**TRƯỜNG ĐẠI HỌC SƯ PHẠM KỸ THUẬT TP. HỒ CHÍ MINH**

**KHOA CÔNG NGHỆ THÔNG TIN**



**BÁO CÁO ĐỒ ÁN CUỐI KỲ**

**Đề tài : PHÂN TÍCH CÁC YẾU TỐ ẢNH HƯỞNG ĐẾN BỆNH TIM**

**SVTH :**

**Hoàng Minh Nhật MSSV: 19133042**

**Trần Nguyên Thái Bảo MSSV: 19133010**

**Đào Thị Cẩm Tiên MSSV: 19133055**

**Đinh Quốc Hùng MSSV: 19133025**

**GVHD :**

**GV. Quách Đình Hoàng**

Tp. Hồ Chí Minh, tháng 12 năm 2021

**Mục Lục**

[**1.** **Tóm tắt** 3](#_Toc90599735)

[**2.** **Giới thiệu** 3](#_Toc90599736)

[2.1. Các yếu tố ảnh hưởng đến bệnh tim có phân bố như thế nào ? 3](#_Toc90599737)

[2.2. Phân tích các yếu tố có ảnh hưởng như thế nào đối với bệnh tim ? đâu là những yếu tố chính gây ra bệnh tim ? 3](#_Toc90599738)

[2.3. Có phải nữ có nguy cơ mắc bệnh tim cao hơn nam hay không ? 4](#_Toc90599739)

[2.4. Ước lượng độ tuổi có nguy cơ mắc bệnh tim cao ? 4](#_Toc90599740)

[2.5. Xây dựng mô hình chuẩn đoán bệnh suy tim dựa vào bộ dữ liệu **heart.csv** 4](#_Toc90599741)

[**3.** **Dữ liệu ( data )** 5](#_Toc90599742)

[**4.** **Trực quan hóa dữ liệu ( data visulization )** 6](#_Toc90599743)

[4.1. Trả lời task 2.1 : Các yếu tố ảnh hưởng đến bệnh tim có phân bố như thế nào ? 6](#_Toc90599744)

[4.2. Trả lời task 2.2 : Phân tích các yếu tố có ảnh hưởng như thế nào đối với bệnh tim ? đâu là những yếu tố chính gây ra bệnh tim ? 7](#_Toc90599745)

[4.3. Trả lời task 2.3 : Có phải nữ có nguy cơ mắc bệnh tim cao hơn nam hay không ? 7](#_Toc90599746)

[4.4. Trả lời task 2.4 : Ước lượng độ tuổi có nguy cơ mắc bệnh tim cao ? 8](#_Toc90599747)

[**5.** **Mô hình hóa dữ liệu ( data modeling )** 9](#_Toc90599748)

[5.1. Mô hình SVM 9](#_Toc90599749)

[5.2. Mô hình Decision tree 10](#_Toc90599750)

[5.3. Mô hình KNN 10](#_Toc90599751)

[**6.** **Thực nghiệm , kết quả , thảo luận** 11](#_Toc90599752)

[6.1. Mô hình SVM 11](#_Toc90599753)

[6.2. Mô hình Decision tree 12](#_Toc90599754)

[6.3. Mô hình KNN 12](#_Toc90599755)

[**7.** **Kết luận** 12](#_Toc90599756)

[**8.** **Phụ lục** 13](#_Toc90599757)

[**9.** **Đóng góp** 13](#_Toc90599758)

[**10.** **Tài liệu tham khảo** 14](#_Toc90599759)

1. **Tóm tắt**

Các bệnh tim mạch (**CVDs**) là nguyên nhân gây tử vong số 1 trên toàn cầu, cướp đi sinh mạng của khoảng 17,9 triệu người mỗi năm, chiếm 31% tổng số ca tử vong trên toàn thế giới. Bốn trong số 5CVD tử vong là do đau tim và đột quỵ, và một phần ba số ca tử vong này xảy ra sớm ở những người dưới 70 tuổi. Suy tim là một sự kiện phổ biến do CVDs gây ra và tập dữ liệu này có 11 yếu tố có thể được sử dụng để dự đoán bệnh tim có thể xảy ra.

Những người bị bệnh tim mạch hoặc có nguy cơ tim mạch cao (do sự hiện diện của một hoặc nhiều yếu tố nguy cơ như tăng huyết áp, tiểu đường, tăng lipid máu hoặc bệnh đã có sẵn) cần được phát hiện và quản lý sớm, trong đó mô hình học máy có thể giúp ích rất nhiều.

1. **Giới thiệu**
   1. Các yếu tố ảnh hưởng đến bệnh tim có phân bố như thế nào ?

- **Biến dự đoán (X)** : 11 biến (Age,Sex,ChestPainType,RestingBP,

Cholesterol,FastingBS,RestingECG,MaxHR,ExerciseAngina,Oldpeak,

ST\_Slope)

- **Biến kết quả (Y)** : count

- **Ý nghĩa** : từ phân tích này giúp chúng ta hiểu được phân bố của các

yếu tố ảnh hưởng đến bệnh tim .

- **Biểu đồ có thể sử dụng** : histogram

- **Phương pháp sử dụng** : thống kê mô tả (Mô tả các đặc trưng của một

phân bố cho biến số , Mô tả các đặc trưng của một phân bố cho biến phân

loại).

* 1. Phân tích các yếu tố có ảnh hưởng như thế nào đối với bệnh tim ? đâu là những yếu tố chính gây ra bệnh tim ?

- **Biến dự đoán (X)** : 11 biến (Age,Sex,ChestPainType,RestingBP,

Cholesterol,FastingBS,RestingECG,MaxHR,ExerciseAngina,Oldpeak,

ST\_Slope )

- **Biến kết quả (Y)** : HeartDisease

- **Ý nghĩa** : từ phân tích này chúng ta thấy được một cách tổng quan về

các yếu tố ảnh hưởng đến bệnh tim , từ biểu đồ rút ra kết luận về các

yếu tố chủ yếu gây nên bệnh tim .

- **Biểu đồ có thể sử dụng** : Bar .

- **Phương pháp sử dụng** : Thống kê mô tả (Mô tả mối quan hệ của hai biến phân loại , Mô tả phân bố của một biến số theo các nhóm của một biến phân loại).

* 1. Có phải nữ có nguy cơ mắc bệnh tim cao hơn nam hay không ?

- **Biến dự đoán (X)** : Sex .

- **Ý nghĩa** : từ kiểm định này chúng ta có thể biết được giới tính có nguy cơ mắc bệnh tim cao hơn .

- **Phương pháp sử dụng** : kiểm định thống kê (two sample t-test for two means (μ1,μ2)).

* 1. Ước lượng độ tuổi có nguy cơ mắc bệnh tim cao ?

- **Biến dự đoán (X)** : Age .

- **Biến kết quả (Y)** : HeartDisease .

- **Ý nghĩa** : từ phân tích này chúng ta biết được khoảng độ tuổi nào là có nguy cơ cao để từ đó có những lời khuyên phù hợp cho bệnh nhân .

- **Phương pháp sử dụng** : Ước lượng population mean

* 1. Xây dựng mô hình chuẩn đoán bệnh suy tim dựa vào bộ dữ liệu **heart.csv**

- Để chuẩn đoán cần **input** là kết quả đo của bác sĩ là 11 biến (Age, Sex,

ChestPainType, RestingBP, Choleserol, FastingBS, RestingECG, MaxHR,

ExerciseAngina, Oldpeak, ST\_Slope).

- **Kết quả** là biến : HeartDisease (1 : có nguy cơ mắc bệnh tim , 0 : không có nguy cơ mắc bệnh tim ) .

**- Ý nghĩa** : Từ mô hình học máy giúp dự đoán bệnh tim nhằm phát hiện sớm để có phương pháp điều trị kịp thời .

- **Phương pháp sử dụng** : học có giám sát SVM (Support vector machine),

tham khảo mô hình KNN , decision tree .

1. **Dữ liệu ( data )**

**Nguồn dữ liệu**: <https://www.kaggle.com/fedesoriano/heart-failure-prediction> . Tập dữ liệu này được tạo bằng cách kết hợp các tập dữ liệu khác nhau đã có sẵn một cách độc lập nhưng chưa được kết hợp trước đó. Tập dữ liệu bao gồm 918 quan sát đến từ :

* + Viện Tim mạch Hungary. Budapest: Andras Janosi, M.D.
  + Bệnh viện Đại học, Zurich, Thụy Sĩ: William Steinbrunn, M.D.
  + Bệnh viện Đại học, Basel, Thụy Sĩ: Matthias Pfisterer, M.D.
  + V.A. Trung tâm Y tế, Long Beach và Tổ chức Phòng khám Cleveland: Robert Detrano, M.D., Ph.D.

**Người tổng hợp** : David W. Aha (aha '@' ics.uci.edu) (714) 856-8779

Tập dữ liệu có 12 biến : trong đó 7 biến phân loại và 5 biến số :

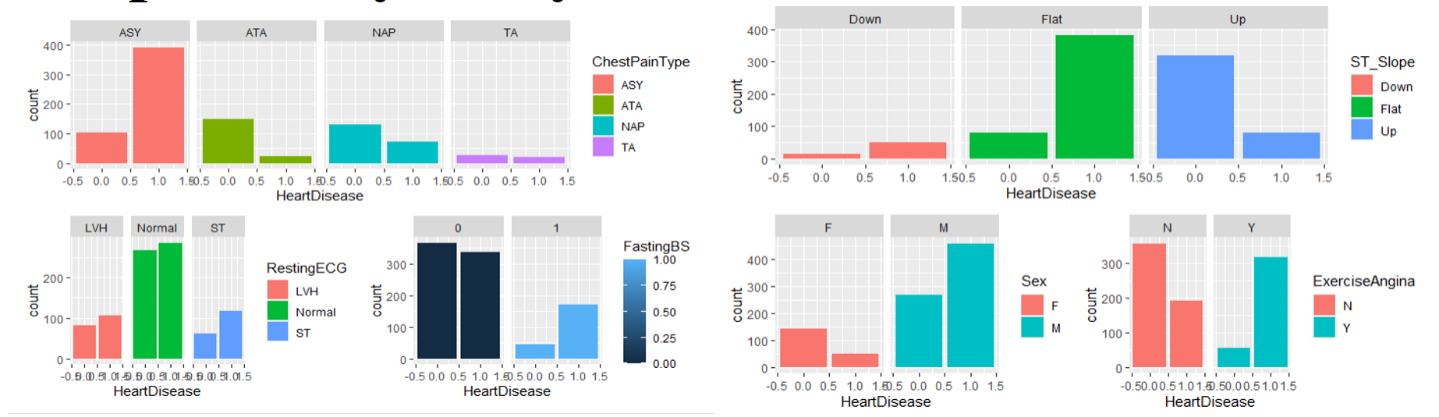
* **Age**: Tuổi của các bệnh nhân (years)
* **Sex**: giới tính của bệnh nhân (M:Male, F: Female)
* **ChestPainType**: các kiểu đau ngực (TA: Typical Angina, ATA: Atypical Angina, NAP: Non-Anginal Pain, ASY: Asymptomatic )
* **RestingBP**: huyết áp nghỉ ngơi (mm Hg)
* **serum Cholesterol**: Mức cholesterol trong huyết thanh của một người đại diện cho tổng lượng cholesterol trong máu của họ (mm/dl)
* **FastingBS**: đường huyết lúc đói (1: nếu BS lúc đói > 120 mg/dl, 0:otherwise)
* **RestingECG**: kết quả điện tâm đồ lúc nghỉ (Normal: bình thường, ST: có bất thường sóng ST-T (đảo ngược sóng T và/hoặc ST chênh lên hoặc giảm> 0,05 mV), LVH: hiển thị phì đại thất trái có thể xảy ra hoặc xác định theo tiêu chí của Estes)
* **MaxHR**: nhịp tim tối đa đạt được (giữa 60 và 202)
* **ExerciseAngina**: đau thắt ngực do tập thể dục (Y: yes, N: No)
* **Oldpeak**: oldpeak =ST(giá trị số được đo bằng độ trầm cảm)
* **ST\_Slope**: nhịp tim lúc tập thể dục(Up: dốc lên, Flat: phẳng, Down: dốc xuống)
* **HeartDisease**: kết quả đầu ra (1: bệnh tim, 0: Bình thường)

**Chuẩn hóa dữ liệu :** Tập dữ liệu nhóm sử dụng không có missing values . Nhóm có tiền xử lý dữ liệu cho mô hình học máy bằng cách đưa các giá trị biến phân loại về các biến nhị phân . Chia 80% để train và 20% để test .

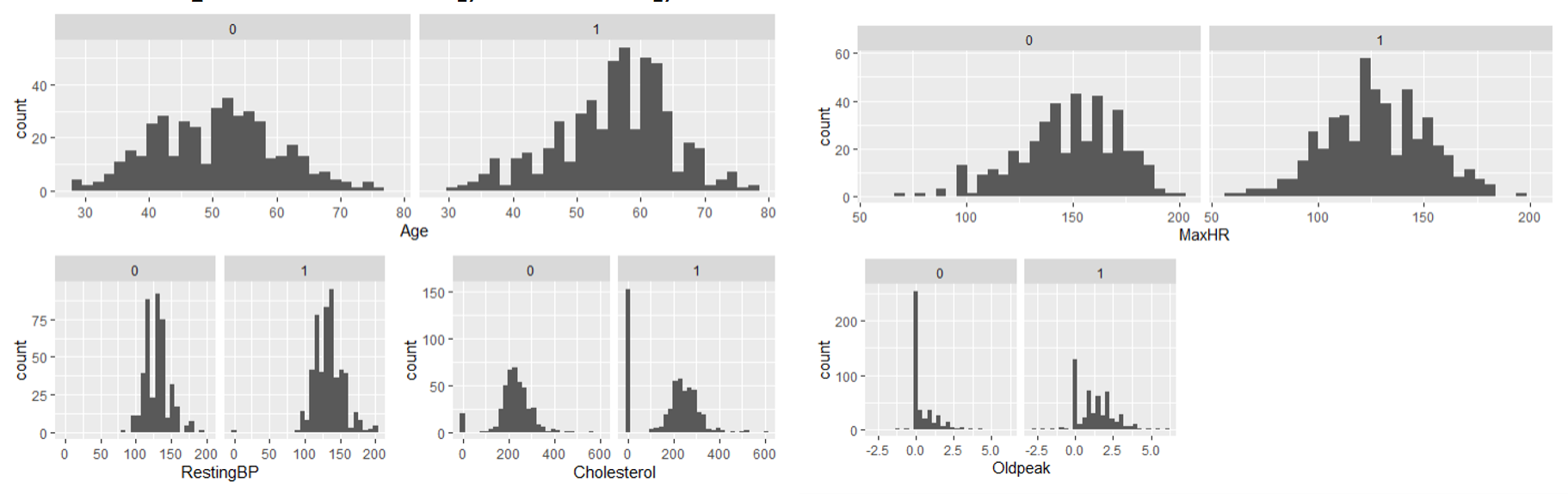
**Thư viện sử dụng :** Các thư viện chính được sử dụng trong phân tích như : tidyverse , dplyr , ggplot2 , e1071 , gbm , caTools , rpart và một số thư viện khác .

1. **Trực quan hóa dữ liệu ( data visulization )**
   1. Trả lời task 2.1 : Các yếu tố ảnh hưởng đến bệnh tim có phân bố như thế nào ?

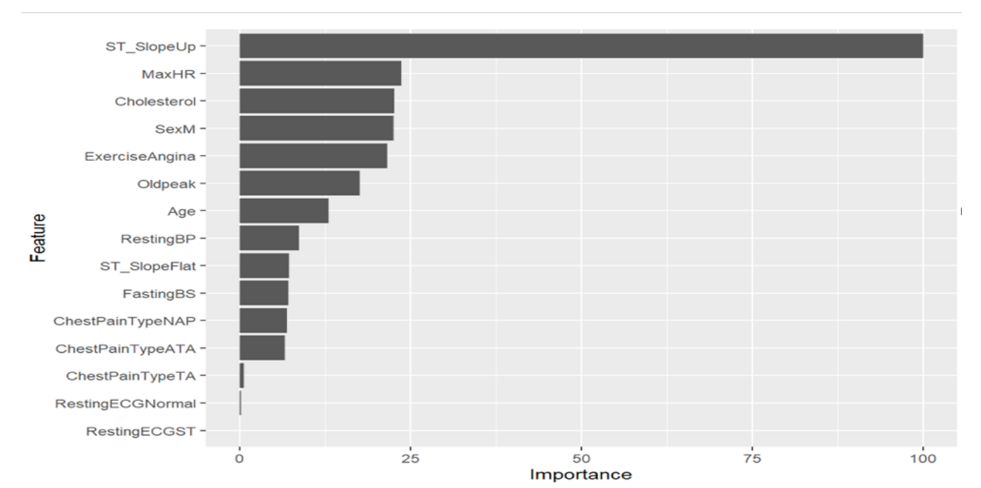
* **Đối với các biến phân loại** :



* **Nhận xét** : Phần lớn dữ liệu là nam giới và số bệnh nhân mắc bệnh tim nhiều hơn số không mắc bệnh. Số phụ nữ bị bệnh ít hơn số phụ nữ không bị bệnh. Hầu hết bệnh nhân bị đau ngực không triệu chứng . Đa số bệnh nhân có kết quả điện tâm đồ bình thường khi nghỉ ngơi Bệnh nhân bị bệnh tim bị đau thắt ngực do gắng sức nhiều hơn (khoảng 300) so với bệnh nhân bị đau thắt ngực do không vận động (khoảng 200). Và phần lớn bệnh nhân có độ dốc nhịp tim bằng phẳng khi tập thể dục.
* **Đối với các biến số :**

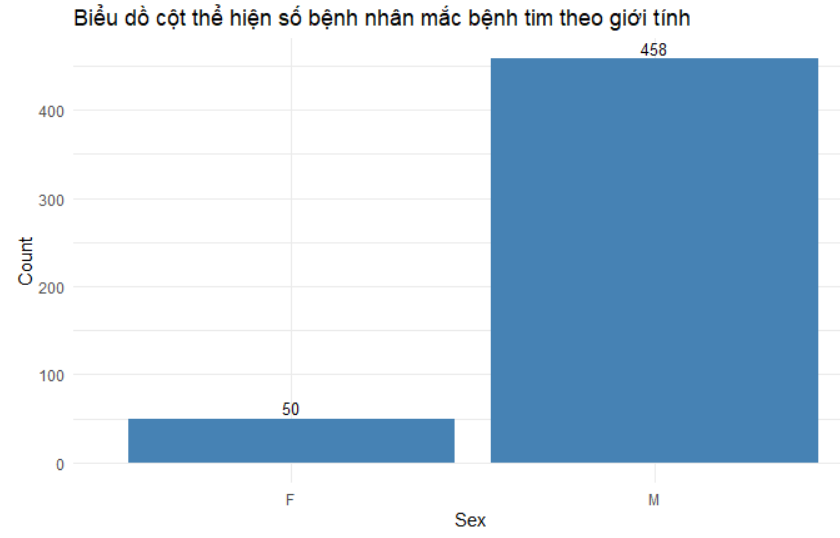


* **Nhận xét** : Về độ tuổi, chúng ta thấy độ tuổi dễ mắc bệnh tim nhất là từ 55 đến 65.Với cholesterol, chúng tôi cũng thấy nhiều quan sát với cholesterol huyết thanh bằng 0.Nhịp tim tối đa được phân phối giữa 60 và 202, không có bất thường quan sát thấy.Và cuối cùng Oldpeak là giá trị số đo bằng độ trầm mặc dù có rất nhiều số 0 nhưng là số đo hợp lý.
  1. Trả lời task 2.2 : Phân tích các yếu tố có ảnh hưởng như thế nào đối với bệnh tim ? đâu là những yếu tố chính gây ra bệnh tim ?



Xem biểu đồ thể hiện tầm quan trọng của các biến đối với biến HeartDisease bằng hàm varImp (một phương pháp chung để tính mức độ quan trọng thay đổi cho các đối tượng được tạo ra bởi các phương pháp cụ thể) .

* **Nhận xét** : Từ biểu đồ có thể thấy biến có ảnh hưởng đến bệnh tim nhất chính là: `ST\_Slope`, tiếp đến là `MaxHR` ,`cholesterol`, `ExerciseAngina`, `Oldpeak` . Một số biến có độ quan trọng thấp như RestingECG , FastingBS,..
  1. Trả lời task 2.3 : Có phải nữ có nguy cơ mắc bệnh tim cao hơn nam hay không ?

Dựa vào số thông kê trên thì ta có 193 nữ và 725 nam, trong đó số bệnh nhân bị bệnh tim là nữ là 50 bệnh nhân và số bệnh nhân bị bệnh tim mà là nam là 458 bệnh nhân.

Gọi a là tỉ lệ số bệnh nhân bị bệnh tim là nữ

Gọi b là tỉ lệ số bệnh nhân bị bệnh tim là nam

Gọi n là số bệnh nhân nữ

Gọi m là số bệnh nhân nam

Ta có tỉ lệ bệnh nhân bị bệnh tim là nữ là p = 50/193

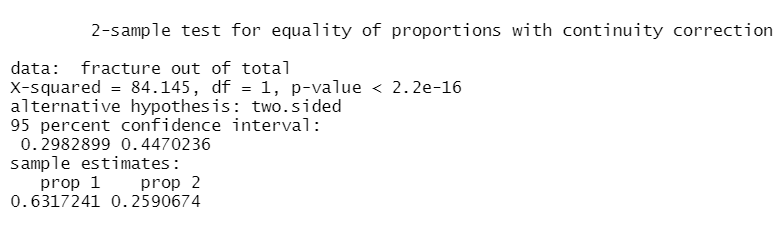
Tỉ lệ bệnh nhân bị bệnh tim là nam là q = 458/725

Để so sánh tị lệ giữa 2 giới tính ta cần kiểm tra giả thiết

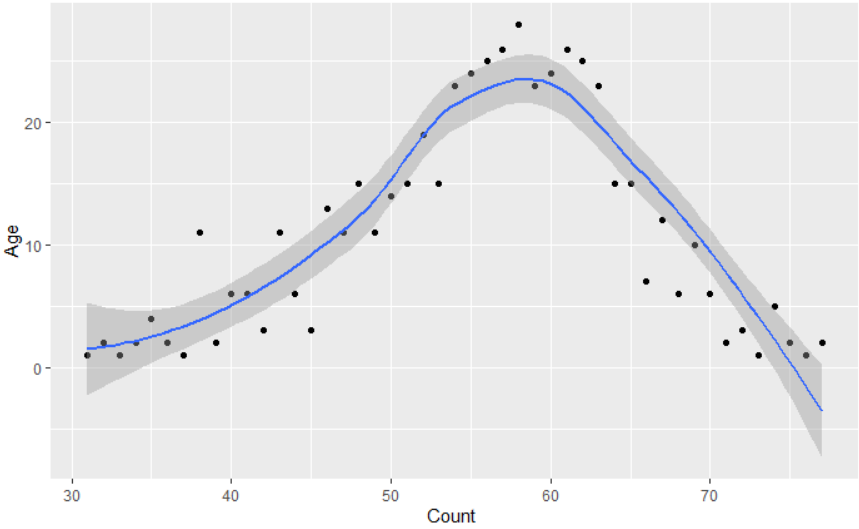
Ho: a >= b

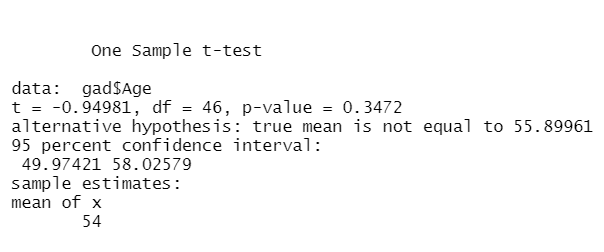
H1: a < b , Ta tính:

Ta có Z =( p - q ) / sqrt(f\*f1\*(1/n + 1/ m))



* **Nhận xét** : Kết quả phân tích trên cho thấy tỉ lệ mắc bệnh đối với nam là 0.6317241 và đối với bệnh nhân nữ là 0.2590674. Phân tích trên còn cho thấy xác suất 95% rằng độ khác biệt giữa hai nhóm có thể 0.299 đến 0.447 (tức 29,9% đến 44,7%). Với trị số p = 2.2e-16, chúng ta có thể nói rằng tỉ lệ mắc bệnh tim của bệnh nhân nam cao hơn bệnh nhân nữ.
  1. Trả lời task 2.4 : Ước lượng độ tuổi có nguy cơ mắc bệnh tim cao ?

 Để kiểm tra thực hiện lọc để đếm bệnh nhân theo độ tuổi và trong 913 bệnh nhân thì có khoảng 505 bệnh nhân mắc bệnh từ độ tuổi 31 đến 77, dựa vào biểu đồ thì chúng ta thấy độ tuổi trung bình khoảng từ 55 đến 58, em sử dụng ước lượng trung bình cho mẫu thì em được kết quả như sau:



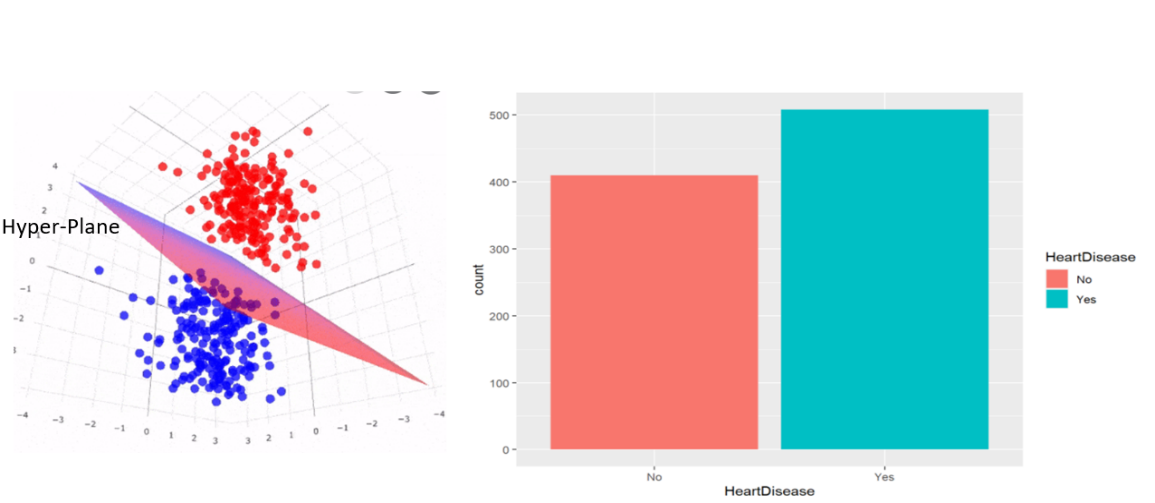
Trong lệnh trên age là biến số chúng ta cần kiểm định, và mu= 55.89 là giá trị giả thiết. R trình bày trị số t = -0.94981 , với 46 bậc tự do, và trị số p < 2.2e-16 (tức rất thấp). R cũng cho biết độ tin cậy 95% của age là từ 49.97421 tuổi đến 58.02579 tuổi (55.89961 tuổi nằm trong khoảng tin cậy này). Nói cách khác, chúng ta có lí do để phát biểu rằng độ tuổi trung bình trong mẫu này có thể nói là tương đương độ tuổi trung bình của quần thể.

1. **Mô hình hóa dữ liệu ( data modeling )**

Trả lời cho task 2.5 : Xây dựng mô hình chuẩn đoán bệnh suy tim dựa vào bộ dữ liệu **heart.csv**

* 1. Mô hình SVM

**SVM** là một thuật toán học có giám sát, nó có thể sử dụng cho cả việc phân loại hoặc hồi quy. Trong phân tích này chúng ta sẽ sử dụng SVM cho phân loại nhị phân .



**Ý tưởng thuật toán** : tìm một siêu phẳng (**hyper lane**) để phân tách các điểm dữ liệu. Siêu phẳng này sẽ chia không gian thành các miền khác nhau và mỗi miền sẽ chứa một loại dữ liệu. Vấn đề đặt ra là có rất nhiều siêu phẳng vậy phải chọn cái nào cho tối ưu nhất . Siêu phẳng tối ưu mà chúng ta cần chọn là siêu phẳng phân tách có margin **[1]** lớn nhất – theo lý thuyết một mặt phẳng như thế sẽ hạn chế thấp nhất lỗi mắc phải .

**Các bước thực hiện :**

**B1** : Phương pháp SVM yêu cầu dữ liệu được diễn tả như các vector của các số thực. Như vậy nếu đầu vào chưa phải là số thì ta cần phải tìm cách chuyển chúng về dạng số của SVM .

**B2** : Tiền xử lý dữ liệu :Thực hiện biến đổi dữ liệu phù hợp cho quá trình tính toán, tránh các số quá lớn mô tả các thuộc tính. Thường nên co giãn (scaling) dữ liệu để chuyển về đoạn [-1, 1] hoặc [0, 1].

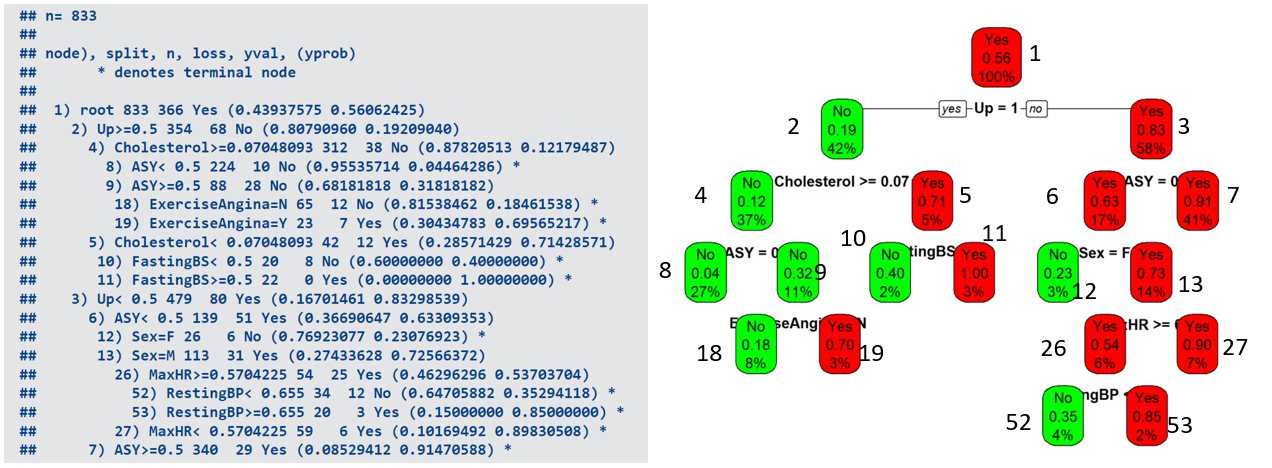
**B3** : Chọn hàm hạt nhân: Lựa chọn hàm hạt nhân phù hợp tương ứng cho từng bài toán cụ thể để đạt đƣợc độ chính xác cao trong quá trình phân lớp.

**B4** : Thực hiện việc kiểm tra chéo để xác định các thám số cho ứng đụng. Điều này cũng quyết định đến tính chính xác của quá trình phân lớp.

**B5** : Sử dụng các tham số cho việc huấn luyện với tập mẫu. Trong quá trình huấn luyện sẽ sử dụng thuật toán tối ưu hóa khoảng cách giữa các siêu phẳng trong quá trình phân lớp, xác định hàm phân lớp trong không gian đặc trưng nhờ việc ánh xạ dữ liệu vào không gian đặc trưng bằng cách mô tả hạt nhân, giải quyết cho cả hai trường hợp dữ liệu là phân tách và không phân tách tuyến tính trong không gian đặc trưng.

* 1. Mô hình 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.



**Ý tưởng thuật toán** : Dựa trên việc rẽ nhánh, chúng ta suy ra được nhãn dự báo cho các quan sát . Đường biên phân chia của cây quyết định sẽ dựa trên kịch bản rẽ nhánh mà chúng ta lựa chọn . Giả sử giá trị đầu vào là ST-Slope up = 1 , cholesterol = 0.1 , FastingBS = 0 . Khi đó xuất phát từ nút gốc , ta lần lượt đặt các câu hỏi :

ST\_Slope Up = 1 là đúng nude gốc sẽ rẽ sang nhánh bên trái (1 -> 2).

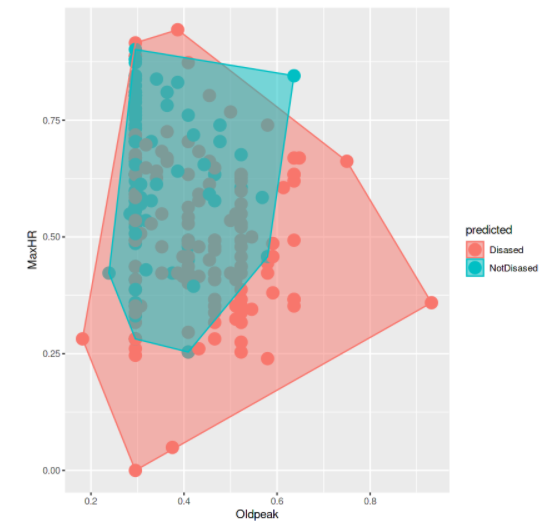
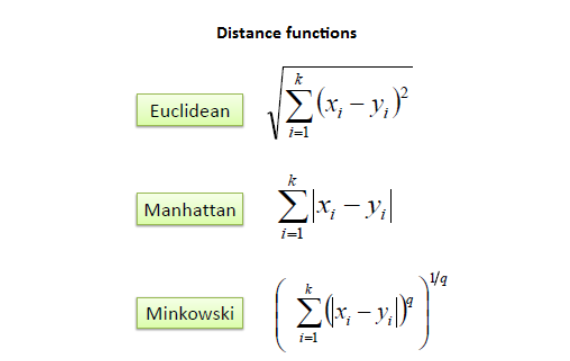
Tiếp theo , cholesterol = 0.1 > 0.07 , rẽ sang phải (2 -> 5) .

Cuối cùng , FastingBS = 0 , rẽ sang trái (5 -> 10) . Kết quả dự đoán ở node lá là No – Không mắc bệnh tim .

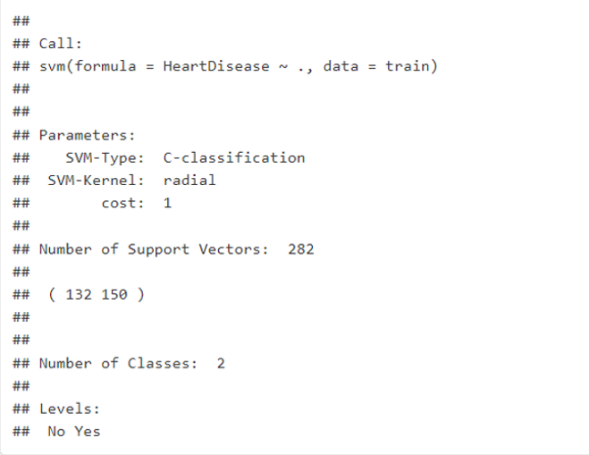
**Khởi tạo cây quyết định** : Để khởi tạo cây chúng ta cần biết được thứ tự câu hỏi là gì và cách đặt câu hỏi như thế nào . Đối với những bộ dữ liệu có số lượng biến đầu vào lớn, việc lựa chọn biến nào sẽ được đặt ở node gốc là điều không dễ dàng. Một kịch bản lựa chọn ngẫu nhiên chỉ có xác suất đúng là 1/d và dường như khả năng chúng ta chọn sai là rất cao. Do đó chúng ta cần phải có một tiêu chí nào đó để lựa chọn biến phù hợp. Xuất phát từ yêu cầu này, chúng ta hình thành nên các độ đo như entropy **[2]** , Gini đo lường mức độ tinh khiến (purity) và vẩn đục (impurity) **[3]** của một biến .

* 1. Mô hình KNN

K-nearest neighbors là thuật toán học máy có giám sát, đơn giản và dễ triển khai. Thường được dùng trong các bài toán phân loại và hồi quy. Trong trường hợp này , chúng ta sử dụng KNN cho phân loại .

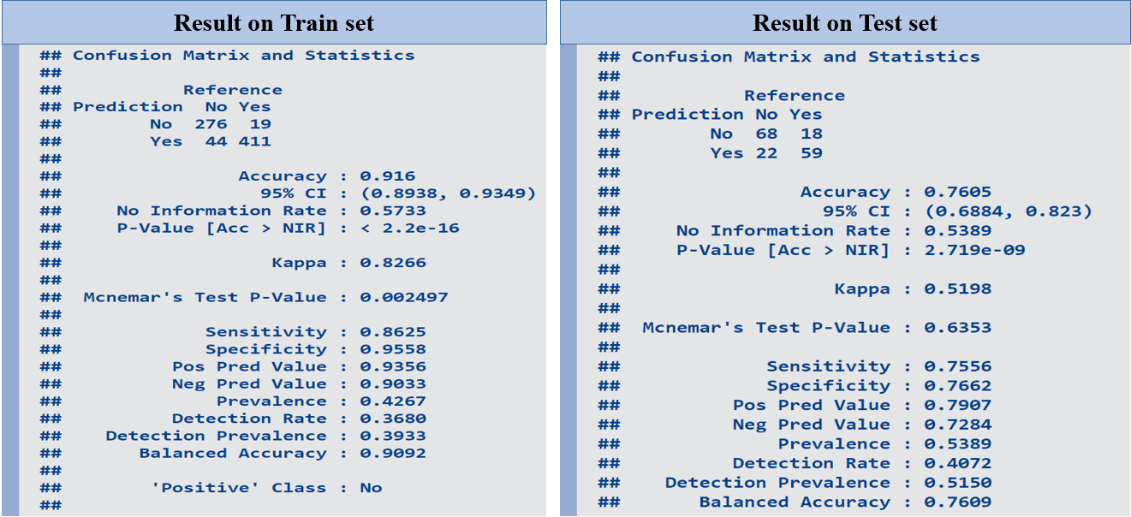
**Ý tưởng thuật toán** : Theo thuật toán KNN những dữ liệu tương tự nhau sẽ tồn tại gần nhau trong một không gian, từ đó công việc của chúng ta là sẽ tìm k điểm gần với dữ liệu cần kiểm tra nhất. Việc tìm khoảng cách giữa 2 điểm củng có nhiều công thức có thể sử dụng, tùy trường hợp mà chúng ta lựa chọn cho phù hợp. Đây là 3 cách cơ bản để tính khoảng cách 2 điểm dữ liệu x, y có k thuộc tính:

1. **Thực nghiệm , kết quả , thảo luận**
   1. Mô hình SVM

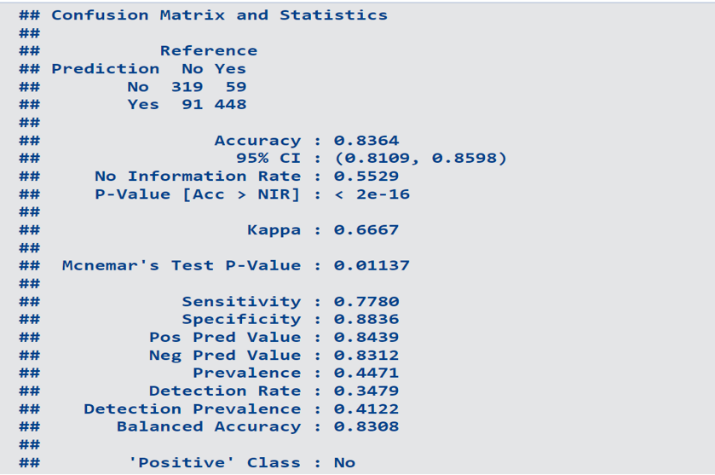
 Các tham số truyền vào : SVM-type : C-classification , kernel : radial , cost = 1 .

Sau khi train model trên tập train (80% ) . Số lớp được phân ra là 2 : No (Không mắc bệnh) , Yes (có mắc bệnh ) .

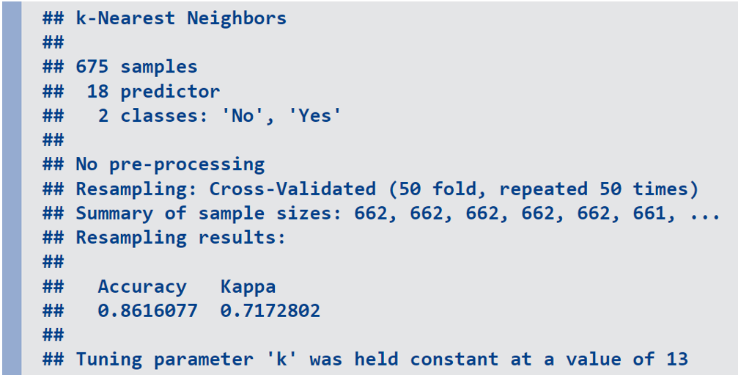
Xác suất dự doán chính xác trên tập train là 91,6% , trên tập test là 76,05% - có sự chênh lệnh lớn . Hơn nữa nhìn vào Confusion matrix của cả hai tập có thể thấy mô hình bị bias và variance .



* 1. Mô hình Decision tree

Xác suất dự đoán của mô hình đạt : 83,64 % . Nhìn vào confusion matrix ta cũng có thể thấy mô hình bị bias và variance .

* 1. Mô hình KNN

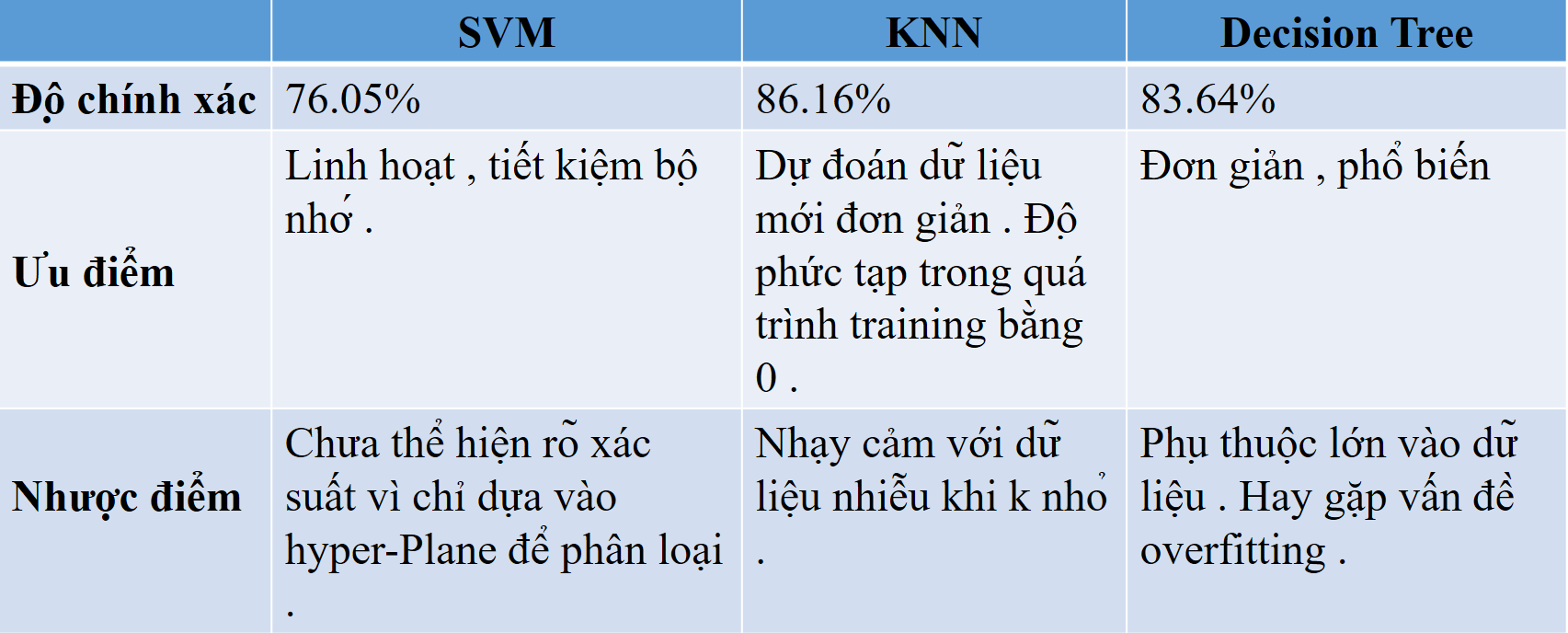
 **Cross-validation** : Nhóm chia làm 50 phần (fold).

**Độ chính xác** dự đoán của mô hình là : 86,16% .

Thử nghiệm với K từ 1 đến 40 . **K = 13** cho kết quả cao nhất .

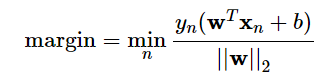
1. **Kết luận**

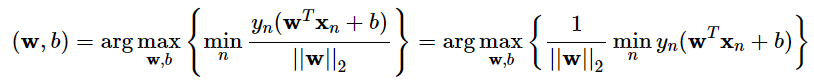
Qua quá trình thực hiện đề tài nhóm em đã học hỏi được thêm kiến thức về học máy ứng dụng ngôn ngữ R cho phân tích dữ liệu . Về tổng quan đề tài đã hoàn thành mức cơ bản , trả lời được các câu hỏi nhóm đã đặt ra với bộ dữ liệu . Tuy nhiên kết quả dự đoán của các mô hình trong đề tài chưa được cao . Sau đây là nhận xét về các thuật toán nhóm đã sử dụng :

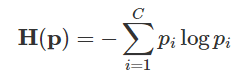


**Hướng phát triển đề tài** : Nhóm em sẽ tiếp tục cải thiện mô hình KNN (*mô hình có hiệu suất cao nhất* ) , cố gắng giảm thiểu avoidable bias bằng cách tăng kích thước quan sát và giảm variance của các mô hình bằng cách loại bỏ bớt một số biến có độ tương quan quá thấp . Hơn nữa qua phần data visualization có một số dữ liệu gây nhiễu , vì thế cần loại bỏ .

1. **Phụ lục**

**[1]** Margin được tính là khoảng cách gần nhất từ 1 điểm tới mặt đó (bất kể điểm nào trong hai classes): Bài toán tối ưu trong SVM chính là bài toán tìm w và b sao cho margin này đạt giá trị lớn nhất:

****

**[2]** Trong thuật toán cây quyết định chúng ta sẽ sử dụng Entropy để đánh giá mức độ tinh khiết của phân phối xác suất của một sự kiện. Giả sử một sự kiện xảy ra với phân phối xác suất là : P (p1,p2,..pc) thỏa mãn  Khi đó hàm entropy đối với sự kiện trên là: 

Trong trường hợp  thì hàm entropy là không xác định do  không tồn tại, tuy nhiên căn cứ vào giá trị hội tụ thì chúng ta sẽ qui ước giá trị của entropy trong trường hợp này là 0. Giá trị lớn nhất đạt được khi p0=p1=1/2. Trong khi giá trị nhỏ nhất đạt được khi một trong hai xác suất bằng 1 và xác suất còn lại bằng 0.

**[3]** Khi đối diện với việc phân loại đó là kết quả trả về tại node lá chỉ thuộc về một lớp. Chúng ta sử dụng một thuật ngữ ngắn gọn để gọi tên trường hợp này là tinh khiết (purity). Trái ngược lại với tinh khiết sẽ là khái niệm vấn đục (impurity), tức phân phối của các nhãn tại node lá còn khá mập mờ, không có xu hướng thiên về một nhãn nào cụ thể. Nếu ra quyết định phân loại dựa trên kịch bản dẫn tới node lá sẽ trả về kết quả không đáng tin cậy.

1. **Đóng góp**

**Nhóm trưởng**: Hoàng Minh Nhật **MSSV**:19133042

Công việc : Câu 5 : Xây dựng mô hình dự đoán bệnh tim dựa trên thuật toán SVM, tham khảo mô hình Decion tree , KNN .Tổng hợp và chỉnh sửa , viết báo cáo .

**Thành viên**: Trần Nguyên Thái Bảo MSSV: 19133010

Công việc : Phần 1. giới thiệu , Câu 2 : Phân tích các yếu tố có ảnh hưởng như thế nào đối với bệnh tim ? đâu là những yếu tố chính gây ra bệnh tim ?

**Thành Viên**: Đào Thị Cẩm Tiên MSSV: 19133055

Công việc : Phần 2 . Dữ liêụ , Câu 1 : Các yếu tố ảnh hưởng đến bệnh tim có phân bố như thế nào ?

**Thành viên**: Đinh Quốc Hùng MSSV: 19133025

Công việc : Câu 3 : Có phải nữ có nguy cơ mắc bệnh tim cao hơn nam hay không ? , Câu 4 : Ước lượng độ tuổi có nguy cơ mắc bệnh tim cao ?

* **Nhận xét** : Các thành viên đều tham gia làm đề tài , thảo luận , thuyết trình .

1. **Tài liệu tham khảo**

Support Vector Machines \* The Interface to libsvm in package e1071 : <https://cran.r-project.org/web/packages/e1071/vignettes/svmdoc.pdf>

<https://data-flair.training/blogs/machine-learning-vs-data-science/>

<https://phamdinhkhanh.github.io/deepai-book/ch_ml/DecisionTree.html>

<https://www.youtube.com/watch?v=PwyiKYiiNJI&list=PLDpRz2wA0qZzTcDLeXP5PSCfmQ96l9-Qr&index=59>