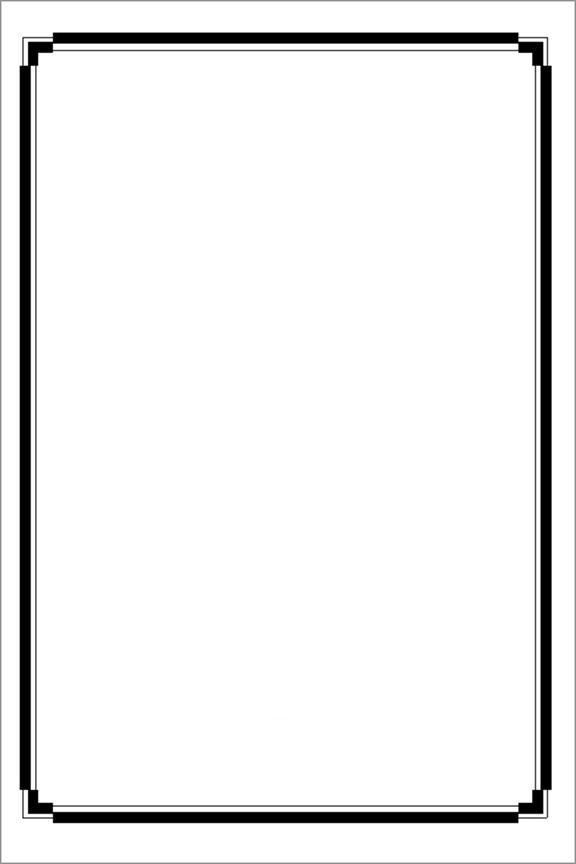
**PHÂN HIỆU TRƯỜNG ĐẠI HỌC THỦY LỢI**

---oOo---



*Báo cáo đồ án môn học*

***MACHINE LEARNING AND DATA MINING***

**Giám sát rủi ro sức khỏe của các bà mẹ**

(Maternal Health Risk Data Set)

**Giảng viên hướng dẫn:** Vũ Thị Hạnh

**Sinh viên:** Võ Nguyên Vinh – 2051067519

Đặng Bảo Trâm – 2054037820

Phạm Duy Tân – 2051067230

Lê Thị Kim Anh – 2051067526

Trần Ngọc Hải - 2051067527

**Năm học: 2022-2023**

MỤC LỤC

[**Lời mở đầu** 3](#_Toc117880663)

[**Lời cảm ơn** 4](#_Toc117880664)

[**DANH MỤC CÁC THUẬT NGỮ** 5](#_Toc117880665)

[**CHƯƠNG I: TỔNG QUAN VỀ ĐỒ ÁN** 6](#_Toc117880666)

[**I.1 Giới thiệu đồ án** 6](#_Toc117880667)

[**I.2 Mục tiêu của đồ án** 9](#_Toc117880668)

[**I.3 Dự kiến kết quả đạt được** 10](#_Toc117880669)

[**I.4 Công cụ thực hiện đồ án** 10](#_Toc117880670)

[**CHƯƠNG 2: CƠ SỞ LÝ THUYẾT** 10](#_Toc117880671)

[**II.1 Giới thiệu về môn học** 10](#_Toc117880672)

[*II.1.1 Machine learning* 10](#_Toc117880673)

[*II.1.2 Data mining* 11](#_Toc117880674)

[**II.2 Kho dữ liệu** 11](#_Toc117880675)

[**II.3 Tổng quan các phương pháp sử dụng** 11](#_Toc117880676)

[***II.3.1 Cây quyết định (Decision Tree)*** 11](#_Toc117880677)

[***II.3.2 Thuật toán DBSCAN*** 12](#_Toc117880678)

[**CHƯƠNG 3: CHƯƠNG TRÌNH THỰC HIỆN** 13](#_Toc117880679)

[**III.1 Quản lý thư viện và dữ liệu** 13](#_Toc117880680)

[**III.2 Danh sách biểu đồ hiển thị tất cả các thuộc tính trong tập dữ liệu** 15](#_Toc117880681)

[**III.3 Bộ chuyển đổi dữ liệu** 19](#_Toc117880682)

[**III.4 Thực hiện thuật toán Cây quyết định (Decision tree)** 20](#_Toc117880683)

[**III.5 Thực hiện thuật toán DBSCAN** 22](#_Toc117880684)

[**CHƯƠNG 4: KẾT LUẬN** 26](#_Toc117880685)

[**TÀI LIỆU THAM KHẢO** 27](#_Toc117880686)

# **Lời mở đầu**

Trong thời đại bùng nổ công nghệ thông tin, vai trò của hệ thống thông tin ngày càng trở nên quan trọng đối với mỗi lĩnh vực từ Doanh nghiệp, cửa hàng, trang trại, y học….

Và hệ thống thông tin y tế (HIS) được thiết kế để quản lý dữ liệu chăm sóc sức khỏe. Nó bao gồm các hệ thống thu thập, lưu trữ, quản lý và truyền hồ sơ y tế điện tử của bệnh nhân; quản lý vận hành của bệnh viện; và hệ thống hỗ trợ các quyết định chính sách chăm sóc sức khỏe.

Phần lớn hệ thống thông tin đòi hỏi phải duy trì toàn bộ dữ liệu từ quá khứ, không ngừng bổ sung thêm các dữ liệu mới được ghi nhận tại hiện tại và dự đoán kết quả sẽ xuất hiện trong tương lai.

Ứng dụng cơ sở dữ liệu và ý nghĩa nhân văn của nó, nhóm tụi em đã thực hiện đề tài phân tích dữ liệu trong tập dữ liệu rủi ro sức khỏe bà mẹ nhầm thống kê, đánh giá và phân tích các yếu tố rủi ro của sức khỏe bà mẹ trong thời kì mang thai thông qua hệ thống giám sát rủi ro trên IoT, có khả năng hỗ trợ và dự báo mức độ rủi ro dựa trên IoT để chăm sóc sức khỏe bà mẹ.

Chúng em đã nghiên cứu về đề tài “ **Giám sát rủi ro sức khỏe của các bà mẹ** ”, để làm báo cáo cho môn học của mình.

# **Lời cảm ơn**

Đầu tiên, nhóm tụi em xin gửi lời cảm ơn chân thành đến tập thể quý Thầy Cô của Phân hiệu trường Đại học Thủy Lợi và quý Thầy Cô khoa Công nghệ thông tin đã giúp cho nhóm tụi em có những kiến thức cơ bản làm nền tảng để xây dựng đề tài này.

Đặc biệt, nhóm tụi em gửi lời cảm ơn sâu sắc và chân thành nhất đến Giảng viên Vũ Thị Hạnh (giảng viên môn Học máy). Cô đã trực tiếp hướng dẫn tận tình, sửa chữa và đóng góp ý kiến quý báo giúp nhóm tụi em hoàn thành tốt báo cáo đồ án của mình. Trong thời gian thực hiện đề tài, nhóm tụi em đã vận dụng những kiến thức nền tảng đã tích lũy đồng thời kết hợp với việc học hỏi và nghiên cứu những kiến thức mới. Từ đó, nhóm tụi em đã vận dụng tối đa những gì đã thu thập được để hoàn thành đồ án một cách tốt nhất.

Tuy nhiên, trong quá trình thực hiện không tránh khỏi những thiếu sót. Chính vì vậy, nhóm tụi em rất mong nhận được những sự góp ý từ quý Thầy Cô nhằm hoàn thiện những kiến thức mà nhóm tụi em đã học tập và cũng là hành trang để nhóm tụi em thực hiện tiếp các đề tài khác trong tương lai.

Xin chân thành cảm ơn các quý Thầy Cô !

# **DANH MỤC CÁC THUẬT NGỮ**

|  |  |  |
| --- | --- | --- |
| **Thuật ngữ viết tắt** | **Thuật ngữ đầy đủ** | **Ý nghĩa** |
| Age |  | Tuổi |
| SystolicBP | Systolic blood pressure | Huyết áp trên |
| DiastolicBP | Diastolic blood pressure | Huyết áp thấp |
| BS | Blood sugar | Nồng độ đường huyết |
| HeartRate |  | Nhịp tim |
| Risk Level |  | Mức độ rủi ro |
| Body temp | Body temperature | Nhiệt độ cơ thể (thân nhiệt) |
| CSDL |  | Cơ sở dữ liệu |
| IoT | Internet of Things | Internet vạn vật |
| LMT | Logistic Model Tree | Mô hình phân loại |
| CNTT |  | Công nghệ thông tin |
| HTTT |  | Hệ thống thông tin |
| CNPM |  | Công nghệ phần mềm |
| AI | artificial intelligence | Trí tuệ nhân tạo |

# **CHƯƠNG I: TỔNG QUAN VỀ ĐỒ ÁN**

## **I.1 Giới thiệu đồ án**

* *Tổng quan bài toán*

Sức khỏe là vốn quý giá của con người, đặc biệt là sức khỏe của các bà mẹ trong quá trình mang thai con nhỏ, việc chăm sóc và quan tâm đến đang được mọi người chú trọng và giám sát chặt chẽ.

Tuy nhiên, nhiều người vẫn ỷ y hoặc chỉ đang chú trọng về mặt dinh dưỡng mà không quan tâm nhiều đến các vấn đề khác, mà vô tình gây ra những hậu quả nghiêm trọng về sau (tỉ lệ tử vong của con, mẹ tăng; sức khỏe mẹ giảm súc, con đề kháng yếu…).

Chính vì vậy, để hiểu rõ và xác định được tình trạng sức khỏe của các bà mẹ (mức rủi ro), nhóm đã thống nhất và quyết định chọn đề tài phân tích, khai thác, đánh giá dữ liệu về các đặc tính trực tiếp ảnh hưởng đến sức khỏe bà mẹ.

Việc chăm sóc sức khỏe cho bà bầu là việc làm ý nghĩa và cần thiết để đảm bảo em bé sinh ra mẹ tròn con vuông.



**Tỷ lệ tử vong bà mẹ ở các nước trên thế giới (2017)**

* *Giới thiệu chung*

+ Dữ liệu đã được thu thập từ các bệnh viện, phòng khám cộng đồng, cơ sở chăm sóc khác nhau đưa ra mức độ rủi ro về sức khỏe bà mẹ thông qua hệ thống giám sát rủi ro dựa trên IoT.

+ IoT là sự đổi mới khéo léo lớn nhất trong kỷ nguyên hiện đại, có thể khai thác cả trong những sứ mệnh quan trọng như ngành chăm sóc sức khỏe.

Bài báo cáo này thể hiện việc giám sát hiệu quả đối với phụ nữ mang thai chủ yếu ở khu vực nông thôn của một nước đang phát triển, với sự trợ giúp của công nghệ hỗ trợ cảm biến có thể đeo, đồng thời thông báo cho phụ nữ mang thai và gia đình về tình trạng sức khỏe. Có nhiều nhà nghiên cứu đã nghiên cứu để giảm tỷ lệ tử vong mẹ và thai nhi nhưng tỷ lệ tử vong không giảm, ở mức không khoan nhượng. Nghiên cứu này nhằm mục đích sử dụng các thuật toán học máy để khám phá mức độ rủi ro trên cơ sở các yếu tố nguy cơ trong thai kỳ.

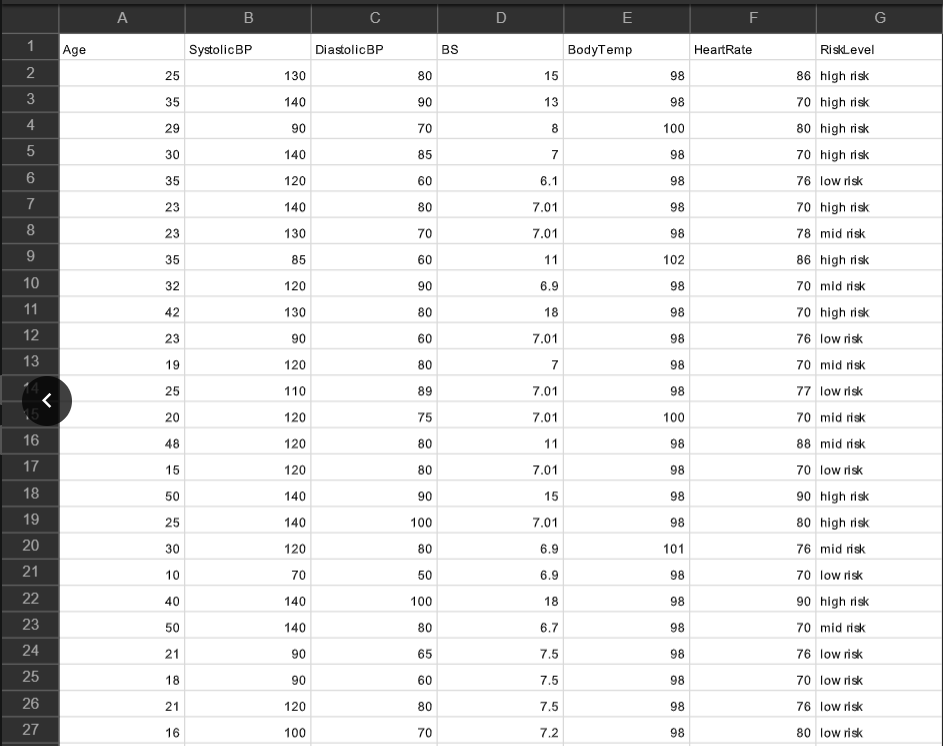
Trong nghiên cứu này, một bộ dữ liệu hiện có (bộ dữ liệu Pima-Indian-ĐTĐ) đã được sử dụng để phân tích yếu tố nguy cơ và so sánh một số thuật toán học máy cho thấy rằng Logistic Model Tree (LMT) cho độ chính xác cao nhất trong trường hợp phân loại và dự đoán mức độ rủi ro.

+ Mặc dù vậy, một số dữ liệu phụ nữ mang thai được chọn đã được thu thập (thông qua các thiết bị hỗ trợ IoT) và quy trình tương tự cũng được áp dụng cho bộ dữ liệu này bằng cách sử dụng LMT. Kết quả so sánh cho thấy dự đoán rủi ro là giống nhau đối với tập dữ liệu hiện có và thực tế.

+ Link data set: [Maternal Health Risk Data Set](https://archive.ics.uci.edu/ml/datasets/Maternal+Health+Risk+Data+Set?fbclid=IwAR2LMAcgFdnf5ABb927AX9idDuZvKH76osf3pebOnQXkdHWjnbrQVX4Nc-8)

+ Bộ dữ liệu này gồm 1014 dòng dữ liệu và 7 thuộc tính.

+ Mỗi hàng đại diện cho một bà mẹ, nó được dán nhãn là ở mức rủi ro cao, ở mức rủi ro giữa và ở mức rủi ro thấp và không có giá trị nào bị thiếu. Các cột là những yếu tố có trách nhiệm và đáng kể sẽ ảnh hưởng đến mức độ rủi ro sức khỏe của các bà mẹ. Các nhãn đã được kiểm tra nó là một yếu tố. Không có giá trị nào hoặc không chính xác được quan sát thấy.



**Bộ dữ liệu**

## **I.2 Mục tiêu của đồ án**

* Hiểu được các thuật toán và vận hành các thuật toán đã học vào thực tế.
* Hệ thống lại kiến thức đã học được trong môn Học máy.
* Tiếp cận, nghiên cứu và vận hành được phần mềm khai phá với bộ dữ liệu có sẵn, từ đó hệ thống, phân tích được dữ liệu, áp dụng vào công việc để mang lại hiệu suất tốt nhất.
* Xây dựng, vận hành các thuật toán như: gom cụm, phân lớp, kết hợp…
* Ngoài ra, chúng ta được biết, hiểu thêm về các kiến thức bên ngoài (các thông tin sơ bộ về sức khỏe của bà mẹ trong quá trình mang thai,…).

## **I.3 Dự kiến kết quả đạt được**

* Dự đoán và giám sát được mức độ rủi ro sức khỏe của các bà mẹ trong quá trình mang thai.
* Chạy được thuật toán Cây quyết định và nêu ra được kết luận.
* Chạy được thuật toán DBscan và nêu ra được kết luận.

## **I.4 Công cụ thực hiện đồ án**

* Nhóm tụi em dùng **Colaboratory** hay còn gọi là **Google Colab** để thực hiện đồ án.
* Google Colab là một sản phẩm từ Google Research, nó cho phép chạy các dòng code python thông qua trình duyệt, đặc biệt phù hợp với Data analysis, machine learning và giáo dục.
* Google Colab được Google phát hành đến với công chúng nhằm để cải thiện chất lượng giáo dục. Và bên cạnh đó là việc nghiên cứu Machine Learning

# **CHƯƠNG 2: CƠ SỞ LÝ THUYẾT**

## **II.1 Giới thiệu về môn học**

### *II.1.1 Machine learning*

* **Machine learning** (ML) hay học máy là một nhánh của trí tuệ nhân tạo (AI), nó là một lĩnh vực nghiên cứu cho phép máy tính có khả năng cải thiện chính bản thân chúng dựa trên dữ liệu mẫu (training data) hoặc dựa vào kinh nghiệm (những gì đã được học). Machine learning có thể tự dự đoán hoặc đưa ra quyết định mà không cần được lập trình cụ thể.
* Học máy cũng là học phần cơ sở ngành bắt buộc cho các ngành CNTT, HTTT và CNPM và là kiến thức cần thiết để học các học phần nâng cao liên quan đến kỹ thuật học và cung cấp kiến thức cơ bản về các mô hình (không giám sát và có giám sát); bài toán phân loại, phân cụm, và bài toán hồi quy; các giải thuật học máy cơ bản như hồi quy tuyến tính, K-mean, Gradient, Học Perceptron, Decision tree, Hồi quy Logistic, SVM, Học kết hợp, và phương pháp đánh giá một hệ thống phân lớp.

### *II.1.2 Data mining*

* **Khai phá dữ liệu** (data mining) là quá trình tính toán để tìm ra các mẫu trong các bộ dữ liệu lớn liên quan đến các phương pháp tại giao điểm của máy học, thống kê và các hệ thống cơ sở dữ liệu. Đây là một lĩnh vực liên ngành của khoa học máy tính.
* Khai phá dữ liệu có thể được sử dụng cho các lĩnh vực y tế, phân tích thị trường, xây dựng ... có thể được xem như là kết quả của sự tiến triển tự nhiên của công nghệ thông tin.

## **II.2 Kho dữ liệu**

* Dữ liệu trong một lĩnh vực trước khi vào hệ thống học máy phải được thu thập và biểu diễn thành dạng cấu trúc với một số đặc tính: đầy đủ, ít nhiễu, nhất quán, có cấu trúc xác định.
* Dữ liệu thu thập cho quá trình học tập là nhỏ, tuy vậy cần phản ánh đầy đủ các mặt vấn đề cần giải quyết.
* Dữ liệu thô sau khi thu thập và tiền xử lý phải giữ được sự đầy đủ các đặc trưng ngữ nghĩa – các đặc trưng ảnh hưởng đến khả năng giải quyết vấn đề.
* Khoa học dữ liệu là một lĩnh vực rộng, ngoài việc sử dụng công cụ áp dụng, nắm vững được các kiến thức cơ bản là điều quan trọng.

## **II.3 Tổng quan các phương pháp sử dụng**

### ***II.3.1 Cây quyết định (Decision Tree)***

*- Khái niệm:*

Cây quyết định (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 tượng có 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.

Một cây quyết định là một đồ thị của các quyết định và các hậu quả có thể của nó.

Tóm lại, cho dữ liệu về các đối tượng gồm các thuộc tính cùng với lớp (classes) của nó, cây quyết định sẽ sinh ra các luật để dự đoán lớp của các dữ liệu chưa biết.

*- Ưu điểm và nhược điểm:*

+ Ưu điểm: Cây quyết định là một thuật toán đơn giản và phổ biến. Thuật toán này được sử dụng rộng rãi với những lợi ích của nó: Mô hình sinh ra các quy tắc dễ hiểu cho người đọc, tạo ra bộ luật với mỗi nhánh lá là một luật của cây. Dữ liệu đầu vào có thể là là dữ liệu missing, không cần chuẩn hóa hoặc tạo biến giả. Có thể làm việc với cả dữ liệu số và dữ liệu phân loại. Có thể xác thực mô hình bằng cách sử dụng các kiểm tra thống kê. Có khả năng là việc với dữ liệu lớn.

+ Nhược điểm: Kèm với đó, cây quyết định cũng có những nhược điểm cụ thể: Mô hình cây quyết định phụ thuộc rất lớn vào dữ liệu của bạn. Thậm chí, với một sự thay đổi nhỏ trong bộ dữ liệu, cấu trúc mô hình cây quyết định có thể thay đổi hoàn toàn. Cây quyết định hay gặp vấn đề overfitting.

- *Ứng dụng*

 Cây quyết định được sử dụng để xây dựng một kế hoạch nhằm đạt được mục tiêu mong muốn. Các cây quyết định được dùng để hỗ trợ quá trình ra quyết định. Cây quyết định là một dạng đặc biệt của cấu trúc cây.

### ***II.3.2 Thuật toán DBSCAN***

*- Khái niệm:*

Thuật toán DBSCAN (Density Based Spatial Clustering of Applications with Noise) là thuật toán gom cụm dựa trên mật độ, hiệu quả với CSDL lớn, có khả năng xử lý nhiễu. DBSCAN thực hiện tốt trên không gian nhiều chiều; thích hợp với cơ sở dữ liệu có mật độ phân bố dày đặc kể cả có phần tử nhiễu.

Thuật toán DBSCAN gom cụm các đối tượng trong cơ sở dữ liệu không gian ứng với  
thông số Eps, MinPts cho trước, DBSCAN xác định một cụm thông qua 2 bước:  
1) Chọn đối tượng bất kỳ thỏa mãn điều kiện đối tượng lõi làm đối tuợng hạt giống;  
2) Tìm các đối tượng tới được theo mật độ từ đối tượng hạt giống.  
  
*- Ưu điểm và nhược điểm:*

+ Ưu điểm: Thuật toán DBSCAN đã khắc phục được vấn đề độ phức tạp tính toán cao và dữ liệu nhiễu.

+ Nhược điểm: Nhưng để có thể tìm ra cụm các đối tượng thì người ta vẫn phải chọn tham số Eps và MinPts để tìm ra cụm chính xác. Các thiết lập tham số như vậy thường khó xác định, đặc biệt trong thế giới thực, khi sự thiết lập có sự khác biệt nhỏ có thể dẫn đến sự phân chia cụm là rất khác nhau.

*-Ứng dụng:*

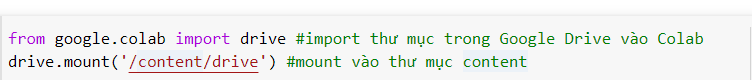
DBSCAN truy cập từng điểm của cơ sở dữ liệu, có thể nhiều lần. DBSCAN tổng quát là sự tổng quát hóa của cùng các tác giả thành các vị từ "lân cận" và "dày đặc" tùy ý.

Còn để song song hóa, ước lượng tham số và hỗ trợ cho dữ liệu không chắc chắn.

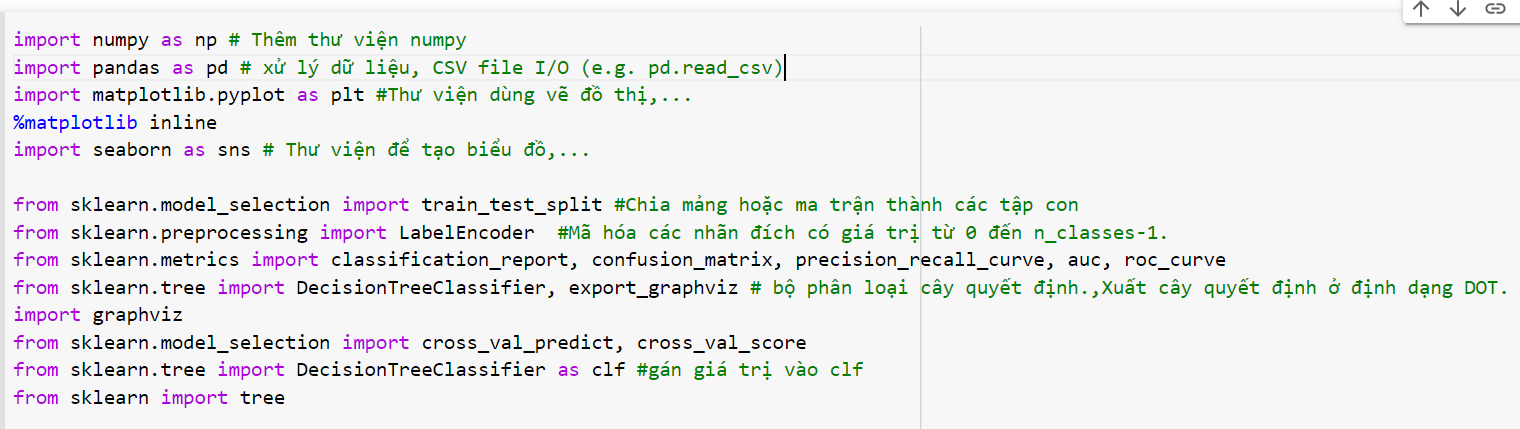
# **CHƯƠNG 3: CHƯƠNG TRÌNH THỰC HIỆN**

## **III.1 Quản lý thư viện và dữ liệu**

- Liên kết thư mục vào gg Colab:



- Thêm thư viện, xử lý dữ liệu gốc, nhập dữ liệu đầu vào:



Comment thêm:

+  **classification\_report :** Xây dựng một báo cáo văn bản hiển thị các chỉ số phân loại.

+  **confusion\_matrix** : Đánh giá độ chính xác của phân loại.

+ **precision\_recall\_curve** : Tính toán độ chính xác cho các ngưỡng xác suất khác nhau.

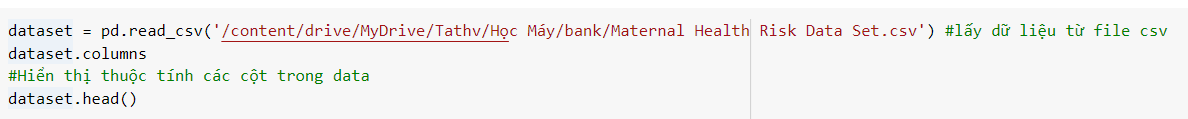
+  **auc** : Tính diện tích dưới đường cong (AUC) bằng cách sử dụng quy tắc hình thang.

+ **roc\_curve**: Khu vực tính toán theo đường cong.

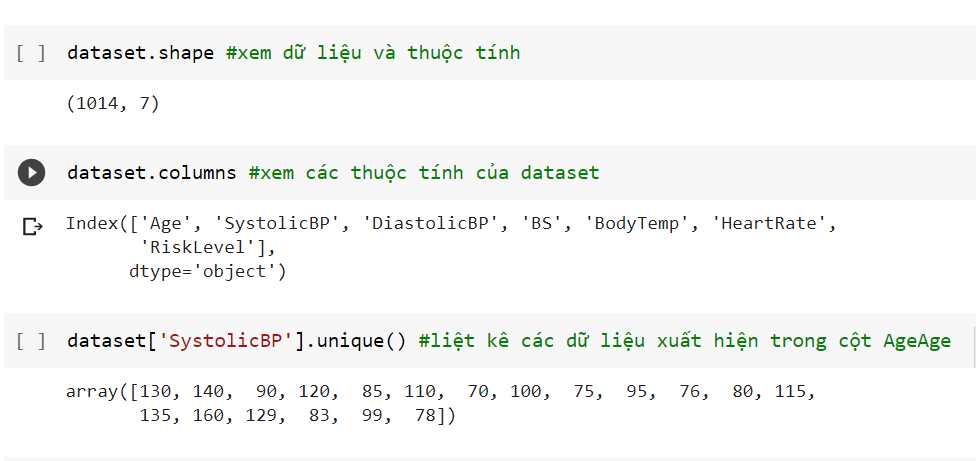
+ **cross\_val\_predict :**Tạo các ước tính được xác thực chéo cho mỗi điểm dữ liệu đầu vào

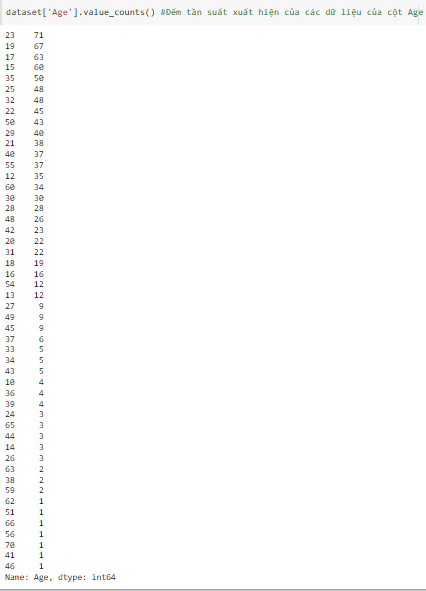
+ **cross\_val\_score**: Đánh giá điểm bằng cách xác nhận chéo.

**-** Liên kết dữ liệu:

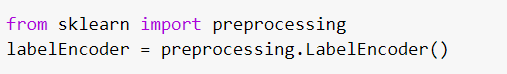


- Làm sạch dữ liệu (Data Cleaning):



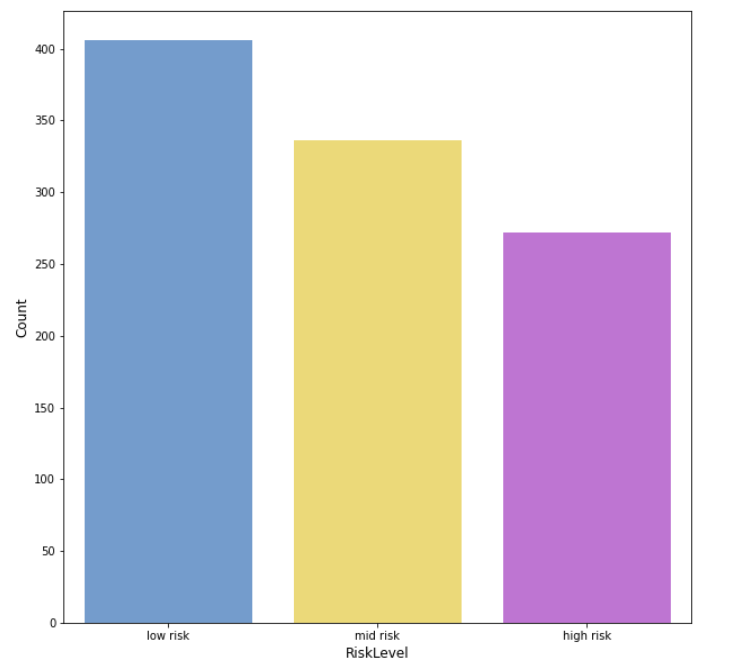


- Tiền xử lý dữ liệu: Thuật toán học được hưởng lợi từ việc chuẩn hóa tập dữ liệu. Nếu có một số ngoại lệ trong tập hợp, các bộ điều chỉnh hoặc biến áp mạnh mẽ sẽ thích hợp hơn.



## **III.2 Danh sách biểu đồ hiển thị tất cả các thuộc tính trong tập dữ liệu**

**-** Biểu đồ đếm số lượng mức độ rủi ro (cao/giữa/thấp):

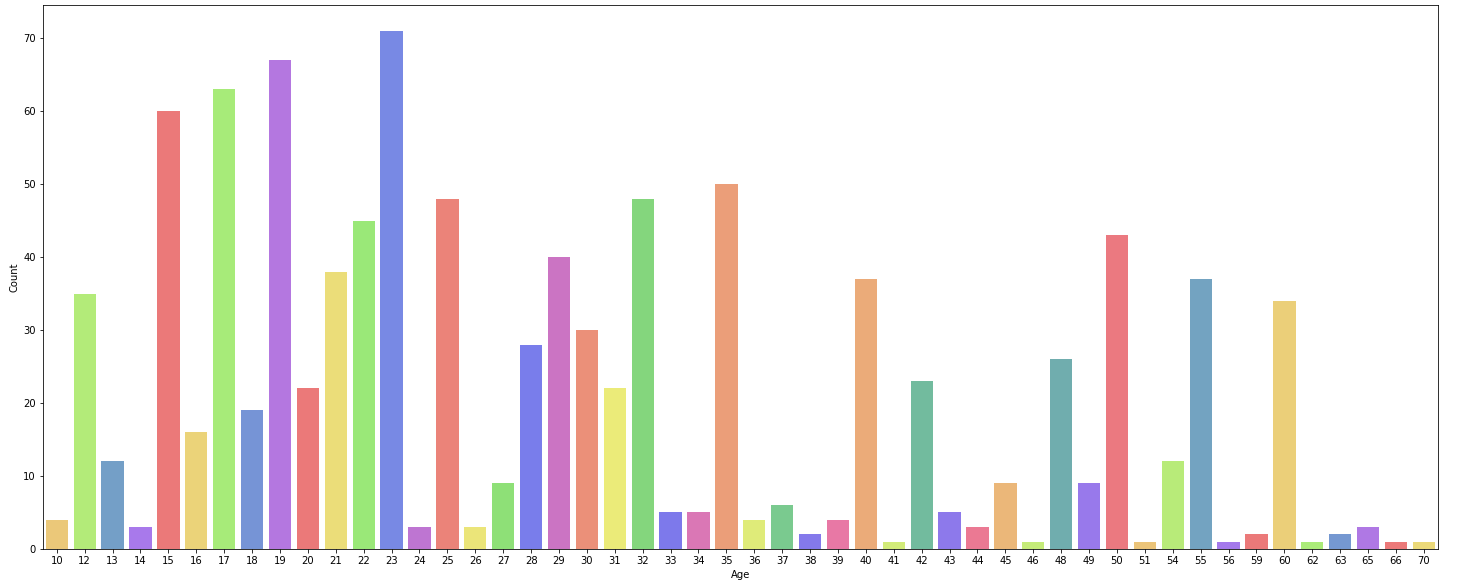


**Biểu đồ đếm số mức độ rủi ro**

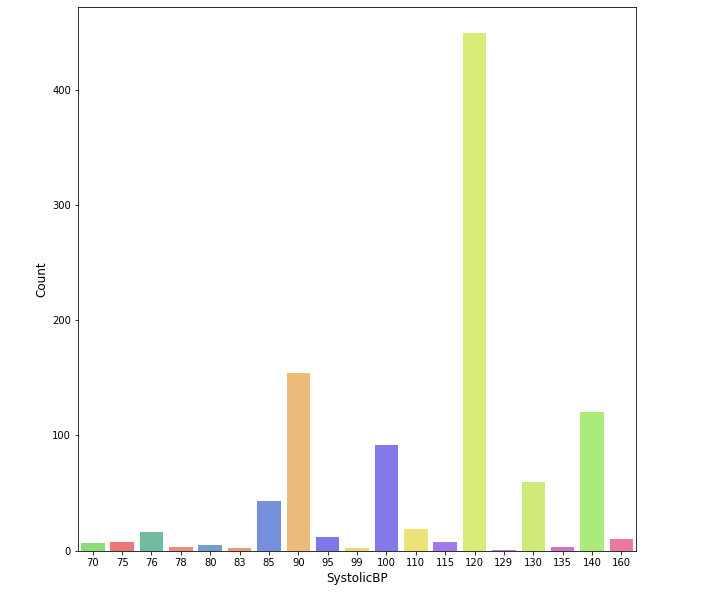
**+** Theo số liệu thu được thì: mức độ rủi ro thấp là 406 (low risk), mức độ rủi ro giữa là 336 (mid risk), mức độ rủi ro cao là 272 (high risk)

+ Nhận xét: Nhìn đồ biểu đồ trên ta thấy không đồng đều và quan trọng nhất là mức rủi ro giữa và cao chiếm tỉ lệ sát nút với rủi ro thấp (tức là tỷ lệ tử vong ở các bà mẹ còn khá cao).

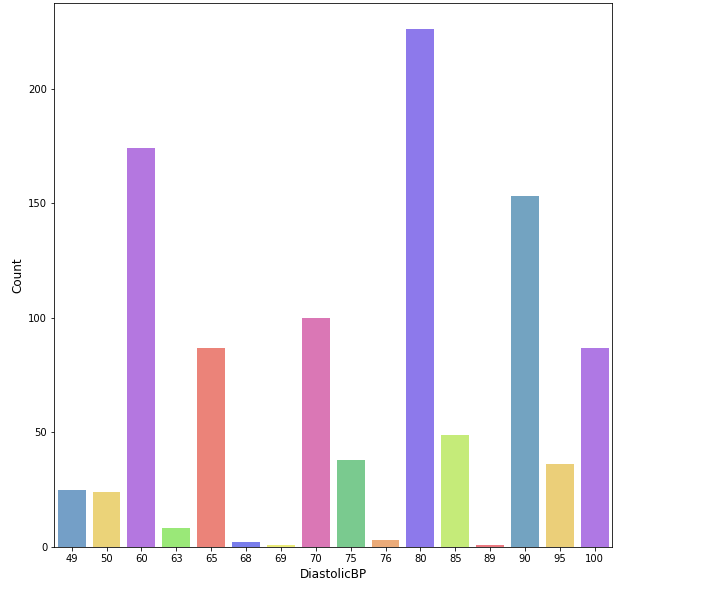
- Mô tả thêm các thuộc tính khác trong bộ dữ liệu:



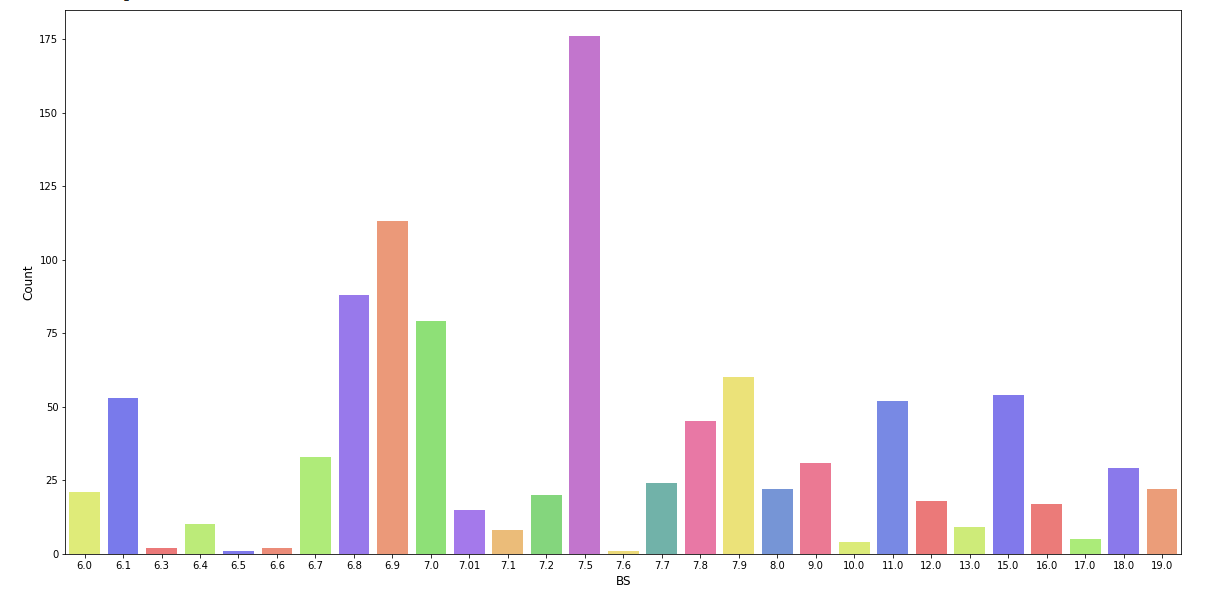
**Biểu đồ đếm số tuổi**



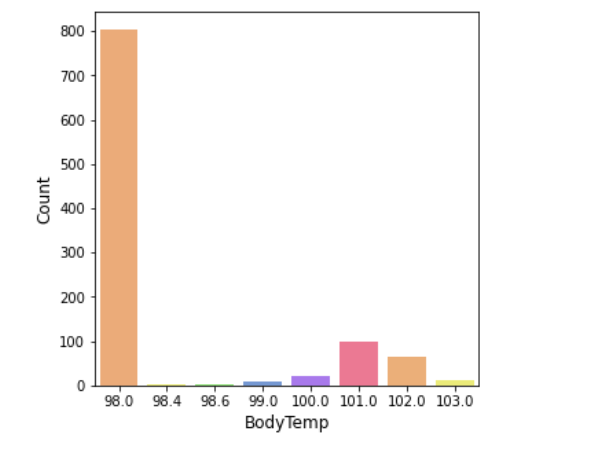
**Biểu đồ đếm số huyết áp trên**



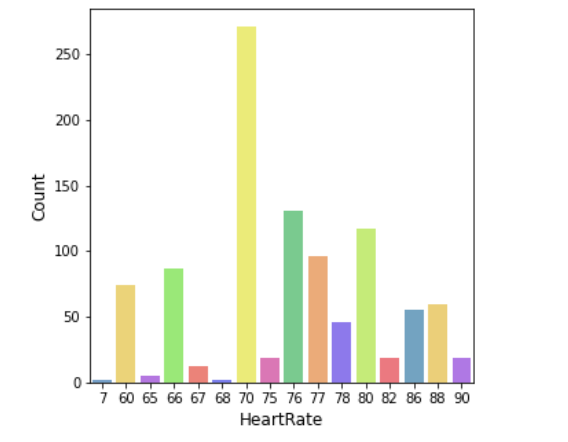
**Biểu đồ đếm số huyết áp thấp**



**Biểu đồ đếm số nồng độ đường huyết**



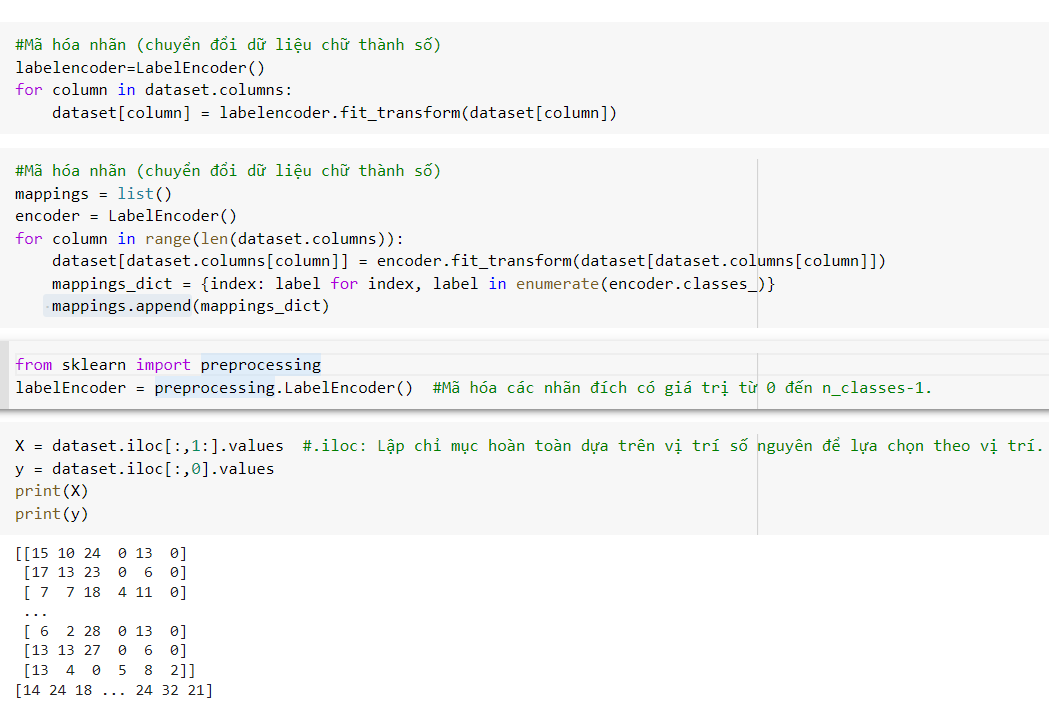
**Biểu đồ đếm số nhiệt độ cơ thể (thân nhiệt)**



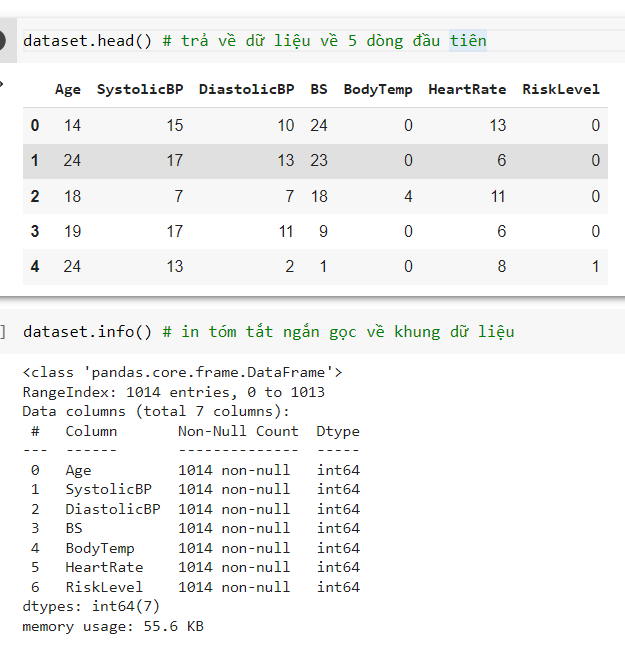
**Biểu đồ đếm số nhịp tim**

## **III.3 Bộ chuyển đổi dữ liệu**

- Mã hóa các nhãn và trả về các nhãn đã được mã hóa:

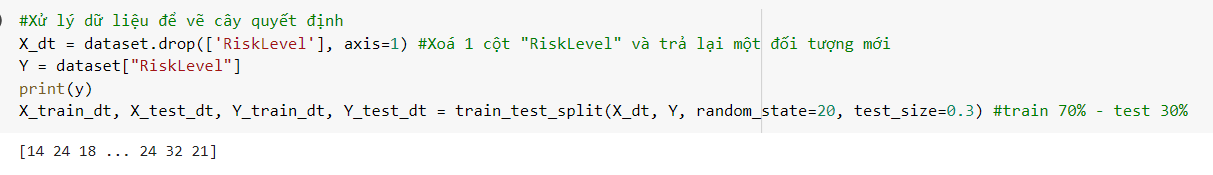


- Xuất dữ liệu:



- Nhận xét: Trích xuất được dữ liệu 5 dòng đầu, hiển thị được khung dữ liệu gồm: 1014 thuộc tính (chạy từ 0 đến 1013), 7 thuộc tính và thuộc kiểu dữ liệu số nguyên.

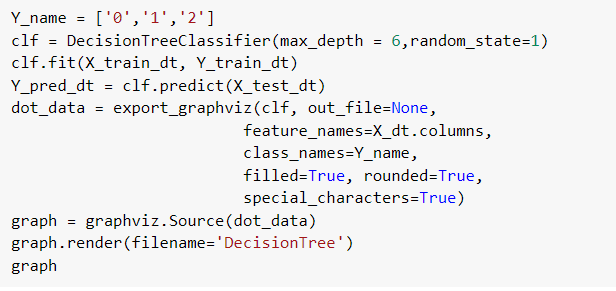
## **III.4 Thực hiện thuật toán Cây quyết định (Decision tree)**

**-** Xử lý dữ liệu:

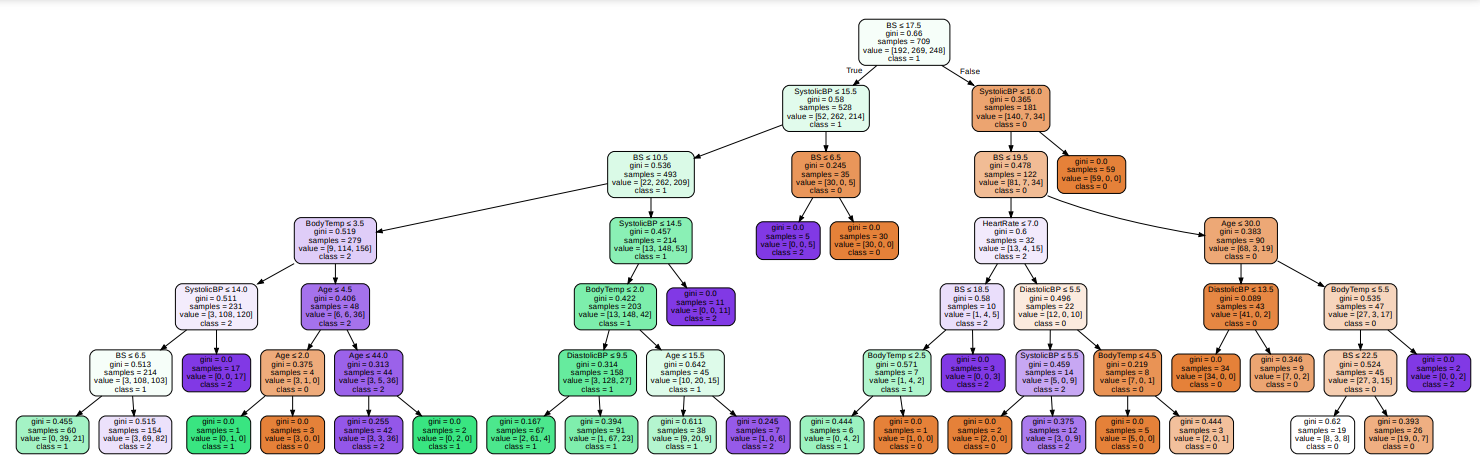
**-**

- Sau khi đã có dữ liệu train và dữ liệu test ta tiến hành xây dựng cây quyết định bằng

đoạn mã bên dưới:



- Sau khi chạy đoạn mã trên ta có được cây quyết định như bên dưới:



**Mô hình cây quyết định**

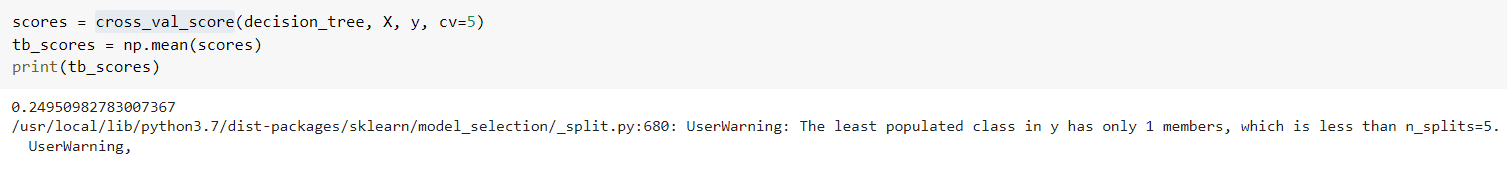
- Nhận xét: Nhìn vào cây quyết định của bộ dữ liệu, ta thấy rằng "BS" là nút gốc của cây.

=> Bằng tất cả các phương pháp đã được kiểm tra bên trên, chúng ta nhận thấy đặc điểm

quan trọng nhất là "BS". Từ đó, chúng ta có thể đưa các giải pháp, hạn chế để giữ cho BS(hay nồng độ đường huyết) được ổn định.

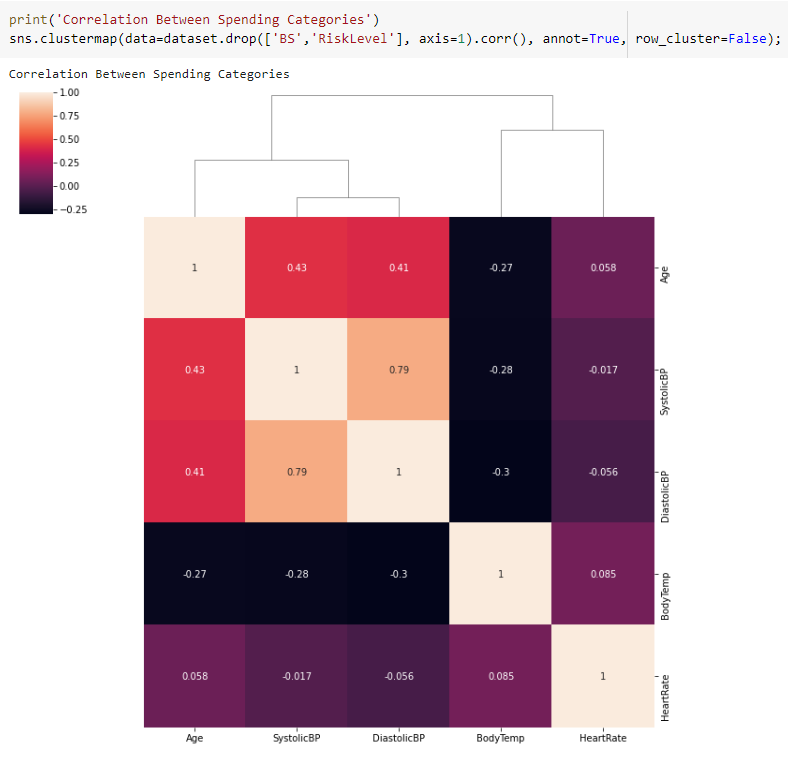
- Ta tiến hành xây dựng **cross\_val\_score** bằng thư viện **sklearn** để đánh giá điểm số (độ chính xác) bằng cách xác nhận chéo:

=> Độ chính xác xấp xỉ bằng :0.25

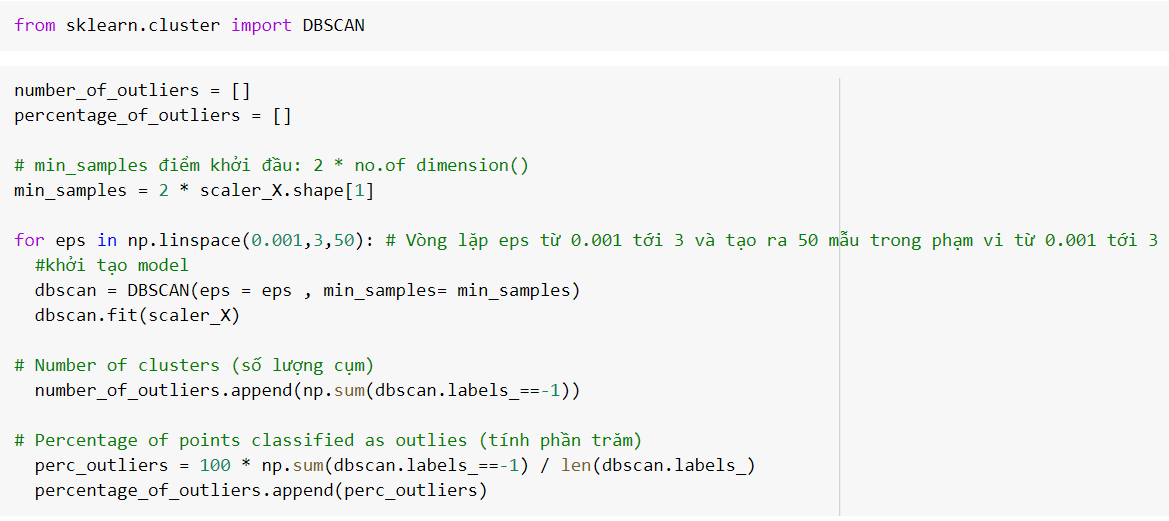


## **III.5 Thực hiện thuật toán DBSCAN**

**-** Vì đây là thuật toán gom cụm dựa trên mật độ nên trước tiên ta xét sự tương quan giữa chúng:



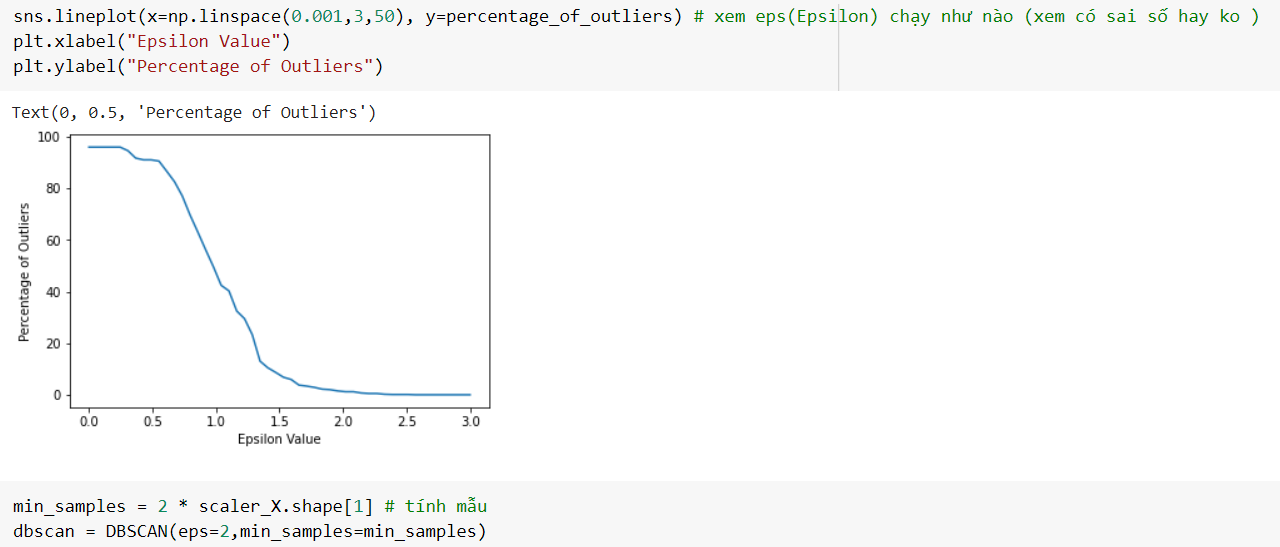
- Sau khi có đủ dữ liệu và chuẩn hóa dữ liệu bằng hàm **scaler** (chia dữ liệu để khớp với model) thì ta tiến hành xây dựng thuật toán DBscan bằng đoạn mã bên dưới:



- Như đã nêu ở trên, DBscan đưa ra 2 tham số là: Eps(bán kính vùng lân cận) và MinPts (số lượng điểm tối thiểu trong vùng lân cận của một điểm trong cụm hay gọi là mẫu).

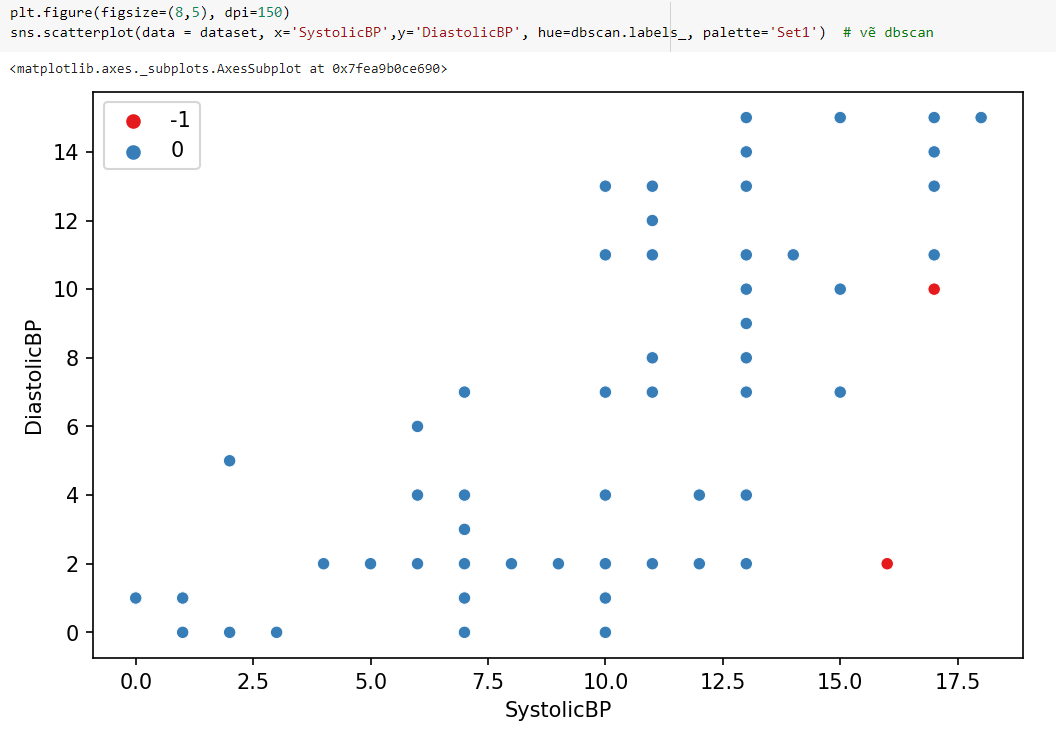
- Vì thuật toán này có thể phân ra rất nhiều cụm nên bắt buộc chúng ta phải tìm ra Eps chính xác nhất để chọn ra cụm chính xác.

Chúng ta sẽ quan sát hình sau để chọn ra Eps:



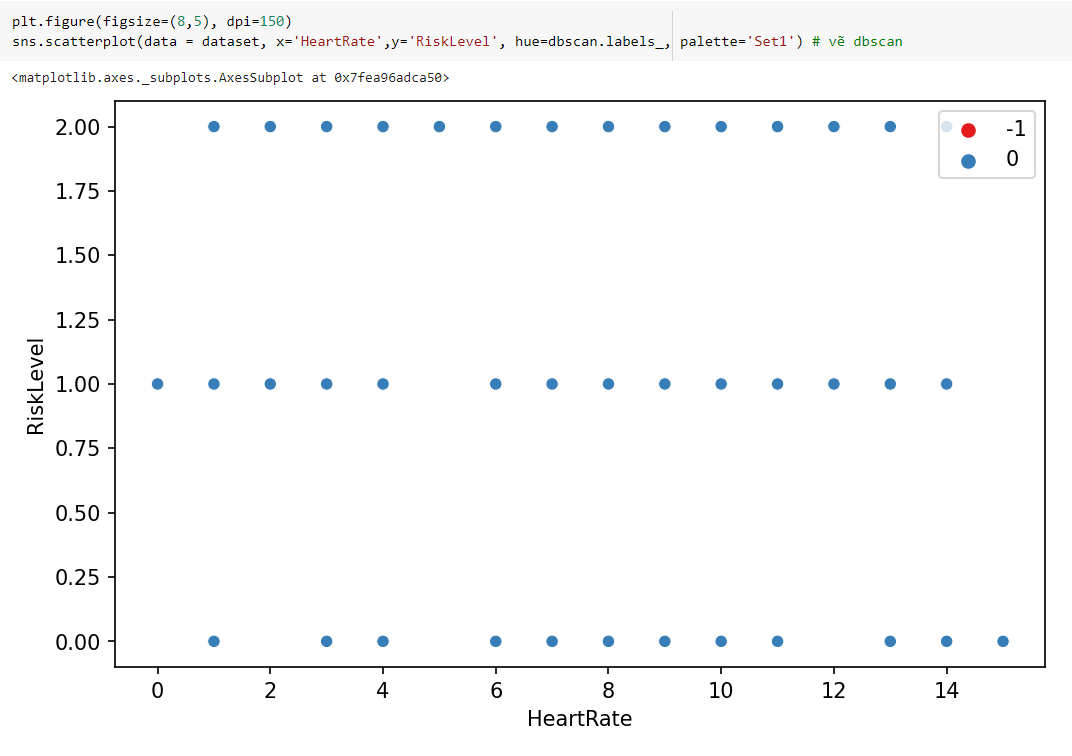
- Theo như hình trên ta thấy được Eps = 2 thì sự giảm Outliers (phần trăm ngoại lệ) gần như không còn nữa, nên nhóm quyết định chọn Eps = 2 và tính được min\_samples=14.

- Sau khi có đủ các thông số, ta đi tìm mật độ phân tán (hay các cụm) của các thuộc tính:



**Biểu đồ mật độ SystolicBP và DiastolicBP**

- Nhận xét: Như hình ở trên thì mật độ phân tán của Systolic BP và Diastolic BP không đều và chủ yếu ở giá trị 0.



**Biểu đồ mật độ của HeartRate và RiskLevel**

- Nhận xét: Như hình ở trên thì mật độ phân tán của HeartRate và RiskLevel khá đều, gần như 3 đường thẳng song song và hầu như ở giá trị 0.

# **CHƯƠNG 4: KẾT LUẬN**

**- *Kết luận:***

**Machine learning** (ML) là lĩnh vực khá quan trọng trong trí tuệ nhân tạo AI 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 đồ án với đề tài “**Giám sát rủi ro sức khỏe của các bà mẹ**”. Nhóm em đã đạt được

một số kết quả như sau:

+ Tìm hiểu tổng quan về machine learning, khai phá dữ liệu, bài toán phân lớp, phân cụm như Cây quyết định, DBscan,... hỗ trợ cho quá trình chăm sóc sức khỏe các bà mẹ trong quá trình mang thai. Thu thập dữ liệu và hiểu thêm về gg calab, tiền xử lý dữ liệu, làm sạch dữ liệu,....

+ Xây dựng nên mô hình phân lớp, phân cụm trên google collaboration.

+ Đánh giá mô hình cây tốt nhất.

+ Đánh giá các mô hình phân cụm khác nhau.

+ Hiểu thêm được các thông tin khác, cụ thể là biết thêm các yếu tố ảnh hưởng đến sức khỏe các bà mẹ, mức độ rủi ro,....

***- Hạn chế:***

Thời gian triển khai ngắn, 1 vài khó khăn về mặt đi lại, đường truyền mạng nên chúng em còn nhiều sơ sót, kiến thức và khả năng lập trình của chúng em còn hạn chế.

***- Hướng phát triển:***

+ Xây dựng, cải tiến mô hình chẩn đoán sức khỏe với phương pháp học máy khác.

+ Tích hợp các mô hình học máy hiện đại, nhằm dự đoán được hết tất cả các yếu tố liên quan đến sức khỏe, không chỉ cho những bà mẹ đang mang thai mà còn những bệnh nhân mang những căn bệnh nguy hiểm.

# **TÀI LIỆU THAM KHẢO**

Link 1: <https://www.youtube.com/watch?app=desktop&v=vOyrfgQDJ4A>

Link 2: <https://scikit-learn.org/stable/>

Link 3:

<https://machinelearningcoban.com/tabml_book/ch_model/decision_tree.html?fbclid=IwAR3nUA8ncu2Oq1UsnQTMjicBCZDpH3ZejBhsUQdTuko6kDn4FtTQg4JOzNw>

Link 4: <https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.info.html?fbclid=IwAR1zaZvGOtbPejVZUCHzXh5F2LaHhsfZWLTDUwvM-Jxz2nbudQQzkhiD7O4>

Link 5: <https://tailieu.vn/doc/thuat-toan-dbscan-1292289.html?fbclid=IwAR1kx5E-MnQCAG24ADzXtyxhAaA6qRDA31iTsAP7r5J6pU1yAtdGIRbBeZw>