

# PHÂN HIỆU TRƯỜNG ĐẠI HỌC THUỶ LỢI BỘ MÔN CÔNG NGHỆ THÔNG TIN

🙞🙞🙞🙞🙞

**KHAI PHÁ DỮ LIỆU**

**ĐỀ TÀI :**

DỰ ĐOÁN BỆNH TIM

|  |  |
| --- | --- |
| **Giảng viên hướng dẫn:** | Vũ Thị Hạnh |
| **Người thực hiện:** | Hoàng Anh Đạt (1951065623) |
| **Lớp :** | S22-61TH2 |

Hồ Chí Minh, ngày tháng năm 2023

MỤC LỤC

[PHÂN HIỆU TRƯỜNG ĐẠI HỌC THUỶ LỢI BỘ MÔN CÔNG NGHỆ THÔNG TIN 1](#_Toc143118188)

[LỜI MỞ ĐẦU 3](#_Toc143118189)

[LỜI CẢM ƠN 5](#_Toc143118190)

[CHƯƠNG 1: TỔNG QUAN VỀ KHAI PHÁ DỮ LIỆU 6](#_Toc143118191)

[1.1. Tại sao lại cần khai phá dữ liệu (datamining) 6](#_Toc143118192)

[1.2. Khai phá dữ liệu là gì? 6](#_Toc143118193)

[1.3. Các chức năng chính của khai phá dữ liệu 7](#_Toc143118194)

[1.4. Ứng dụng của khai phá dữ liệu 8](#_Toc143118195)

[1.5. Các bước thực hiện khai phá dữ liệu của một tập dữ liệu 8](#_Toc143118196)

[CHƯƠNG 2: TIỀN XỬ LÝ DỮ LIỆU 10](#_Toc143118197)

[2.1. Thông tin tập dữ liệu: 10](#_Toc143118198)

[2.2. Thông tin thuộc tính: 10](#_Toc143118199)

[2.3 Giải quyết bài toán phân loại được đặt ra: 24](#_Toc143118200)

[CHƯƠNG 3: GIỚI THIỆU VỀ RANDOM FOREST CLASSIFIER 26](#_Toc143118201)

[3.1 Giới thiệu về Decision Tree 26](#_Toc143118202)

[3.2 Ưu nhược điểm 26](#_Toc143118203)

[3.3 Thực hiện: 26](#_Toc143118204)

[CHƯƠNG 4: GIỚI THIỆU VỀ RANDOM FOREST CLASSIFIER 29](#_Toc143118205)

[4.1 Giới thiệu về Random Forest Classifier 29](#_Toc143118206)

[4.2 Ưu nhược điểm 29](#_Toc143118207)

[4.3 Thực hiện: 29](#_Toc143118208)

[CHƯƠNG 5: SO SÁNH KẾT LUẬN 31](#_Toc143118209)

[5.1. So Sánh: 31](#_Toc143118210)

[5.2. Kết Luận: 32](#_Toc143118211)

[TÀI LIỆU THAM KHẢO 33](#_Toc143118212)

# LỜI MỞ ĐẦU

Công nghệ ngày càng phổ biến, không ai có thể phủ nhận được tầm quan trọng và những hiệu quả mà nó đem lại cho cuộc sống chúng ta. Bất kỳ trong lĩnh vực nào, sự góp mặt của trí tuệ nhân tạo sẽ giúp con người làm việc và hoàn thành tốt công việc hơn. Và gần đây, một thuật ngữ “Data Mining” rất được nhiều người quan tâm.Thay vì phải code phần mềm với cách thức thủ công theo một bộ hướng dẫn cụ thể nhằm hoàn thành một nhiệm vụ đề ra thì máy tính sẽ tự “học hỏi” bằng cách sử dụng một lượng lớn dữ liệu cùng những thuật toán cho phép nó thực hiện các tác vụ.

Đây là một lĩnh vực khoa học tuy không mới, nhưng cho thấy lĩnh vực trí tuệ nhân tạo đang ngày càng phát triển và có thể tiến xa hơn trong tương lai. Đồng thời, thời điểm này nó được xem là một lĩnh vực “nóng” và dành rất nhiều mối quan tâm để phát triển nó một cách mạnh mẽ, bùng nổ hơn.

Hiện nay, việc quan tâm machine learning càng ngày càng tăng lên là vì nhờ có machine learning giúp gia tăng dung lượng lưu trữ các loại dữ liệu sẵn, việc xử lý tính toán có chi phí thấp và hiệu quả hơn rất nhiều.

Những điều trên được hiểu là nó có thể thực hiện tự động, nhanh chóng để tạo ra những mô hình cho phép phân tích các dữ liệu có quy mô lớn hơn và phức tạp hơn đồng thời đưa ra những kết quả một cách nhanh và chính xác hơn.

*Lý do chọn đề tài:*

Việc lựa chọn đề tài "Bệnh tim" trong khai phá dữ liệu là một sự quyết định có logic và quan trọng vì nó liên quan trực tiếp đến sức khỏe của con người và có thể mang lại nhiều giá trị trong việc phát triển y học và chăm sóc sức khỏe. Dưới đây là một số lý do mà bạn có thể xem xét khi chọn đề tài này:

* Vấn đề sức khỏe cộng đồng: Bệnh tim là một trong những nguyên nhân hàng đầu gây tử vong trên toàn cầu. Nghiên cứu khai phá dữ liệu về bệnh tim có thể giúp tìm ra các mô hình, xu hướng và yếu tố nguy cơ liên quan, từ đó hỗ trợ cải thiện chẩn đoán, dự đoán, và phòng ngừa bệnh tim.
* Phát hiện yếu tố nguy cơ: Dữ liệu về bệnh tim chứa đựng rất nhiều thông tin quan trọng như tiền sử bệnh, thói quen ăn uống, hoạt động thể chất, di truyền, và nhiều yếu tố khác có thể góp phần vào sự phát triển của bệnh tim. Khai phá dữ liệu có thể giúp xác định những yếu tố nguy cơ quan trọng và mô hình hóa mối liên hệ giữa chúng.
* Dự đoán và phòng ngừa: Dự đoán bệnh tim dựa trên dữ liệu có thể giúp bác sĩ và người bệnh đưa ra quyết định đúng lúc, giúp điều trị hiệu quả hơn và giảm nguy cơ tử vong. Các mô hình dự đoán có thể giúp nhận biết sớm các dấu hiệu tiềm tàng của bệnh tim, từ đó giúp phòng ngừa kịp thời.
* Phân tích tương quan và tác động: Khai phá dữ liệu cung cấp cơ hội để phân tích tương quan giữa các yếu tố và hiểu rõ hơn về cách chúng tác động lẫn nhau trong việc phát triển bệnh tim. Điều này có thể dẫn đến việc khám phá các mối quan hệ mới và cung cấp thông tin hữu ích cho việc đưa ra quyết định.

# LỜI CẢM ƠN

Chúng em xin chân thành gửi lời cảm ơn tới các thầy cô giáo trong Trường Đại học Thủy Lợi Cơ Sở 2 nói chung và các thầy cô giáo trong Khoa Công nghệ thông tin nói riêng đã tận tình giảng dạy, truyền đạt cho chúng em những kiến thức cũng như kinh nghiệm quý báu trong suốt quá trình học. Đặc biệt, em gửi lời cảm ơn đến cô Vũ Thị Hạnh đã tận tình theo sát giúp đỡ, trực tiếp chỉ bảo hướng dẫn trong suốt quá trình nghiên cứu và học tập của chúng em.

# CHƯƠNG 1: TỔNG QUAN VỀ KHAI PHÁ DỮ LIỆU

## Tại sao lại cần khai phá dữ liệu (datamining)

Khoảng hơn một thập kỷ trở lại đây, lượng thông tin được lưu trữ trên các thiết bị điện tử (đĩa cứng, CD-ROM, băng từ, .v.v.) không ngừng tăng lên. Sự tích lũy dữ liệu này xảy ra với một tốc độ bùng nổ. Người ta ước đoán rằng lượng thông tin trên toàn cầu tăng gấp đôi sau khoảng hai năm và theo đó số lượng cũng như kích cỡ của các cơ sở dữ liệu (CSDL) cũng tăng lên một cách nhanh chóng. Nói một cách hình ảnh là chúng ta đang “ngập” trong dữ liệu nhưng lại “đói” tri thức. Câu hỏi đặt ra là liệu chúng ta có thể khai thác được gì từ những “núi” dữ liệu tưởng chừng như “bỏ đi” ấy không ? “Necessity is the mother of invention” - Data Mining ra đời như một hướng giải quyết hữu hiệu cho câu hỏi vừa đặt ra ở trên. Khá nhiều định nghĩa về Data Mining và sẽ được đề cập ở phần sau, tuy nhiên có thể tạm hiểu rằng Data Mining như là một công nghệ tri thức giúp khai thác những thông tin hữu ích từ những kho dữ liệu được tích trữ trong suốt quá trình hoạt động của một công ty, tổ chức nào đó.

## Khai phá dữ liệu là gì?

Khai phá dữ liệu (datamining) được định nghĩa như một quá trình chắt lọc hay khai phá tri thức từ một lượng lớn dữ liệu. Một ví dụ hay được sử dụng là việc khai phá vàng từ đá và cát, Datamining được ví dụ như công việc “Đãi cát tìm vàng” trong một tập hợp lớn các dữ liệu cho trước. Thuật toán Datamining ám chỉ việc tìm kiếm một tập hợp nhỏ có giá trị từ một số lượng lớn các dữ liệu thô. Có nhiều thuật ngữ hiện được dùng cũng có nghĩa tương tự với từ Datamining như Knowledge Mining (khai phá tri thức), knowledge extraction(chắt lọc tri thức), data/patern analysis(phân tích dữ liệu/mẫu), data archaeoloogy (khảo cổ dữ liệu), datadredging(nạo vét dữ liệu),...

Định nghĩa: Khai phá dữ liệu là một tập hợp các kỹ thuật được sử dụng để tự động khai thác và tìm ra các mối quan hệ lẫn nhau của dữ liệu trong một tập hợp dữ liệu khổng lồ và phức tạp, đồng thời cũng tìm ra các mẫu tiềm ẩn trong tập dữ liệu đó.



## Các chức năng chính của khai phá dữ liệu

Data Mining được chia nhỏ thành một số hướng chính như sau:

* + - Mô tả khái niệm (concept description): thiên về mô tả, tổng hợp và tóm tắt khái niệm. Ví dụ: tóm tắt văn bản
    - Luật kết hợp (association rules): là dạng luật biểu diễn tri thức ở dạng khá đơn giản. Ví dụ: “60% nam giới vào siêu thị nếu mua bia thì có tới 80% trong số họ sẽ mua thêm thịt bò khô”. Luật kết hợp được ứng dụng nhiều trong lĩnh vực kinh doanh, y học, tin -sinh học, tài chính & thị trường chứng khoán v.v.
    - Phân lớp và dự đoán (classification & prediction): xếp một đối tượng vào một trong những lớp đã biết trước. Ví dụ: phân lớp vùng địa lý theo dữ liệu thời tiết. Hướng tiếp cận này thường sử dụng một số kỹ thuật của machine learning như cây quyết định (decision tree), mạng nơ ron nhân tạo (neural network), .v.v. Người ta còn gọi phân lớp là học có giám sát (học có thầy).
    - Phân cụm (clustering): xếp các đối tượng theo từng cụm (số lượng cũng như tên của cụm chưa được biết trước. Người ta còn gọi phân cụm là học không giám sát (học không thầy).
    - Khai phá chuỗi (sequential/temporal patterns): tương tự như khai phá luật kết hợp nhưng có thêm tính thứ tự và tính thời gian. Hướng tiếp cận này được ứng dụng nhiều trong lĩnh vực tài chính và thị trường chứng khoán vì nó có tính dự báo cáo.

## Ứng dụng của khai phá dữ liệu

Data Mining tuy là một hướng tiếp cận mới nhưng thu hút được rất nhiều sự quan tâm của các nhà nghiên cứu và phát triển nhờ vào những ứng dụng thực tiễn của nó. Chúng ta có thể liệt kê ra đây một số ứng dụng điển hình:

* Phân tích dữ liệu và hỗ trợ ra quyết định (data analysis & decision support)
* Điều trị y học (medical treatment)
* Text mining & Web mining
* Tin-sinh (bio-informatics)
* Tài chính và thị trường chứng khoán (finance & stock market)
* Bảo hiểm (insurance)
* Nhận dạng (pattern recognition) .v.v

## Các bước thực hiện khai phá dữ liệu của một tập dữ liệu

* Thu thập và đánh giá dữ liệu: Thu thập dữ liệu từ các nguồn khác nhau, sau đó đánh giá chất lượng dữ liệu bằng cách kiểm tra tính đầy đủ, tính chính xác và tính xác thực của dữ liệu.
* Tiền xử lý dữ liệu: Tiền xử lý dữ liệu là bước quan trọng để làm sạch và chuẩn hóa dữ liệu trước khi khai thác. Điều này bao gồm loại bỏ các giá trị rỗng, điền giá trị rộng(nếu cần), loại bỏ ngoại lệ, chuẩn hóa dữ liệu và mã hóa dữ liệu( nếu cần).
* Phân tích dữ liệu: Đây là bước khai thác và phân tích dữ liệu để tìm ra các mẫu, liên hệ và kiểu dữ liệu. Có thể sử dụng nhiều kỹ thuật khác nhau để phân tích dữ liệu như
* phân tích nhân tố, phân tích định lượng, phân tích hồi quy, phân tích nhóm và phân tích chuỗi thời gian.
* Mô hình hóa dữ liệu: Dữ liệu được mô hình hóa để xác định các mẫu, liên hệ và kiểu dữ liệu. Có thể sử dụng nhiều kỹ thuật khác nhau để mô hình hóa dữ liệu như cây quyết định, mạng neuron nhân tạo, SVM(Support Vector Machine), phân tích trung bình động và phân tích tương tự.
* Kiểm tra và đánh giá mô hình: Mô hình được kiểm tra và đánh giá để đảm bảo tính chính xác và độ tin cậy của mô hình. Có nhiều phương pháp khác nhau để kiểm tra và đánh giá mô hình như phân chia dữ liệu, độ chính xác, độ tin cậy, độ nhạy cảm và độ đặc hiệu.
* Sử dụng mô hình để dự đoán hoặc phân loại: Sau khi xây dựng mô hình, ta sử dụng mô hình để dự đoán hoặc phân loại dữ liệu mới.

# CHƯƠNG 2: TIỀN XỬ LÝ DỮ LIỆU

## Thông tin tập dữ liệu:

Dataset "Heart Disease" là một tập dữ liệu gồm các thông tin liên quan đến các tin tuyển dụng việc làm. Mục đích của dataset này là để dự đoán xem một tin tuyển dụng việc làm có phải là tin đăng tuyển dụng việc làm thật sự hay là tin giả mạo. Tập dữ liệu này bao gồm 18 thuộc tính bao gồm các thông tin về công ty đăng tuyển dụng, mô tả công việc, yêu cầu công việc, lương, địa điểm, số lượng tuyển dụng và các thuộc tính khác. Tập dữ liệu này đã được phân lớp sẵn với 0 tương ứng với tin giả mạo và 1 tương ứng với tin tuyển dụng việc làm thật sự.

## Thông tin thuộc tính:

|  |  |
| --- | --- |
| **Tên dữ liệu gốc** | **Ý nghĩa** |
| Heartdisease | Bệnh tim |
| BMI | Chỉ số khối cơ thể |
| Smoking | Hút thuốc |
| Alcoholdrinking | Uống rượu |
| Stroke | Đột quỵ |
| Physicalhealth | Sức khoẻ thể chất |
| Mentalhealth | Sức khỏe tinh thần |
| Diffwalking | Đi bộ |
| Sex | Gới tính |
| Agecategory | Danh mục tuổi |
| Race | Dân tộc |
| Diabetic | Tiểu đường |
| Physicalactivity | Hoạt động thể chất |

|  |  |
| --- | --- |
| Genhealth | Gen Sức khỏe |
| Sleeptime | Giờ ngủ |
| Asthma | hen suyễn |
| KidneyDisease | Bệnh thận |
| SkinCancer | Ung thư da |

Để xử lý tiền dữ liệu cho dataset Heart diseasetrên trang Kaggle, chúng ta có thể thực hiện các bước sau:

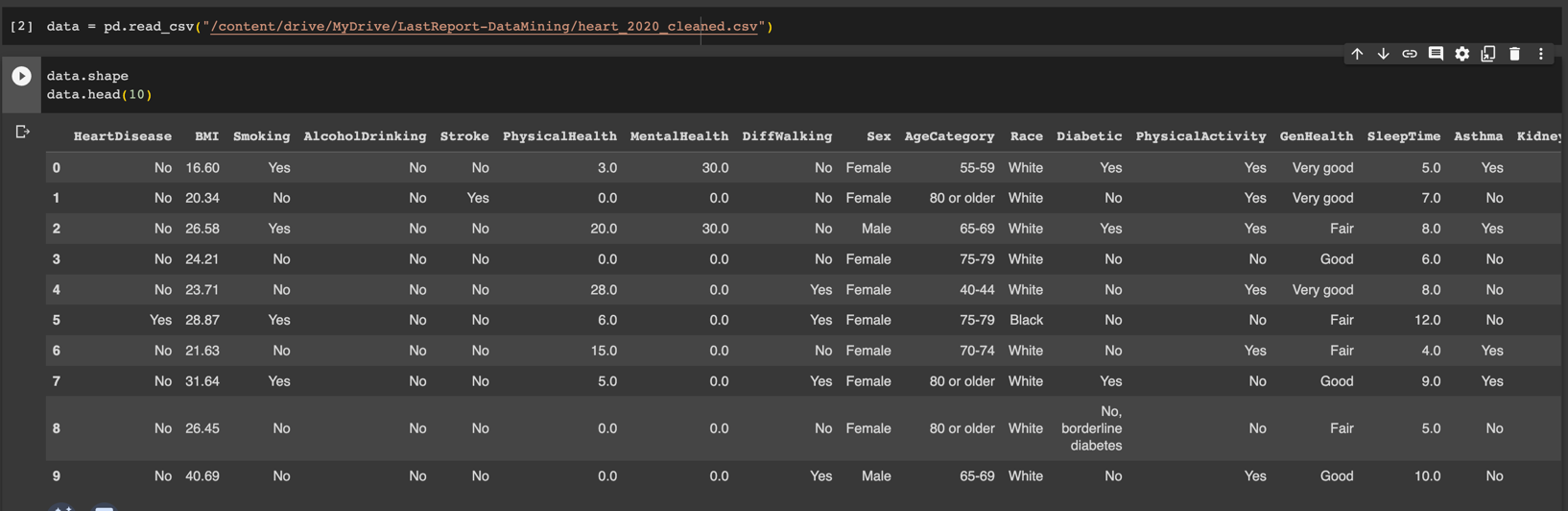
Đọc file CS, đưa dữ liệu ra màn hình và kiểm tra tập dữ liệu có giá trị NaN/Null không.

Thay thế giá trị thiếu (NaN) bằng giá trị trung bình (đối với các thuộc tính số) hoặc giá trị phổ biến nhất (đối với các thuộc tính dạng chuỗi) hoặc khoảng trắng để đảm bảo tính toàn vẹn dữ liệu.

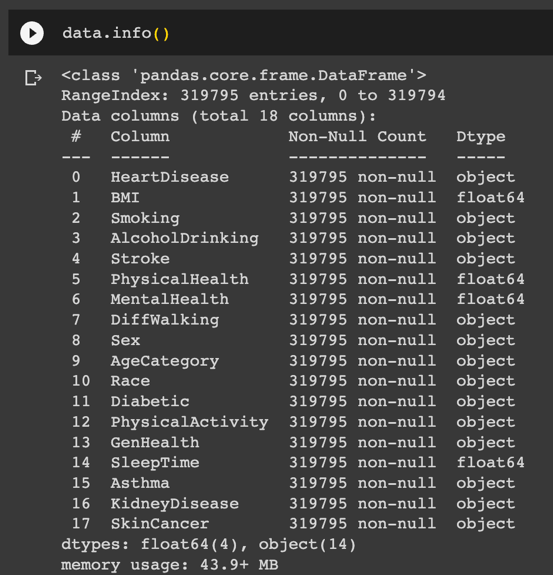
Chuyển đổi các thuộc tính dạng chuỗi sang dạng số bằng phương pháp mã hóa one-hot (nếu số lượng giá trị duy nhất của thuộc tính là nhỏ).

Tiền xử lý dữ liệu chi tiết :

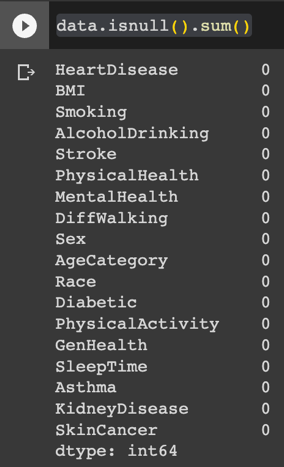
+ Đọc bộ dữ liệu Heart Disease và xuất 10 giá trị đầu.



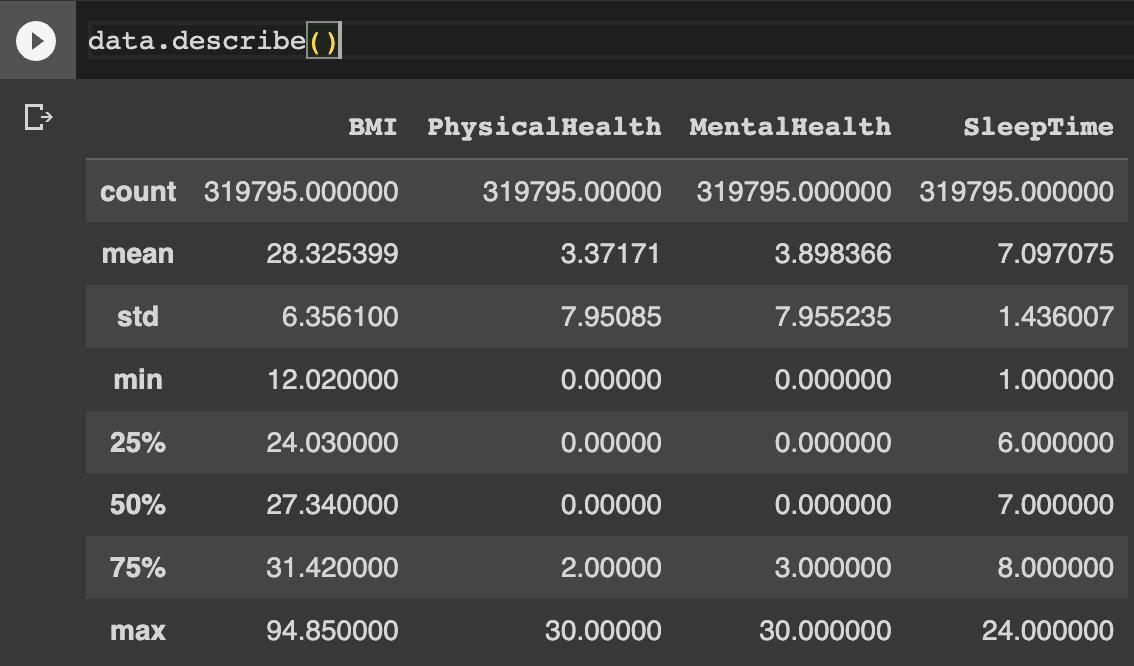
+ Kiểm tra thống kê cơ bản của dữ liệu



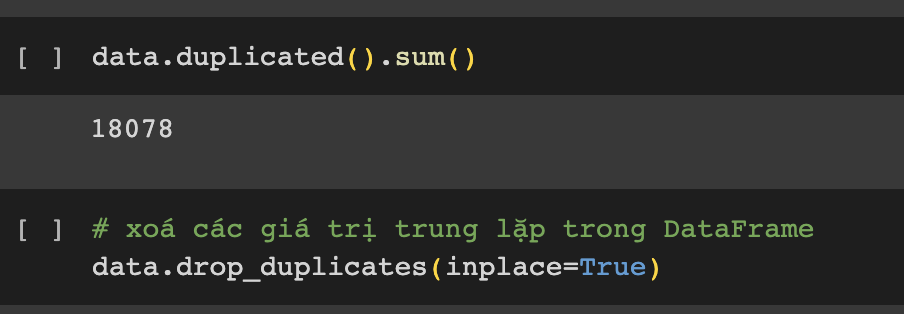
+ Đếm số lượng giá trị thiếu (null) trong từng cột của tập dữ liệu.



+ Tạo ra một tóm tắt thống kê của các biến số trong tập dữ liệu

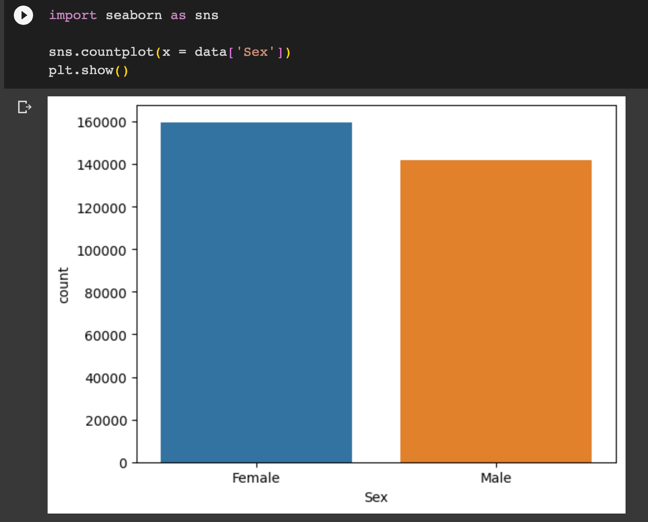


+ Kiểm tra các giá trị trùng lặp trong data và xoá nếu có



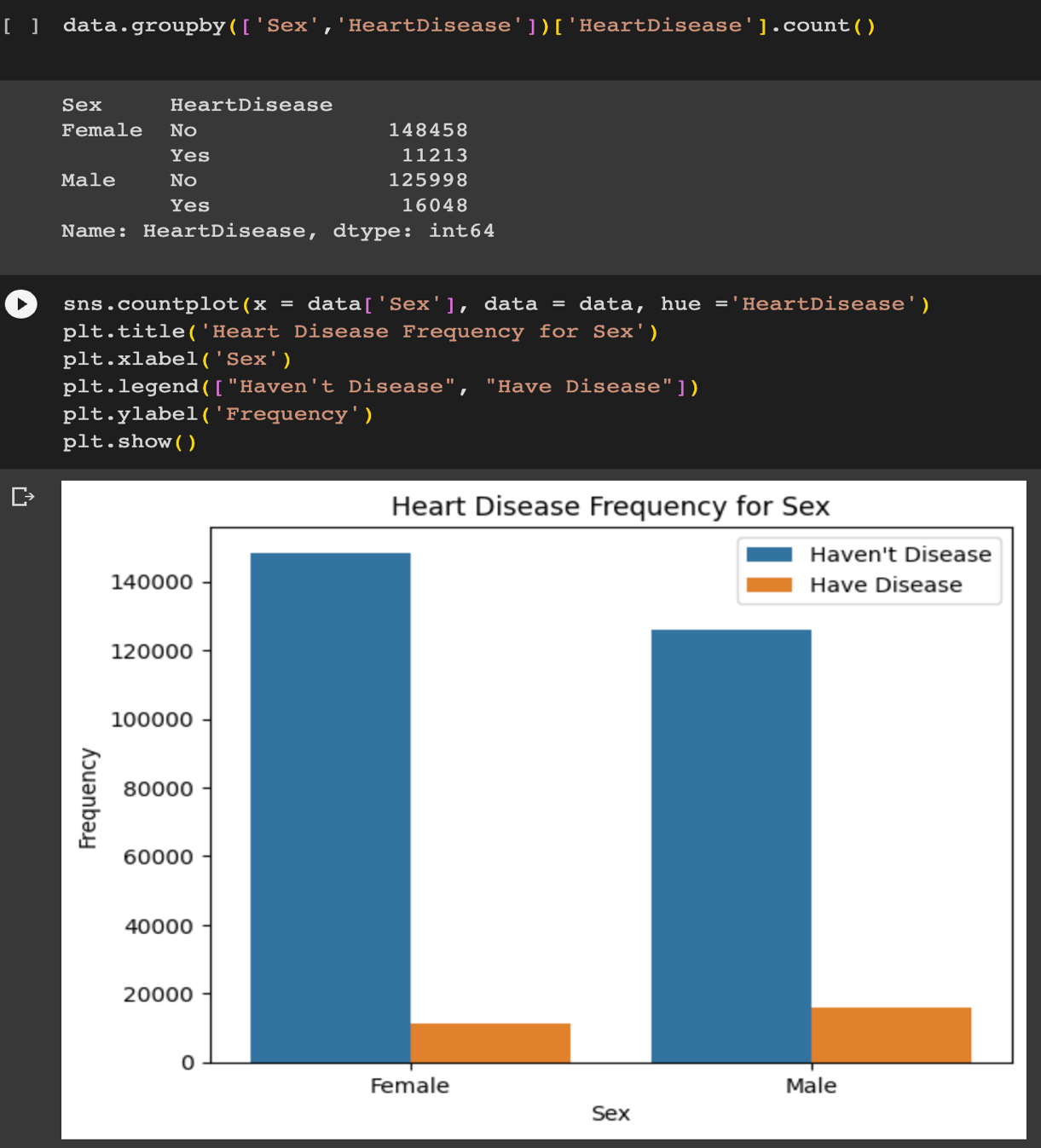
+ Vẽ biểu đồ thanh để trực quan hóa Heart Disease.



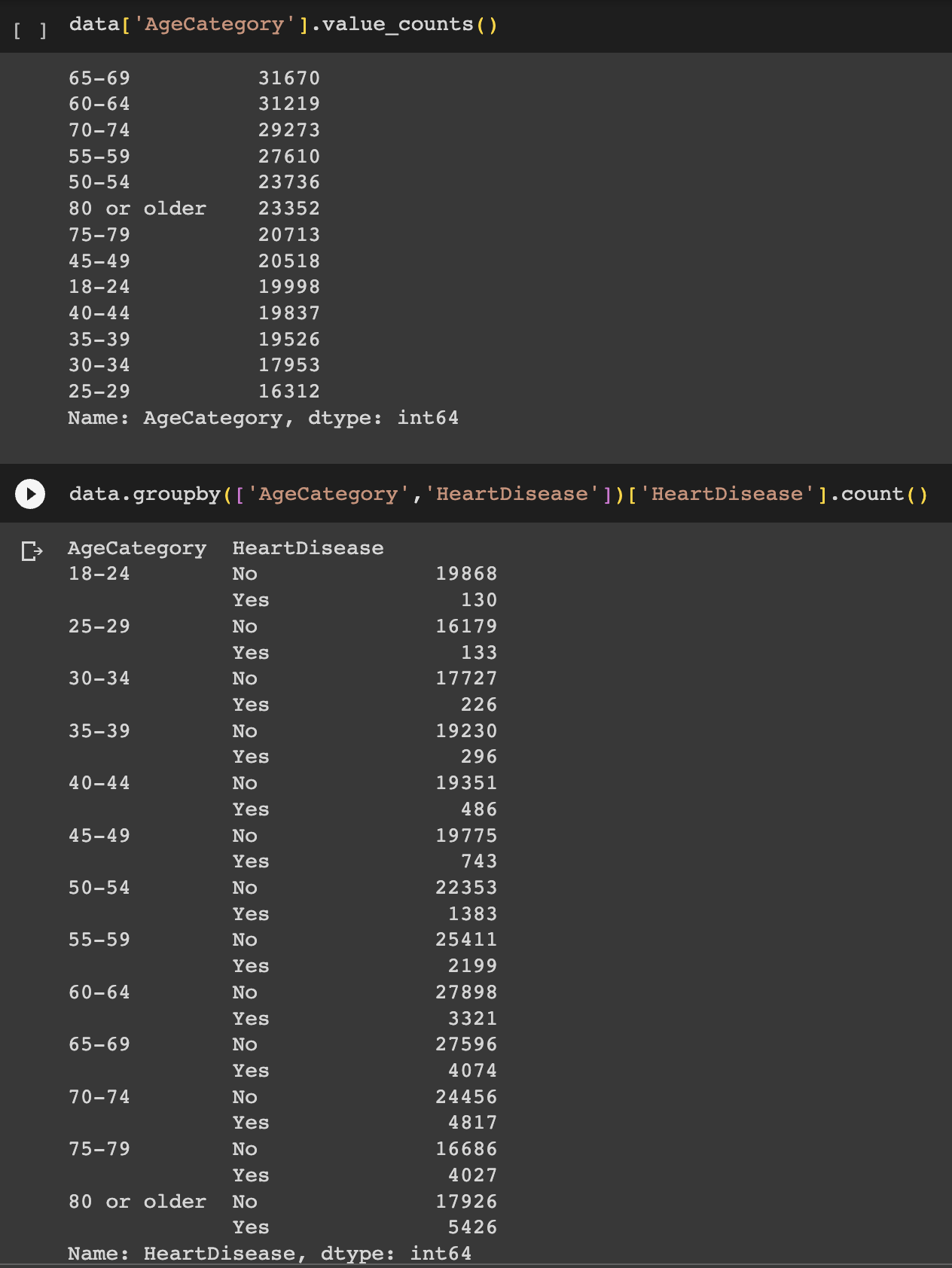


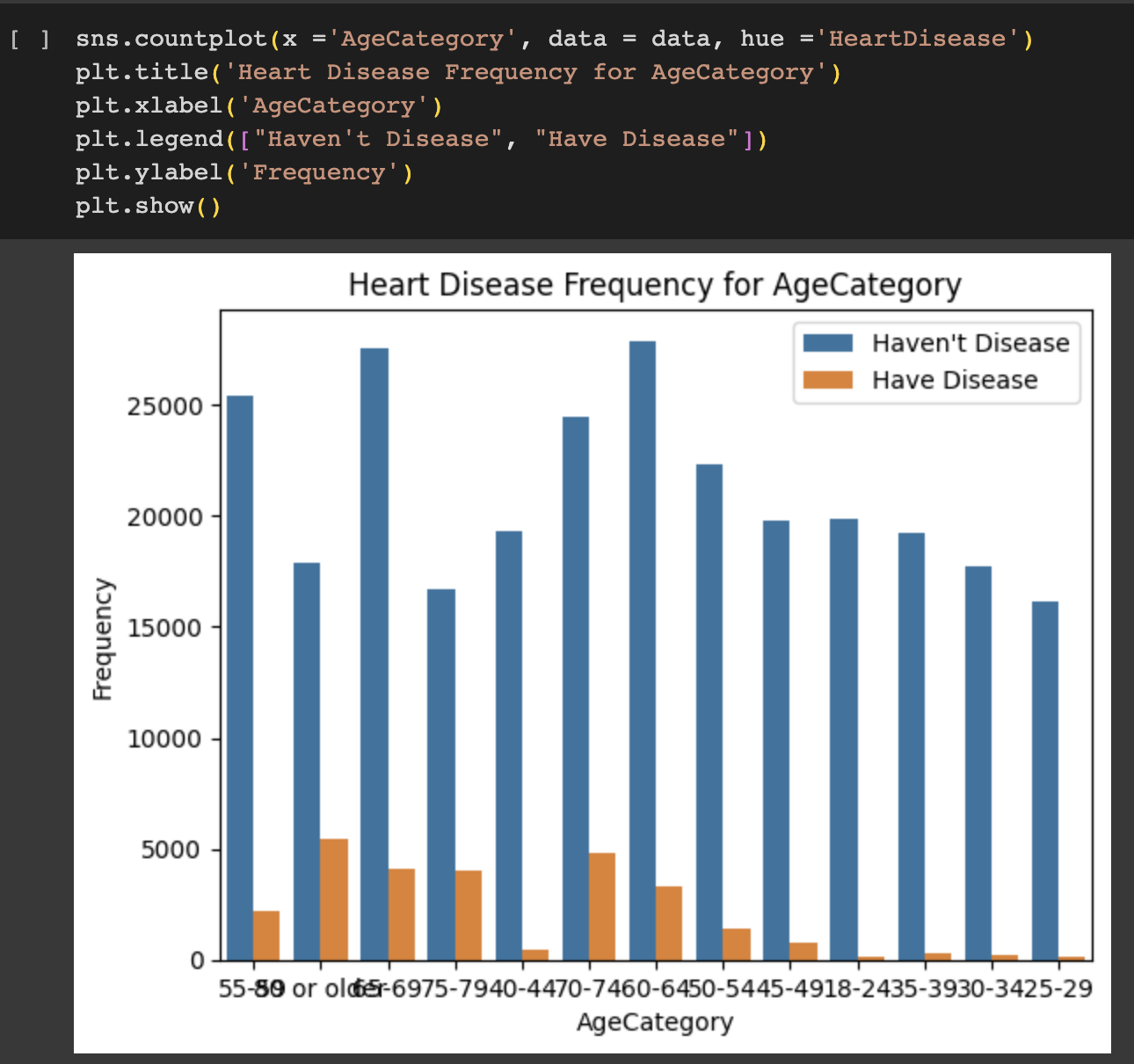
* Nhóm và đếm

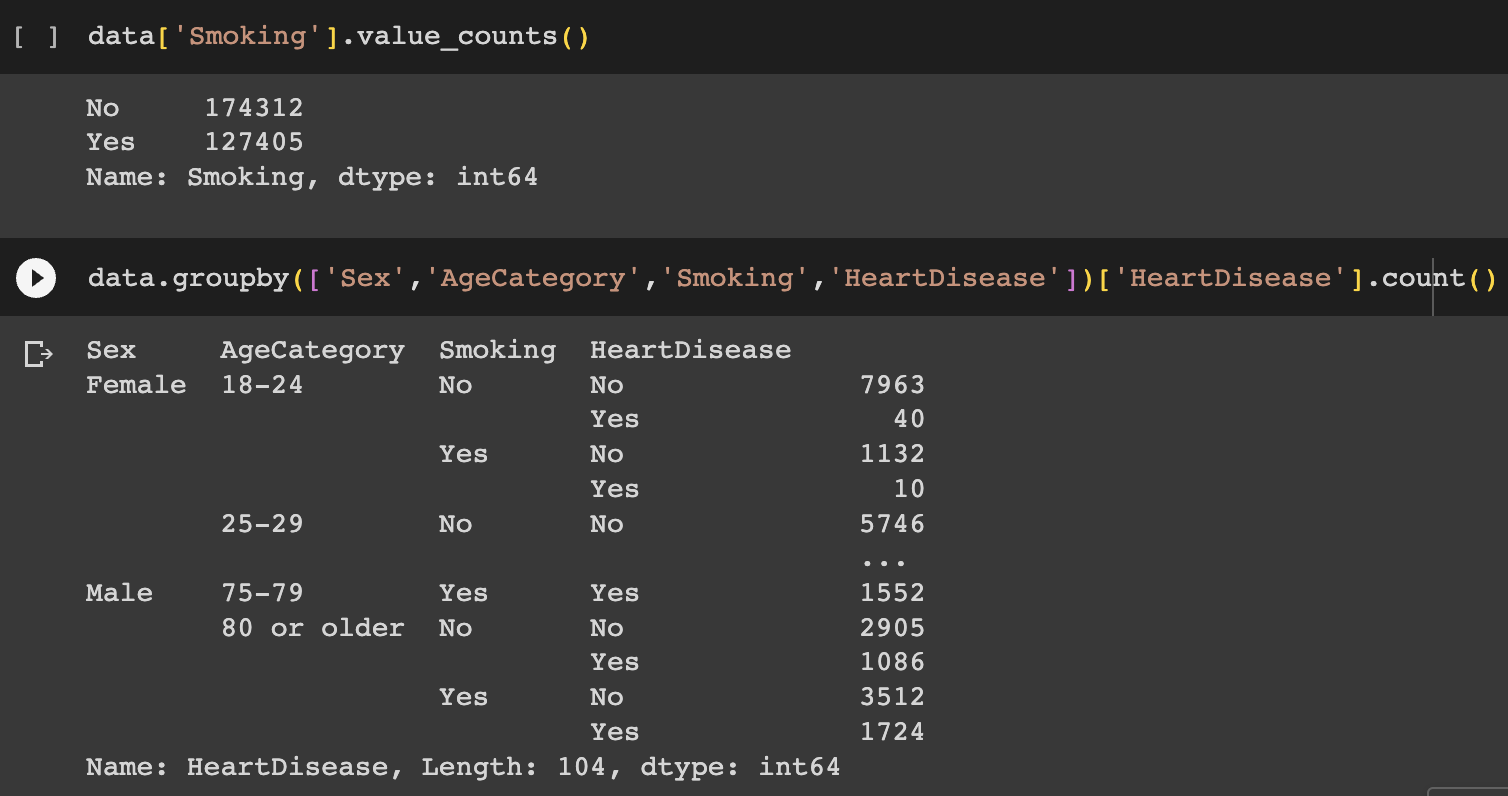
+ Nhóm các cột có giá trị là “Sex” và “HeartDisease”, đếm giá trị các thuộc tính và vẽ biểu đồ thể hiện

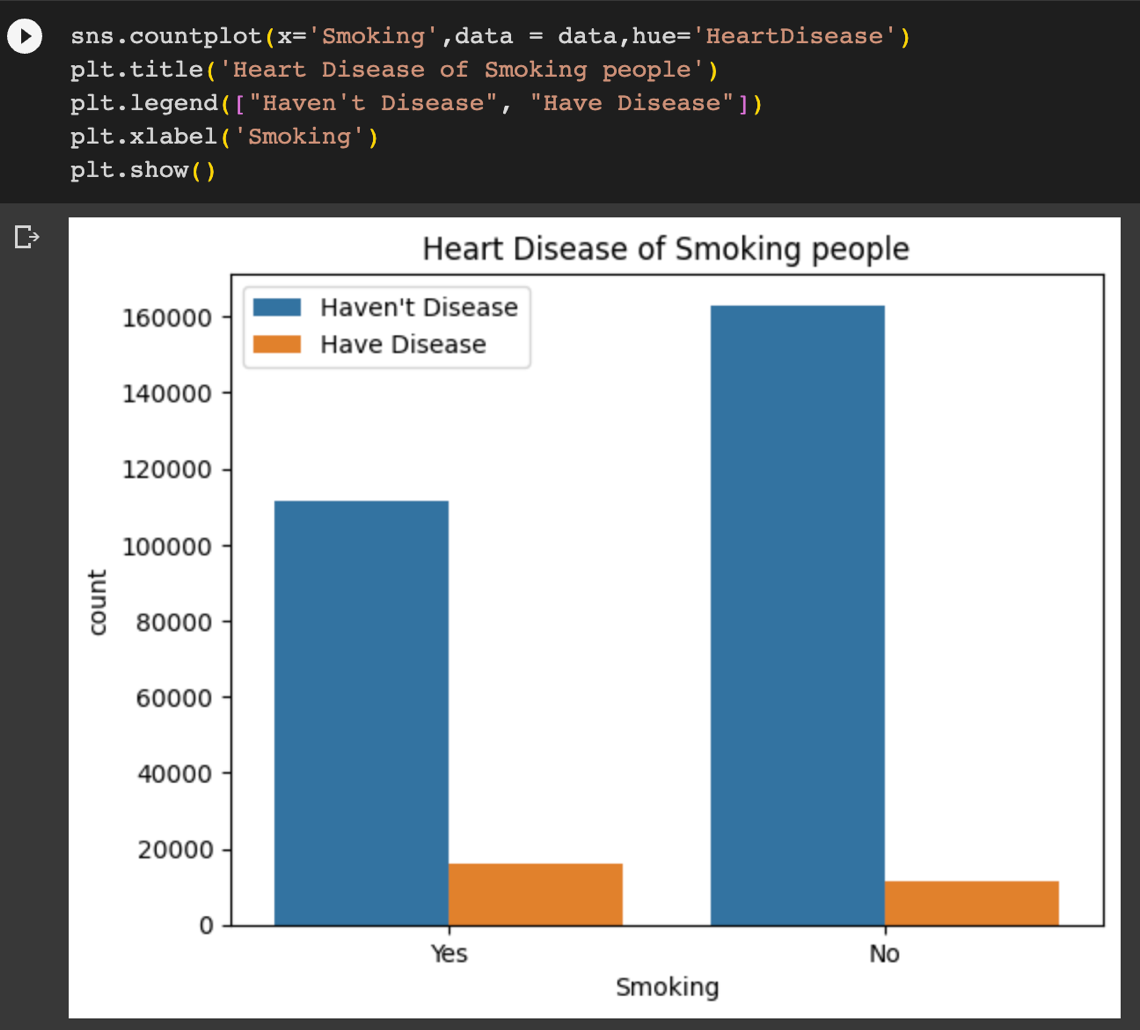


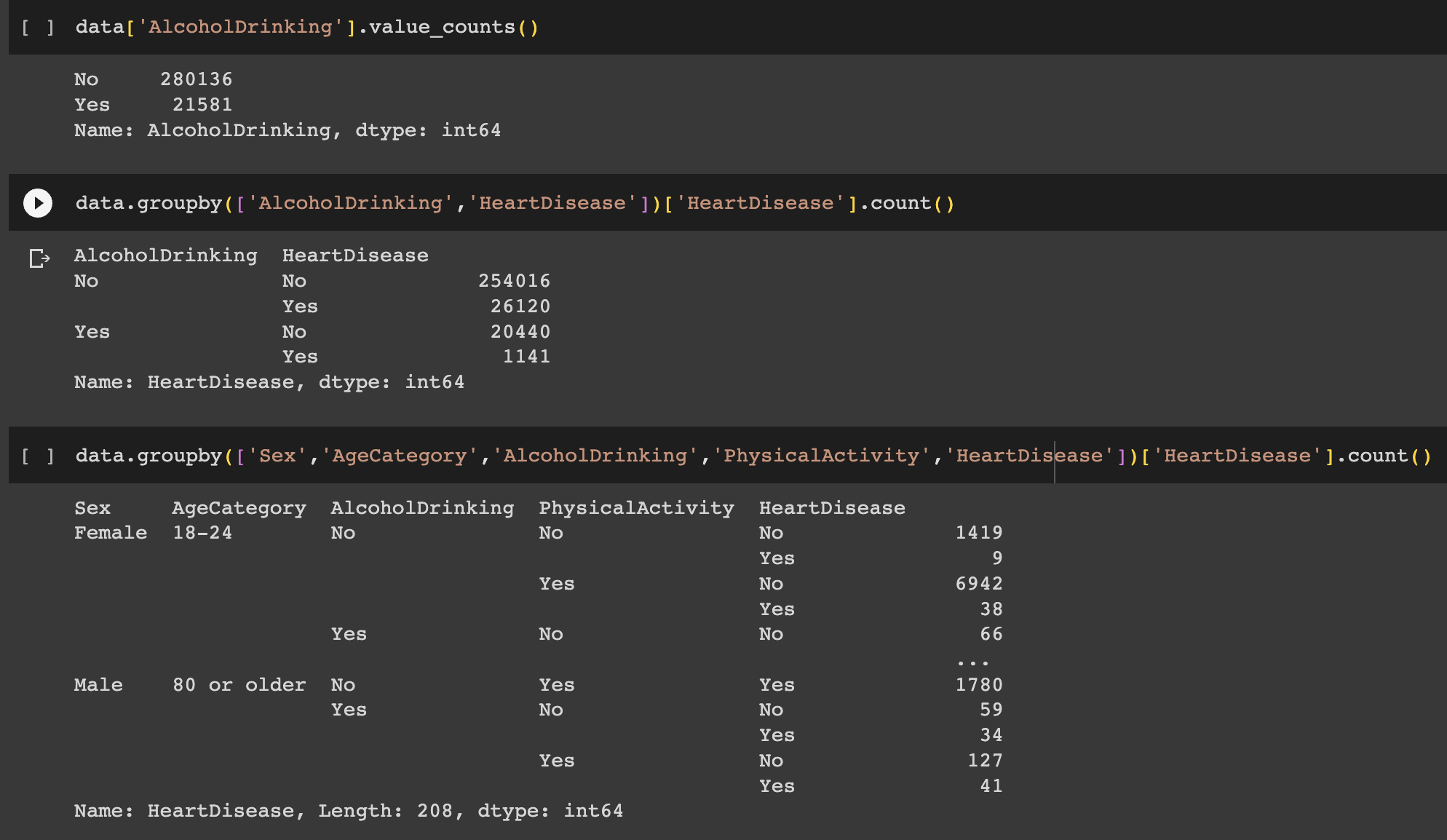
+ Đưa ra số nhóm độ tuổi và số lượng từng nhóm, gom các cột có giá trị là “AgeCategory” và “HeartDisease”, đếm giá trị các thuộc tính và vẽ biểu đồ thể hiện

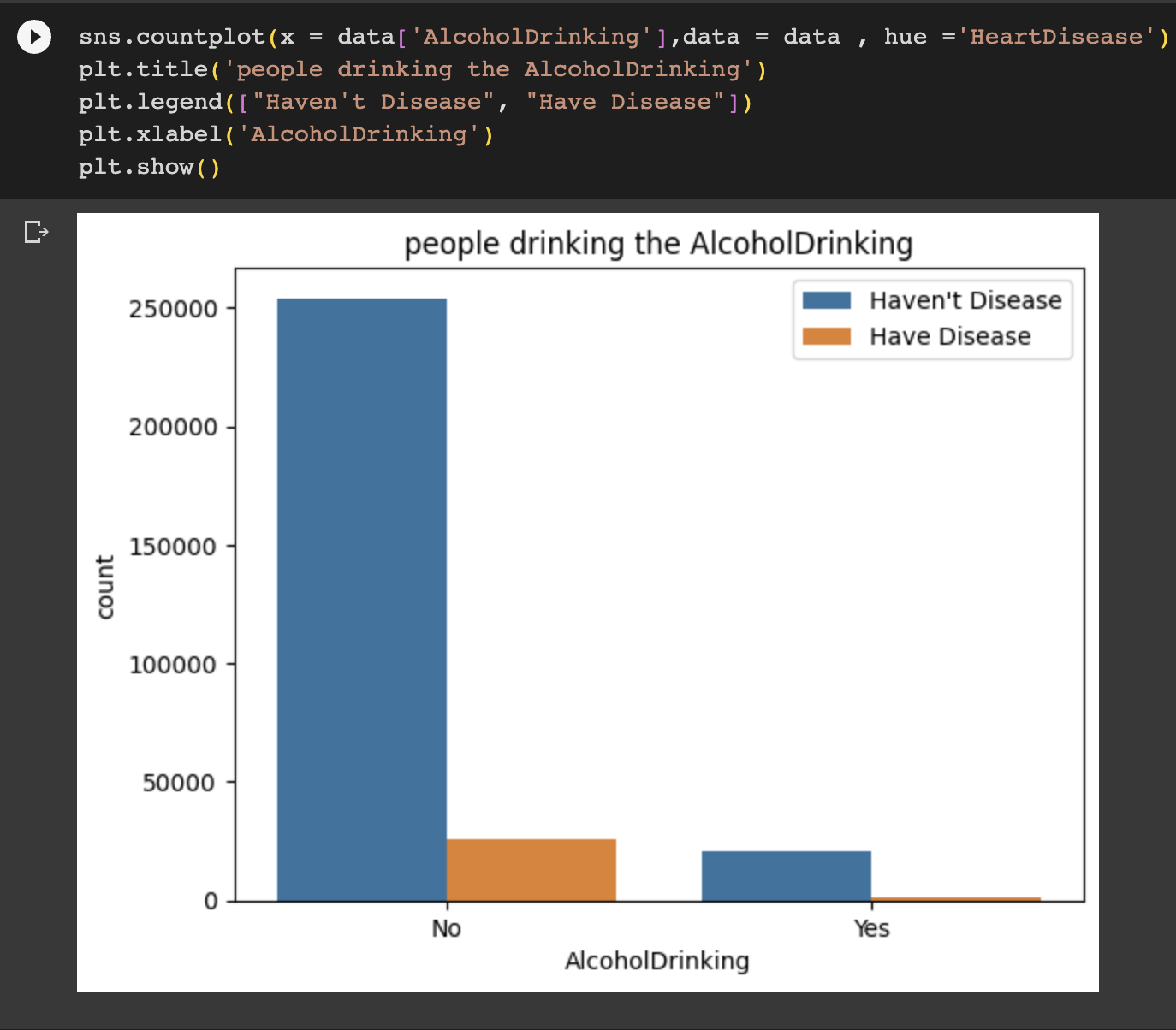


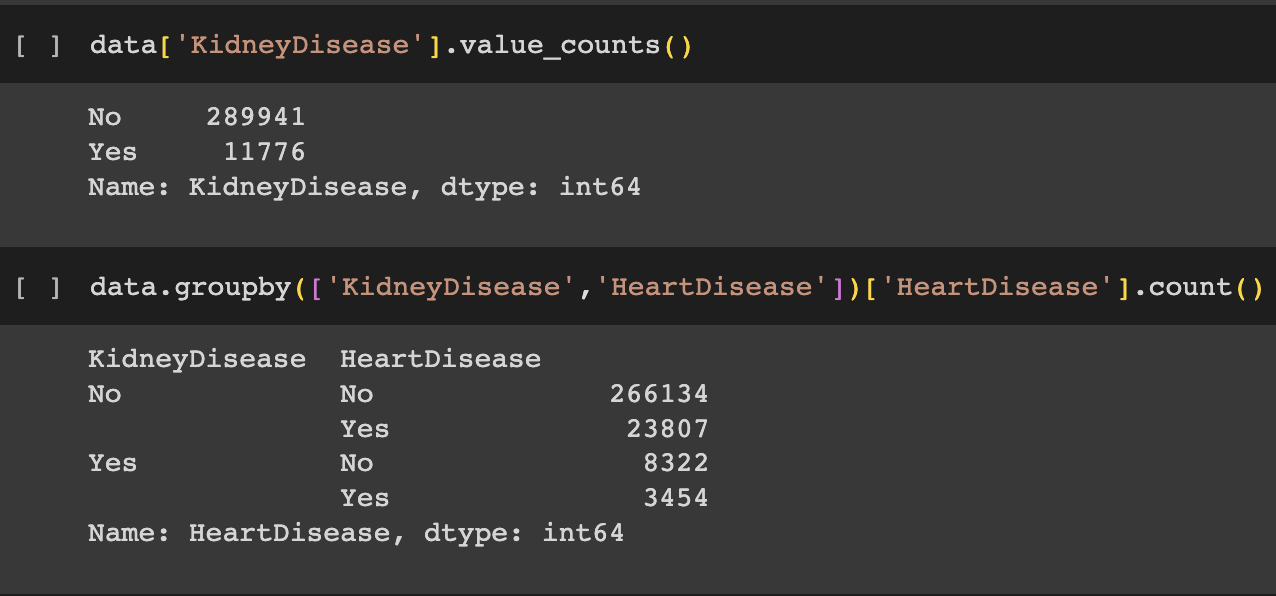




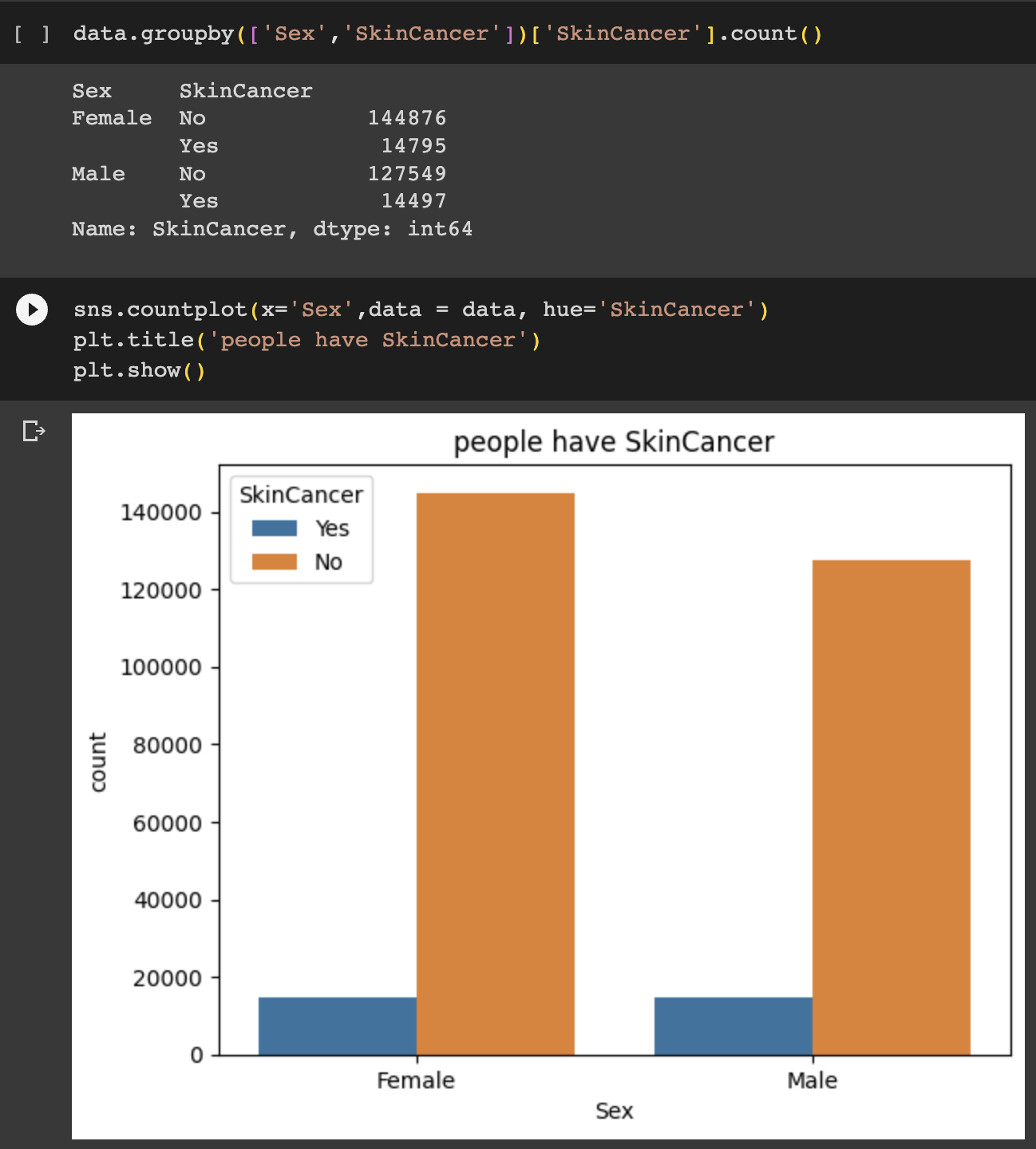


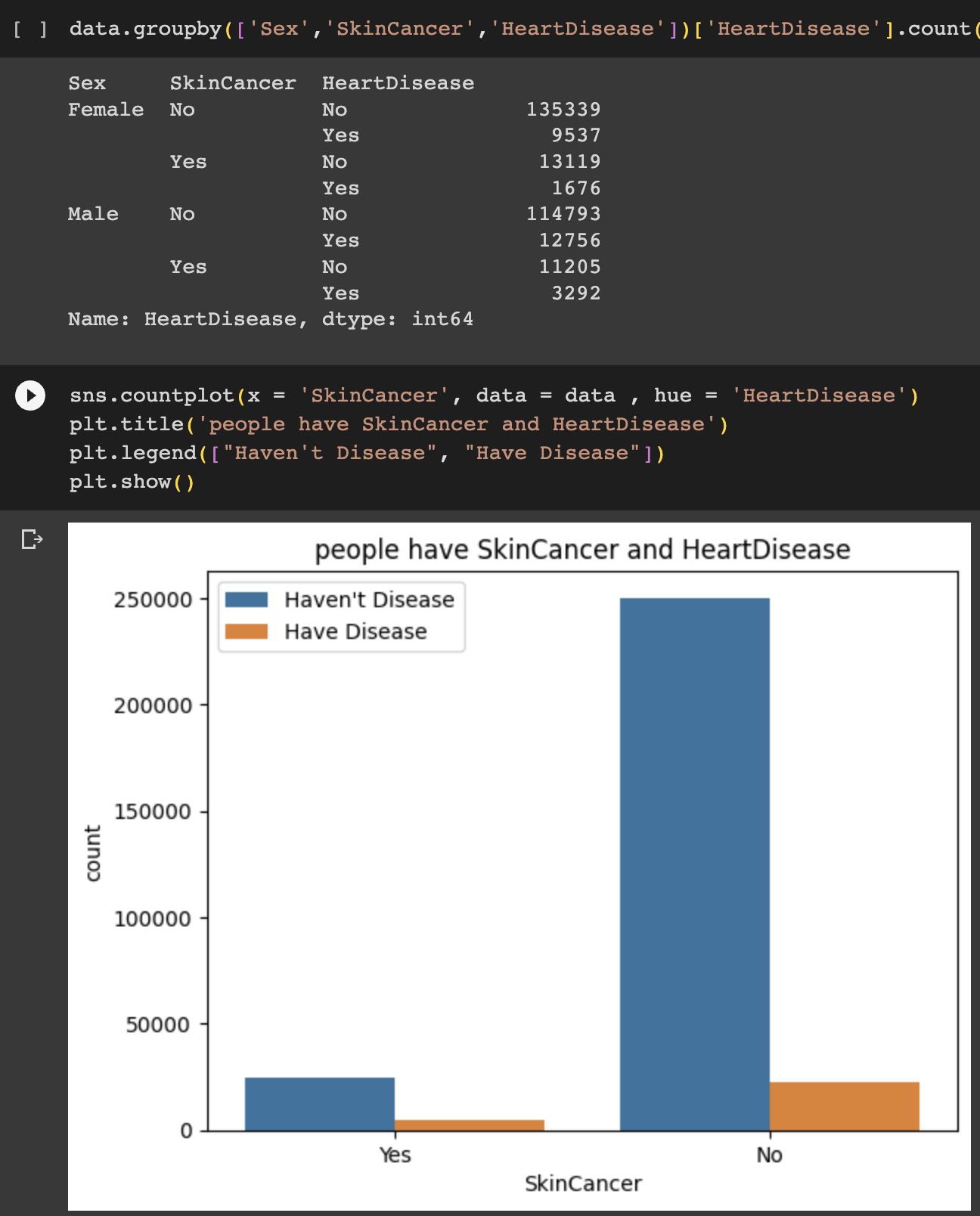


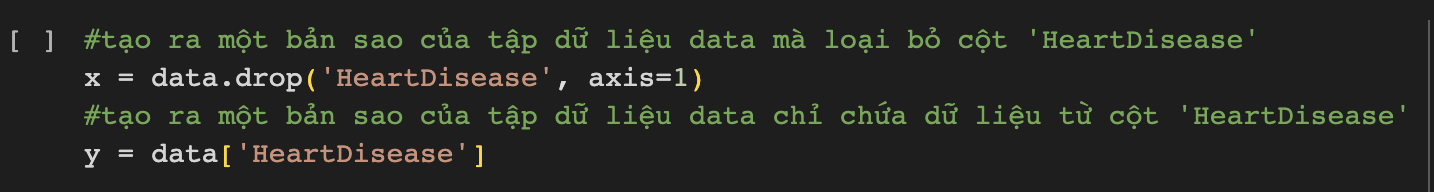




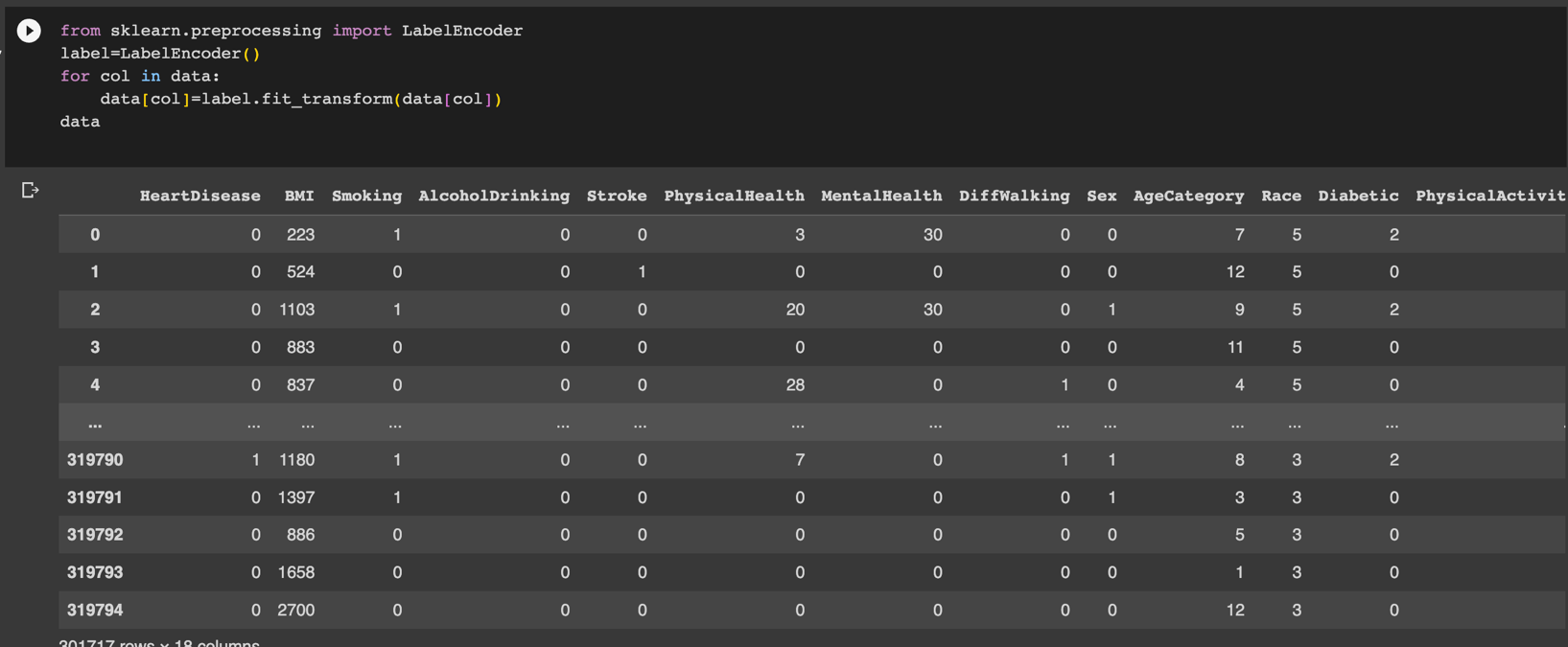




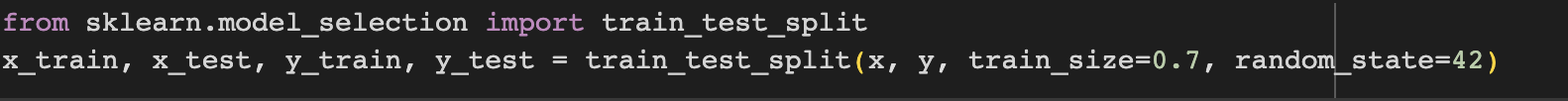


+ Phân chia data

+ Thực hiện mã hóa nhãn.



+ Thực hiện chia tập dữ liệu thành hai phần: tập huấn luyện và tập kiểm tra.



## Giải quyết bài toán phân loại được đặt ra:

* Xây dựng Cây quyết định: Decision Tree là một mô hình dự đoán dựa trên việc xây dựng cây dựa trên các quy tắc logic từ dữ liệu huấn luyện. Cây này sẽ phân chia dữ liệu thành các nhóm dựa trên các đặc trưng của chúng.
* Random Forest Classifier: RandomForestClassifier là một mô hình dựa trên một tập hợp của các cây quyết định ngẫu nhiên. Mỗi cây sẽ được huấn luyện trên một phần của dữ liệu và sẽ đưa ra dự đoán riêng. Kết quả cuối cùng sẽ dựa trên sự kết hợp của các dự đoán từ các cây con.
* Sử dụng K-Nearest Neighbors (KNN): KNN là một thuật toán phân loại dựa trên việc tìm kiếm các điểm dữ liệu gần nhất trong không gian đặc trưng. Chúng ta có thể huấn luyện mô hình KNN bằng cách cung cấp số lân cận (k) thích hợp. Mô hình sẽ phân loại một công việc là bệnh tim hay không dựa trên kết quả của các thuộc tính tương tự.
* Đánh giá mô hình: Sau khi huấn luyện các mô hình, chúng ta cần đánh giá hiệu suất của chúng trên tập dữ liệu kiểm tra. Điều này giúp đảm bảo rằng mô hình không chỉ tốt trên dữ liệu huấn luyện mà còn có khả năng tổng quát hóa cho dữ liệu mới.
* Trong bài toán này, việc đánh giá mô hình là một bước quan trọng. Điều này có thể được thực hiện bằng cách tính toán độ chính xác (accuracy), ma trận nhầm lẫn (confusion matrix) để xác định số lượng dự đoán đúng và sai, cũng như các độ đo khác như độ chính xác cân bằng (balanced accuracy) hoặc độ đo F

# CHƯƠNG 3: GIỚI THIỆU VỀ RANDOM FOREST CLASSIFIER

## 3.1 Giới thiệu về Decision Tree

Cây quyết định ([Decision Tree](https://trituenhantao.io/tu-dien-thuat-ngu/decision-tree/)) là một cây phân cấp có cấu trúc được dùng để phân lớp các đối tượng dựa vào dãy các luật. Các thuộc tính của đối tượngncó thể thuộc các kiểu dữ liệu khác nhau như Nhị phân (Binary) , Định danh (Nominal), Thứ tự (Ordinal), Số lượng (Quantitative) trong khi đó thuộc tính phân lớp phải có kiểu dữ liệu là Binary hoặc Ordinal.

## 3.2 Ưu nhược điểm

* Ưu điểm

+ Dễ hiểu và dễ giải thích:

+ Khả năng xử lý dữ liệu không hoàn hảo

+ Khả năng làm việc với các loại dữ liệu

+ Tính tương đối nhanh trong việc dự đoán

* Nhược điểm:

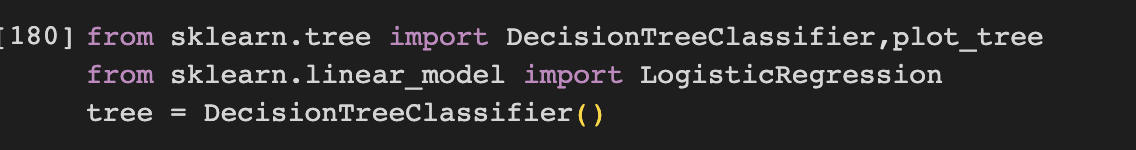
+ Không ổn định

+ Không tạo ra dự đoán tốt trong các tình huống phức tạp

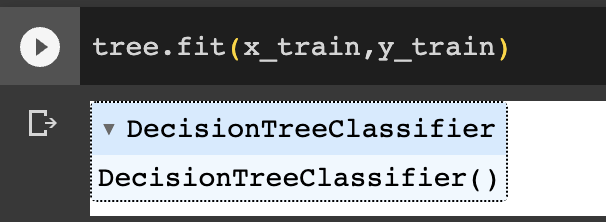
+ Khả năng tạo ra cây quá sâu

## 3.3 Thực hiện:

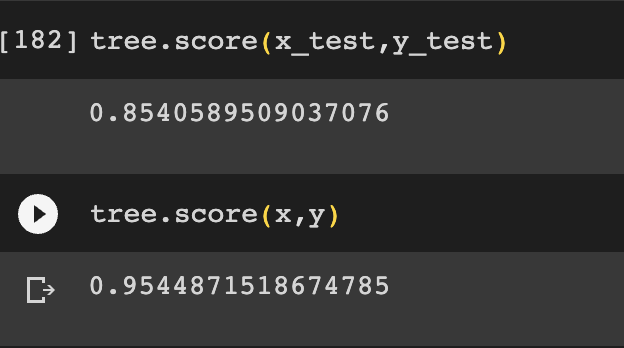
* Import thư viện cho thuật toán



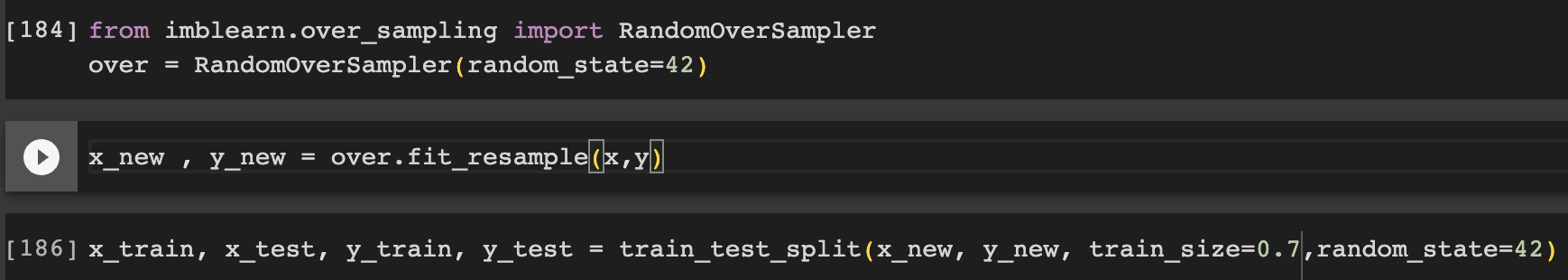
* Huấn luyện tập mô hình

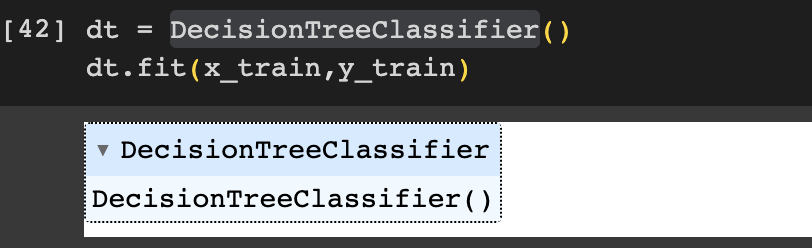


* Đánh giá hiệu xuất tận huấn luyện dựa trên tập huấn luyện mô hình và tạp huấn luyện kiểm tra



* Sữ lý dữ liệu mất cân bằng

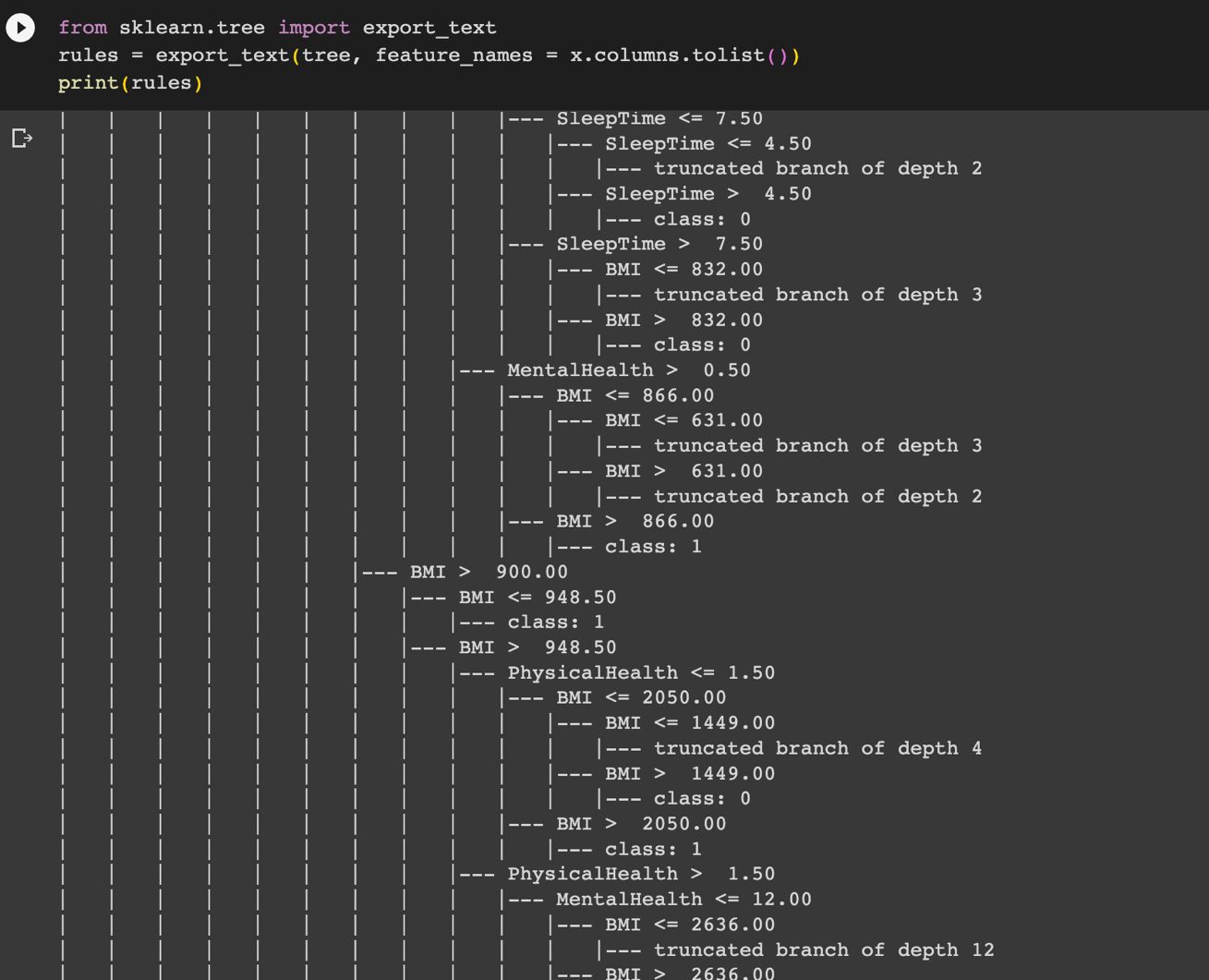




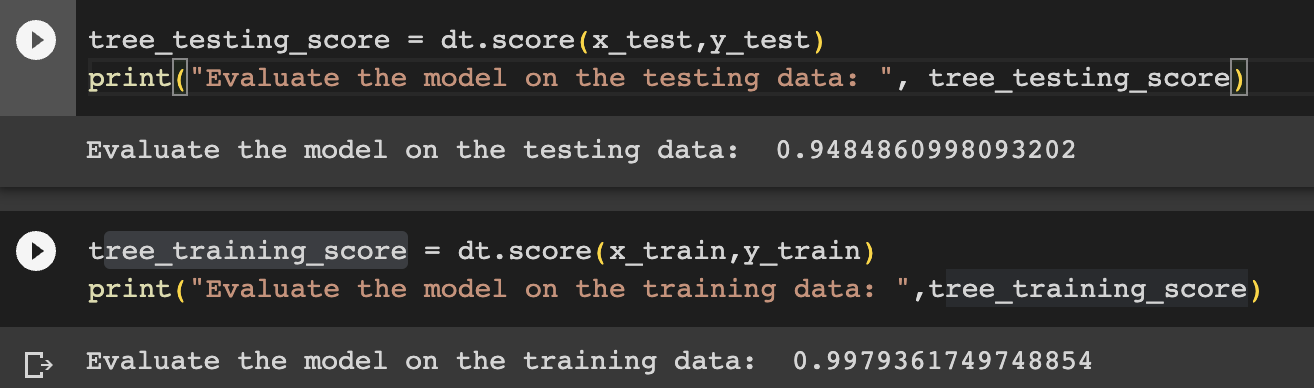
+ dt = DecisionTreeClassifier(): Tạo một đối tượng của lớp DecisionTreeClassifier. Tạo ra một mô hình Decision Tree chưa được cấu hình.

+ dt.fit(x\_train, y\_train): thực hiện việc huấn luyện mô hình Decision Tree bằng cách sử dụng dữ liệu huấn luyện. x\_train chứa các đặc trưng của dữ liệu huấn luyện và y\_train chứa nhãn tương ứng của dữ liệu huấn luyện. Quá trình này sẽ giúp mô hình học cách xây dựng các quy tắc phân loại dựa trên đặc trưng và nhãn của dữ liệu huấn luyện

* Mô tả cách mô hình Decision Tree đưa ra các quyết định dựa trên giá trị của các đặc trưng



* Sử dụng mô hình Decision Tree đã được huấn luyện để đánh giá hiệu suất của mô hình trên tập dữ liệu kiểm tra



# CHƯƠNG 4: GIỚI THIỆU VỀ RANDOM FOREST CLASSIFIER

## 4.1 Giới thiệu về Random Forest Classifier

Thuật toán Random Forest Classifier dựa trên cơ sở của cây quyết định (decision trees). Một cây quyết định là một cấu trúc cây có các nút đại diện cho các quyết định và các lá lá của cây đại diện cho các lớp phân loại. Tuy nhiên, một cây quyết định đơn lẻ có thể dễ dàng bị overfitting (quá khớp) cho dữ liệu huấn luyện và không thể tổng quát hóa tốt cho dữ liệu mới.

Random Forest Classifier giải quyết vấn đề này bằng cách tạo ra nhiều cây quyết định ngẫu nhiên, sau đó kết hợp dự đoán từ các cây này để đưa ra kết quả phân loại cuối cùng. Quá trình này giúp làm giảm sự quá khớp và tăng tính tổng quát hóa của mô hình.

## 4.2 Ưu nhược điểm

* Ưu điểm

+ Tính tổng quát hóa tốt

+ Khả năng xử lý dữ liệu nhiễu

+ Tính đa dạng trong dự đoán

+ Xác định tầm quan trọng của thuộc tính

+ Khả năng làm việc với dữ liệu lớn

* Nhược điểm:

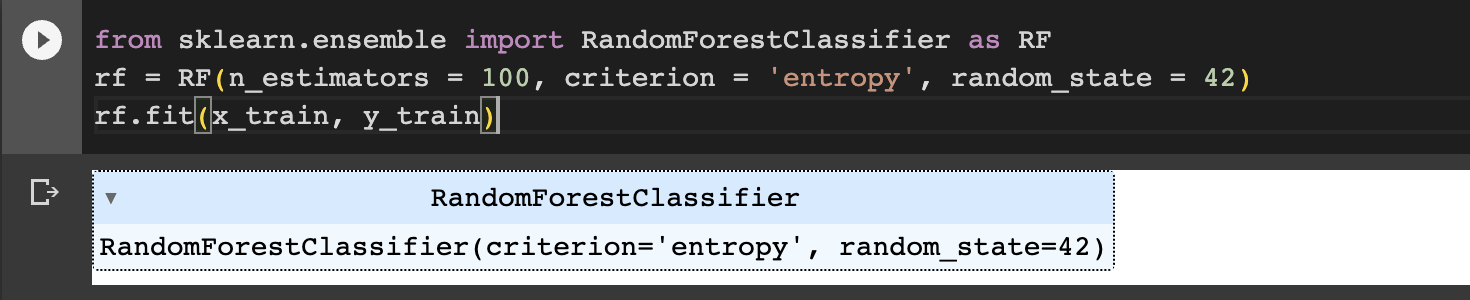
+ Tốn thời gian huấn luyện

+ Phức tạp về hiểu và định cấu hình

+ Khó khả năng mở rộng

## 4.3 Thực hiện:

* Import thư viện RandomForestClassifier cho thuật toán

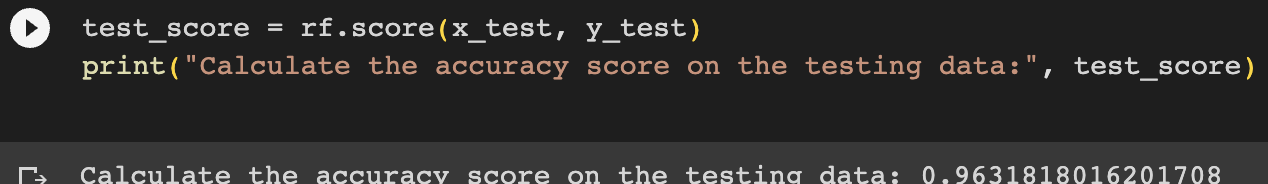


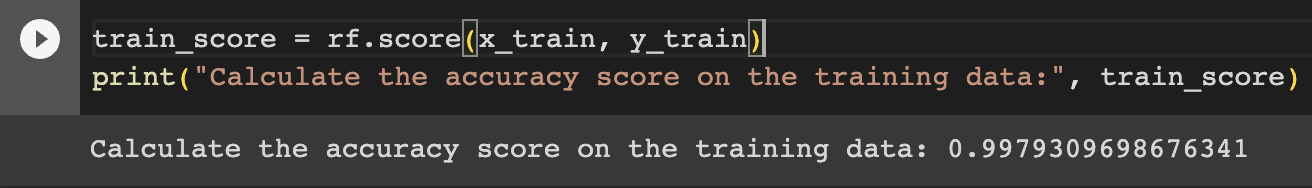
+ n\_estimators: Số lượng cây quyết định trong rừng. Trong trường hợp này, sử dụng 100 cây.

+ criterion: Tiêu chí sử dụng để đo lường chất lượng của việc chia nút trong cây quyết định

+ random\_state: Seed để kiểm soát việc tạo số ngẫu nhiên giúp đảm bảo kết quả của mô hình.

* Sử dụng mô hình RandomForestClassifier đã được huấn luyện để đánh giá hiệu suất của mô hình trên tập dữ liệu kiểm tra





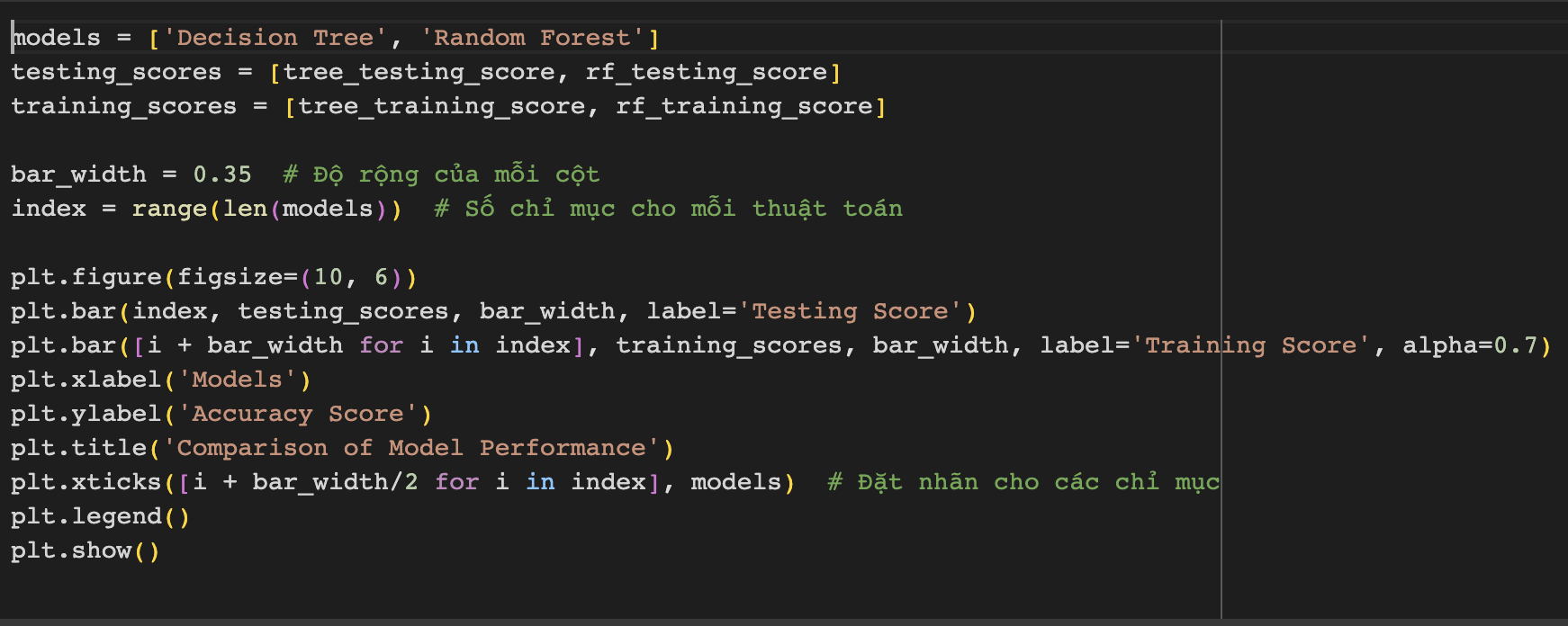
# CHƯƠNG 5: SO SÁNH KẾT LUẬN

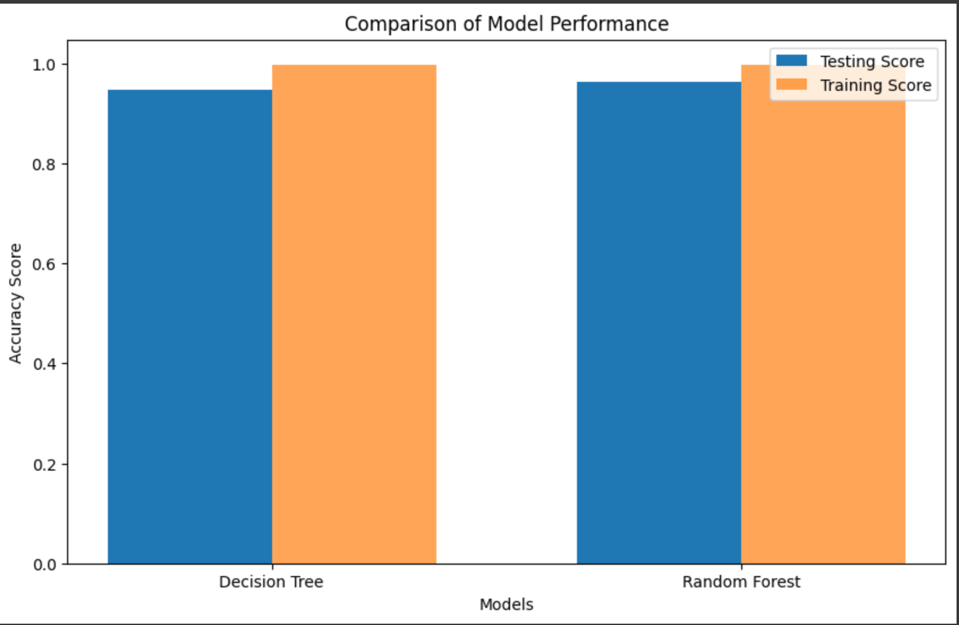
## 5.1. So Sánh:

|  |  |  |
| --- | --- | --- |
| Thuật toán | testing\_score | training\_score |
| DecisionTreeClassifier | 0.9484860998093202 | 0.9979361749748854 |
| RandomForestClassifier | 0.9631818016201708 | 0.9979309698676341 |

Cả hai thuật toán đều có độ chính xác cao trên cả tập dữ liệu kiểm tra và huấn luyện, với điểm training\_score cận 1.

* RandomForestClassifier có điểm testing\_score cao hơn so với DecisionTreeClassifier, tức là nó có khả năng tổng quát hóa tốt hơn cho các dữ liệu không được sử dụng trong quá trình huấn luyện.
* Sự chênh lệch giữa điểm training\_score và testing\_score của RandomForestClassifier thấp hơn so với DecisionTreeClassifier, cho thấy mô hình RandomForest có xu hướng tránh overfitting hơn.





Bottom of Form

5.2. Kết Luận:

Phân lớp dữ liệu là lĩnh vực khá quan trọng trong khai phá dữ liệu nhằm đưa ra những dự đoán, xu hướng trong tương lai, nó được ứng dụng trong nhiều ngành như y tế, thương mại, ngân hàng, giáo dục, nông nghiệp ... Trong quá trình tìm hiểu và hoàn thành bài tập lớn với đề tài “Khai phá dữ liệu phân loại Bệnh Tim”. Nhóm em đã đạt được một số kết quả như sau:

* Tìm hiểu tổng quan về khai phá dữ liệu, bài toán phân lớp,… hỗ trợ phân loại các bài đăng tuyển dụng thật hay giả.
* Thu thập dữ liệu trên kaggle, tiền xử lý dữ liệu bằng thư viện pandas trong python.
* Xây dựng nên mô hình phân lớp trên google collaboration.
* So sánh kết quả tỷ lệ train/test để lựa chọn tỷ lệ đánh giá mô hình tốt nhất.
* Đánh giá các mô hình phân lớp khác nhau.
* Tìm hiểu về ngôn ngữ Python các các thư viện hỗ trợ

**Hạn chế:**

* Thời gian triển khai ngắn, nên chúng em còn nhiều sơ sót và kiến thức của chúng em về mặt toán học và khả năng lập trình còn hạn chế.

# TÀI LIỆU THAM KHẢO

1. Th.S Vũ Thị Hạnh, (2021), Slide Bài giảng Khai Phá dữ liệu - Datamining.
2. <https://cuongndh.blogspot.com/p/khai-pha-du-lieu.html>
3. <https://viblo.asia/p/gioi-thieu-ve-support-vector-machine-svm-6J3ZgPVElmB>
4. <https://codelearn.io/sharing/thuat-toan-k-nearest-neighbors-knn>
5. https:[//www.kagg](http://www.kaggle.com/code)l[e.co](http://www.kaggle.com/code)m[/code](http://www.kaggle.com/code)