PHÂN LOẠI MÃ ĐỘC PE SỬ DỤNG MACHINE LEARNING

Lê Xuân Hiếu

[1] Trường ĐH Công nghệ Thông tin
ĐHQG TPHCM

[2] University of Information Technology
HCMC, Vietnam

What ?

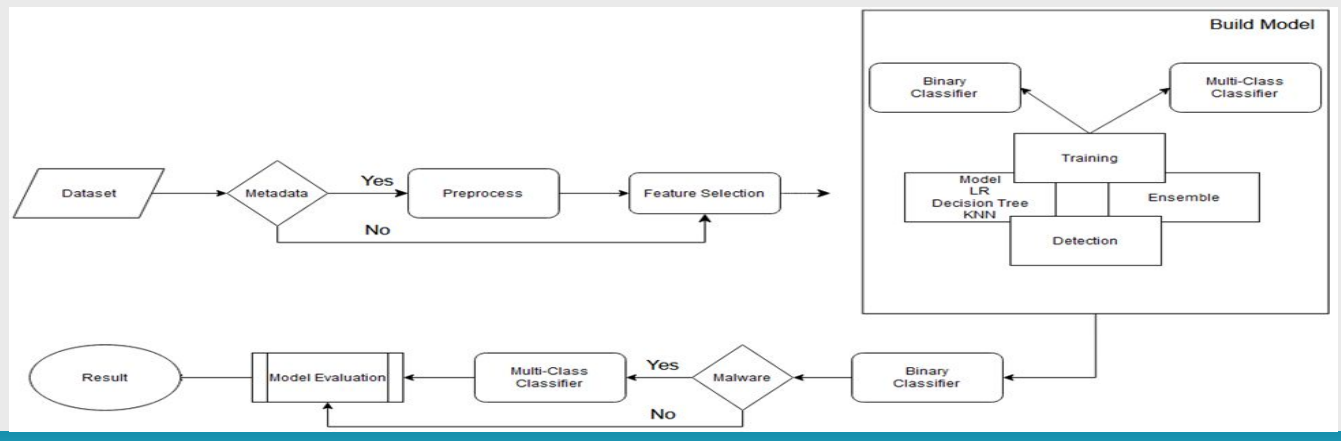
Giới thiệu về việc phân loại mã độc sử dụng machine learning

- Model machine learning sử dụng trong việc phát hiện và phân loại mã độc (malware)
- Hướng phát triển phân loại malware tiềm năng trong tương lai
- Tối ưu hơn khi sử dụng phần mềm antivirus truyền thống

Why ?

- Số lượng mối đe dọa từ mã độc (malware) ngày càng gia tăng khiến các phương pháp phát hiện dựa theo signatures truyền thống không đủ khả năng để chống lại các cuộc tấn công mới và phức tạp.
- Tập trung vào hiệu suất phát hiện các mô hình, so sánh đánh giá tham số metadata (timestamp,family)

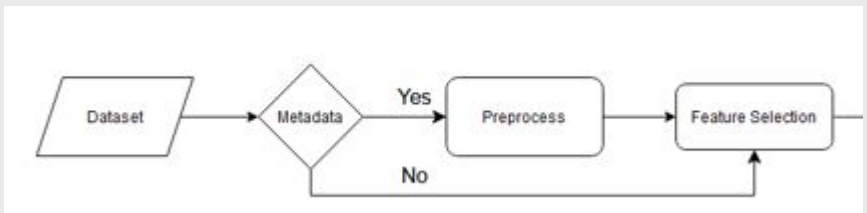
Overview



Description

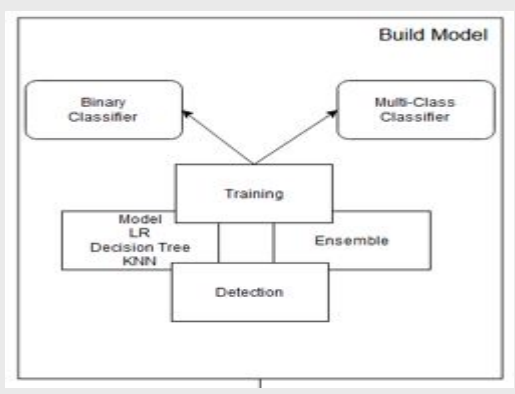
1. Dataset Collection and Feature Selection

- Sử dụng Dataset BODMAS Yang Limin chứa các tập tin malware và benign (lành tính)
- Tạo thêm file metadata có timesptamp, family
- Chuyển các timestamp thành giá trị số để dễ dàng tính toán và huấn luyện
- Trích xuất đặc trưng dựa trên dự án LIEF, kỹ thuật giống dự án Ember



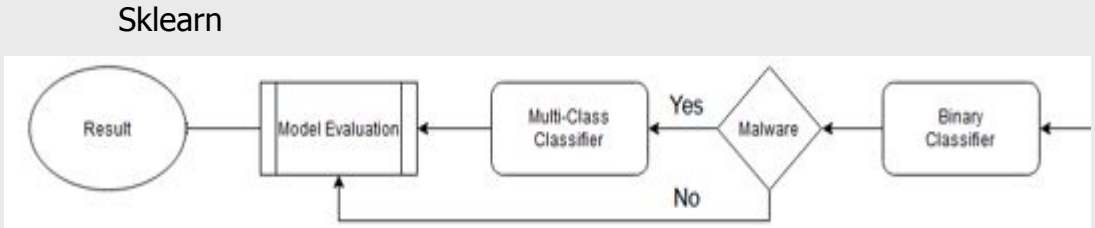
2. Build Model

- Dataset: 80% sử dụng cho huấn luyện, 20% kiểm thử
- Sử dụng model Logistic Regeression, K_Nearest Neighbor, Decision Tree và Ensemble (kết hợp 3 mô hình)
- Phân loại nhị phân và phân loại đa nhãn (có và không sử dụng metadata)



3. Model Evalation

- Sử dụng các chỉ số đánh giá Accuracy score, precision, recall
- Sử dụng hàm đánh giá của thư viện Sklearn



4. Kết quả mong đợi

- Phát hiện và phân biệt đợc tệp tin malware và benign
- Hiệu suất mô hình machine learning tốt nhất trong các mô hình trên

NII

Lê Xuân Hiếu– Trường Đại học Công nghệ Thông tin, ĐHQG TP HCM
TEL : 0706321991 Email : hieulx.18@grad.uit.edu.vn