Shape, rectangle

Description automatically generatedLogo

Description automatically generatedĐẠI HỌC QUỐC GIA TP. HỒ CHÍ MINH

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN

**KHOA HỆ THỐNG THÔNG TIN**



**BÁO CÁO ĐỒ ÁN MÔN HỌC: Khai Thác Dữ Liệu**

**ĐỀ TÀI:** **CHUẨN ĐOÁN BỆNH TIM**

**Giảng viên hướng dẫn**

Cao Thị Nhạn

Vũ Minh Sang

**Sinh viên thực hiện:** Dương Bảo Tâm - 20521865

Nguyễn Trần Lưỡng Hà - 20521271

Phạm Thái Mỹ Duyên - 20521255

Lê Thanh Giang - 20521260

Contents

[CHƯƠNG 1 : Tổng Quan Đề Tài 3](#_Toc137548380)

[1. Tổng quan về đề tài 3](#_Toc137548381)

[1.1 Giới thiệu đề tài 3](#_Toc137548382)

[1.2 Lý do chọn đề tài 3](#_Toc137548383)

[2. Dữ liệu và các thuộc tính 4](#_Toc137548384)

[2.1 Giới thiệu về dataset 4](#_Toc137548385)

[2.2 Tiền xử lý dữ liệu 6](#_Toc137548386)

[CHƯƠNG 2 : Công Cụ Khai Thác Dữ Liệu và Các Thư Viện 6](#_Toc137548387)

[1. Giới thiệu về công cụ khai thác dữ liệu 6](#_Toc137548388)

[1.1 Google Colaboratory 6](#_Toc137548389)

[1.2 Jupyter Notebook 7](#_Toc137548390)

[2. Cài đặt các thư viện đi kèm và import file csv. 9](#_Toc137548391)

[CHƯƠNG 3 : Tiền Xử Lý Dữ Liệu 10](#_Toc137548392)

[1. Kiểm tra dữ liệu có phù hợp để sử dụng thuật toán 10](#_Toc137548393)

[2. Chuyển đổi dữ liệu có thuộc tính Yes No về dạng 0 1 và các thuộc tính đặc biệt dạng chữ về dạng số 13](#_Toc137548394)

[CHƯƠNG 4 : Phân Tích và Trực Quan Hóa Dữ Liệu 18](#_Toc137548395)

[1. Phân tích sự phân phối dữ liệu 18](#_Toc137548396)

[2. Phân tích sự ảnh hưởng giữa các biến 22](#_Toc137548397)

[CHƯƠNG 5 : Ứng Dụng Các Giải Thuật DataMining và Đánh Giá 30](#_Toc137548398)

[1. Import thư viện và chia dữ liệu thành 2 phần Train và Test 30](#_Toc137548399)

[2. Thuật toán Random Forest 31](#_Toc137548400)

[3. Thuật toán Decision Tree 34](#_Toc137548401)

[3.1 Độ đo Gini: 34](#_Toc137548402)

[3.2 Độ đo Entropy 37](#_Toc137548403)

[4. Thuật toán Naïve Bayes 40](#_Toc137548404)

[5. Thuật toán Logistic Regression 45](#_Toc137548405)

[6. Đánh giá mô hình 49](#_Toc137548406)

[CHƯƠNG 6 : Kết luận 50](#_Toc137548407)

[1. Ưu điểm 50](#_Toc137548408)

[2. Nhược điểm 50](#_Toc137548409)

[3. Hướng phát triển 50](#_Toc137548410)

[CHƯƠNG 7 : Tài Liệu Tham Khảo 51](#_Toc137548411)

# 

# CHƯƠNG 1 : Tổng Quan Đề Tài

## Tổng quan về đề tài

### Giới thiệu đề tài

Ở đề tài này, nhóm đã tiến hành phân tích bộ dữ liệu Personal Key Indicators of Heart Disease có sẵn để xây dựng một mô hình chẩn đoán bệnh tim ở người nhanh nhất và chính xác nhất có thể dựa vào các thông tin về triêu chứng và chỉ số y học.

Input

Triệu chứng

Các chỉ số y học

Model

Output

Dương tính

Âm tính

*Mô tả bài toán.*

Toàn bộ bài đề tài đều sử dụng ngôn ngữ lập trình Python với các công cụ là Jupyter Notebook và Google Colaboratory. Trong quá trình phân tích, nhóm đã áp dụng các kiến thức đã học về tiền xử lí, phân tích, trực quan dữ liệu.

Từ kết quả sau khi phân tích, nhóm tiến hành thực nghiệm trên các mô hình gồm RandomForest, DecisionTree, K-Neighbors, Naïve Bayes, LogisticRegression. Các mô hình thực nghiệm được đánh giá dựa trên độ đo chính là Recall để tìm ra mô hình chẩn đoán bệnh tim tốt nhất.

### Lý do chọn đề tài

Bệnh lý tim mạch là nguyên nhân hàng đầu gây tử vong trên toàn thế giới, nhiều hơn cả bệnh lý ung thư, dù là ở các nước phát triển hay đang phát triển. Tại Việt Nam, theo thống kê của Bộ Y tế mỗi năm có khoảng 200.000 người tử vong vì bệnh tim mạch, chiếm 31% ca tử vong. Theo thống kê của Viện Tim Mạch năm 2015, tỷ lệ tăng huyết áp ở người trưởng thành trong độ tuổi 18- 65 chiếm 25%, vậy cứ 4 người trưởng thành thì có một người tăng huyết áp. Tăng huyết áp làm tăng nguy cơ tử vong do đột quỵ lên gấp 4 lần và tăng nguy cơ tử vong do bệnh lý tim mạch lên gấp 3 lần so với người không mắc bệnh.

A picture containing text, screenshot, font, circle

Description automatically generated

Sơ đồ các nguyên nhân gây tử vong tại Việt Nam (WHO 2016).

Bệnh lý tim mạnh thường được cho là chỉ gặp ở người lớn tuổi. Nhưng trên thực tế, tần suất mắc bệnh ở người trẻ và trung niên chiếm tỷ lệ khá cao. Nó có thể xảy ra với bất kỳ ai và bất kỳ lứa tuổi nào. Những người mắc mới căn bệnh này cũng ngày càng trẻ hóa.

Như những ý đã nêu trên, nhóm đã chọn đề tài này với mục đích xây dựng một mô hình chẩn đoán bệnh tim mạch một cách hiệu quả và chính xác nhất, đồng thời cũng đóng góp một phần nhỏ vào công cuộc giảm thiểu bệnh tim mạch tại Việt Nam.

## Dữ liệu và các thuộc tính

### Giới thiệu về dataset

Tên Dataset: Personal Key Indicators of Heart Disease

Đường dẫn: [Personal Key Indicators of Heart Disease | Kaggle](https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease)

Thông tin dữ liệu:

* + - * Format: \*.csv

Các thành phần có liên quan đến dataset:

**Dữ liệu có 18 cột và 319.795 dòng**

**HeartDisease:** những người trả lời đã từng mắc bệnh tim, với Yes là có và No là không mắc bệnh tim.

**BMI:** mô tả chỉ số cơ thể, người bình thường sẽ giao động từ 18,5 - 24,9

**Smoking:** đã từng hút thuốc hay chưa, với Yes là đã từng hút trên 100 điếu, và No là chưa từng hút, 100 điếu tương đương với 5 gói thuốc.

**AlcoholDrinking:** có phải là người nghiện rượu nặng hay không, với Yes là có, và No là không phải (Ước tính nghiện rượu nặng đối với nam giới là 14 ly mỗi tuần và nữ giới là 7 ly mỗi tuần.

**Stroke:** có từng bị đột quỵ hay chưa, với Yes là có và No là chưa từng bị đột quỵ.

**PhysicalHealth:** mô tả sức khỏe thể chất, khoảng từ 0 đến 30 ngày, có bao nhiêu ngày thể chất bị ảnh hưởng xấu, ví dụ bệnh tật và thương tích.

**MentalHealth:** mô tả sức khỏe tinh thần, khoảng từ 0 đến 30 ngày, có bao nhiêu ngày tinh thần bị ảnh hưởng xấu, ví dụ trầm cảm và cô đơn.

**DiffWalking:** có gặp khó khăn khi đi bộ hoặc leo cầu thang hay không, với Yes là có và No là không gặp khó khăn.

**Sex:** thuộc loại giới tính nào, với Male là nam và Female là nữ

**AgeCategory:** thuộc nhóm tuổi nào, 65-69, 60-64, 70-74, 55-59, 50-54

**Race:** thuộc chủng tộc nào, White, Hispanic, Black, Other, Asia.

**Diabetic:** đã từng bị bệnh tiểu đường bao giờ chưa, với Yes là có, và No là chưa từng bị.

**PhysicalActivity:** có hoạt động thể chất trong 30 ngày qua hay không, Yes là có và No là không có hoạt động thể chất.

**GenHealth:** tình hình sức khỏe hiện tại là gì, Very good, Good, Excellent, Fair, Poor.

**SleepTime:** ngủ bao nhiêu giờ một ngày, có giá trị từ 1 đến 24

**Asthma:** đã bao giờ mắc bệnh hen suyễn hay chưa, với Yes là có và No là chưa từng mắc

**KidneyDisease:** đã bao giờ mắc bệnh về thận hay chưa, với Yes là có và No là chưa từng mắc

**SkinCancer:** đã bao giờ mắc bệnh ung thư da hay chưa, với Yes là có và No là chưa từng mắc

### Tiền xử lý dữ liệu

# CHƯƠNG 2 : Công Cụ Khai Thác Dữ Liệu và Các Thư Viện

## Giới thiệu về công cụ khai thác dữ liệu

### Google Colaboratory

Việc cài đặt công cụ khai thác dữ liệu dựa trên Google Colaboratory vô cùng đơn giản, có thể thực hiện một số bước sau để có thể sử dụng:

**Bước 1:** Tạo thư mục chứa file Google Colaboratory trên drive

Graphical user interface, application, Teams

Description automatically generated

*Hình. Tạo thư mục chứa file Google Colaboratory*

**Bước 2:** Tạo file Google Colabratory như hình bên dưới

Graphical user interface, application

Description automatically generated

*Hình. Tạo file Google Colaboratory*

**Bước 3:** Đổi tên file và ta có thể sử dụng

Graphical user interface, application, email

Description automatically generated

*Hình. Thực hiện đổi tên file (tùy chọn)*

### Jupyter Notebook

**Bước 1:** Mở cmd và chạy jupyter notebook

A screenshot of a computer

Description automatically generated with medium confidence

*Hình. Chạy jupyter notebook*

**Bước 2:** Tạo file Python 3 (ipykernel)

A screenshot of a computer

Description automatically generated with medium confidence

*Hình. Tạo file Python 3*

**Bước 3:** Đã có thể sử dụng

A screenshot of a computer

Description automatically generated with medium confidence

*Hình. Màn hình chạy code*

## Cài đặt các thư viện đi kèm và import file csv.

Cài đặt các thư viện.

A screenshot of a computer code

Description automatically generated with low confidence

Import file dataset

A white rectangle with red and blue text

Description automatically generated with medium confidence

Kiểm tra file dataset

A picture containing text, number, font, line

Description automatically generated

# CHƯƠNG 3 : Tiền Xử Lý Dữ Liệu

## Kiểm tra dữ liệu có phù hợp để sử dụng thuật toán

Thêm dữ liệu

A screenshot of a computer

Description automatically generated with low confidence

Kiểm tra thông tin dữ liệu

A screenshot of a computer

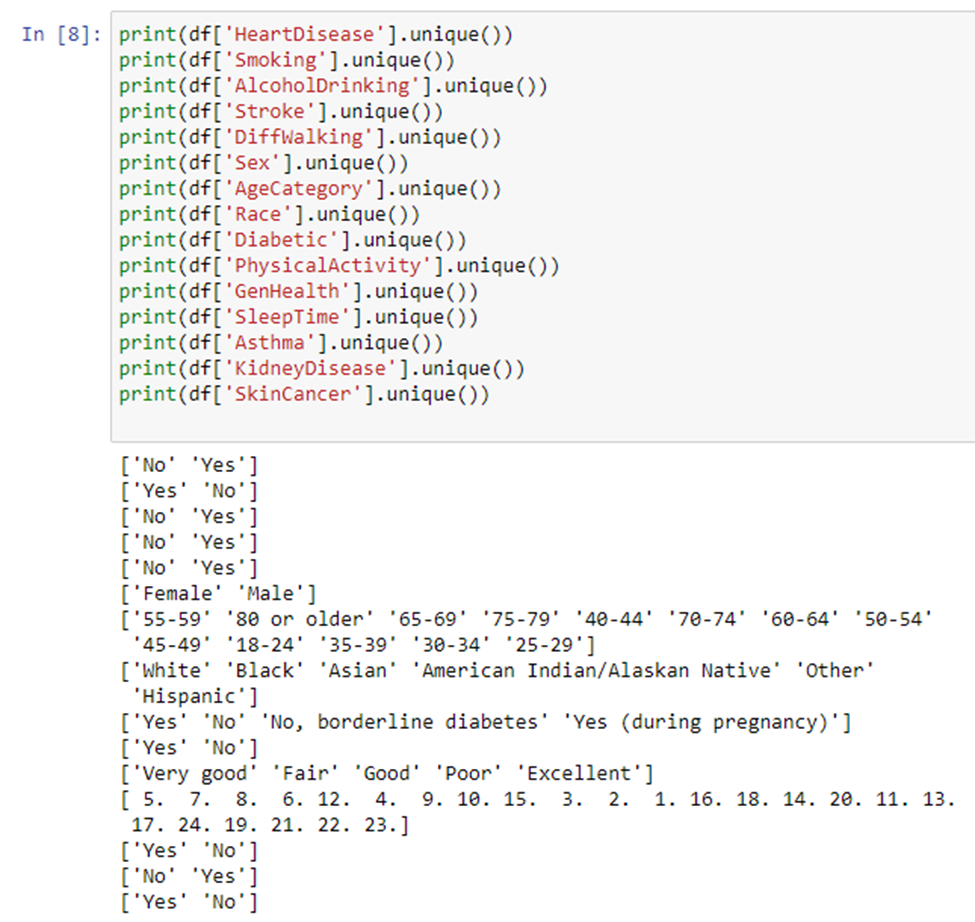
Description automatically generated

Tìm các giá trị Null, tính tổng số lượng và sắp xếp tăng dần

A screenshot of a computer

Description automatically generated

Kiểm tra các thuộc tính của mỗi cột dữ liệu



## Chuyển đổi dữ liệu có thuộc tính Yes No về dạng 0 1 và các thuộc tính đặc biệt dạng chữ về dạng số để phù hợp với các thuật toán Machine Learning.

Import các thư viện cần thiết

A screenshot of a computer program

Description automatically generated with low confidence

Xác định các thuộc tính đặc biệt cần phải chuyển đổi

A screenshot of a computer program

Description automatically generated with medium confidence

Sử dụng thư viện để chuyển đổi dữ liệu

A screenshot of a computer program

Description automatically generated with medium confidence

Dữ liệu mới sau khi đã được chuyển đổi

A screenshot of a computer

Description automatically generated with low confidence

Kiểm tra thông tin dữ liệu sau khi chuyển đổi

A screenshot of a computer

Description automatically generated

Xuất dữ liệu ra file csv



Sau khi xuất dữ liệu ra file csv, nhóm đã loại bỏ các dòng giữ liệu bị trùng (giúp giảm dung lượng và tối ưu hóa quá trình xử lý) và chỉ lấy 16570 dòng dữ liệu để hoàn thành báo cáo.

Dữ liệu sau khi đã được lọc trùng trong excel

A screenshot of a calendar

Description automatically generated with low confidence

# CHƯƠNG 4 : Phân Tích và Trực Quan Hóa Dữ Liệu

## Phân tích sự phân phối dữ liệu

* Nhóm sẽ có thao tác chia nhỏ thuộc tính BMI và các thuộc tính thuộc kiểu dữ liệu categorical qua việc tham khảo việc chia nhỏ bên y học:

A screen shot of a computer program

Description automatically generated with low confidence

* Đối với các biến rời rạc (“Smoking”, “AlcoholDrinking”, “Stroke”, “DiffWalking”, ”Sex”, “Race”, “Diabetic”, “PhysicalActivity”, “GenHealth”, “Asthma”, ”KidneyDisease”, “SkinCancer”):

A picture containing text, screenshot, display, software

Description automatically generatedA picture containing text, screenshot, display, rectangle

Description automatically generatedA picture containing text, screenshot, display, software

Description automatically generatedA picture containing text, screenshot, display, software

Description automatically generatedA picture containing text, screenshot, display, software

Description automatically generatedA picture containing text, screenshot, display, rectangle

Description automatically generatedA picture containing text, screenshot, display, software

Description automatically generatedA graph with red and blue squares

Description automatically generated with low confidenceA picture containing text, screenshot, display, rectangle

Description automatically generatedA picture containing text, screenshot, display, number

Description automatically generatedA picture containing text, screenshot, display, software

Description automatically generatedA picture containing text, screenshot, display, software

Description automatically generated

*Thống kê số lượng người mắc bệnh tim và không mắc bệnh tim theo các giá trị của các biến rời rạc.*

Dựa vào kết quả hình trên, ta có thể nhận thấy một vài dấu hiệu của người có khả năng cao mắc bệnh tim:

* Có thói quen hút thuốc thường xuyên (trên 100 điếu tương đương 5 gói thuốc).
* Có tiền sử bị đột quỵ.
* Có giới tính là nam (“sex” = 1).
* Có gặp khó khăn khi đi bộ hoặc leo cầu thang.
* Người da trắng và người Hispanic có nguy cơ mắc bệnh cao hơn các chủng tộc còn lại (Với “Race” = 5 là White và 3 là Hispanic).
* Có tiền sử bị bệnh tiểu đường hoặc tiền tiểu đường (với “Diabetic” = 2 là tiểu đường và 1 là tiền tiểu đường).
* Không thường xuyên hoạt động thể chất (trong 30 ngày qua).
* Người có tình trạng sức khỏe hiện tại là Poor và Fair (với “GenHealth” = 3 là Poor và 1 là Fair).
* Người có tiền sử mắc bệnh hen suyễn.
* Người có tiền sử mắc bệnh về thận.
* Người có tiền sử mắc bệnh ung thư da
* Đối với các biến liên tục (“AgeCategory”, “BMI”, “PhysicalHealth”, “MentalHealth”, “SleepTime”):

A picture containing text, screenshot, plot, line

Description automatically generatedA picture containing text, screenshot, plot, diagram

Description automatically generatedA picture containing text, screenshot, plot, line

Description automatically generatedA picture containing text, screenshot, plot, number

Description automatically generatedA picture containing text, screenshot, plot, line

Description automatically generated

Nhìn vào hình dạng các phân phối ở hình trên, ta thấy phân phối thuộc tính “AgeCategory” có xu hướng tăng dần trái sang phải, phân phối thuộc tính “oldpeak” có xu hướng lệch phải phải, phân phối các thuộc tính “PhysicalHealth”, “MentalHealth” có xu hướng tập trung ở đầu và ít hơn ở cuối.

Và đặc biệt ở thuộc tính “BMI”, dựa vào kiến thức nhóm tìm hiểu được ở một tài liệu y học [1], BMI đươc chia thành 4 loại Underweight (<18.5), Healthy Weight (18.5-24.9), Overweight (25-30), Obese (>30) nên nhóm đã quyết định chia giỏ thuộc tính này theo tài liệu trên.

A picture containing text, screenshot, diagram, plot

Description automatically generated

*Thống kê số lượng người mắc bệnh tim và không mắc bệnh tim theo các giá trị của biến “BMI”.*

Sau khi chia thành 4 nhóm như trên thì nhận thấy rằng những người có BMI trong ngưỡng Overweight và Obesity thì khả năng cao sẽ mắc bệnh.

## Phân tích sự ảnh hưởng giữa các biến

A screenshot of a computer

Description automatically generated with medium confidence

Nhìn vào ma trận tương quan, nhận thấy:

- Độ tương quan giữa các thuộc tính với biến mục tiêu tương đối thấp.

- Thuộc tính “AgeCategory” tương quan mạnh với biến mục tiêu.

- Thuộc tính “Stroke” tương quan yếu nhất với biến mục tiêu.

- Các cặp thuộc tính “PhysicalHealth” – “DiffWalking”, “MentalHealth” – “PhysicalHealth”, “PhysicalActivity” – “DiffWalking” tương quan tương đối mạnh với nhau.

* Cặp thuộc tính “PhysicalHealth” – “DiffWalking”

A picture containing text, screenshot, diagram, plot

Description automatically generated

Nhận xét:

+ Tại cặp tương quan này, những người bị ảnh hưởng về thể chất lâu dài (PhysicalHealth) sẽ càng có tỉ lệ mắc bệnh tim cao hơn.

+ Tương tự như vậy, với những người bị khó khăn với việc đi lại (DiffWakking) thì phần lớn họ đều có khả năng sẽ mắc bệnh.

+ Khi PhysicalHealth càng tăng tại DiffWalking ở giá trị 1, thì số trường hợp HeartDisease ở giá trị 1 xuất hiện càng dày đặc.

à Kết luận: Cặp tương quan “PhysicalHealth” – “DiffWalking” ảnh hưởng đến biến mục tiêu khá nhiều, khiến cho việc chuẩn đoán bệnh trở nên chuẩn xác hơn.

* Cặp thuộc tính “MentalHealth” – “PhysicalHealth”

A picture containing screenshot, text, line, pattern

Description automatically generated

Nhận xét:

+ Những người gặp vấn đề về tâm lý (MentalHealth) nhưng lại có thể trạng khoẻ mạnh (PhysicaHealth) thì sẽ có ít khả năng bị bệnh tim.

+ Ngược lại, những người có sức khoẻ yếu thường sẽ dễ bị mắc bệnh hơn.

+ Và bệnh tim sẽ càng dễ xuất hiện ở những người có thể trạng yếu và tâm lý không ổn định trong thời gian dài.

à Kết luận: Cặp tương quan “MentalHealth” – “PhysicalHealth” ảnh hưởng lên biến mục tiêu tuy không nhiều như cặp tương quan trên nhưng vẫn góp phần quan trọng trong việc chuẩn đoán bệnh.

* Cặp thuộc tính “PhysicalActivity” – “DiffWalking”:

A picture containing text, screenshot, diagram, line

Description automatically generated

Nhận xét:

+ Ở cặp tương quan này, biến PhysicalActivity (Hoạt động thể chất) không có ảnh hưởng nhiều đến biến mục tiêu nên khá khó để sử dụng biến này để dự đoán.

+ Ngược lại đối với DiffWalking, nếu như một người gặp khó khăn trong việc đi lại, thì người đó sẽ có khả năng mắc bệnh rất cao.

à Kết luận: Tại cặp tương quan “PhysicalActivity” – “DiffWalking”, chúng ta có thể thấy rằng biến PhysicalActivity đóng góp không nhiều để chúng ta chuẩn đoán bệnh. Nó cần phải kết hợp với biến DiffWalking để cho ra kết quả khả quan hơn.

- Tiếp theo nhóm sẽ phân tích Anova cho các biến rời rạc

* Phân tích giữa biến HeartDisease ~ Sex

A screenshot of a computer

Description automatically generated with medium confidence

* Phân tích giữa biến HeartDisease ~ Smoking

A screenshot of a computer

Description automatically generated with medium confidence

* Phân tích HeartDisease ~ AlcoholDrinking

A screenshot of a computer

Description automatically generated with medium confidence

* Phân tích HeartDisease ~ Stroke

A screenshot of a computer

Description automatically generated with medium confidence

* Phân tích HeartDisease ~ DiffWalking

A screenshot of a computer

Description automatically generated with medium confidence

* Phân tích HeartDisease ~ Race

A screenshot of a computer

Description automatically generated with low confidence

* Phân tích HeartDisease ~ Diabetic

A screenshot of a computer

Description automatically generated with medium confidence

* Phân tích HeartDisease ~ PhysicalActivity

A screenshot of a computer

Description automatically generated with medium confidence

* Phân tích HeartDisease ~ GenHealth

A screenshot of a computer

Description automatically generated with medium confidence

* Phân tích HeartDisease ~ Asthma

A screenshot of a computer

Description automatically generated with medium confidence

* Phân tích HeartDisease ~ KidneyDisease

A screenshot of a computer

Description automatically generated with medium confidence

* Phân tích HeartDisease ~ SkinCancer

A screenshot of a computer

Description automatically generated with medium confidence

|  |  |
| --- | --- |
| Features | P-value |
| Sex | 3.182991e-62 |
| Smoking | 1.640728e-106 |
| AlcoholDrinking | 8.475492e-12 |
| Stroke | 5.622912e-155 |
| DiffWalking | 5.965833e-263 |
| Race | 7.516386e-129 |
| Diabetic | 1.052856e-239 |
| PhysicalActivity | 3.098540e-79 |
| GenHealth | 0.0 |
| Asthma | 5.757905e-18 |
| KidneyDisease | 1.571051e-115 |
| SkinCancer | 1.593548e-74 |

Dựa vào kết quả phân tích Anova, nhóm nhận thấy thuộc tất cả thuộc tính đều có giá trị p-value < 0.05 => tất cả các thuộc tính đều có ý nghĩa

# CHƯƠNG 5 : Ứng Dụng Các Giải Thuật DataMining và Đánh Giá

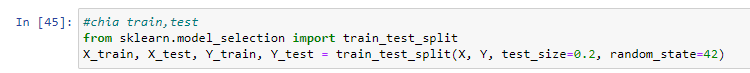
## Import thư viện và chia dữ liệu thành 2 phần Train và Test

Import thư viện

A picture containing text, screenshot, font, line

Description automatically generated

Chia 2 phần train và test cho dữ liệu



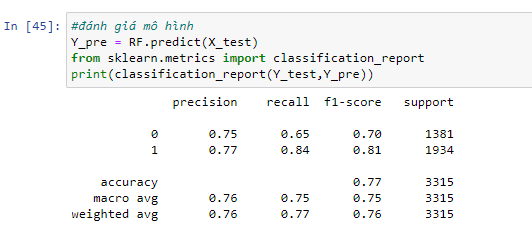
## Thuật toán Random Forest

**Huấn luyện mô hình** :

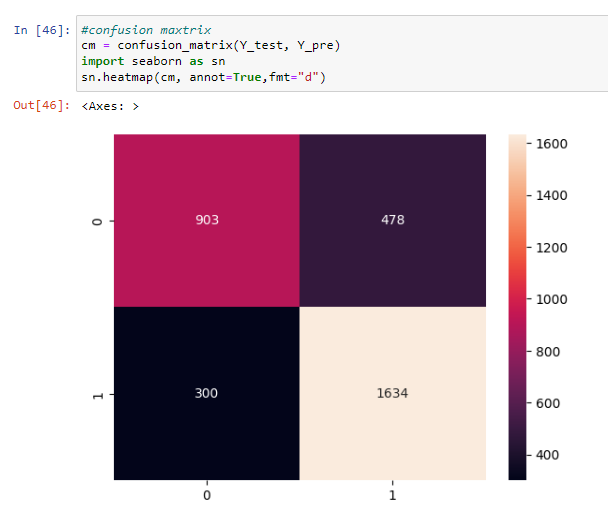
A screenshot of a computer code

Description automatically generated with low confidence

**Đánh giá mô hình**:



**Ma trận hỗn loạn**:



* True Negative = 903 (Với TN (True Negative)):903 bản ghi không phải bị bệnh tim đã được dự đoán chính xác bởi mô hình
* True Positive = 1634 (Với TP (True Positive): 1634 bản ghi bị bệnh tim đã được mô hình dự báo chính xác bởi mô hình
* False Negative = 300 (Với FN (False Negative)): kết quả dự báo / phân loại mô hình là “Negative” sai với giá trị thực tế là “Positive”) hay còn gọi là sai lầm loại II (Type II Error).
* False Positive = 478 (Với FP (False Positive)): kết quả dự báo / phân loại của mô hình là “Positive” sai với giá trị thực là “Negative”) hay còn gọi là sai lầm loại I (Type I Error).

**Các chỉ số (metrics):**

Precision (Positive Predicted Value) là tỷ lệ các trường hợp Positive của mô hình dự báo được phân loại đúng.

Sensitivity (Recall) là tỷ lệ các trường hợp Positive thực sự (actual) được phân loại đúng.

Negative Predictive Value là tỷ lệ các trường hợp Negative được phân loại đúng.

Specificity là tỷ lệ các trường hợp Negative thực sự (actual) đã được phân loại đúng.

Accuracy Rate là tỷ lệ dự báo đúng cho tất cả các trường hợp Negative và Positive trên thực tế và dự báo.

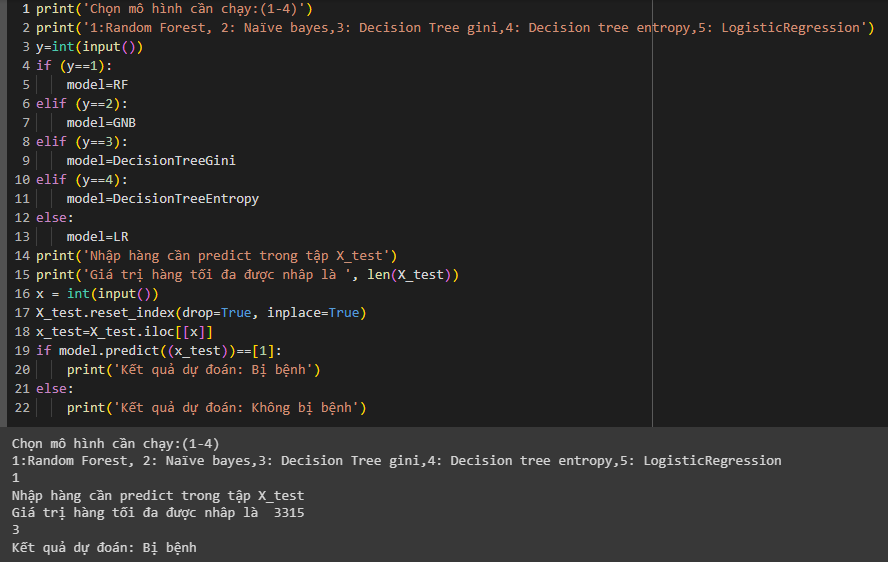
Error Rate bằng 1 - Accuracy Rate là tỷ lệ dự báo sai cho tất cả các trường hợp Negative và Positive trên thực tế và dự báo.

F1-score, là harmonic mean của precision và recall (giả sử rằng hai đại lượng này khác không).

A picture containing text, receipt, font, screenshot

Description automatically generated

**Dự đoán kết quả thuật toán Random Forest**



## Thuật toán Decision Tree

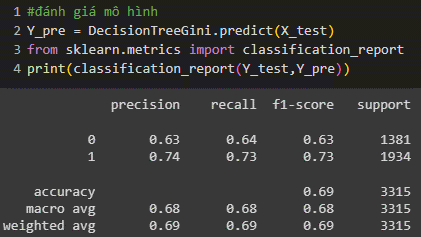
### Độ đo Gini:

* Huấn luyện mô hình:

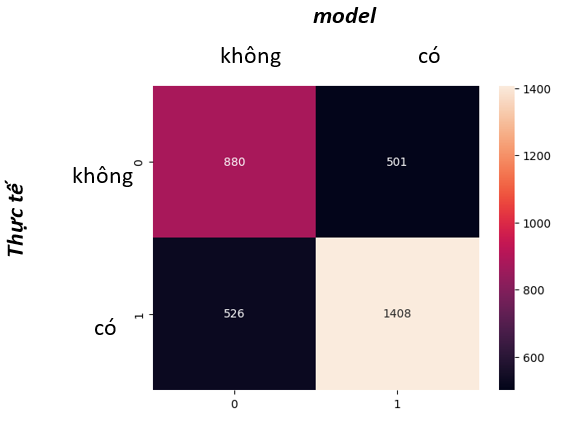
A screenshot of a computer program

Description automatically generated with medium confidence

* Đánh giá mô hình:



* Ma trận hỗn loạn:



* True Negative = 880 (Với TN (True Negative)):880 bản ghi không phải bị bệnh tim đã được dự đoán chính xác bởi mô hình
* True Positive = 1408 (Với TP (True Positive): 1408 bản ghi bị bệnh tim đã được mô hình dự báo chính xác bởi mô hình
* False Negative = 523 (Với FN (False Negative)): kết quả dự báo / phân loại mô hình là “Negative” sai với giá trị thực tế là “Positive”) hay còn gọi là sai lầm loại II (Type II Error).
* False Positive = 501 (Với FP (False Positive)): kết quả dự báo / phân loại của mô hình là “Positive” sai với giá trị thực là “Negative”) hay còn gọi là sai lầm loại I (Type I Error).

**Các chỉ số (metrics):**

Precision (Positive Predicted Value) là tỷ lệ các trường hợp Positive của mô hình dự báo được phân loại đúng.

Sensitivity (Recall) là tỷ lệ các trường hợp Positive thực sự (actual) được phân loại đúng.

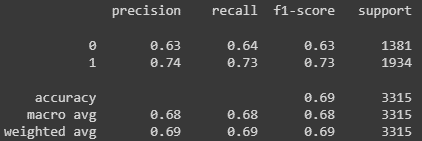
Negative Predictive Value là tỷ lệ các trường hợp Negative được phân loại đúng.

Specificity là tỷ lệ các trường hợp Negative thực sự (actual) đã được phân loại đúng.

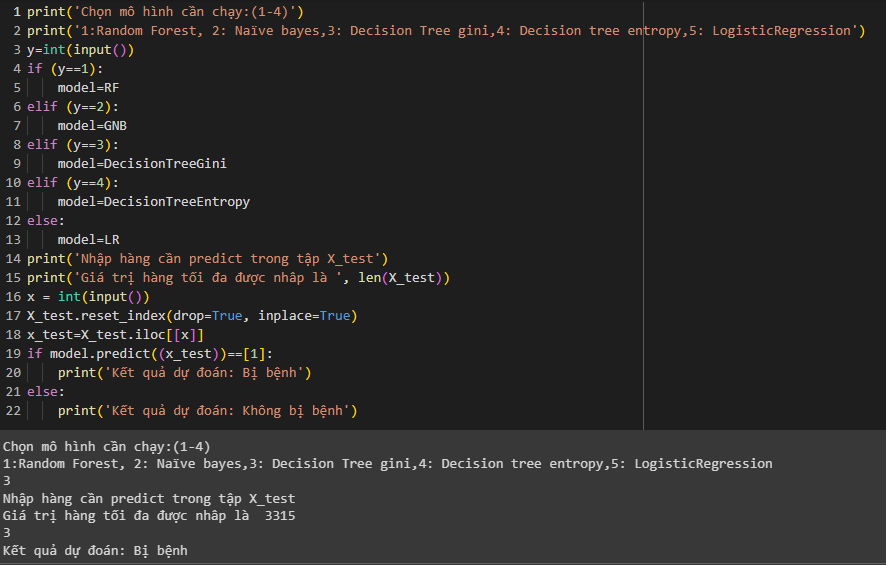
Accuracy Rate là tỷ lệ dự báo đúng cho tất cả các trường hợp Negative và Positive trên thực tế và dự báo.

Error Rate bằng 1 - Accuracy Rate là tỷ lệ dự báo sai cho tất cả các trường hợp Negative và Positive trên thực tế và dự báo.

F1-score, là harmonic mean của precision và recall (giả sử rằng hai đại lượng này khác không).

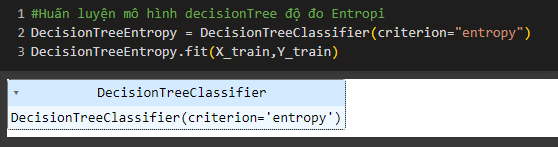


* Dự đoán kết quả từ thuật toán:

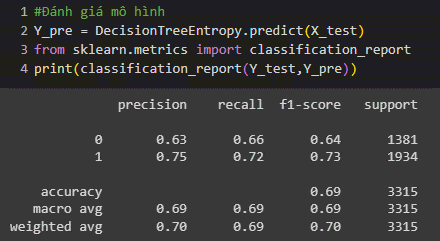


### Độ đo Entropy

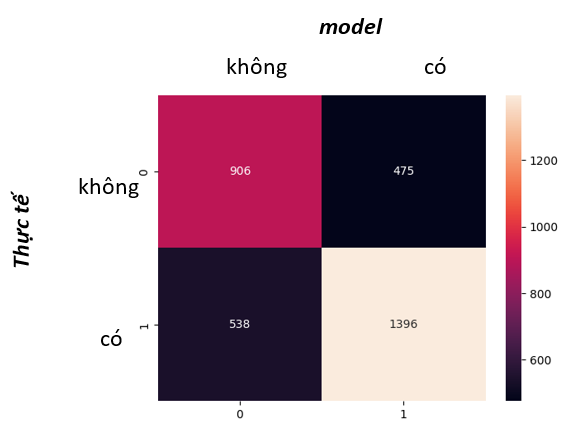
* Huấn luyện mô hình:



* Đánh giá mô hình:



* Ma trận hỗn loạn:



* True Negative = 906 (Với TN (True Negative)):906 bản ghi không phải bị bệnh tim đã được dự đoán chính xác bởi mô hình
* True Positive = 1396 (Với TP (True Positive): 1396 bản ghi bị bệnh tim đã được mô hình dự báo chính xác bởi mô hình
* False Negative = 538 (Với FN (False Negative)): kết quả dự báo / phân loại mô hình là “Negative” sai với giá trị thực tế là “Positive”) hay còn gọi là sai lầm loại II (Type II Error).
* False Positive = 475 (Với FP (False Positive)): kết quả dự báo / phân loại của mô hình là “Positive” sai với giá trị thực là “Negative”) hay còn gọi là sai lầm loại I (Type I Error).

**Các chỉ số (metrics):**

Precision (Positive Predicted Value) là tỷ lệ các trường hợp Positive của mô hình dự báo được phân loại đúng.

Sensitivity (Recall) là tỷ lệ các trường hợp Positive thực sự (actual) được phân loại đúng.

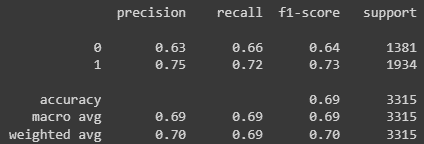
Negative Predictive Value là tỷ lệ các trường hợp Negative được phân loại đúng.

Specificity là tỷ lệ các trường hợp Negative thực sự (actual) đã được phân loại đúng.

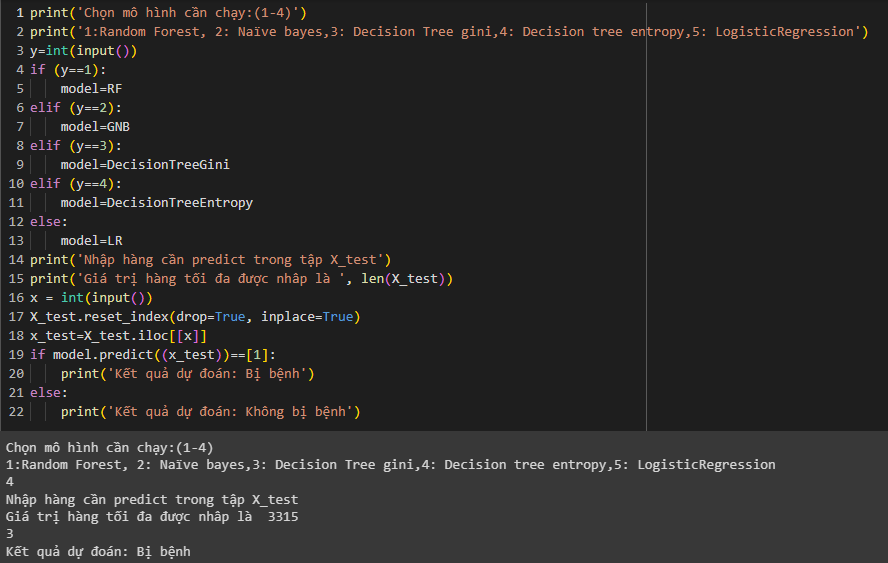
Accuracy Rate là tỷ lệ dự báo đúng cho tất cả các trường hợp Negative và Positive trên thực tế và dự báo.

Error Rate bằng 1 - Accuracy Rate là tỷ lệ dự báo sai cho tất cả các trường hợp Negative và Positive trên thực tế và dự báo.

F1-score, là harmonic mean của precision và recall (giả sử rằng hai đại lượng này khác không).



* Dự đoán kết quả từ thuật toán:



## Thuật toán Naïve Bayes

* Huấn luyện mô hình:

A picture containing text, screenshot, font

Description automatically generated

* Đánh giá mô hình:

A screenshot of a computer

Description automatically generated with medium confidence

* Ma trận hỗn loạn:

A screenshot of a computer

Description automatically generated with medium confidence

* True Negative = 812 (Với TN (True Negative)): 812 bản ghi không phải bị bệnh tim đã được dự đoán chính xác bởi mô hình
* True Positive = 1684 (Với TP (True Positive): 1684 bản ghi bị bệnh tim đã được mô hình dự báo chính xác bởi mô hình
* False Negative = 250 (Với FN (False Negative)): kết quả dự báo / phân loại mô hình là “Negative” sai với giá trị thực tế là “Positive”) hay còn gọi là sai lầm loại II (Type II Error).
* False Positive = 569 (Với FP (False Positive)): kết quả dự báo / phân loại của mô hình là “Positive” sai với giá trị thực là “Negative”) hay còn gọi là sai lầm loại I (Type I Error).

**Các chỉ số (metrics):**

Precision (Positive Predicted Value) là tỷ lệ các trường hợp Positive của mô hình dự báo được phân loại đúng.

Sensitivity (Recall) là tỷ lệ các trường hợp Positive thực sự (actual) được phân loại đúng.

Negative Predictive Value là tỷ lệ các trường hợp Negative được phân loại đúng.

Specificity là tỷ lệ các trường hợp Negative thực sự (actual) đã được phân loại đúng.

Accuracy Rate là tỷ lệ dự báo đúng cho tất cả các trường hợp Negative và Positive trên thực tế và dự báo.

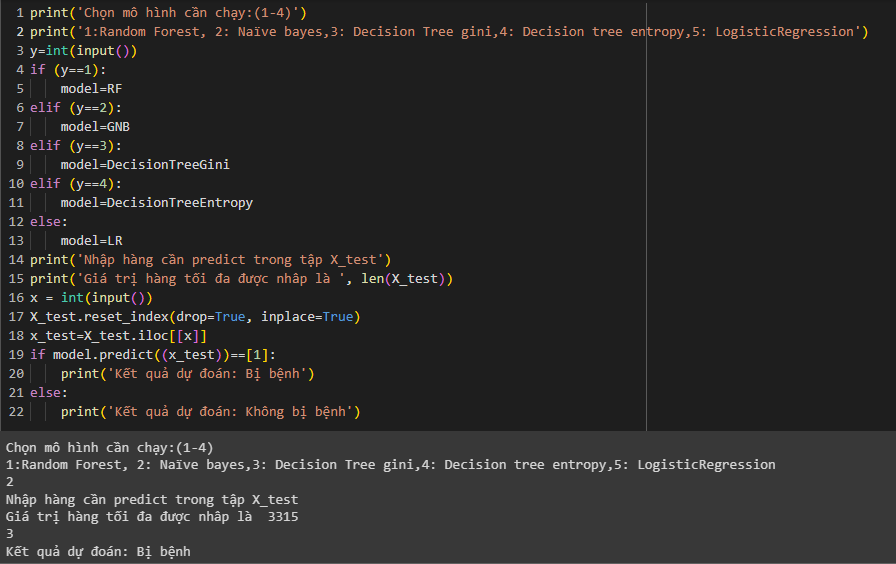
Error Rate bằng 1 - Accuracy Rate là tỷ lệ dự báo sai cho tất cả các trường hợp Negative và Positive trên thực tế và dự báo.

F1-score, là harmonic mean của precision và recall (giả sử rằng hai đại lượng này khác không).

A picture containing text, screenshot, font, receipt

Description automatically generated

* Dự đoán kết quả từ thuật toán:



## Thuật toán Logistic Regression

* Huấn luyện mô hình :

A screenshot of a computer

Description automatically generated with low confidence

A close-up of a logistic

Description automatically generated with low confidence

* Đánh giá mô hình:

A screenshot of a computer

Description automatically generated with medium confidence

* Ma trận hỗn loạn:

A screenshot of a computer

Description automatically generated with medium confidence

* True Negative = 943 (Với TN (True Negative)):943 bản ghi không phải bị bệnh tim đã được dự đoán chính xác bởi mô hình
* True Positive = 1650 (Với TP (True Positive): 1650 bản ghi bị bệnh tim đã được mô hình dự báo chính xác bởi mô hình
* False Negative = 284 (Với FN (False Negative)): kết quả dự báo / phân loại mô hình là “Negative” sai với giá trị thực tế là “Positive”) hay còn gọi là sai lầm loại II (Type II Error).
* False Positive = 438 (Với FP (False Positive)): kết quả dự báo / phân loại của mô hình là “Positive” sai với giá trị thực là “Negative”) hay còn gọi là sai lầm loại I (Type I Error).

**Các chỉ số (metrics):**

Precision (Positive Predicted Value) là tỷ lệ các trường hợp Positive của mô hình dự báo được phân loại đúng.

Sensitivity (Recall) là tỷ lệ các trường hợp Positive thực sự (actual) được phân loại đúng.

Negative Predictive Value là tỷ lệ các trường hợp Negative được phân loại đúng.

Specificity là tỷ lệ các trường hợp Negative thực sự (actual) đã được phân loại đúng.

Accuracy Rate là tỷ lệ dự báo đúng cho tất cả các trường hợp Negative và Positive trên thực tế và dự báo.

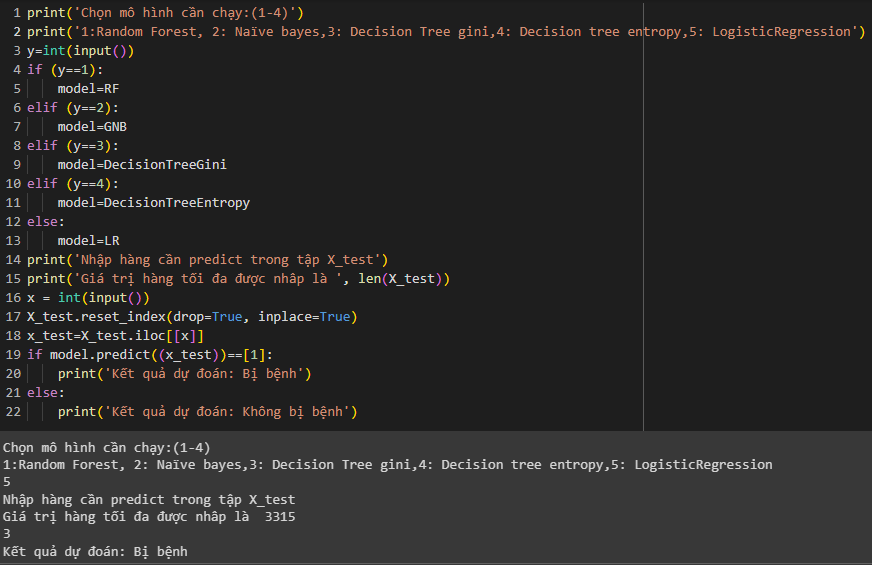
Error Rate bằng 1 - Accuracy Rate là tỷ lệ dự báo sai cho tất cả các trường hợp Negative và Positive trên thực tế và dự báo.

F1-score, là harmonic mean của precision và recall (giả sử rằng hai đại lượng này khác không).

A picture containing text, screenshot, font, number

Description automatically generated

**Dự đoán kết quả thuật toán Logistic Regression**



## Đánh giá mô hình

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| STT | Tên mô hình | Độ đo | | | |
| Accuracy | Recall | Precision | F-1 |
| 1 | Random Forest | 0,76 | 0,84 | 0,77 | 0,7 |
| 4 | Decision Tree: Gini | 0,69 | 0,73 | 0,74 | 0,63 |
| 5 | Decision Tree: Entropy | 0,69 | 0,72 | 0,75 | 0,64 |
| 6 | Naïve bayes | 0,75 | 0,87 | 0,75 | 0,80 |
| 7 | Logistic Regression | 0,78 | 0,85 | 0,79 | 0,72 |

Trong báo cáo này, nhóm đã tiến hành sử dụng các phương pháp, thuật toán phân tích dữ liệu đã học để phân tích bộ dữ liệu Heart Disease qua từng nhóm thuộc tính, cũng như đi sâu vào một số thuộc tính cụ thể. Sau đó thực nghiệm trên một số mô hình máy học như Random Forest, Decision Tree, Naïve bayes, Logistic Regression… để tìm ra mô hình chẩn đoán bệnh tim tốt nhất.

Kết quả mô hình được chọn cuối cùng là Logistic Regression, được xây dựng trên bộ dữ liệu kết hợp giữa dữ liệu gốc và dữ liệu bên ngoài, các tham số mô hình được tối ưu với Recall = 0,85 và Accuracy = 0,78. Đây là kết quả tương đối tốt cho một mô hình hồi qui, tuy nhiên đối với một bài toán trong lĩnh vực y học, chúng ta cần phải nghiên cứu, thực nghiệm nhiều hơn nữa trong tương lai để có thể nâng cao hiệu suất mô hình tốt hơn và áp dụng vào thực tế.

# CHƯƠNG 6 : Kết luận

## 1. Ưu điểm

Nhóm đã biết cách tìm hiểu và các bước khai thác dữ liệu (tiền xử lý, phân lớp bằng các thuật toán đã học) đánh giá kết quả, làm việc nhóm, xây dựng mô hình dự báo hỗ trợ cho công tác điều tra về tình hình bệnh tim mạch. Đưa ra được nhiều thuật toán dự báo. Tính thời gian chạy cũng như độ chính xác của các giải thuật. Đưa ra được những phân tích đánh giá giữa các giải thuật với nhau.

## 2. Nhược điểm

Bởi vì thời gian trong một học kì có hạn, nên nhóm chưa kịp tìm ra cách để tối ưu độ chính xác cao nhất, chưa am hiểu rành mạch về các giải thuật giẫn đến nhiều nhầm lẫn trong khi thực hiện.

## 3. Hướng phát triển

Thời gian tới trong tương lai. Nhóm sẽ nghiên cứu các thuật toán để giảm khả năng sai sót cũng như thời gian tính toán của dữ liệu. Loại bỏ cột hợp lý trong việc tiền xử lý dữ liệu để chọn ra cột nào làm cho thuật toán có độ chính xác cao nhất và tốt nhất.

# CHƯƠNG 7 : Tài Liệu Tham Khảo

[1] “All About Adult BMI,” *Centers for Disease Control and Prevention*, Jun. 03, 2022. https://www.cdc.gov/healthyweight/assessing/bmi/adult\_bmi/index.html (accessed Jun. 12, 2023).

[2] “Know Your Risk for Heart Disease | cdc.gov,” *Centers for Disease Control and Prevention*, Mar. 21, 2023. https://www.cdc.gov/heartdisease/risk\_factors.htm (accessed Jun. 12, 2023).

[3] “Heart Disease Prediction App · Streamlit.” https://kamilpytlak-heart-condition-checker-app-2r42q4.streamlit.app/ (accessed Jun. 12, 2023).