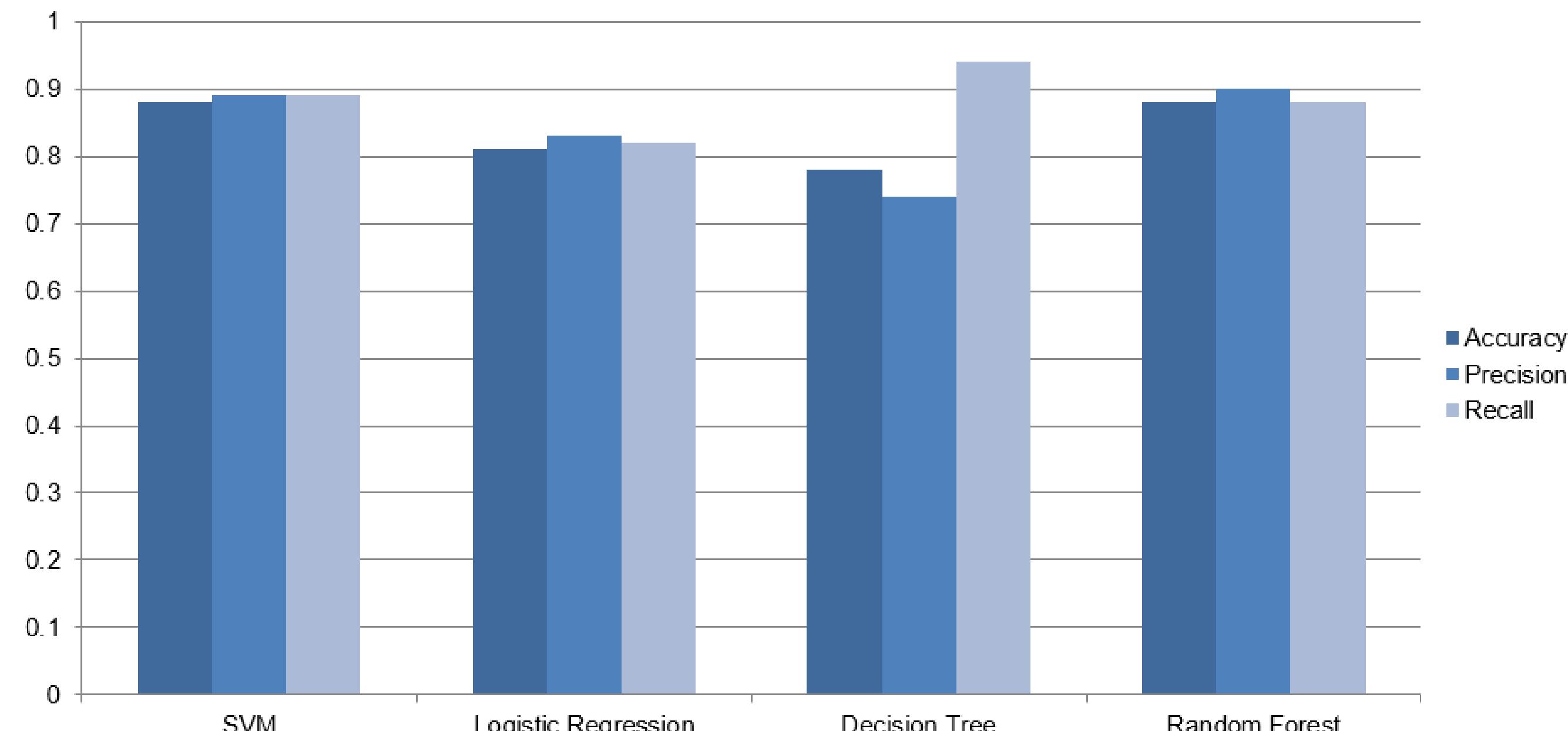
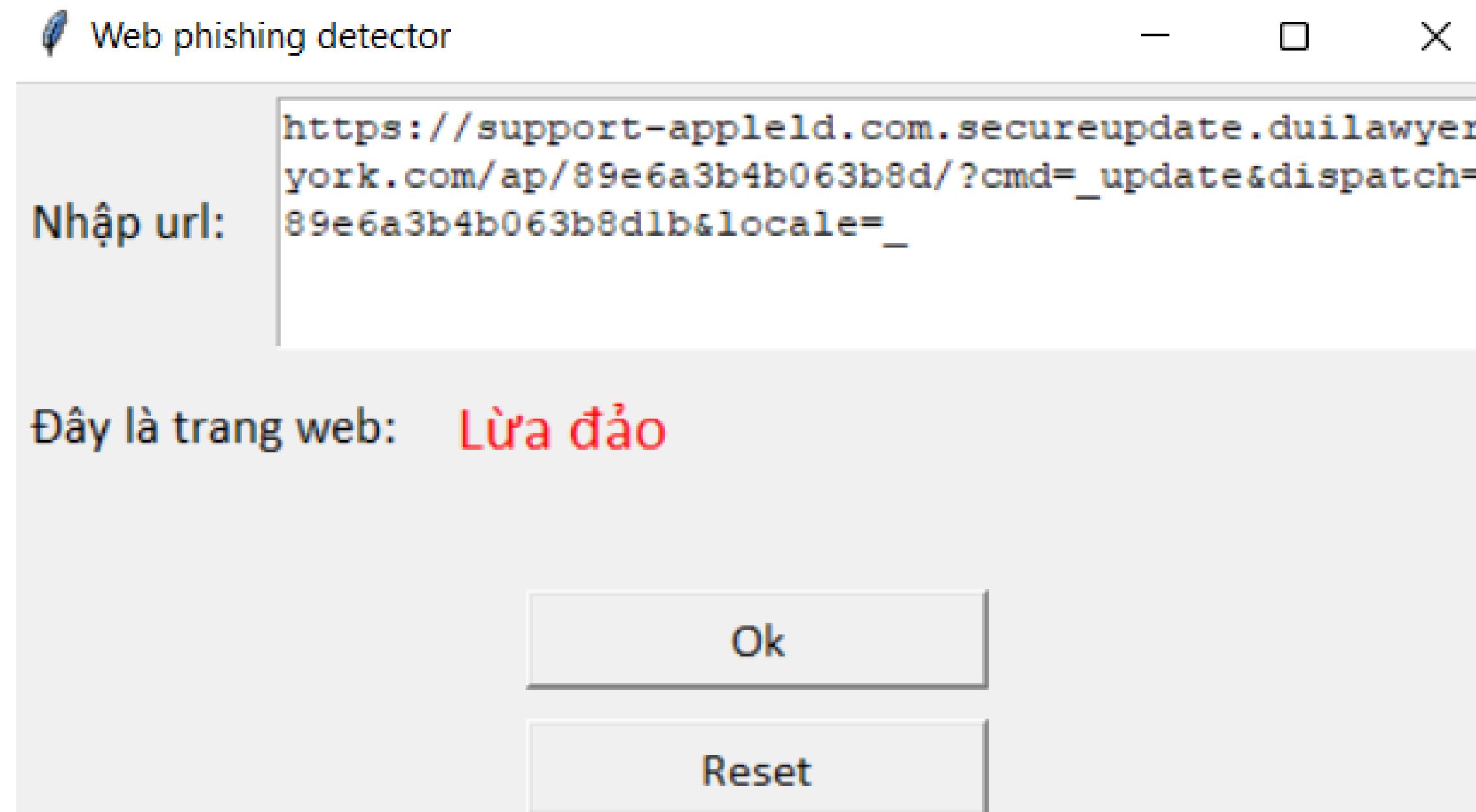


Chart 1. Model Evaluation

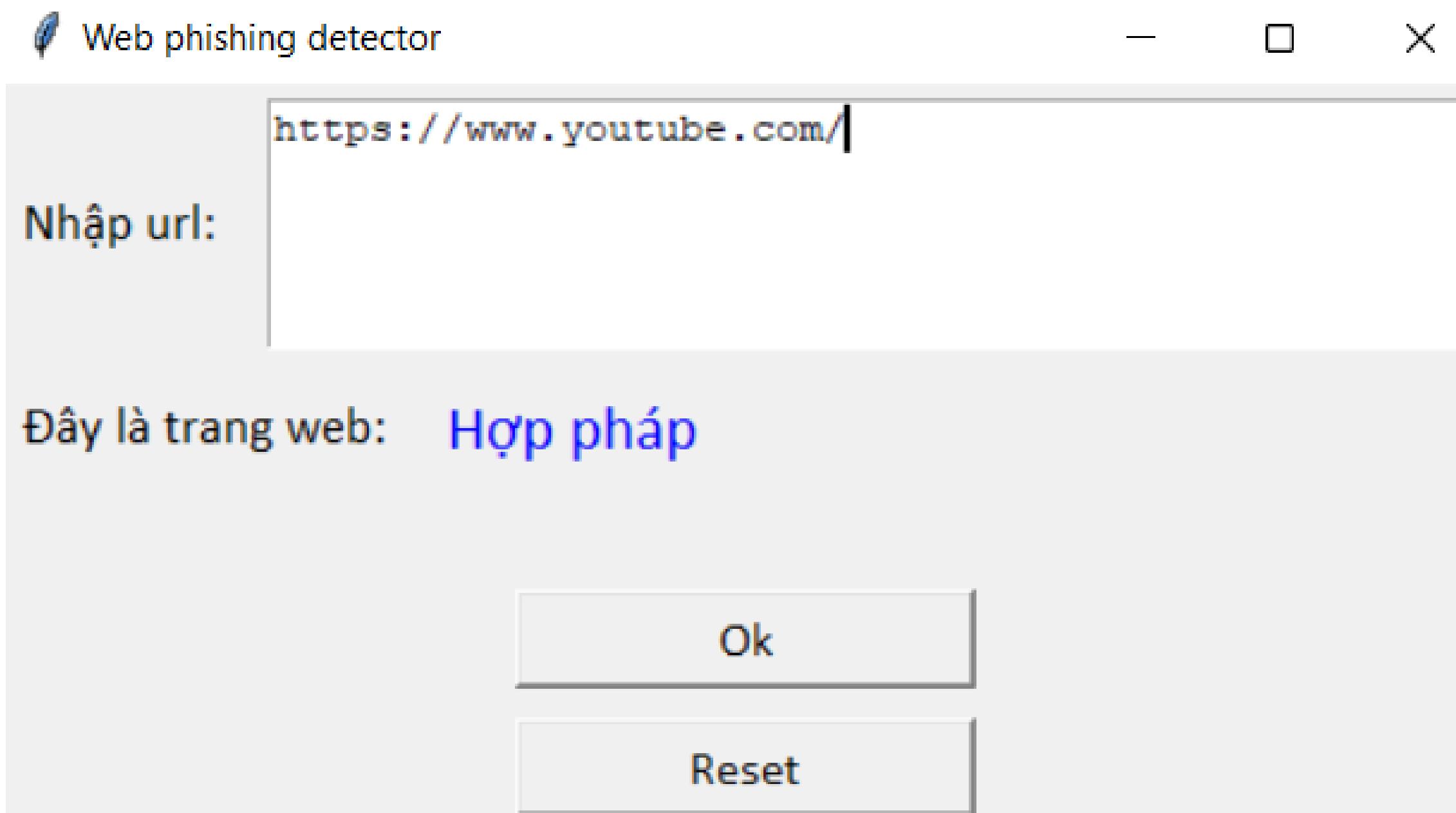
# KẾT QUẢ DEMO



# Web lừa đảo



# Web hợp pháp





# GIÁ TRỊ KHOA HỌC

---





# Mục tiêu nghiên cứu

- Mục tiêu nhận thức
- Mục tiêu sáng tạo
- Mục tiêu kinh tế





# KẾT LUẬN



## Đạt được

- Sử dụng thuật toán SVM phát triển mô hình
  - Phát triển một ứng dụng đơn giản để phát hiện web giả mạo
- .....

## Cần cải thiện

- Phát triển tập dữ liệu lớn hơn
- Cải thiện ứng dụng thành một extension sử dụng trên web
- Cải thiện mô hình chính xác hơn



# TÀI LIỆU THAM KHẢO

- Dalia Shihab Ahmed, Assist. Prof. Dr. Karim Q. Hussein, Hanan Abed Alwally Abed Allah. (2022). Turkish Journal of Computer and Mathematics Education. Vol. 13 No. 01. 100 - 107.
- Jian Mao, Jingdong Bian, Wenqian Tian, Shishi Zhu, Tao Wei, Aili Liand Zhenkai Liang (2019). Phishing page detection via learning classifiers from pagelayout feature. Truy cập từ: <https://jwcn-erasipjournals.springeropen.com/articles/10.1186/s13638-019-1361-0>
- Abdelhakim Hannousse, Salima Yahiouche. (2021). Web page phishing detection. Truy cập từ:  
<https://data.mendeley.com/datasets/c2gw7fy2j4/3>
- Rami M. Mohammad, Fadi Thabtah, Lee McCluskey. (2022). Phishing Websites Features.
- Huynh Chi Trung. (2020). Giới thiệu về Support Vector Machine (SVM).  
Truy cập từ: <https://viblo.asia/p/gioi-thieu-ve-support-vector-machine-svm-6J3ZgPVElmB>



Thank you!