Demo cho việc "Phát hiện website phishing bằng Học máy (Machine Learning)".

**1. Về Dataset (Dữ liệu):**

* Demo này sử dụng file urldata.csv được giả định là chứa cả domain phishing và legitimate. File này cần có sự cân bằng giữa hai loại domain để mô hình học được cách phân biệt chúng hiệu quả.
* Sử dụng các file riêng biệt phishing.csv và legitimate.csv. Tuy nhiên, cần đảm bảo tổng hợp cả hai file này để tạo thành một dataset đủ lớn và cân bằng cho việc huấn luyện.
* Nguồn dữ liệu từ Phishtank

**2. Quá trình Training (Huấn luyện):**

* **Bước 1: Đọc dữ liệu**
  + Thư viện pandas được sử dụng để đọc dữ liệu từ file urldata.csv vào dataframe data.
* **Bước 2: Mã hóa biến mục tiêu**
  + Thư viện sklearn.preprocessing cung cấp công cụ LabelEncoder để chuyển đổi các nhãn "Label" (ví dụ: phishing, legitimate) thành dạng số (ví dụ: phishing = 0, legitimate = 1). Điều này giúp thuật toán học máy xử lý dữ liệu dễ dàng hơn.
* **Bước 3: Chia tách dữ liệu**
  + Thư viện sklearn.model\_selection cung cấp hàm train\_test\_split để chia tách dữ liệu thành hai phần:
    - Dữ liệu huấn luyện (X\_train, y\_train): chiếm 70% dung lượng, dùng để huấn luyện mô hình.
    - Dữ liệu kiểm tra (X\_test, y\_test): chiếm 30% dung lượng, dùng để đánh giá hiệu suất của mô hình sau khi huấn luyện.
* **Bước 4: Huấn luyện mô hình**
  + Thư viện sklearn.ensemble cung cấp thuật toán RandomForestClassifier để xây dựng mô hình. Mô hình này được huấn luyện dựa trên dữ liệu huấn luyện (X\_train, y\_train).
* **Bước 5: Đánh giá mô hình**
  + Sau khi huấn luyện, mô hình được sử dụng để dự đoán nhãn cho dữ liệu kiểm tra (X\_test). Kết quả dự đoán (y\_pred) được so sánh với nhãn thật (y\_test) để tính toán các指標 (chỉ tiêu) đánh giá hiệu suất như: độ chính xác (accuracy), báo cáo phân loại (classification report).

**3. Kết luận:**

* Dựa trên kết quả đánh giá (độ chính xác, báo cáo phân loại), ta có thể đánh giá được mức độ hiệu quả của mô hình trong việc phân biệt giữa domain phishing và legitimate.
* Ngoài ra, demo còn tính toán tầm quan trọng của từng đặc trưng trong việc dự đoán. Điều này giúp hiểu đặc trưng nào ảnh hưởng mạnh nhất đến kết quả dự đoán của mô hình.

TỪ KHÚC NÀY ĐỔ ĐI KHÔNG ĐỌC NỮA, PHẦN DƯỚI LÀ ĐỂ CÔ HỎI THÌ DỰA TRÊN ĐỂ TRẢ LỜI!!!!!

**Lưu ý:**

* Hiệu quả của mô hình phụ thuộc rất nhiều vào chất lượng và độ cân bằng của dữ liệu.
* Cần cập nhật dữ liệu thường xuyên để mô hình thích ứng với các kỹ thuật phishing mới.
* Trong thực tế, có thể kết hợp thêm các thuật toán học máy khác để nâng cao hiệu quả tổng thể.

**Tóm lại:** Đoạn code này cung cấp một ví dụ cơ bản về việc ứng dụng Học máy để phát hiện website phishing. Để triển khai thực tế, cần chú trọng đến việc xây dựng dataset chất lượng, huấn luyện và tối ưu hóa mô hình phù hợp.

**Vai trò của Học máy (Machine Learning) trong demo phát hiện website phishing**

**1. Huấn luyện mô hình:**

* Thuật toán RandomForestClassifier (Học máy) được sử dụng để huấn luyện mô hình dựa trên dữ liệu urldata.csv. Dữ liệu này cần chứa cả domain phishing và legitimate để mô hình học được cách phân biệt chúng.
* Quá trình huấn luyện bao gồm:
  + Chia tách dữ liệu thành tập huấn luyện và tập kiểm tra.
  + Áp dụng thuật toán RandomForestClassifier để học hỏi mối liên hệ giữa các đặc trưng (từ các cột trong urldata.csv) và nhãn "Label" (phishing hoặc legitimate).
  + Điều chỉnh tham số của mô hình (ví dụ: số lượng cây quyết định) để đạt hiệu quả tối ưu.

**2. Dự đoán tính hợp pháp của domain:**

* Sau khi huấn luyện, mô hình có thể dự đoán nhãn "Label" (phishing hoặc legitimate) cho domain mới.
* Quá trình dự đoán bao gồm:
  + Chuẩn bị dữ liệu đầu vào cho domain mới (bao gồm các đặc trưng tương tự như trong urldata.csv).
  + Sử dụng mô hình đã huấn luyện để dự đoán nhãn "Label" cho domain mới.
* Kết quả dự đoán thể hiện khả năng domain mới là phishing hay legitimate.

**3. Đánh giá hiệu quả:**

* Hiệu quả của mô hình được đánh giá dựa trên khả năng dự đoán chính xác domain phishing và legitimate trong tập kiểm tra.
* Các chỉ tiêu đánh giá thường dùng bao gồm:
  + Độ chính xác (Accuracy): Tỷ lệ dự đoán đúng trên toàn bộ dữ liệu kiểm tra.
  + Báo cáo phân loại (Classification Report): Cung cấp thông tin chi tiết về hiệu suất dự đoán cho từng lớp (phishing và legitimate), bao gồm độ chính xác, độ nhạy (Recall), độ đặc hiệu (Precision), F1-Score.

**4. Phân tích tầm quan trọng của đặc trưng:**

* Thuật toán RandomForestClassifier có khả năng tính toán tầm quan trọng của từng đặc trưng trong việc dự đoán.
* Thông tin này giúp hiểu rõ đặc trưng nào ảnh hưởng mạnh nhất đến kết quả dự đoán, từ đó có thể:
  + Chọn lọc các đặc trưng quan trọng để xây dựng mô hình hiệu quả hơn.
  + Tập trung vào việc thu thập và xử lý dữ liệu cho các đặc trưng quan trọng.

**Kết luận:**

Học máy đóng vai trò quan trọng trong việc phát hiện website phishing bằng cách:

* **Huấn luyện mô hình:** Học hỏi mối liên hệ giữa các đặc trưng và nhãn "Label" để dự đoán domain mới là phishing hay legitimate.
* **Dự đoán tính hợp pháp:** Cung cấp kết quả dự đoán cho domain mới giúp người dùng đưa ra quyết định truy cập an toàn.
* **Đánh giá hiệu quả:** Giúp đánh giá mức độ chính xác và hiệu quả của mô hình trong việc phân biệt domain phishing và legitimate.
* **Phân tích tầm quan trọng của đặc trưng:** Hiểu rõ đặc trưng nào ảnh hưởng mạnh nhất đến kết quả dự đoán để tối ưu hóa mô hình và thu thập dữ liệu hiệu quả.

**Lưu ý:**

* Hiệu quả của mô hình phụ thuộc vào chất lượng và độ cân bằng của dữ liệu.
* Cần cập nhật dữ liệu thường xuyên để mô hình thích ứng với các kỹ thuật phishing mới.
* Có thể kết hợp thêm các thuật toán học máy khác để nâng cao hiệu quả tổng thể.