



Phát hiện URL website độc hại dựa trên học máy

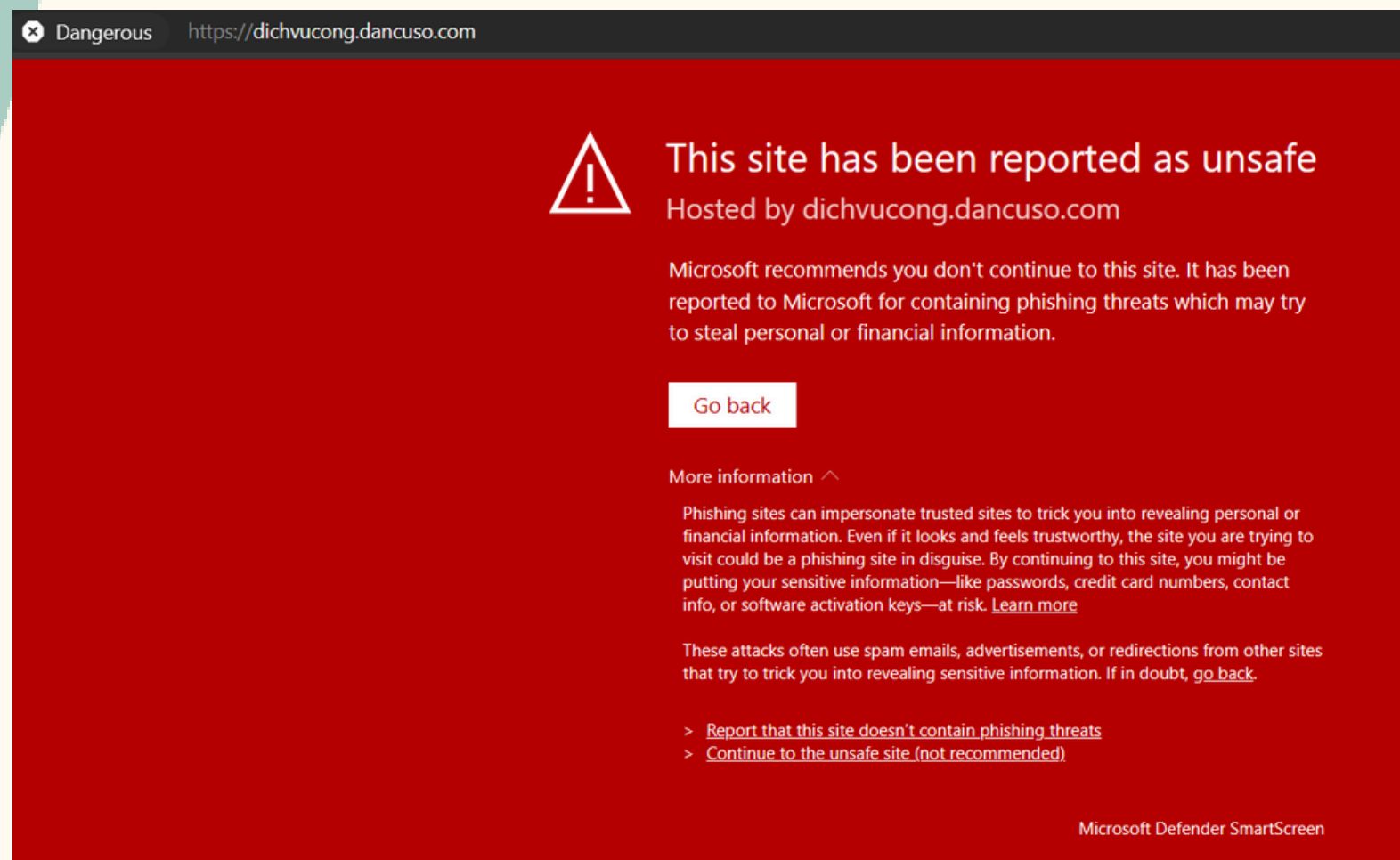
Phạm Văn Tiến - B21DCCN708



Giới thiệu

Trong thời đại công nghệ số, internet là một phần không thể thiếu trong cuộc sống. Tuy nhiên, internet cũng mang lại nhiều mối đe dọa về an toàn thông tin, đặc biệt là các trang web độc hại. Những trang này có thể chứa mã độc, đánh cắp thông tin cá nhân, lừa đảo tài chính...

Để giải quyết vấn đề này, nghiên cứu và phát hiện sớm các URL độc hại trước khi truy cập trang web là rất cần thiết. Sử dụng các mô hình học máy để phân loại URL độc hại là một phương pháp hiệu quả. Báo cáo này sẽ áp dụng một số thuật toán phân loại học máy để đánh giá URL và thiết kế một tiện ích mở rộng trên trình duyệt dựa trên mô hình hiệu quả nhất.

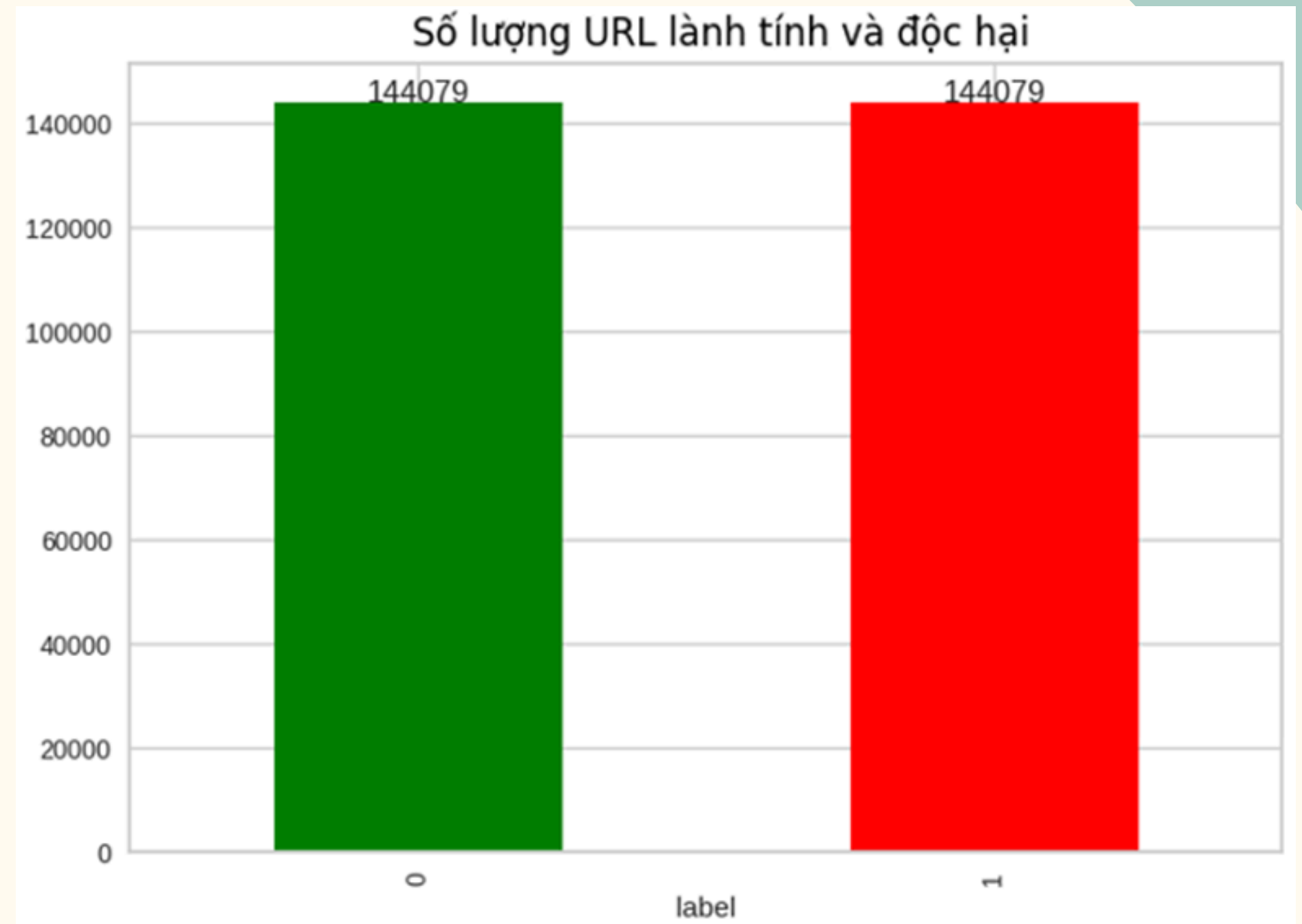


Data Collection

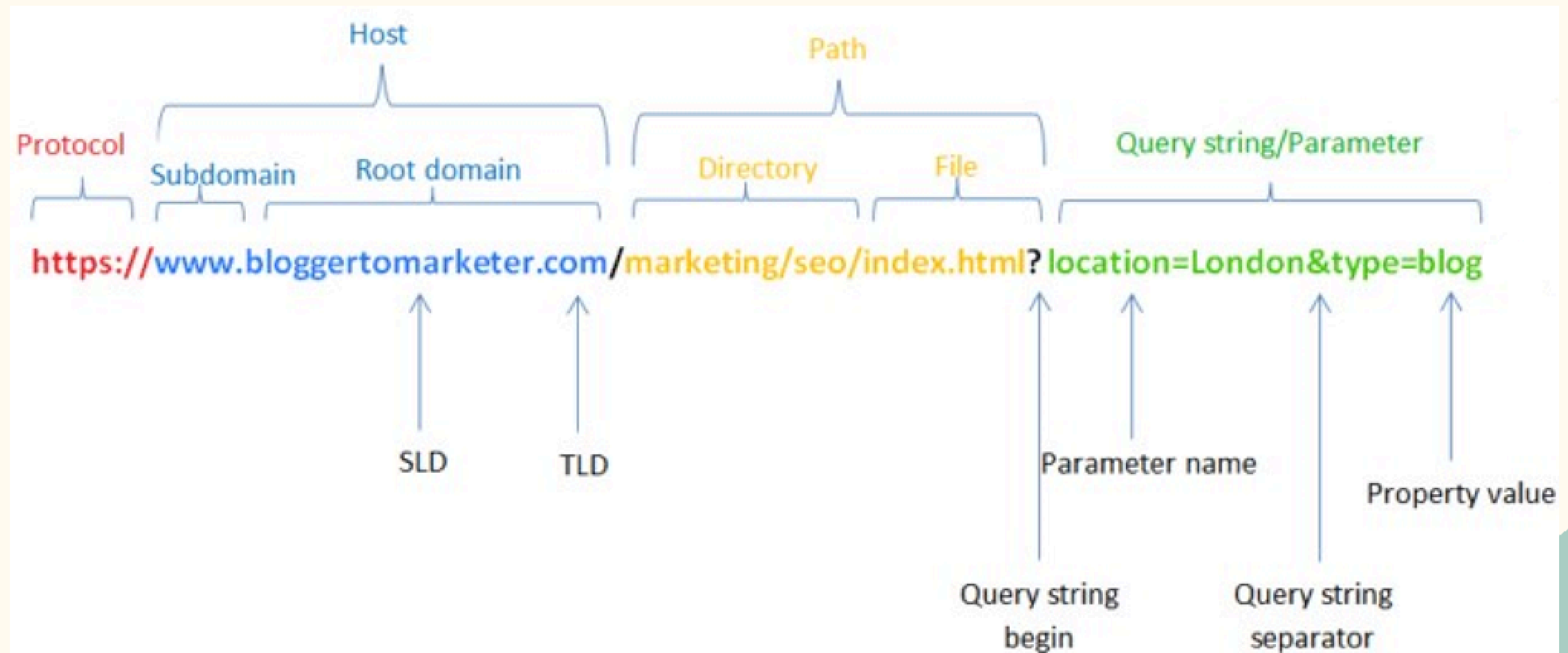
Các URL độc hại được thu thập từ các trang OpenPhish, PhishTank và các URL an toàn được thu thập từ Kaggle và UCI Machine Learning Repository.

Danh sách các URL được tổng hợp lại và dán nhãn 1 cho các URL độc hại và 0 cho URL an toàn.

Ta thu được một danh sách 288158 URL gồm 144079 URL an toàn và 144079 URL độc hại.



Cấu trúc cơ bản của URL



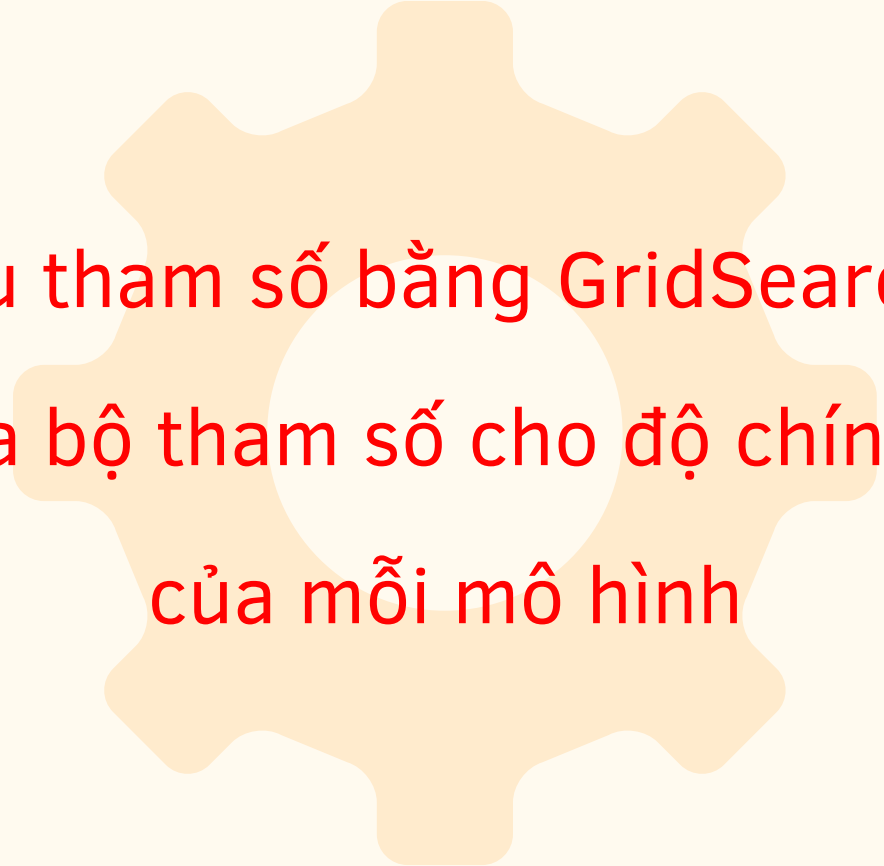
Feature Extraction

Sử dụng các thư viện urllib, tldextract và tld để trích chọn dữ liệu các thành phần của URL. Các đặc trưng được sử dụng gồm 26 đặc trưng từ vựng:

STT	Value	Feature
1	0/1	IpAddress, UseShortService, SusTlds, AtSymbol, TildeSymbol, DoubleSlashInPath, DomainInSubdomains, DomainInPaths, rank_host
2	integer	UrlLength, HostnameLength, PathLength, QueryLength NumSensitiveWords, NumNumericChars, NumDots, NumDash, NumDashInHostname, NumUnderscore, NumPercent, NumAmpersand, NumHash, NumQueryComponents SubDomainLevel, PathLevel
3	float	EntropyDomainName

Models Tested

- Logistic regression
- Decision Tree
- Random Forest
- Gradient Boosting
- XGBoost



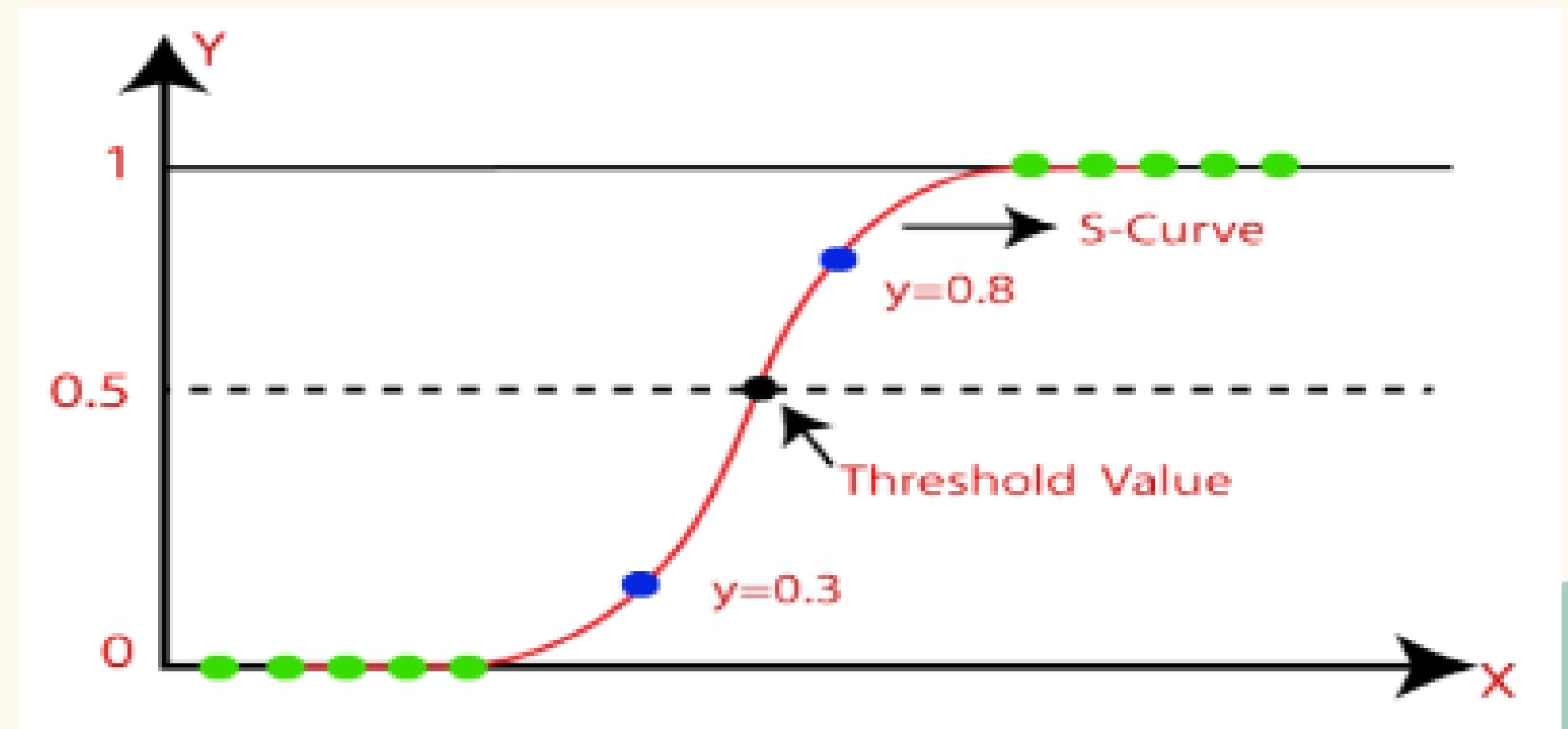
Điều chỉnh siêu tham số bằng GridSearchCV (tìm kiếm lưới) để tìm ra bộ tham số cho độ chính xác tốt nhất của mỗi mô hình

Logistic regression

Logistic Regression là một thuật toán thường được sử dụng trong phân loại (thường là phân loại nhị phân) dù tên gọi có chứa từ "Regression".

Logistic Regression dự đoán xác suất của một sự kiện nhất định, hoạt động bằng cách sử dụng hàm logit (hay sigmoid) để biến đổi đầu ra của một mô hình hồi quy tuyến tính thành một giá trị xác suất từ 0 đến 1.

Giá trị đầu ra dự đoán được xác định bằng cách so sánh với 1 ngưỡng.

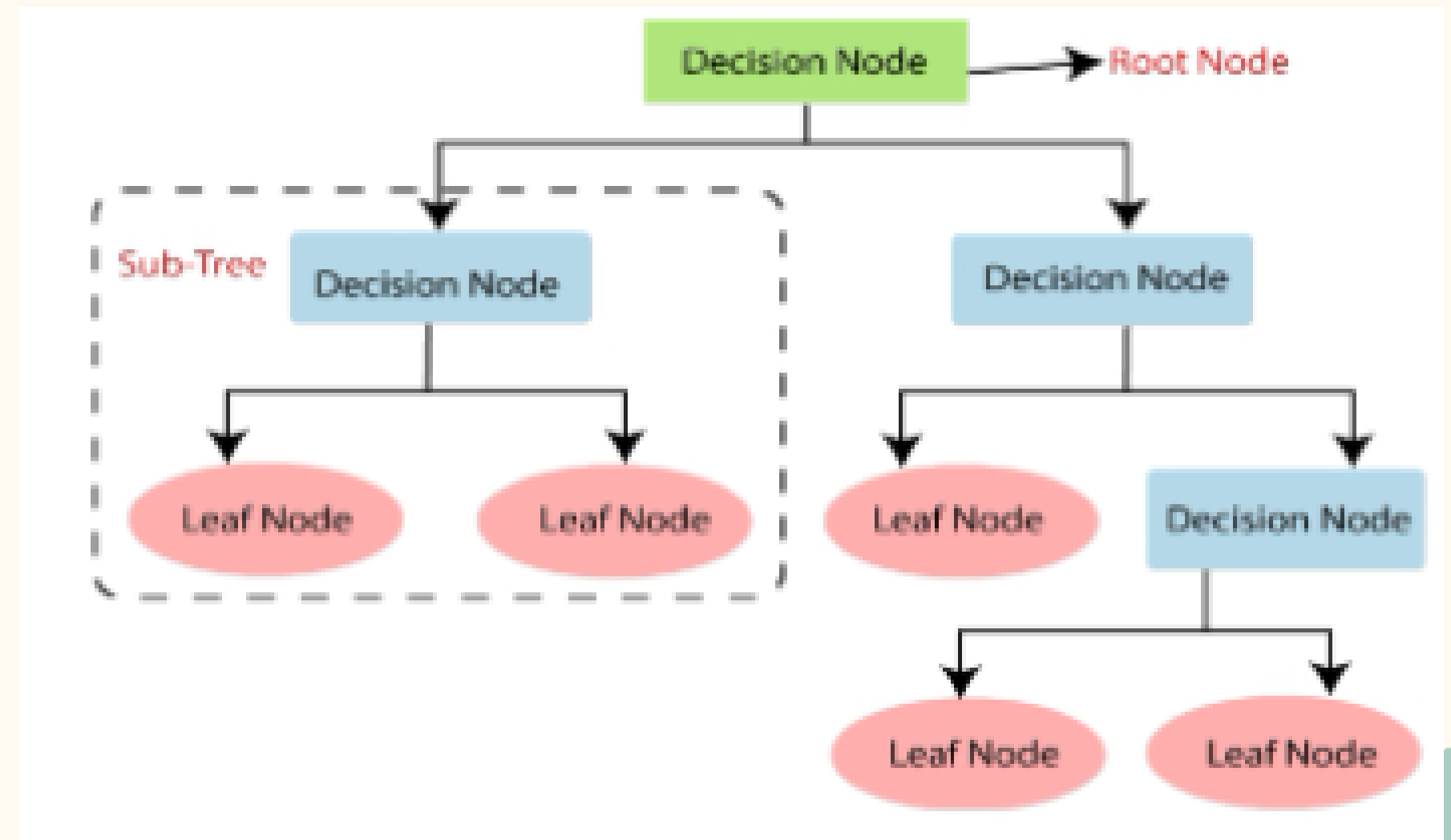


Decision Tree

Decision Tree là một phương pháp học máy phổ biến cho cả phân loại và hồi quy.

Cấu trúc của nó bao gồm nút gốc, nút quyết định và nút lá. Mỗi nút quyết định tương ứng với một thuộc tính của dữ liệu, và các nhánh từ nút đó tương ứng với các giá trị hoặc khoảng giá trị của thuộc tính đó. Nút lá chứa nhãn hoặc giá trị dự đoán.

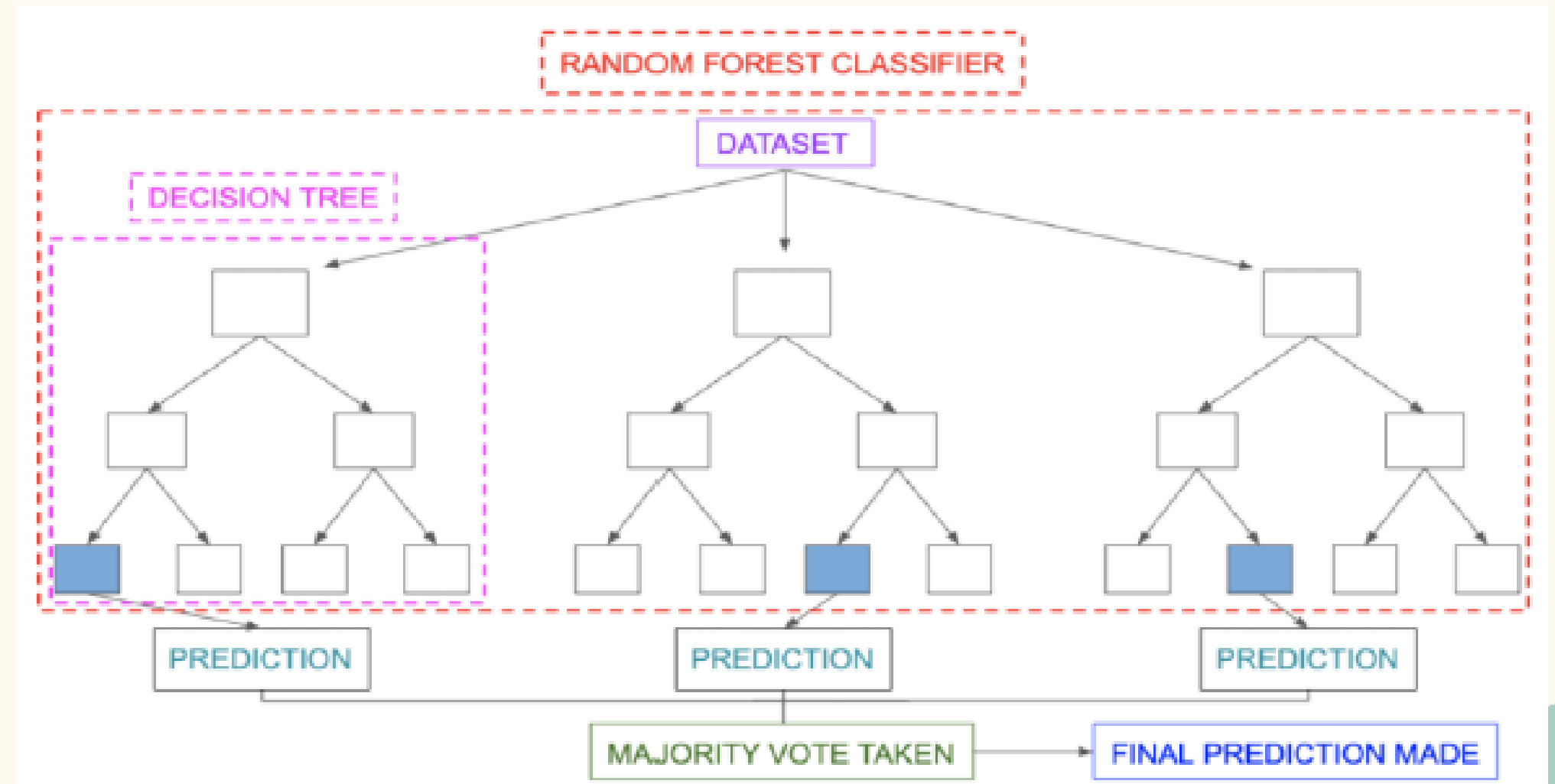
Việc chọn thuộc tính để chia dữ liệu tại mỗi nút sẽ thông qua các thuật toán như ID3, C4.5 và CART, dựa trên các thông số như Entropy, Information Gain (độ tăng thông tin) hay Gini Index để đo lường chất lượng của một phép chia.



Random Forest

Random Forest là một trong những kỹ thuật học máy tập hợp (ensemble learning) phổ biến nhất, thuộc nhóm phương pháp Bagging. Mô hình Random Forest là sự kết hợp của rất nhiều mô hình học Decision Tree để áp dụng vào các bài toán phân loại hoặc hồi quy.

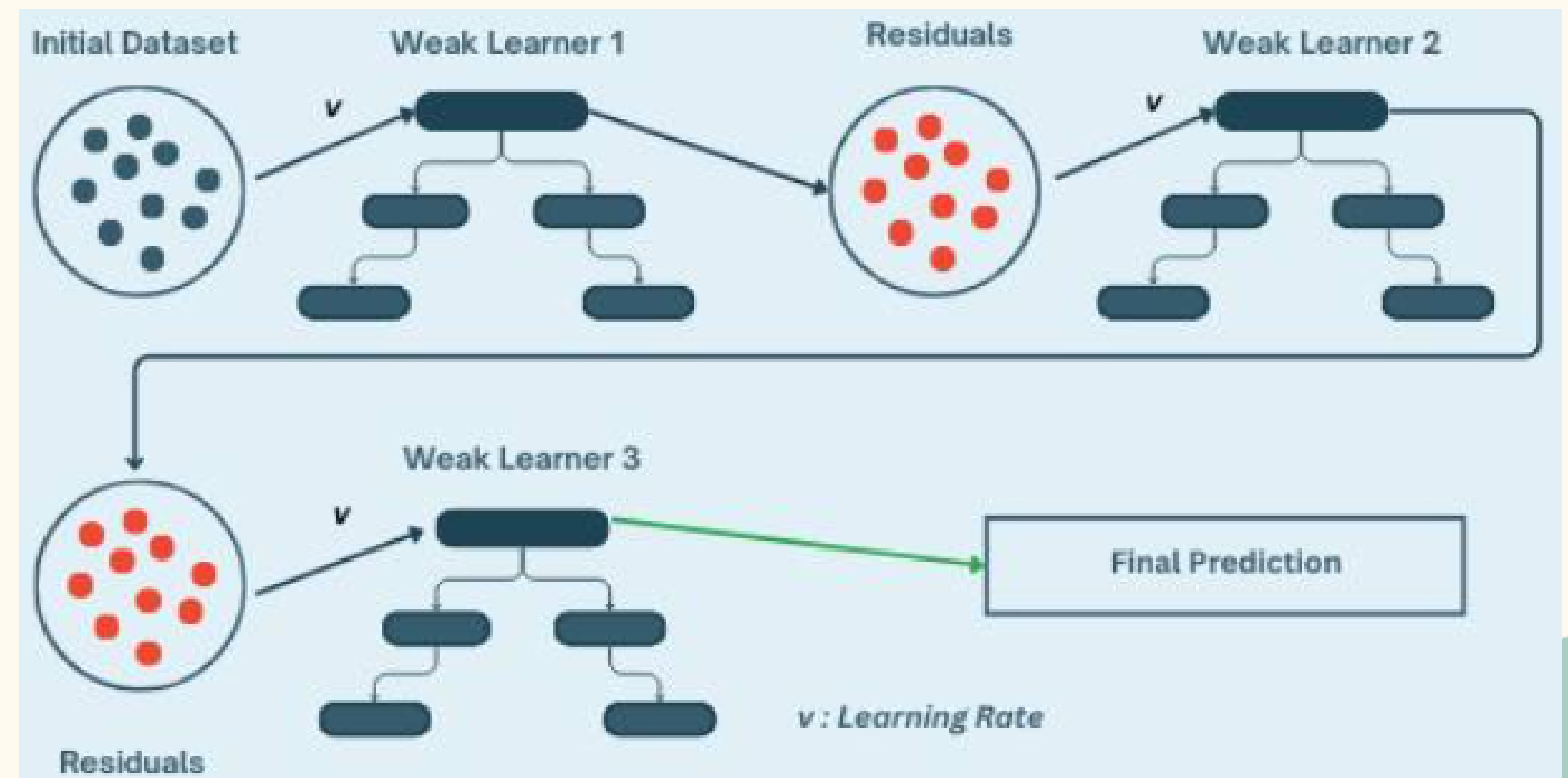
Hoạt động bằng cách tạo ra một tập hợp các cây quyết định trong quá trình huấn luyện, sau đó đưa ra dự đoán dựa trên kết quả đầu ra của mỗi cây quyết định. Việc lựa chọn kết quả dự đoán cuối cùng thường tuân theo nguyên tắc đa số.



Gradient Boosting

Gradient Boosting cũng là một trong những kỹ thuật học máy tập hợp nhưng thuộc nhóm phương pháp Boosting.

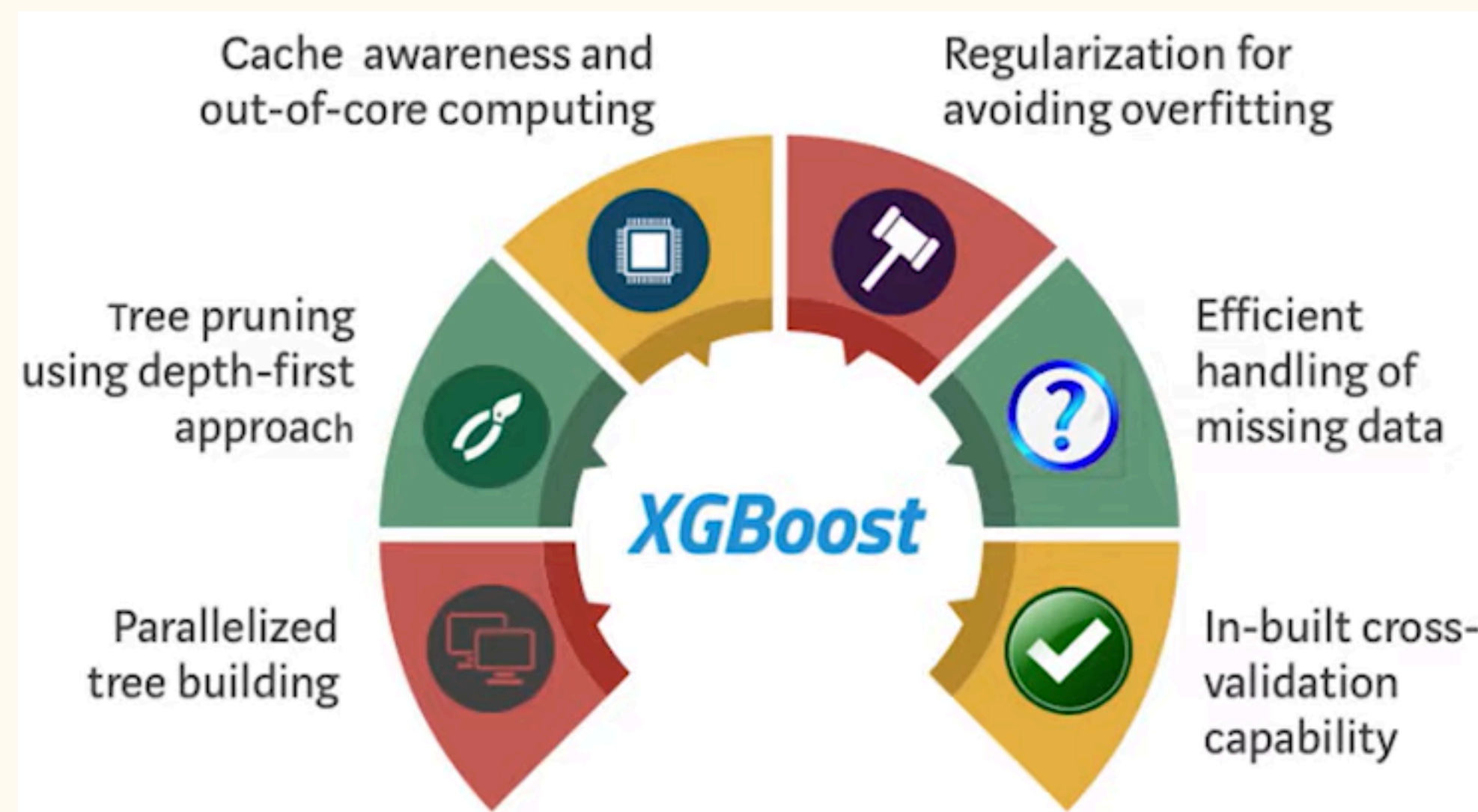
Việc huấn luyện mỗi mô hình mới (thường là một cây quyết định) dựa trên các residuals (sự chênh lệch giữa giá trị dự đoán và giá trị thực tế), từng mô hình mới sẽ cố gắng giảm thiểu sai số của mô hình trước đó. Quá trình này sử dụng thuật toán Gradient Descent để tối ưu hóa mô hình, điều chỉnh dự đoán theo hướng giảm thiểu sai số.



XGBoost

XGBoost (eXtreme Gradient Boosting) là phiên bản cải tiến của Gradient Boosting. Ưu điểm:

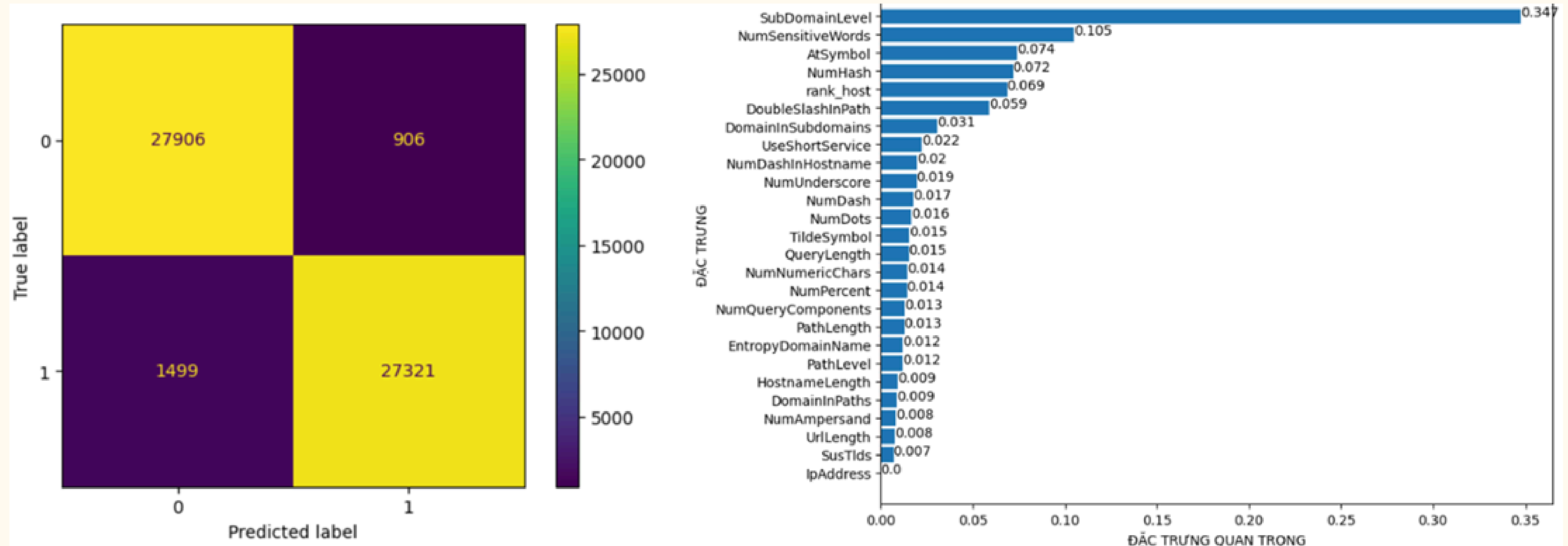
- Tốc độ xử lý có thể gấp 10 lần GBM với tính toán song song
- Áp dụng cơ chế Regularization nên hạn chế đáng kể hiện tượng overfitting
- Linh hoạt trong sử dụng hàm tối ưu
- Tự động xử lý missing value
- Tự động cắt tỉa cây (auto pruning): tự động bỏ qua những leaves, nodes không mang giá trị tích cực trong quá trình mở rộng tree.



Model Evaluation

Model	Train Accuracy	Test Accuracy	F1 Score	Recall	Precision
XGBoost	98.8%	95.8%	95.8%	94.8%	96.8%
Gradient Boosting	99.4%	94.6%	94.6%	93.6%	95.6%
Random Forest	98.5%	93.6%	93.5%	92.1%	94.8%
Decision Tree	96.6%	91.7%	91.6%	90.2%	92.9%
Logistic Regression	86.9%	86.8%	86.5%	84.5%	88.5%

XGBoost



Extension

Malicious URL Detector

Đường dẫn này có tỉ lệ 92.05% dẫn đến trang web an toàn

Đường dẫn này có tỉ lệ 71.70% dẫn đến trang web độc hại

Malicious URL Detector

Đường dẫn này có tỉ lệ 51.45% dẫn đến trang web an toàn

Đường dẫn này có tỉ lệ 99.33% dẫn đến trang web độc hại

Kết luận

Tính hiệu quả của phương pháp

Từ độ chính xác và thử nghiệm, có thể thấy tính hiệu quả của việc áp dụng các mô hình học máy để phát hiện URL độc hại

Thực tế

Tuy nhiên, việc chỉ sử dụng các đặc trưng từ vựng được trích xuất từ URL là không đủ để đối mặt với tất cả tình huống thực tế hiện nay

Hướng phát triển thêm

- Sử dụng các bộ dữ liệu lớn hơn và mới hơn.
- Nghiên cứu và chọn ra thêm những đặc trưng khác từ URL để tăng cường độ chính xác.
- Tối ưu hoá thêm các tham số cho mô hình.
- Thử nghiệm và triển khai các mô hình học máy tiên tiến khác.



Xin cảm ơn!