TRƯỜNG ĐẠI HỌC ĐIỆN LỰC

**KHOA CÔNG NGHỆ THÔNG TIN**



**BÁO CÁO CHUYÊN ĐỀ HỌC PHẦN**

**KHAI PHÁ DỮ LIỆU**

**ĐỀ TÀI:**

**XÂY DỰNG ỨNG DỤNG DỰ ĐOÁN TUỔI THỌ CỦA CÁC NƯỚC SỬ DỤNG THUẬT TOÁN RANDOM FOREST**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  | | --- | --- | | **Sinh viên thực hiện** | **: TRỊNH NHẬT HUY**  **NGUYỄN TUẤN ANH**  **TRẦN HẢI LONG** | | **Giảng viên hướng dẫn** | **: NGUYỄN THỊ THANH TÂN** | | | **Ngành** | **: CÔNG NGHỆ THÔNG TIN** | | | **Chuyên ngành** | **: CÔNG NGHỆ PHẦN MỀM** | | | **Lớp** | **: D13CNPM4** | | | **Khóa** | **: 2018-2023** | | |  |
|  |  |

***Hà Nội, tháng 6 năm 2021***

**PHIẾU CHẤM ĐIỂM**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **STT** | **Họ và tên sinh viên** | **Nội dung thực hiện** | **Điểm** | **Chữ ký** |
| 1 | TRỊNH NHẬT HUY |  |  |  |
| 2 | NGUYỄN TUẤN ANH |  |  |  |
| 3 | TRẦN HẢI LONG |  |  |  |

|  |  |  |
| --- | --- | --- |
| **Họ và tên giảng viên** | **Chữ ký** | **Ghi chú** |
| Giảng viên chấm 1: |  |  |
| Giảng viên chấm 2: |  |  |

**MỤC LỤC**

[LỜI MỞ ĐẦU 1](#_Toc76053688)

[CHƯƠNG 1. TỔNG QUAN VỀ KHAI PHÁ DỮ LIỆU 2](#_Toc76053689)

[**1.1. Khái niệm về khai phá dữ liệu** 2](#_Toc76053690)

[**1.1.1. Khai phá dữ liệu là gì?** 2](#_Toc76053691)

[**1.1.2. Các bước trong khai phá dữ liệu** 2](#_Toc76053692)

[**1.1.3. Ứng dụng** 3](#_Toc76053693)

[**1.2. Các phương pháp khai phá dữ liệu** 3](#_Toc76053694)

[**1.2.1. Phân lớp dữ liệu** 3](#_Toc76053695)

[**1.2.2. Phân cụm** 4](#_Toc76053696)

[**1.2.3. Luật kết hợp** 4](#_Toc76053697)

[**1.2.4. Hồi quy tuyến tính** 4](#_Toc76053698)

[**1.3. Bài toán phân lớp** 5](#_Toc76053699)

[**1.3.1. Phân lớp nhị phân** 5](#_Toc76053700)

[**1.3.2. Phân lớp nhiều lớp** 5](#_Toc76053701)

[**1.3.3. Quá trình phân lớp dữ liệu** 5](#_Toc76053702)

[CHƯƠNG 2 : THUẬT TOÁN RANDOM FOREST 7](#_Toc76053703)

[**2.1.** **Random Forest là gì?** 7](#_Toc76053704)

[**2.2.** **Giới thiệu về thuật toán Random Forest** 7](#_Toc76053705)

[**2.3.** **Xây dựng thuật toán Random Forest** 8](#_Toc76053706)

[CHƯƠNG 3: ỨNG DỤNG CỦA THUẬT TOÁN VÀO CHƯƠNG TRÌNH 10](#_Toc76053707)

[**3.1.** **Đọc dữ liệu** 10](#_Toc76053708)

[**3.2.** **Khai phá dữ liệu** 11](#_Toc76053709)

[**3.3.** **Tiền xử lý dữ liệu** 11](#_Toc76053710)

[**3.4.** **Chia dữ liệu thành x và y** 12](#_Toc76053711)

[**3.5.** **Tạo mô hình** 12](#_Toc76053712)

[**3.6.** **Xử lý thuật toán và đưa ra dự đoán** 13](#_Toc76053713)

[CHƯƠNG 4: KẾT LUẬN 15](#_Toc76053714)

[TÀI LIỆU THAM KHẢO 16](#_Toc76053715)

# LỜI MỞ ĐẦU

Trong thời đại ngày nay, yếu tố quyết định thành công trong mọi lĩnh vực luôn gắn liền với việc nắm bắt, thống kê và khai thác thông tin hiệu quả. Dữ liệu ngày càng lớn nên việc tìm ra những thông tin tiềm ẩn trong chúng càng khó khăn hơn. Khai phá tri thức là một lĩnh vực nghiên cứu mới, mở ra  một  thời  kỳ trong  việc tìm ra thông tin hữu ích. Nhiệm vụ cơ bản của lĩnh vực này là khai phá tri thức trong cơ sở dữ liệu, khai phá dữ liệu trong cơ sở dữ liệu không phải là một hệ thống  phân tích tự động mà là một quá trình tương tác thường xuyên giữa con người với cơ sở dữ liệu được sự trợ giúp của nhiều phương pháp và công cụ tin học.

Ứng dụng trong khai phá dữ liệu nhằm xây dựng hệ thống chuẩn đoán là một trong những hướng nghiên cứu chính của đề tài. Sau khi phân tích một số thuật toán cũng như đặc điểm của dữ liệu thu nhập được về số lượng của sản phẩm và phân loại hình, đề tài đề xuất ứng dụng mô hình phân lớp dữ liệu bằng thuật toán Random Forest để tìm ra quy luật tiềm ẩn trong dữ liệu.

Nhóm chúng em  xin được gửi lời cảm ơn của mình tới cô Nguyễn Thị Thanh Tân, người đã trực tiếp hướng  dẫn, chỉ bảo tận tình, cung cấp tài liệu để nhóm chúng em hoàn thành bản báo cáo này.

Trong suốt quá trình nghiên cứu, mặc dù đã hết sức cố gắng nhưng chắc chắn bài báo cáo không tránh khỏi những thiếu sót, chúng em rất  mong thầy cô góp ý để báo cáo được hoàn chỉnh hơn. Chúng em xin chân thành cảm ơn!

# CHƯƠNG 1. TỔNG QUAN VỀ KHAI PHÁ DỮ LIỆU

## **1.1. Khái niệm về khai phá dữ liệu**

### **1.1.1. Khai phá dữ liệu là gì?**

Data mining – khai phá dữ liệu là quá trình phân loại, sắp xếp các tập hợp dữ liệu lớn để xác định các mẫu và thiết lập các mối liên hệ nhằm giải quyết các vấn đề nhờ phân tích dữ liệu. Các MCU khai phá dữ liệu cho phép các doanh nghiệp có thể dự đoán được xu hướng tương lai.

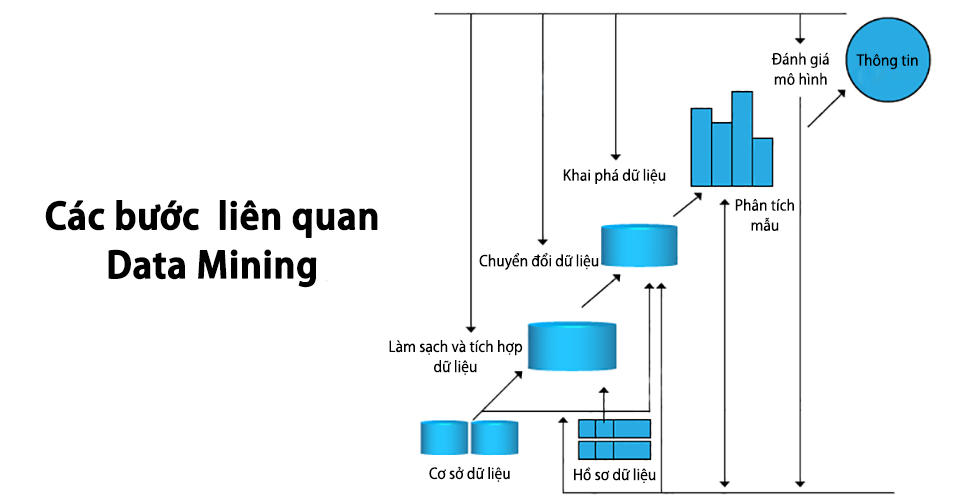
Quá trình khai phá dữ liệu là một quá trình phức tạp bao gồm kho dữ liệu chuyên sâu cũng như các công nghệ tính toán. Hơn nữa, Data Mining không chỉ giới hạn trong việc trích xuất dữ liệu mà còn được sử dụng để chuyển đổi, làm sạch, tích hợp dữ liệu và phân tích mẫu.

Có nhiều tham số quan trọng khác nhau trong Data Mining, chẳng hạn như quy tắc kết hợp, phân loại, phân cụm và dự báo. Một số tính năng chính của Data Mining:

* Dự đoán các mẫu dựa trên xu hướng trong dữ liệu.
* Tính toán dự đoán kết quả
* Tạo thông tin phản hồi để phân tích
* Tập trung vào cơ sở dữ liệu lớn hơn.
* Phân cụm dữ liệu trực quan

### **1.1.2. Các bước trong khai phá dữ liệu**

* **Bước 1**: Làm sạch dữ liệu – Trong bước này, dữ liệu được làm sạch sao cho không có tạp âm hay bất thường trong dữ liệu.
* **Bước 2**: Tích hợp dữ liệu – Trong quá trình tích hợp dữ liệu,  nhiều nguồn dữ liệu sẽ kết hợp lại thành một.
* **Bước 3**: Lựa chọn dữ liệu – Trong bước này, dữ liệu được trích xuất từ cơ sở dữ liệu.
* **Bước 4**: Chuyển đổi dữ liệu – Trong bước này, dữ liệu sẽ được chuyển đổi để thực hiện phân tích tóm tắt cũng như các hoạt động tổng hợp.
* **Bước 5**: Khai phá dữ liệu – Trong bước này, chúng tôi trích xuất dữ liệu hữu ích từ nhóm dữ liệu hiện có.
* **Bước 6**: Đánh giá mẫu – Chúng tôi phân tích một số mẫu có trong dữ liệu.
* **Bước 7**: Trình bày thông tin – Trong bước cuối cùng, thông tin sẽ được thể hiện dưới dạng cây, bảng, biểu đồ và ma trận.



*Hình 1.1 Các bước của Khai phá dữ liệu*

### **1.1.3. Ứng dụng**

* Phân tích thị trường và chứng khoán
* Phát hiện gian lận
* Quản lý rủi ro và phân tích doanh nghiệp
* Phân tích giá trị trọn đời của khách hàng
* Khám phá thêm 10 ứng dụng khai phá dữ liệu

## **1.2. Các phương pháp khai phá dữ liệu**

Khai phá dữ liệu là một chuyên ngành rất rộng và có rất nhiều hướng nghiên cứu bài toán khác nhau. Tuy nhiên, chúng được tiếp cận theo các hướng chính sau:

### **1.2.1. Phân lớp dữ liệu**

Phân lớp dữ liệu là việc xây dựng một mô hình mà có thể phân các đối tượng thành những lớp để dự đoán giá trị bị mất tại một số thuộc tính của dữ liệu hay tiêu đoán giá trị của dữ liệu sẽ xuất hiện trong tương lai.

Quá trình gồm 2 bước:

* Bước học (giai đoạn huấn luyện): xây dựng bộ phân loại (classifier) bằng việc phân tích/ học tập huấn luyện.
* Bước phân loại (classification): phân loại dữ liệu/đối tượng mới nếu mức độ chính xác của bộ phận phân loại được đánh giá là có thể chấp nhận được.

Các giải thuật phân loại dữ liệu:

* Phân loại dữ liệu với cây quyết định (decision tree)
* Phân loại dữ liệu với mạng Bayesian
* Phân loại dữ liệu với mạng neural
* Phân loại dữ liệu với k phần tử gần nhất (k-nearest neighbor)
* Phân loại dữ liệu với SVM…

### **1.2.2. Phân cụm**

Phân cụm là việc nhóm một tập các đối tượng có cùng đặc điểm giống nhau hay gần giống nhau vào cùng một nhóm.

Các đối tượng trong cùng một nhóm tương tự với nhau hơn so với đối tượng ở cụm khác. Phương pháp phân cụm hỗ trợ giai đoạn tiền xử lý dữ liệu, mô tả sự phân bố dữ liệu/đối tượng…

Các phương pháp phân cụm phổ biến:

* Phân hoạch(partitioning): các phân hoạch được tạo ra và đánh giá theo một tiêu chí nào đó.
* Phân cấp(hierarchical): phân rã tập dữ liệu/đối tượng có thứ tự phân cấp theo một tiêu chí nào đó.
* Dựa trên cấp độ (density-based): dựa trên connectivity and density functions
* Dựa trên lưới (grid-based): dựa trên multiple-level granularity structure
* Dựa trên mô hình (model-based): một mô hình giả thuyết được tạo ra cho mỗi cụm; sau đó hiệu chỉnh các thông số để mô hình phù hợp với cụm dữ liệu/đối tượng nhất……

### **1.2.3. Luật kết hợp**

Là quá trình khám phá các tập giá trị thuộc tính xuất hiện phổ biến trong các đối tượng dữ liệu. Từ tập phổ biến có thể tạo ra các luật kết hợp giữa các giá trị thuộc tính trong tập các đối tượng.

### **1.2.4. Hồi quy tuyến tính**

Phương pháp hồi quy được sử dụng để đưa ra các dự báo dựa trên các dữ liệu đang tồn tại bằng các áp dụng các công thức. Một hàm sẽ được học ra từ bộ dữ liệu hiện có bằng cách sử dụng các kỹ thuật hồi quy và tuyến tính từ việc thống kê. Sau đó, dữ liệu mới sẽ căn cứ vào hàm này để đưa ra những dự đoán.

Trong bài báo cáo này bọn em chọn phương pháp phân lớp bởi vì nó là một thuật toán tuy đơn giản nhưng lại khá hiệu quả và được sử dụng rộng rãi.

## **1.3. Bài toán phân lớp**

Bài toán phân lớp (classification) và bài toán phân cụm (cluster) là hai bài toán lớn trong lĩnh vực Machine Learnig (ML). Bài toán phân lớp là quá trình phân lớp một đối tượng dữ liệu vào một hay nhiều lớp đã cho trước nhờ một mô hình phân lớp (model). Mô hình này được xây dựng dựa trên một tập dữ liệu được xây dựng trước đó có gán nhãn (hay còn gọi là tập huấn luyện). Quá trình phân lớp là quá trình gán nhãn cho đối tượng dữ liệu.

Như vậy, nhiệm vụ của bài toán phân lớp là cần tìm một mô hình phần lớp để khi có dữ liệu mới thì có thể xác định được dữ liệu đó thuộc vào phân lớp nào.

Có nhiều bài toán phân lớp dữ liệu như phân lớp nhị phân (binary), phân lớp đa lớp (multiclass), phân lớp đa trị.

### **1.3.1. Phân lớp nhị phân**

Bài toán phân lớp nhị phân là bài toán gắn nhãn dữ liệu cho đối tượng vào một trong hai lớp khác nhau dựa vào việc dữ liệu đó có hay không có các đặc trưng (feature) của bộ phân lớp

### **1.3.2. Phân lớp nhiều lớp**

Bài toán phân lớp đa lớp là quá trình phân lớp dữ liệu với số lượng lớp lớn hơn hai. Như vậy với từng dữ liệu chúng ta phải xem xét và phân lớp chúng vào những lớp khác nhau chứ không phải là hai lớp như bài toán phân lớp nhị phân. Và thực chất bài toán phân lớp nhị phân là một bài toán đặt biệt của phân lớp đa lớp.

### **1.3.3. Quá trình phân lớp dữ liệu**

**Bước 1: Chuẩn bị tập dữ liệu huấn luyện (dataset) và rút trích đặc trưng (feature extraction)**

Công đoạn này được xem là công đoạn quan trọng trong các bài toán về Machine Learning. Vì đây là input cho việc học dể tìm ra mô hình của bài toán. Chúng ta phải biết cần chọn ra những đặt trưng tốt (good feature) của dữ liệu, lược bỏ những đặc trưng không tốt của dữ liệu, gây nhiễu (noise). Uớc lượng số chiều của dữ liệu bao nhiêu là tốt hay nói cách khác là chọn bao nhiêu feature. Nếu số chiều quá lớn gây khó khăn cho việc tính toán thì phải giảm số chiều của dữ liệu nhưng vẫn giữ được độ chính xác của dữ liệu (reduce demension).

Ở bước này chúng ta cũng chuẩn bị bộ dữ liệu để test trên mô hình. Thông thường sẽ sử dụng cross-validation (kiểm tra chéo) để chia tập datasets thành hai phần, một phần phục vụ cho training (training datasets) và phần còn lại phục vụ cho mục đích testing trên mô hình (testing dataset). Có hai cách thường sử dụng trong cross-validation là splitting và k-fold.

**Bước 2: Xây dựng mô hình phân lớp (classifier model)**

Mục đích của mô hình huấn luyện là tìm ra hàm f(x) và thông qua hàm f tìm được để chúng ta gán nhãn cho dữ liệu. Bước này thường được gọi là học hay training.

*f(x) = y*

Trong đó:

* x là các feature hay input đầu vào của dữ liệu
* y là nhãn lớp hay output đầu ra.

Thông thường để xây dựng mô hình phân lớp cho bài toán này chúng ta sử dụng các thuật toán học giám sát (supervised learning) như KNN, Neural Network, SVM, Decision Tree, Navie Bayers

**Bước 3: Kiểm tra dữ liệu với mô hình (make prediction)**

Sau khi đã tìm được mô hình phân lớp ở bước 2, thì ở bước này chúng ta sẽ đưa vào các dữ liệu mới để kiểm tra trên mô hình phân lớp.

**Bước 4: Đánh giá mô hình phân lớp và chọn ra mô hình tốt nhất**

Bước cuối cùng chúng ta sẽ đánh giá mô hình bằng cách đánh giá mức độ lỗi của dữ liệu testing và dữ liệu traning thông qua mô hình tìm được. Nếu không đạt được kết quả mong muốn của chúng ta thì phải thay đổi các tham số (turning parameter) của các thuật toán học để tìm ra các mô hình tốt hơn và kiểm tra, đánh giá lại mô hình phân lớp. Và cuối cùng chọn ra mô hình phân lớp tốt nhất cho bài toán của chúng ta.

# CHƯƠNG 2 : THUẬT TOÁN RANDOM FOREST

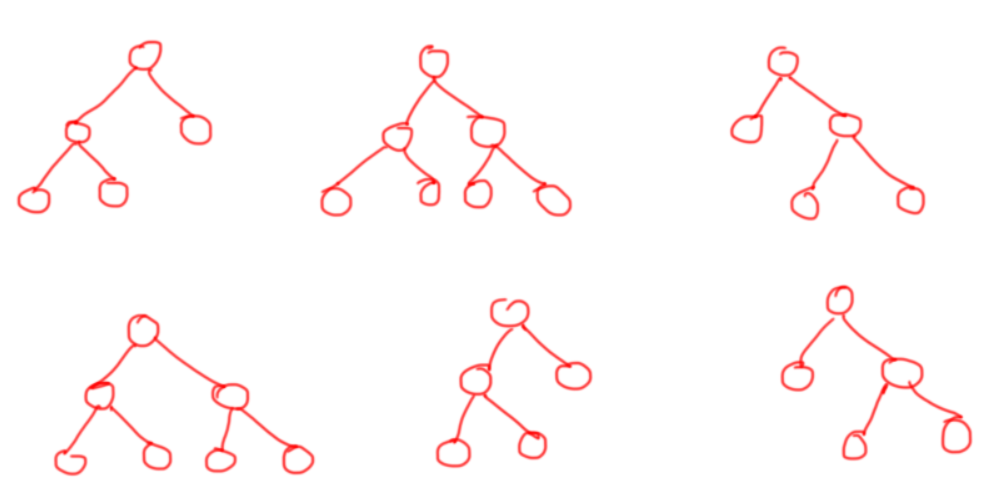
## **Random Forest là gì?**

- Là thuật toán supervised learning, có thể giải quyết cả bài toán regression và classification.

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

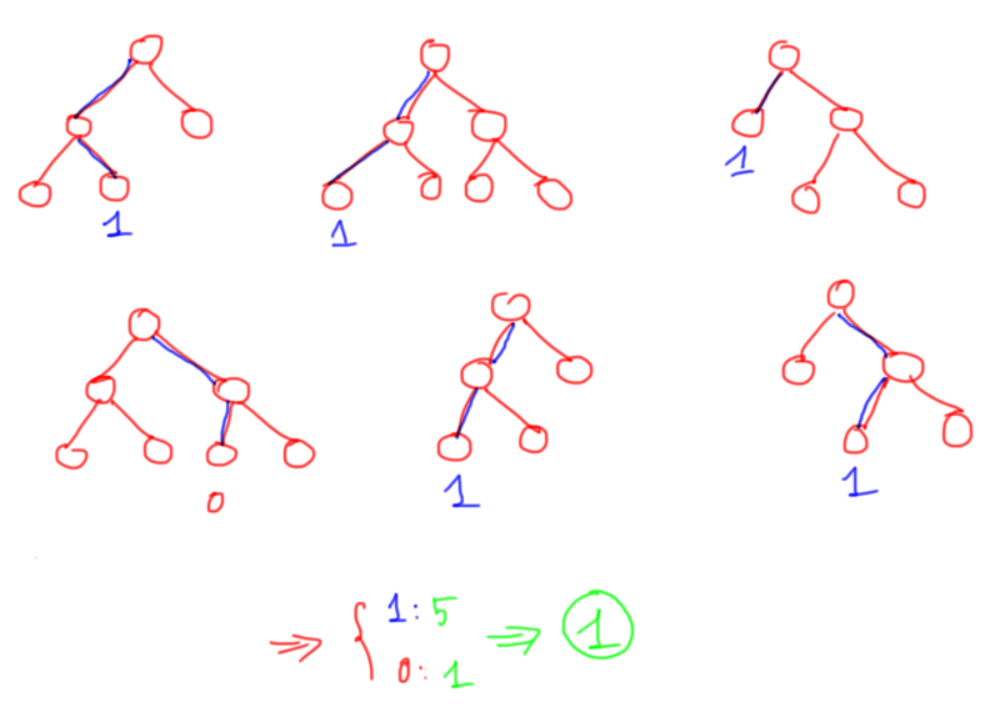
- Random là ngẫu nhiên, Forest là rừng, nên ở thuật toán Random Forest mình sẽ xây dựng nhiều cây quyết định bằng thuật toán Decision Tree, tuy nhiên mỗi cây quyết định sẽ khác nhau (có yếu tố random). Sau đó kết quả dự đoán được tổng hợp từ các cây quyết định.

- Ở bước huấn luyện thì mình sẽ xây dựng nhiều cây quyết định, các cây quyết định có thể khác nhau (phần sau mình sẽ nói mỗi cây được xây dựng như thế nào).



*Hình 2.2.1: Giới thiệu thuật toán*

- Sau đó ở bước dự đoán, với một dữ liệu mới, thì ở mỗi cây quyết định mình sẽ đi từ trên xuống theo các node điều kiện để được các dự đoán, sau đó kết quả cuối cùng được tổng hợp từ kết quả của các cây quyết định.



*Hình 2.2.2: Ví dụ thuật toán*

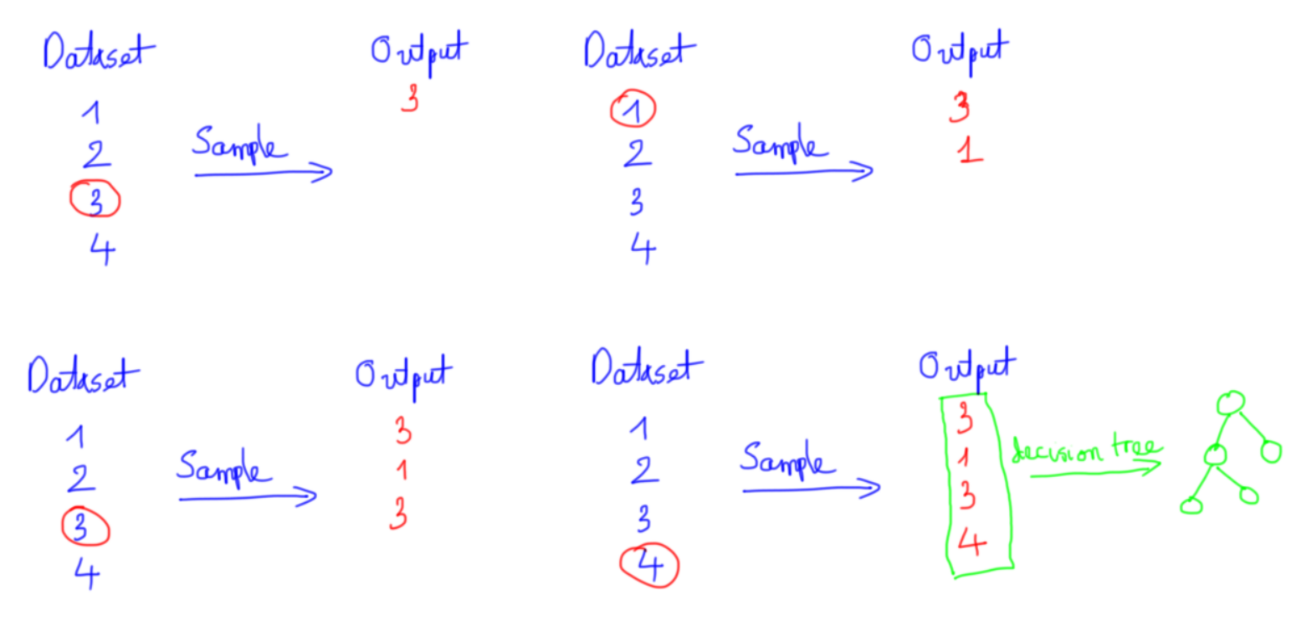
Ví dụ như trên, thuật toán Random Forest có 6 cây quyết định, 5 cây dự đoán 1 và 1 cây dự đoán 0, do đó mình sẽ vote là cho ra dự đoán cuối cùng là 1.

## **Xây dựng thuật toán Random Forest**

- Giả sử bộ dữ liệu của mình có n dữ liệu (sample) và mỗi dữ liệu có d thuộc tính (feature).

- Để xây dựng mỗi cây quyết định mình sẽ làm như sau:

* Lấy ngẫu nhiên n dữ liệu từ bộ dữ liệu với kĩ thuật [Bootstrapping](https://en.wikipedia.org/wiki/Bootstrapping_(statistics)), 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.



* Sau khi sample được n dữ liệu từ bước 1 thì mình chọn ngẫu nhiên ở k thuộc tính (k < n). Giờ mình được bộ dữ liệu mới gồm n dữ liệu và mỗi dữ liệu có k thuộc tính.
* Dùng thuật toán Decision Tree để xây dựng cây quyết định với bộ dữ liệu ở bước 2.

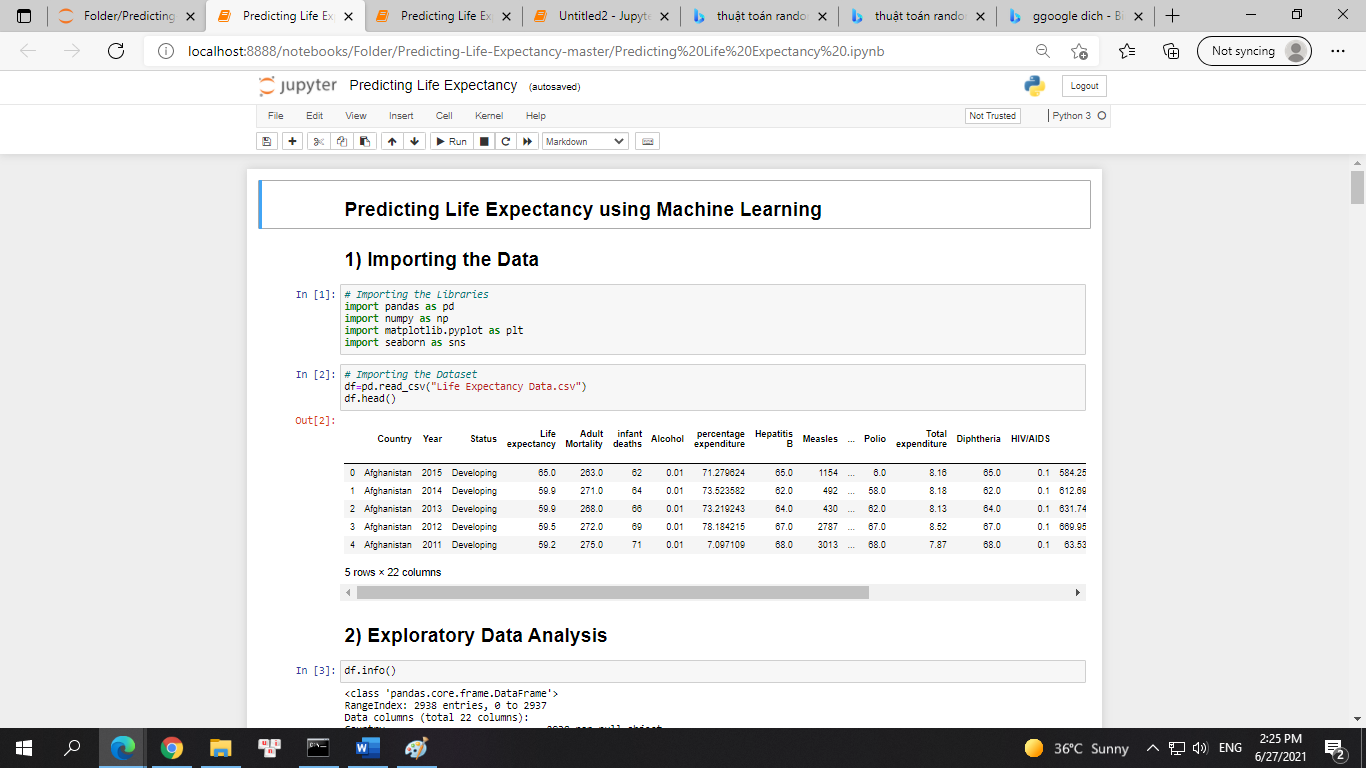
- Do quá trính xây dựng mỗi cây quyết định đều có yếu tố ngẫu nhiên (random) nên kết quả là các cây quyết định trong thuật toán Random Forest có thể khác nhau.

- Thuật toán Random Forest sẽ bao gồm nhiều cây quyết định, mỗi cây được xây dựng dùng thuật toán Decision Tree 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.

- Khi dùng thuật toán Random Forest, mình hay để ý các thuộc tính như: số lượng cây quyết định sẽ xây dựng, số lượng thuộc tính dùng để xây dựng cây. Ngoài ra, vẫn có các thuộc tính của thuật toán Decision Tree để xây dựng cây như độ sâu tối đa, số phần tử tối thiểu trong 1 node để có thể tách.

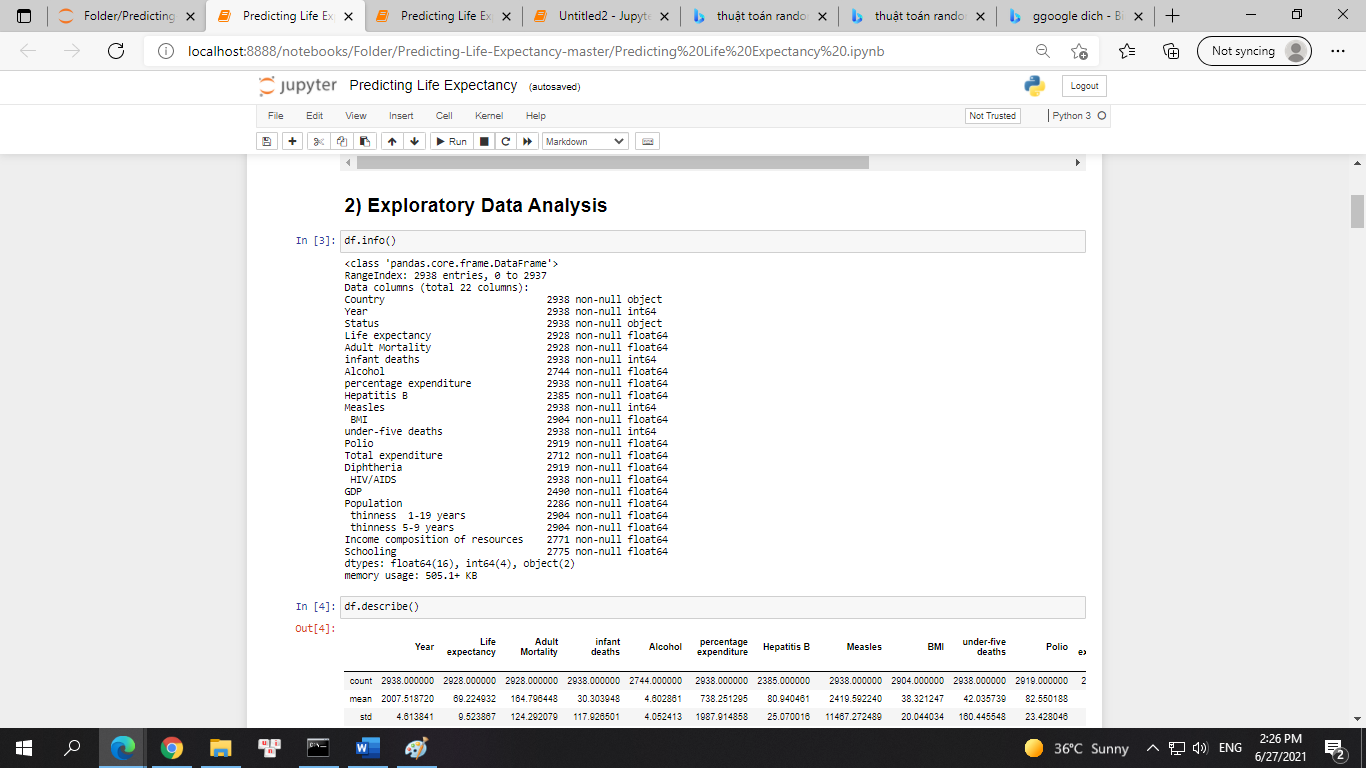
# CHƯƠNG 3: ỨNG DỤNG CỦA THUẬT TOÁN VÀO CHƯƠNG TRÌNH

## **Đọc dữ liệu**



*Hình 3.1: Đọc dữ liệu*

## **Khai phá dữ liệu**



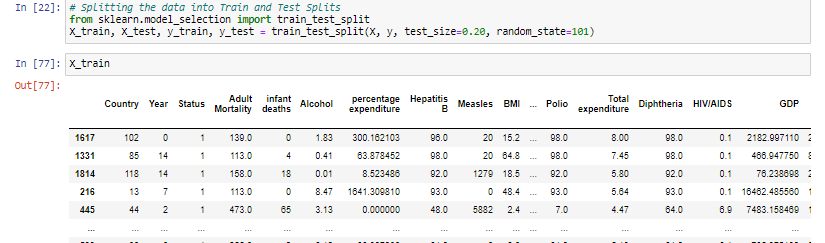
*Hình 3.2: Khai phá dữ liệu*

## **Tiền xử lý dữ liệu**



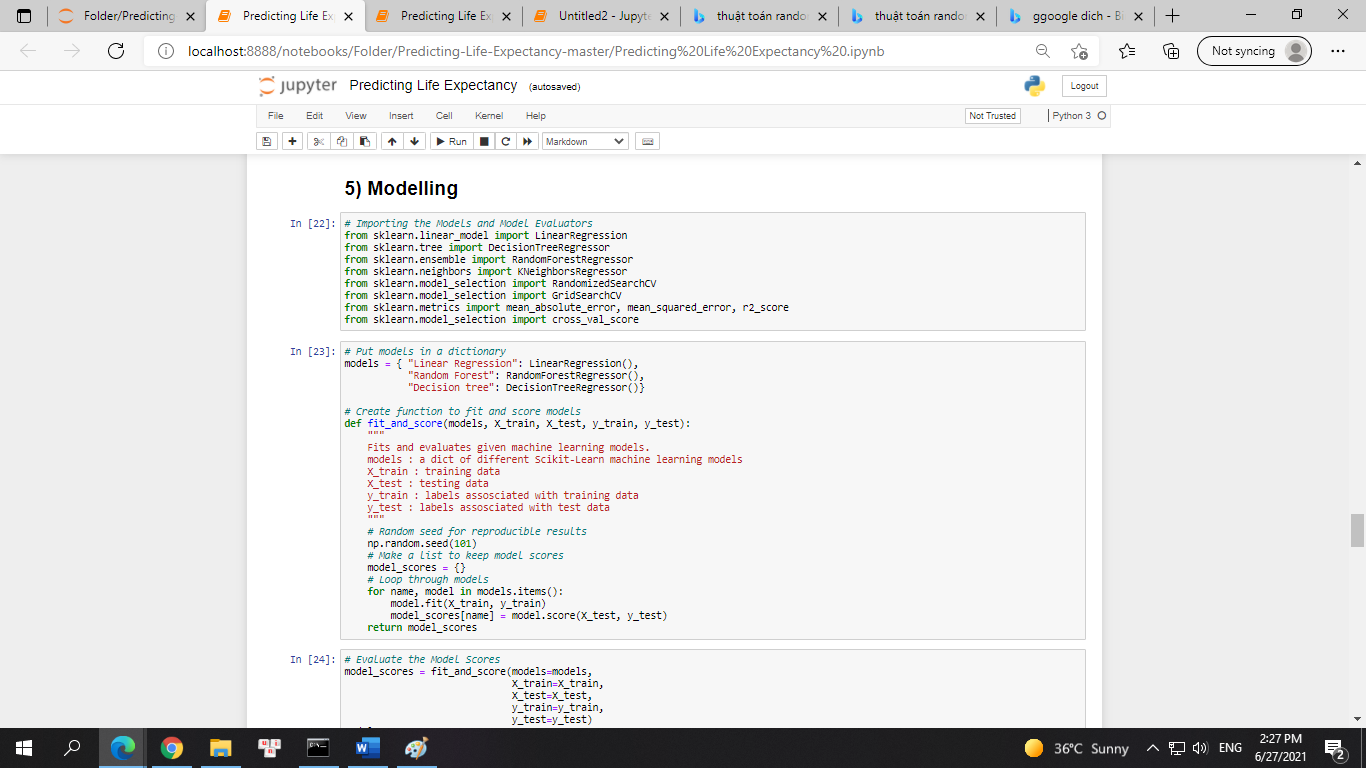
*Hình 3.3: Tiền xử lý dữ liệu*

## **Chia dữ liệu thành x và y**



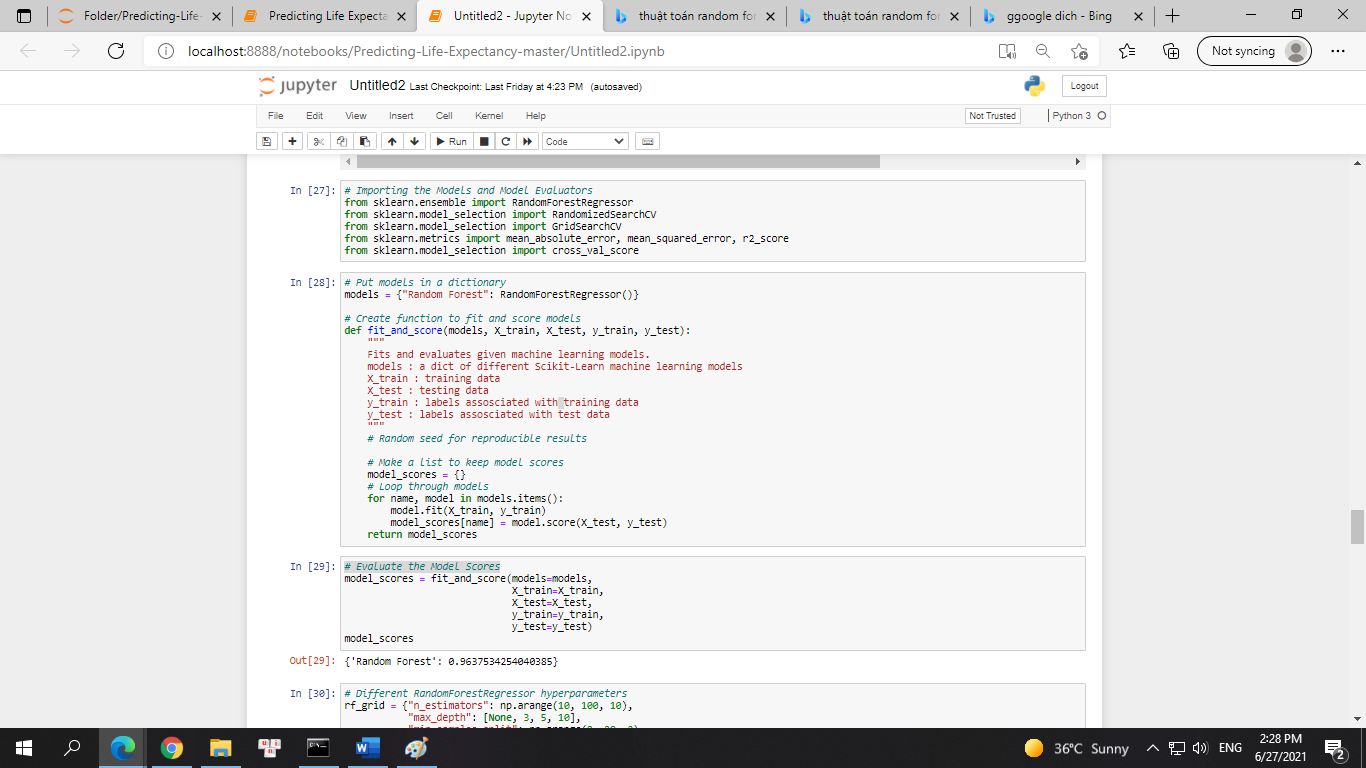
*Hình 3.4: Chia dữ liệu*

## **Tạo mô hình**



*Hình 3.5: Chia dữ liệu*

## **Xử lý thuật toán và đưa ra dự đoán**

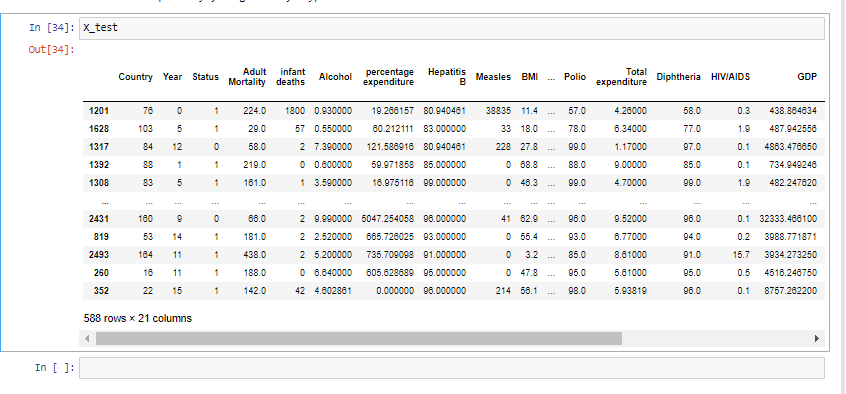


*Hình 3.6.1: Xử lý thuật toán*

* Giải thích:
* Tuấn Anh: Đọc và phân tích dữ liệu, biến dữ liệu thô thành 1 dạng dễ đọc
* Huy Trịnh: Lấy dataset, đọc dataset
* Hải Long: Kiểm tra dự đoán

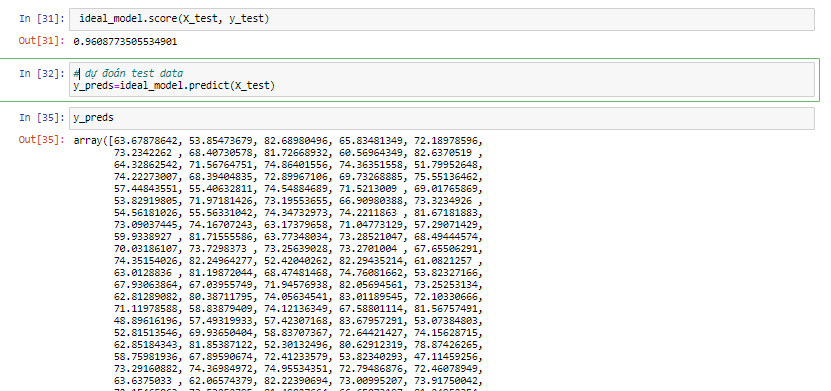
- Chia tập x và y thành x\_train-test, y\_train, test là huấn luyện tập x, y rồi test dữ liệu -> Lắp vào thuật toán và đưa ra dự đoán [Python: Tách tập dữ liệu của bạn với train\_test\_split() của scikit-learning | V1Study](https://v1study.com/python-tham-khao-tach-tap-du-lieu-cua-ban-voi-train_test_split-cua-scikit-learning.html)

- Dựa vào dataset ta có hơn 20000 tập liệu và sau khi chuẩn hóa và test dữ liệu cho ra 588



*Hình 3.6.2: Kiểm thử dữ liệu*

- Hàm predict lấy X\_test để đưa ra dự đoán



*Hình 3.6.3: Kết quả dự đoán*

# CHƯƠNG 4: KẾT LUẬN

Như vậy, đã minh họa cho các bạn thấy cách thức thực hiện thuật toán Random Forest ứng dụng trong dự đoán tuổi thọ. Tuy trong thực tế, chúng ta phải xét rất nhiều biến khác nhau tác động lên kết quả, và phân tích một khối lượng lớn dữ liệu, sử dụng nhiều phương pháp khác nhau để có được kết quả chính xác nhưng thuật toán vẫn được coi là một trong những cách tiếp cận hiệu quả.

**Những kết quả đạt được:**

* Sự hiểu biết về thuật toán cơ bản tương đối tốt
* Hiểu biết thêm về ngôn ngữ lập trình Python
* Tìm hiểu được kiến thức về khai phá dữ liệu và biết được ứng dụng của khai phá dữ liệu với học máy

**Hạn chế:**

* Mô hình học vẫn còn chưa hiệu quả, tỉ lệ đúng vẫn còn ở mức trung bình
* Kết quả đưa ra có thể có sự thay đổi
* Vì là code thuần nên code sẽ kém tối ưu, kém hiệu quả

# TÀI LIỆU THAM KHẢO

[1] Thư Viện Scikit-learn

[2] Kaggle

[3] Machine learning

[4] Bài giảng Khai Phá Dữ Liệu