

PHÂN LOẠI THƯ RÁC SỬ DỤNG RANDOM FOREST

Lê Anh Cường - 230202002

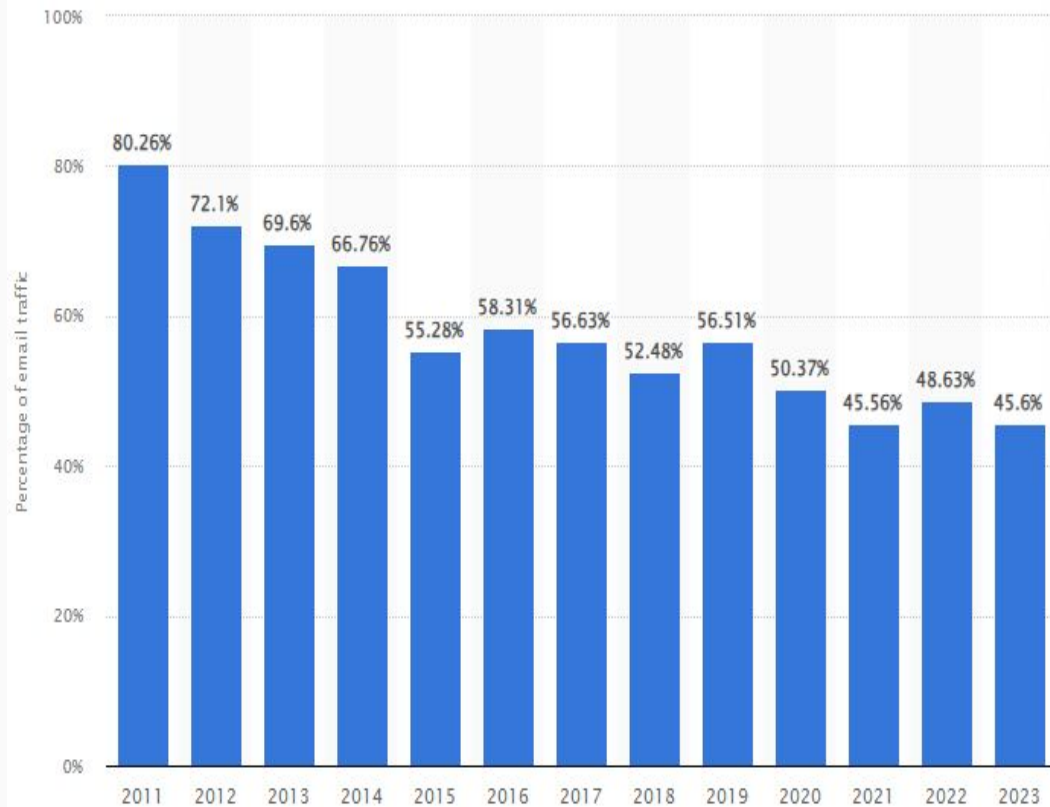
Tóm tắt

- Lớp: CS2205.CH181
- Link Github: [mrbordean/CS2205.CH181:230202002-LeAnhCuong \(github.com\)](https://github.com/mrbordean/CS2205.CH181:230202002-LeAnhCuong)
- Link YouTube video:
<https://youtu.be/t7a6kTIFAQc>
- Họ và Tên: Lê Anh Cường



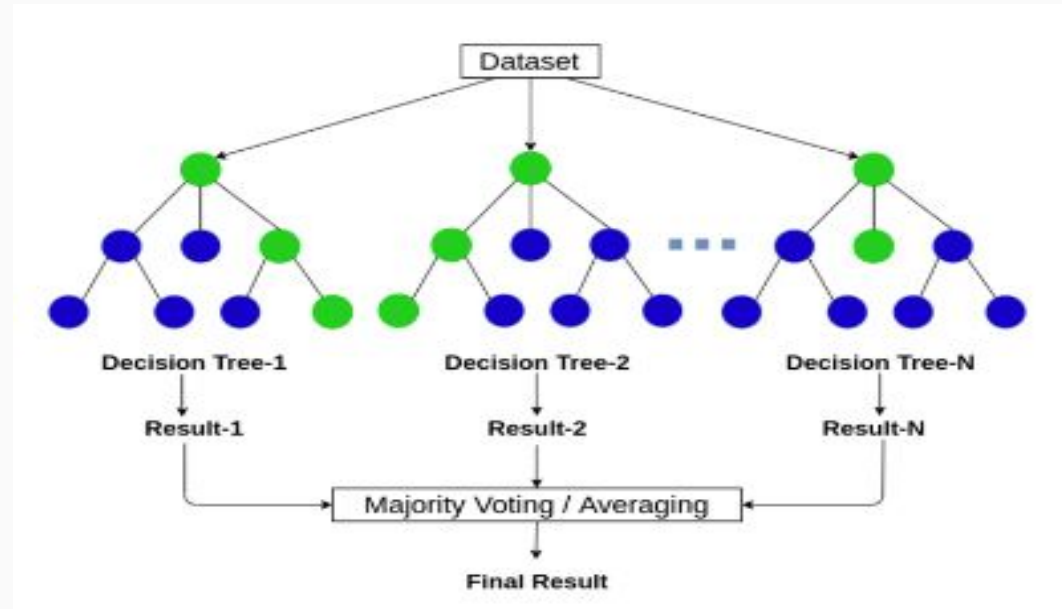
Giới thiệu

- Email Spam là các Email không mong muốn, được gửi hàng loạt với mục đích quảng cáo, lừa đảo hoặc phát tán mã độc.
- Có tới 45.6% lưu lượng Email toàn cầu vào năm 2023 được đánh giá là Spam.[1]
- Email Spam có 4 đặc điểm và 4 mục tiêu chính.



Giới thiệu

- Random Forest là thuật toán sử dụng nhiều cây quyết định (decision trees) để cải thiện độ chính xác và độ ổn định của mô hình dự đoán.
- Có những ưu điểm hơn so với các thuật toán khác khi áp dụng vào chủ đề lọc Email Spam.




Mục tiêu

- Tính ứng dụng: Xây dựng một mô hình phân loại Email bằng thuật toán Random Forest.
- Kết quả: Huấn luyện mô hình để có thể cho ra kết quả tốt nhất từ tập dataset Spam Email của Kaggle.

Spam Email Dataset

Classify Spam and ham data



Data Card Code (0) Discussion (0) Suggestions (0)

About Dataset

This is a dataset collected at Hewlett-Packard Labs by Mark Hopkins, Erik Reeber, George Forman, and Jaap Suermondt and shared with the [UCI Machine Learning Repository](#). The dataset classifies 4601 e-mails as spam or non-spam, with additional variables indicating the frequency of certain words and characters in the e-mail.

Data Dictionary

spam.csv

Usability

9.41

License

Other (specified in description)

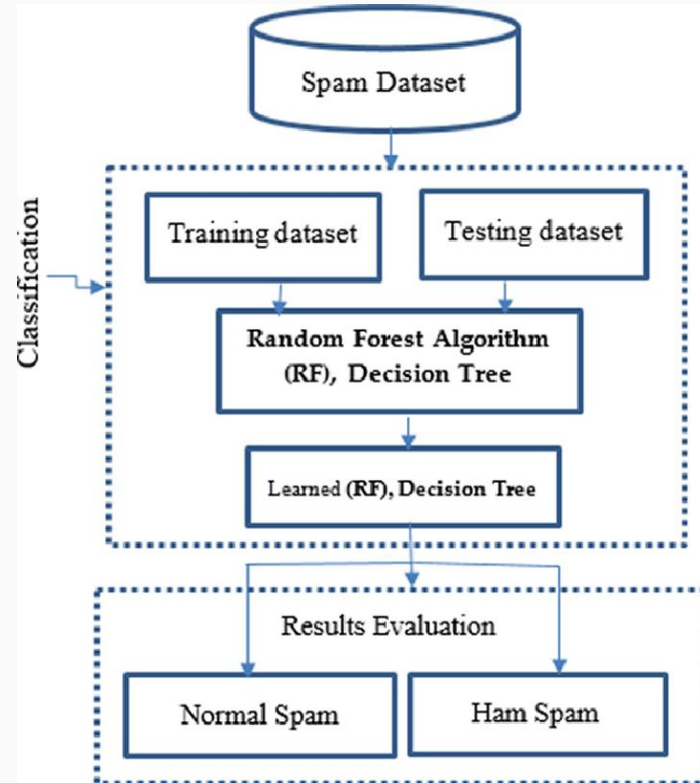
Expected update frequency

Never

Tags

Email and Messaging

Nội dung và Phương pháp



Nội dung và Phương pháp

1. Thu thập và chuẩn bị dữ liệu:

- a. Thu thập dữ liệu: tập hợp nguồn dữ liệu Email bao gồm hợp lệ và Spam. Có thể lấy từ nguồn cá nhân, cơ sở dữ liệu công khai và các tập dữ liệu trong các cuộc thi học máy.
- b. Chuẩn bị dữ liệu: làm sạch dữ liệu bằng cách loại bỏ các ký tự đặc biệt, dấu câu và các phần không cần thiết khác. Biến đổi các Email thành dạng có thể xử lý được.

2. Trích xuất đặc trưng:

- a. Vector hóa: chuyển đổi văn bản Email thành các vector đặc trưng. Có thể sử dụng các kỹ thuật sau: Bag of Words (BoW), Embedding.

Nội dung và Phương pháp

b. Lựa chọn đặc trưng: lựa chọn các đặc trưng quan trọng để giảm kích thước của vector đặc trưng, giúp mô hình hoạt động hiệu quả hơn.

3. Chia dữ liệu thành tập huấn luyện và tập kiểm tra:

- a. Chia tập dữ liệu đã được chuẩn bị thành hai phần, một phần để huấn luyện mô hình và một phần để kiểm tra hiệu quả của mô hình.

4. Xây dựng và đánh giá mô hình Random Forest:

- b. Xây dựng: mô hình sẽ được sử dụng thuật toán Random Forest. Huấn luyện mô hình trên tập huấn luyện.
- c. Kiểm tra mô hình: sử dụng tập kiểm tra để đánh giá hiệu suất của mô hình dựa trên các chỉ số như: độ chính xác, độ nhạy, độ đặc hiệu và điểm F1.

Nội dung và Phương pháp

PHƯƠNG PHÁP THỰC HIỆN

1. Sử dụng tập dataset đến từ Kaggle.
2. Sử dụng thuật toán Random Forest để xây dựng mô hình.
3. Huấn luyện mô hình lọc Email Spam và đánh giá kết quả của mô hình.

Kết quả dự kiến

- Mô hình cho ra kết quả khả quan với những chỉ báo như độ chính xác cao, khả năng xử lý dữ liệu mất cân bằng tốt và thời gian huấn luyện hợp lý. Có khả năng phân loại chính xác các Email thành thư rác hay không.
- Có thể áp dụng cho người dùng thực tế.

Tài liệu tham khảo

- [1]. [Spam e-mail traffic share 2023 | Statista](#)
- [2]. Intelligent Security Schema for SMS Spam Message Based on Machine Learning Algorithms - Scientific Figure on ResearchGate. Available from: https://www.researchgate.net/figure/Processes-of-SMS-spam-classification_fig2_354076063 [accessed 31 May, 2024]
- [3]. [\(PDF\) Classification of Spam Messages using Random Forest Algorithm \(researchgate.net\)](#)
- [4]. [Basic Comparison Between RandomForest, SVM, and XGBoost | by Nattapoj Apichardsilkij | Medium](#)
- [5]. [Spam Email Dataset \(kaggle.com\)](#)