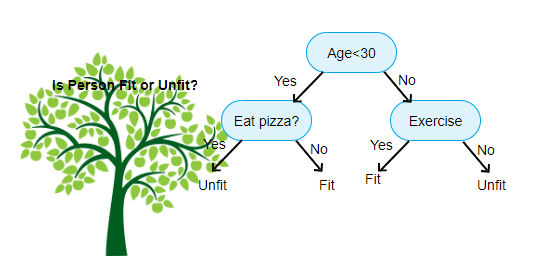
# **I. Random Forest:**

## **Tổng quan:**

### **Giới thiệu về Decision Tree (Cây quyết định):**

Trước tiên, để hiểu Random Forest là gì thì ta cần hiểu về khái niệm Decision Tree. Decision Tree là một thuật toán học máy thuộc nhóm supervised learning (học có giám sát), được sử dụng cho cả bài toán phân loại (classification) và hồi quy (regression). Mô hình này biểu diễn các quyết định và kết quả của chúng dưới dạng một cấu trúc cây, trong đó mỗi nút (node) đại diện cho một thuộc tính (feature), mỗi nhánh (branch) là một kết quả của phép kiểm tra, và mỗi lá (leaf) biểu thị một nhãn hoặc giá trị dự đoán.



### **Giới thiệu về thuật toán Random Forest:**

Random Forest là một thuật toán học máy thuộc nhóm ensemble learning (học tập tổng hợp), được phát triển dựa trên ý tưởng kết hợp nhiều cây quyết định (decision trees) để cải thiện độ chính xác và độ ổn định của mô hình. Thay vì sử dụng một cây quyết định duy nhất, Random Forest tạo nhiều cây quyết định độc lập và lấy kết quả theo số đông (đối với bài toán phân loại) hoặc trung bình (đối với bài toán hồi quy).

### 1.1.3. **Một số ứng dụng của thuật toán Random Forest:**

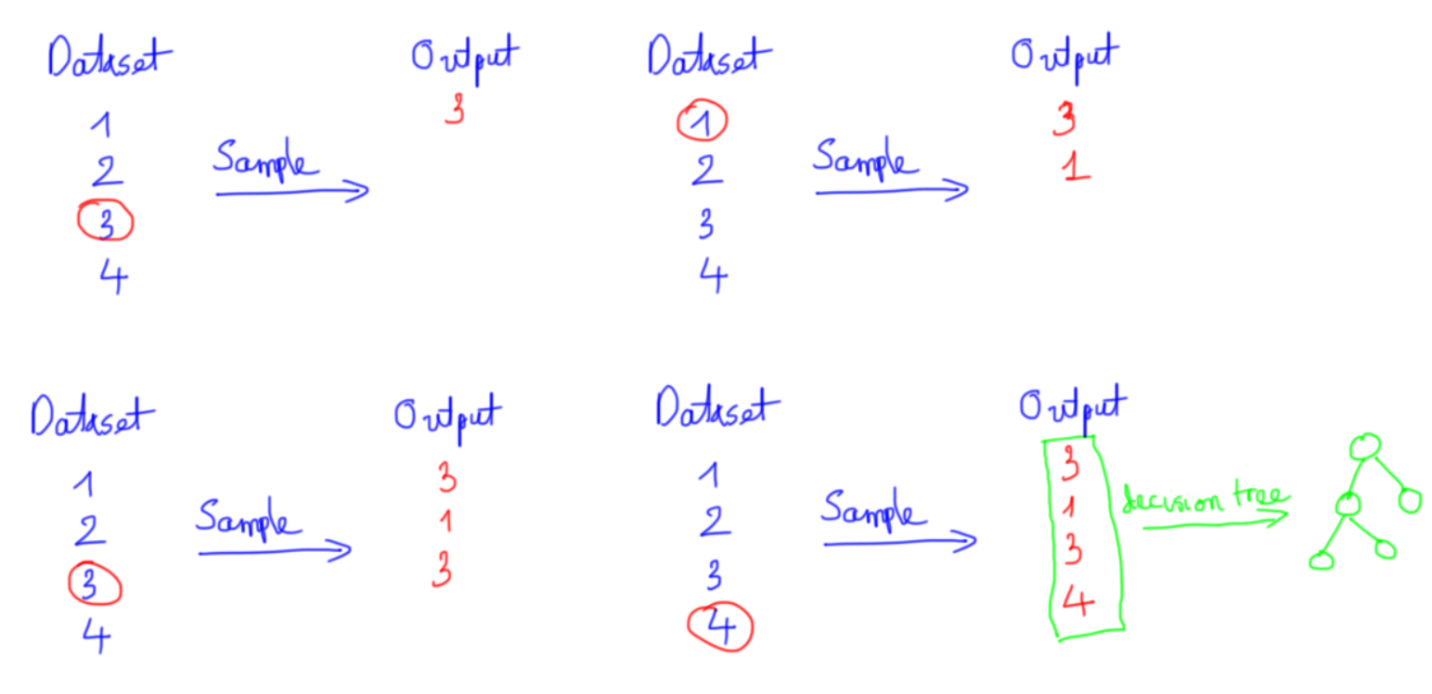
* Phân loại hình ảnh: Nhận diện đối tượng hoặc phân loại cảm xúc trong xử lý ảnh (ví dụ: phát hiện ung thư qua hình ảnh y tế).
* Dự đoán tài chính: Dự báo giá cổ phiếu, đánh giá rủi ro tín dụng, hoặc phát hiện gian lận tài chính.
* Y học: Chẩn đoán bệnh dựa trên dữ liệu bệnh nhân, phân tích dữ liệu gen để xác định nguy cơ bệnh.
* Marketing: Phân khúc khách hàng, dự đoán hành vi mua sắm, hoặc tối ưu hóa chiến dịch quảng cáo.
* Phân loại email: Xác định email là spam hay không spam
* Phân tích cảm xúc: Phân loại đánh giá sản phẩm hoặc bài viết trên mạng là tích cực, tiêu cực
* Dự đoán ô nhiễm: Dự đoán mức độ ô nhiễm không khí hoặc nước dựa trên các yếu tố môi trường

## **Cơ chế hoạt động của thuật toán Random Forest và hướng dẫn sử dụng:**

### **Cơ chế hoạt động:**

Từ bộ dữ liệu có n dữ liệu (sample) và mỗi dữ liệu có d thuộc tính (feature), việc xây dựng các cây quyết định bao gồm các bước như sau:

Bước 1: Lấy ngẫu nhiên n dữ liệu từ bộ dữ liệu với kĩ thuật Bootstrapping, hay còn gọi là random sampling with replacement. Tức khi mình sample được 1 dữ liệu thì mình không bỏ dữ liệu đấy ra mà vẫn giữ lại trong tập dữ liệu ban đầu, rồi tiếp tục sample cho tới khi sample đủ n dữ liệu. Khi dùng kĩ thuật này thì tập n dữ liệu mới của mình có thể có những dữ liệu bị trùng nhau.



Bước 2: Chọn ngẫu nhiên k thuộc tính từ d (k < d) được bộ dữ liệu mới bao gồm n dữ liệu và mỗi dữ liệu có k thuộc tính.

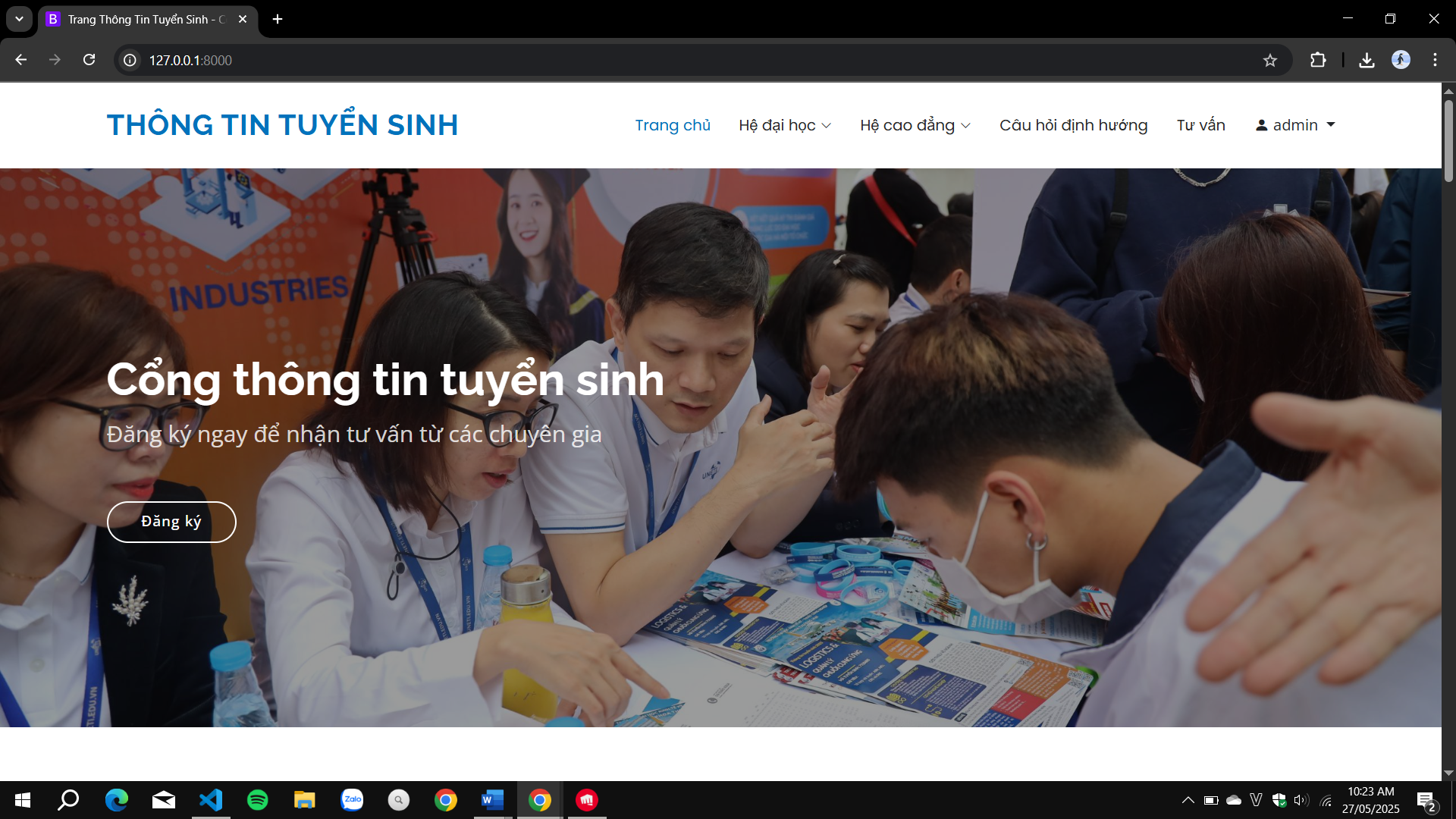
Bước 3: Dùng thuật toán Decision Tree để xây dựng cây quyết định với bộ dữ liệu ở bước 2.

Do quá trính xây dựng mỗi cây quyết định đều có yếu tố ngẫu nhiên (random) nên kết quả là các cây quyết định trong thuật toán Random Forest có thể khác nhau.

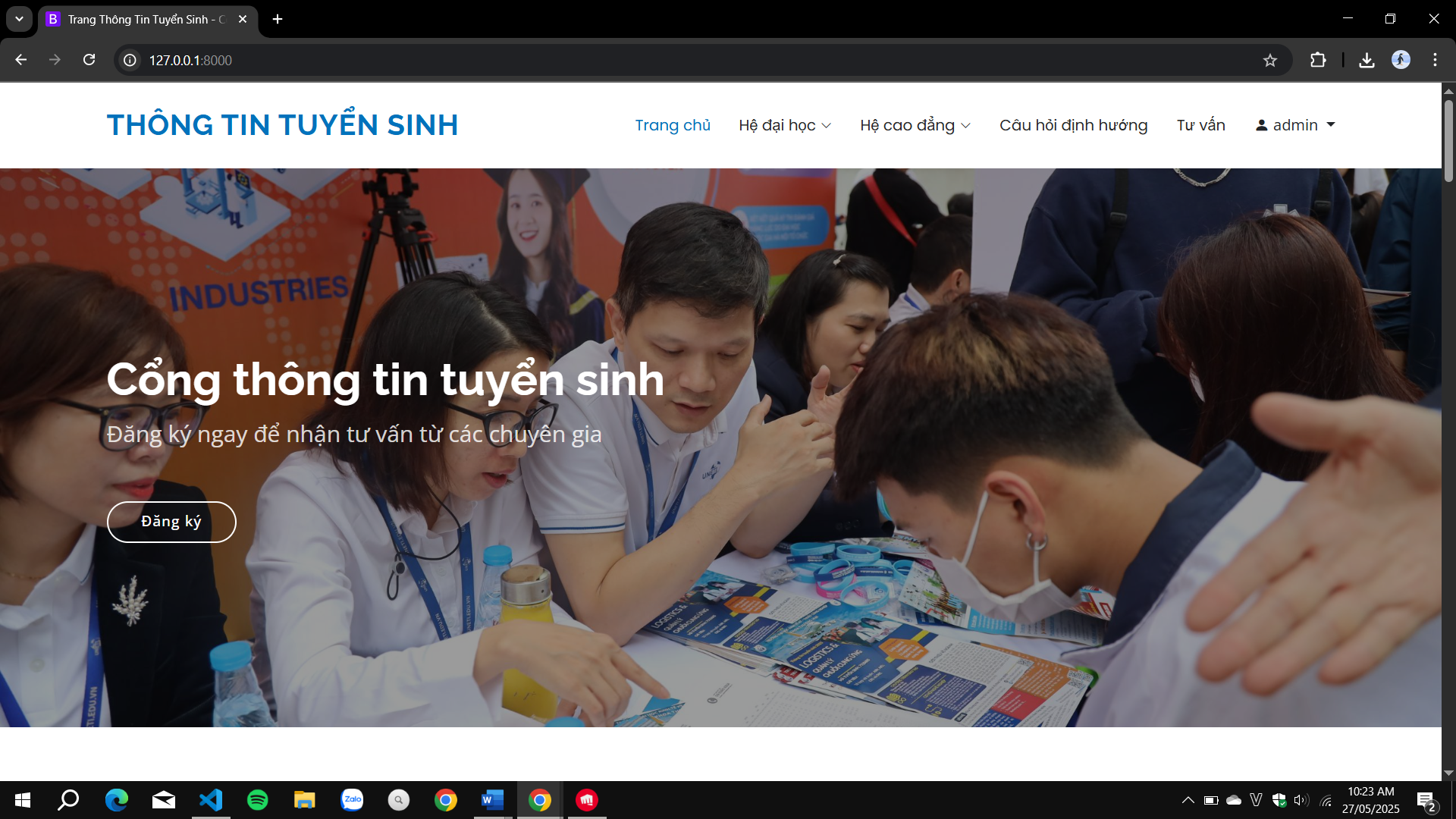
Thuật toán Random Forest sẽ bao gồm nhiều cây quyết định, mỗi cây được xây dựng trên tập dữ liệu khác nhau và dùng tập thuộc tính khác nhau. Sau đó kết quả dự đoán của thuật toán Random Forest sẽ được tổng hợp từ các cây quyết định.

### **Hướng dẫn sử dụng:**

Bước 1: Truy cập vào trang web tư vấn tuyển sinh:



Bước 2: Vào phần câu hỏi định hướng:



Bước 3: Trả lời hết các câu hỏi và hệ thống sẽ hiển thị gợi ý ngành học dựa trên câu trả lời:

