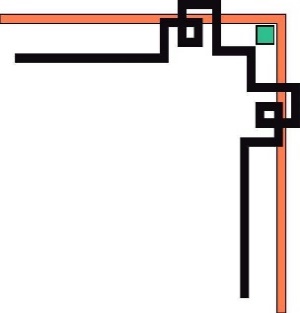
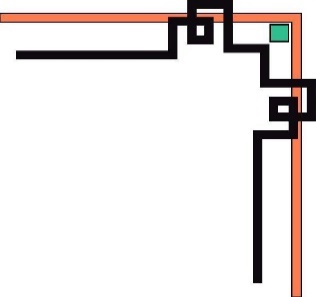
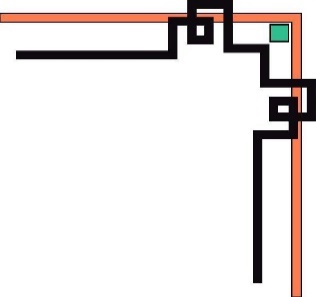
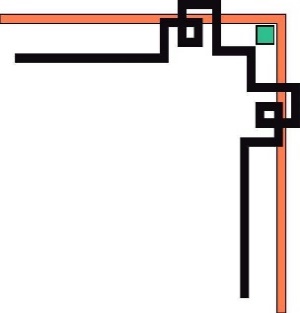
**ĐẠI HỌC HUẾ**



**KHOA KỸ THUẬT VÀ CÔNG NGHỆ**

🙠🙟🕮🙝🙢

**Icon

Description automatically generated**

**BÁO CÁO**

**ĐỒ ÁN**

**Học kỳ II, năm học 2021 – 2022**

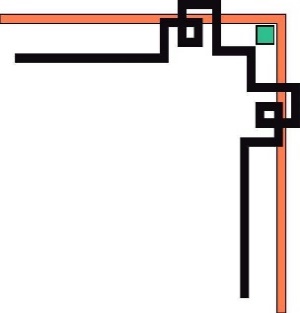
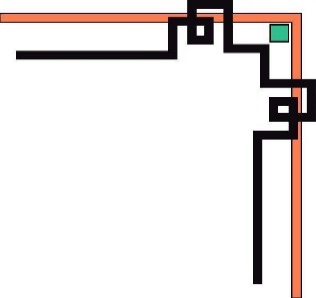
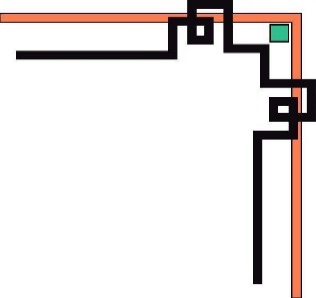
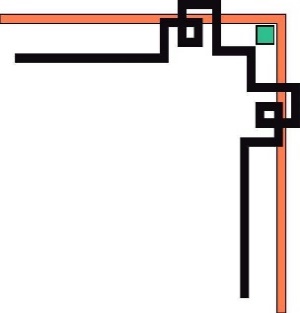
**HỌC PHẦN: ĐỒ ÁN 1**

**Đề tài: Xây dựng mô hình hồi quy tuyến tính để chẩn đoán xơ vữa động mạch**

|  |
| --- |
| **Số phách**  *(Do hội đồng chấm thi ghi)* |

**Thừa Thiên Huế, ngày….tháng….năm 2022**

**ĐẠI HỌC HUẾ**



**KHOA KỸ THUẬT VÀ CÔNG NGHỆ**

🙠🙟🕮🙝🙢

**Icon

Description automatically generated**

**(Bìa phụ 2)**

**BÁO CÁO**

**ĐỒ ÁN**

**Học kỳ II, năm học 2021 – 2022**

**HỌC PHẦN: ĐỒ ÁN 1**

**Đề tài: Xây dựng mô hình hồi quy tuyến tính để chẩn đoán xơ vữa động mạch**

**Giảng viên hướng dẫn: Nguyễn Đình Hoa Cương**

**Lớp: Khoa học dữ liệu & Trí tuệ nhân tạo – Khóa 1**

**Sinh viên thực hiện: Tôn Thất Bảo Văn**

*(ký và ghi rõ họ tên)*

|  |
| --- |
| **Số phách**  *(Do hội đồng chấm thi ghi)* |

**Thừa Thiên Huế, ngày….tháng….năm 2022**

**ĐẠI HỌC HUẾ**

**KHOA KỸ THUẬT VÀ CÔNG NGHỆ**



**PHIẾU ĐÁNH GIÁ ĐỒ ÁN/TIỂU LUẬN/BÀI TẬP LỚN Học kỳ I, năm học 2021 - 2022**

|  |  |
| --- | --- |
| **Cán bộ chấm thi 1** | **Cán bộ chấm thi 2** |
| **Nhận xét:**  ..............................................................  ..............................................................  ..............................................................  ..............................................................  ..............................................................  ..............................................................  ..............................................................  ..............................................................  ..............................................................  ..............................................................  **Điểm đánh giá của CBCT1:**  Bằng số: .........................................  Bằng chữ: ....................................... | **Nhận xét:**  ..............................................................  ..............................................................  ..............................................................  ..............................................................  ..............................................................  ..............................................................  ..............................................................  ..............................................................  ..............................................................  ..............................................................  **Điểm đánh giá của CBCT2:**  Bằng số: .........................................  Bằng chữ: ....................................... |

Điểm kết luận:

Bằng số:

Bằng chữ:

*Thừa Thiên Huế, ngày…tháng…năm 2022*

**Cán bộ chấm thi 1 Cán bộ chấm thi 2**

*(Ký và ghi rõ họ và tên) (Ký và ghi rõ họ và tên)*

# MỤC LỤC

[MỤC LỤC I](#_Toc106176105)

[DANH MỤC HÌNH VẼ II](#_Toc106176106)

[CHƯƠNG 1: CƠ SỞ LÝ THUYẾT 3](#_Toc106176107)

[1. Hồi quy tuyến tính 3](#_Toc106176108)

[2. Phương trình hồi quy tuyến tính 3](#_Toc106176109)

[3. Hàm mất mát 4](#_Toc106176110)

[4. Tìm nghiệm của mô hình hồi quy tuyến tính 4](#_Toc106176111)

[5. Mô tả dữ liệu 5](#_Toc106176112)

[CHƯƠNG 2: XÂY DỰNG MÔ HÌNH HỒI QUY 6](#_Toc106176113)

[1. Đọc dữ liệu từ file .csv 6](#_Toc106176114)

[2. Thống kê mô tả 7](#_Toc106176115)

[2.1. Tính toán các chỉ số cần thiết 7](#_Toc106176116)

[2.2. Vẽ biểu đồ 8](#_Toc106176117)

[3. Chia tập dữ liệu và xây dựng mô hình hồi quy 10](#_Toc106176118)

[3.1. Chia tập dữ liệu 10](#_Toc106176119)

[3.2. Huấn luyện mô hình 10](#_Toc106176120)

[CHƯƠNG 3. THÍ NGHIỆM VÀ KẾT LUẬN 13](#_Toc106176121)

[1. Kết quả 13](#_Toc106176122)

[2. So sánh các mô hình hồi quy 14](#_Toc106176123)

[2.1. Hồi quy ridge 14](#_Toc106176124)

[2.2. Hồi quy lasso 14](#_Toc106176125)

[2.3. So sánh các kết quả 15](#_Toc106176126)

[3. Kết luận 15](#_Toc106176127)

[TÀI LIỆU THAM KHẢO 16](#_Toc106176128)

[KẾT QUẢ KIỂM TRA ĐẠO VĂN 17](#_Toc106176129)

# DANH MỤC HÌNH VẼ

[Hình 1. Các thuộc tính của bộ dữ liệu 6](#_Toc106176130)

[Hình 2. Đọc dữ liệu từ file csv 6](#_Toc106176131)

[Hình 3. Các chỉ số cần thiết 7](#_Toc106176132)

[Hình 4. Biểu đồ histogram của các cột dữ liệu 8](#_Toc106176133)

[Hình 5. Biểu đồ Boxplot của các cột dữ liệu 9](#_Toc106176134)

[Hình 6. Biểu đồ sau khi đã loại bỏ giá trị ngoại lai 10](#_Toc106176135)

[Hình 7. Kết quả dự đoán của mô hình hồi quy tuyến tính 13](#_Toc106176136)

[Hình 8. Mức độ lỗi của mô hình tuyến tính 13](#_Toc106176137)

[Hình 9. Bảng biểu so sánh giá trị y dự đoán và y thực tế 14](#_Toc106176138)

[Hình 10. Kết quả dự đoán của mô hình ridge 14](#_Toc106176139)

[Hình 11. Mức độ lỗi của mô hình ridge 14](#_Toc106176140)

[Hình 12. Kết quả dự đoán của mô hình lasso 14](#_Toc106176141)

[Hình 13. Mức độ lỗi của mô hình lasso 15](#_Toc106176142)

[Hình 14. So sánh mức độ lỗi của các mô hình 15](#_Toc106176143)

[Hình 15. Kiểm tra đạo văn 17](#_Toc106176144)

# CHƯƠNG 1: CƠ SỞ LÝ THUYẾT

## Hồi quy tuyến tính

Hồi quy tuyến tính la một phương pháp thống kê để hồi quy dữ liệu với biến phụ thuộc có giá trị liên tục trong khi các biến độc lập có thể có một trong hai giá trị liên tục hoặc là giá trị phân loại. Hồi quy tuyến tính là một trong hai dạng lớn của học có giám sát (supervised learning) dựa trên tập dữ liệu mẫu.

Nói cách khác "Hồi quy tuyến tính" là một phương pháp để dự đoán biến phụ thuộc (Y) dựa trên giá trị của biến độc lập (X). Nó có thể được sử dụng cho các trường hợp chúng ta muốn dự đoán một số lượng liên tục. Ví dụ, dự đoán giao thông ở một cửa hàng bán lẻ, dự đoán thời gian người dùng dừng lại một trang nào đó hoặc số trang đã truy cập vào một website nào đó v.v...

## Phương trình hồi quy tuyến tính

Hồi quy tuyến tính có phương trình đạng :

F(x) = w0 + w1x1 + w2x2 + … + wnxn (1)

Trong đó, w1, w2, wn, w0 là các hằng số, w0 còn được gọi là bias hay sai số. Mối quan hệ giữa y f(x) bên trên là một mối quan hệ tuyến tính (linear). Bài toán chúng ta đang làm là một bài toán thuộc loại regression. Bài toán đi tìm các hệ số tối ưu { w1, w2, wn, w0 } chính vì vậy được gọi là bài toán Linear Regression (Hồi quy tuyến tính).

Trong phương trình (1) nếu chúng ta đặt **w = [**w0, w1, w2, wn**]**T là một vecter (cột) hệ số cần phải tối ưu và = [1, x1, x2, xn] (đọc là x bar trong tiếng Anh) là vector (hàng) dữ liệu đầu vào mở rộng. Số 1 ở đầu được thêm vào để phép tính đơn giản hơn và thuận tiện cho việc tính toán. Khi đó, phương trình (1) có thể được viết lại dưới dạng:

y ( trong đó là một vecter hàng)

## Hàm mất mát

Máy học từ giá trị trung bình của một hàm mất mát. Đây là một phương pháp đánh giá độ hiệu quả của một thuật toán nào đó trên bộ dữ liệu cho trước. Nếu kết quả dự đoán chênh lệch quá nhiều so với kết quả thực tế, hàm mất mát sẽ là một số rất lớn. Điều tương tự xảy ra với tất cả các cặp (xi, yi), i = 1, 2, 3, …, N với N là số lượng dữ liệu quan sát được. Để hàm mất mát nhỏ nhất khi đó tổng sai số là nhỏ nhất tương đương với việc tìm **w** để hàm số sau đạt giá trị nhỏ nhất:

J(w) = (2)

Hàm số J(w) được gọi là hàm mất mát (loss function) của bài toán Linear Regression. Chúng ta luôn mong muốn rằng sự mất mát (sai số) là nhỏ nhất, điều đó đồng nghĩa với việc tìm vector hệ số **w** sao cho giá trị của hàm mất mát này càng nhỏ càng tốt.

Trước khi đi tìm lời giải, chúng ta đơn giản hóa phép toán trong phương trình hàm mất mát (2). Đặt là một vector cột chứa tất cả các output của training data; = là ma trận dữ liệu đầu vào (mở rộng) mà mỗi hàng của nó là một điểm dữ liệu.

## Tìm nghiệm của mô hình hồi quy tuyến tính

Để tìm nghiệm cho một bài toán tối ưu chúng ta thường giải phương trình đạo hàm J(w) = bằng 0.

Đạo hàm theo **w** của hàm mất mát là:



Phương trình đạo hàm bằng 0 tương đương với:

Đặt và b ( ) khi đó ta có :

A.W = b

(với I là ma trận đơn vị)

là nghiệm của mô hình hồi quy tuyến tính.

Trên thực tế A có thể không khả nghịch nên ta sẽ dùng ma trận giả nghịch đảo nên ta có W = hay W = Đây chính là nghiệm tổng quát của hồi quy tuyến tính.

## Mô tả dữ liệu

Bộ dữ liệu xây dựng là một bộ dữ liệu được xây dựng bởi một bệnh viện trong quá trình khảo sát xơ vữa động mạch ở 100 bệnh nhân khác nhau. Bộ dữ liệu gồm có 6 thuộc tính bao gồm: tuổi, chỉ số BMI, chỉ số huyết áp tối đa (mmHg), nồng độ glucose trong máu (mmol/l), nồng dộ cholesterol trong máu (mmol/l) và bề dày thành mạch (mm).

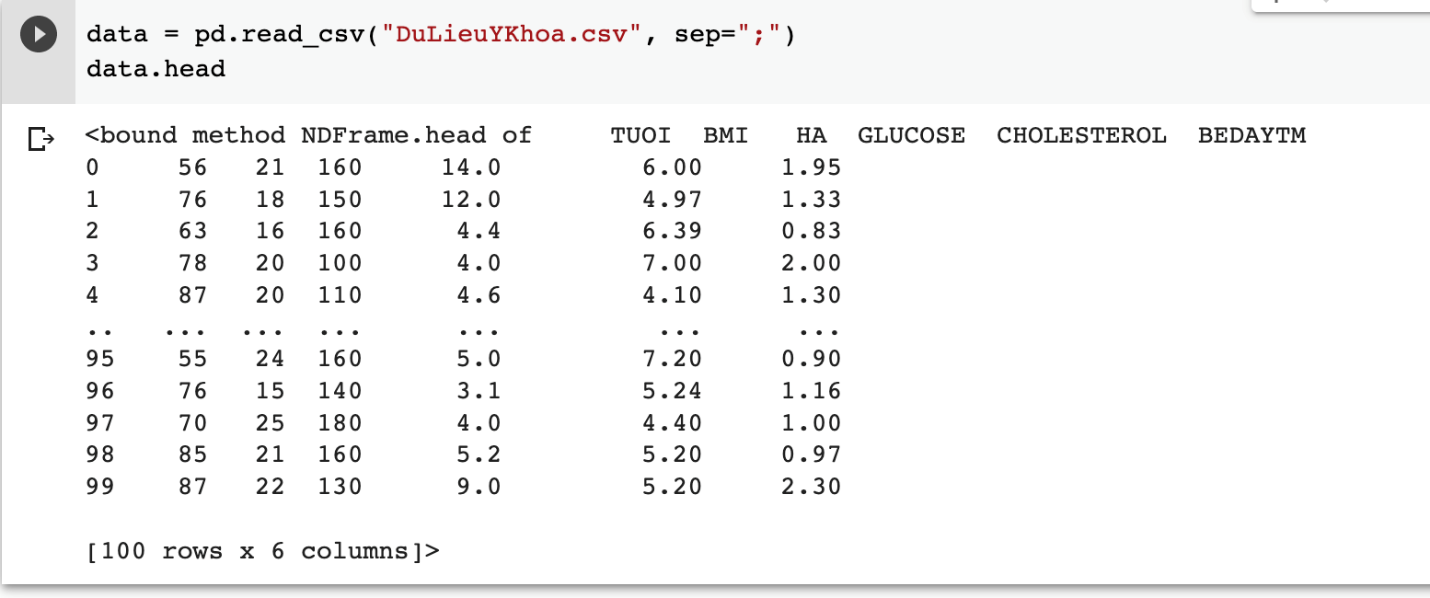
Bộ dữ liệu sẽ chia thành 2 phần trong đó 80% dữ liệu được làm bộ dữ liệu huấn luyện mô hình, 20% còn lại làm bộ dữ liệu thử nghiệm độ chính xác của mô hình. Dữ liệu về độ dày thành mạch làm biến mục tiêu, những bộ dữ liệu còn lại bao gồm (tuổi, BMI, nồng độ glucose máu, nồng độ cholesterol máu) .

|  |  |
| --- | --- |
| **Thuộc tính** | **Ý Nghĩa** |
| Tuổi | Tuổi |
| BMI | Chỉ số sức khỏe |
| HA | Huyết áp tối đa |
| Glucose | Nồng độ Glucose |
| Cholesterol | Nồng độ Cholesterol |
| Bedaytm | Bề dày thành mạch |

Hình . Các thuộc tính của bộ dữ liệu

# CHƯƠNG 2: XÂY DỰNG MÔ HÌNH HỒI QUY

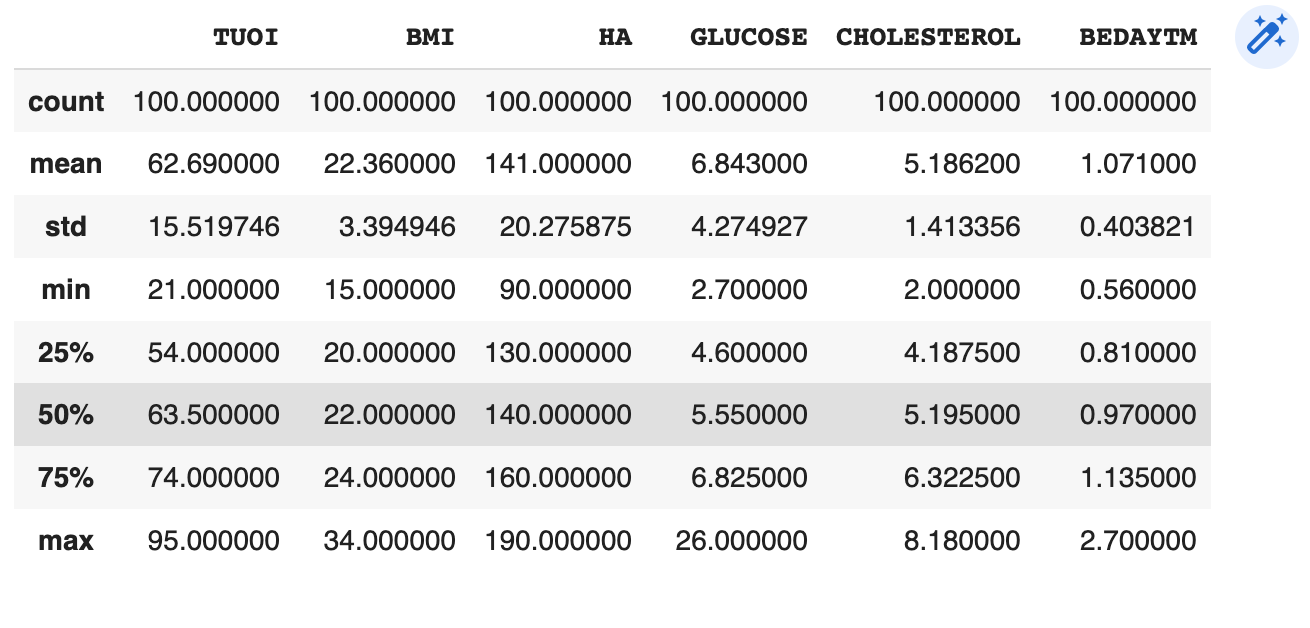
## Đọc dữ liệu từ file .csv



Hình . Đọc dữ liệu từ file csv

## Thống kê mô tả

### Tính toán các chỉ số cần thiết

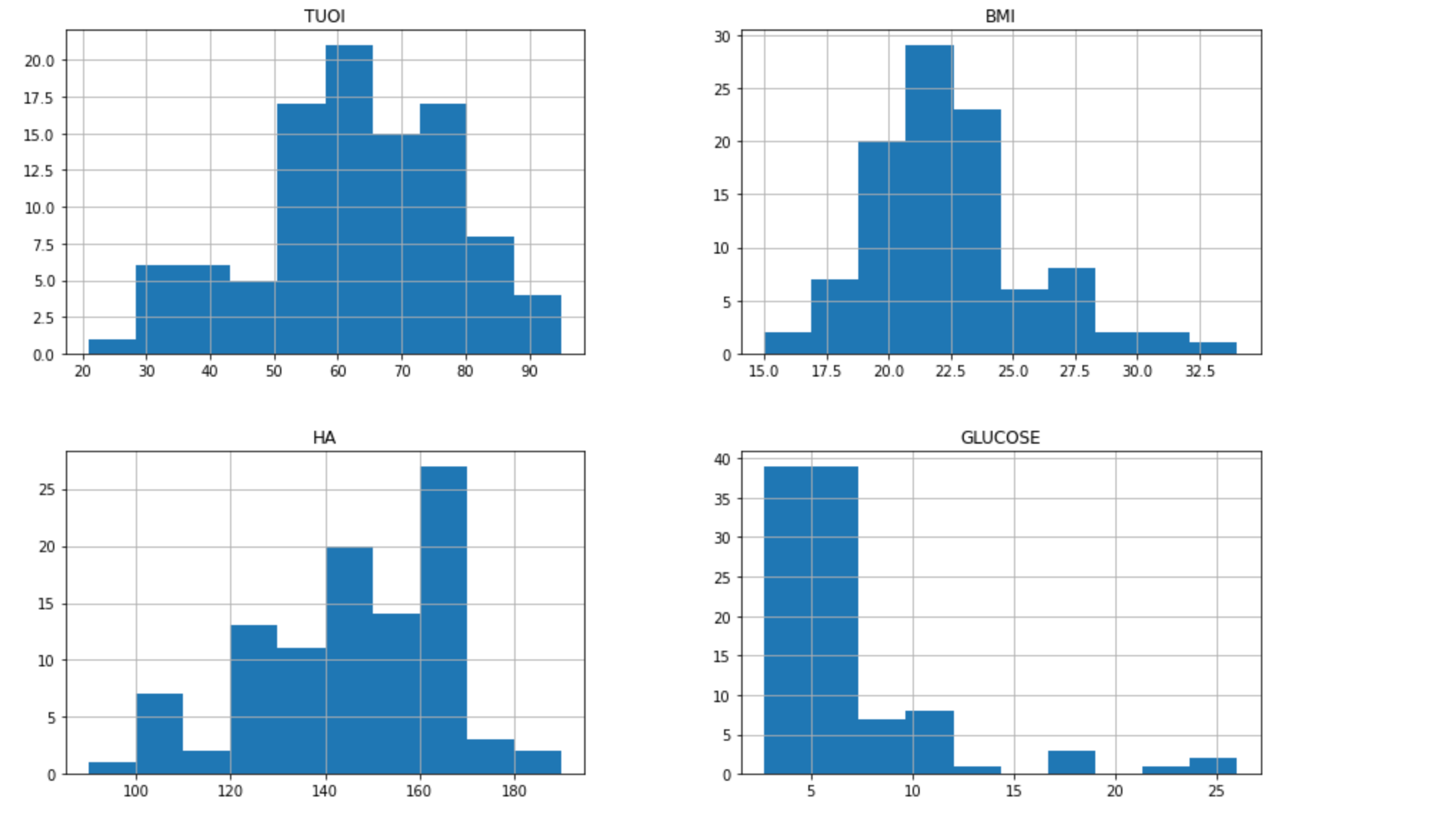


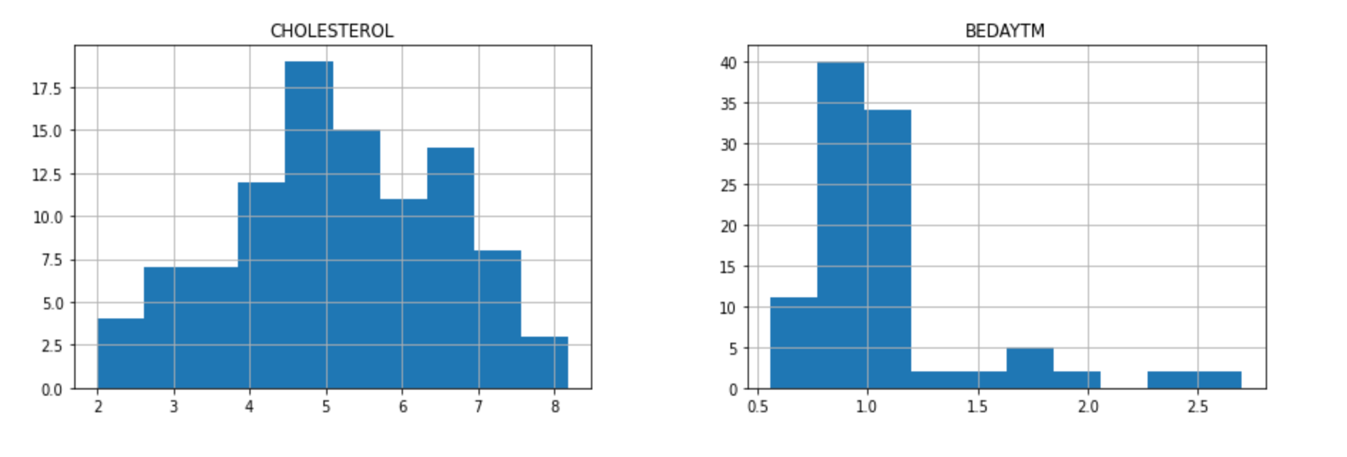
Hình . Các chỉ số cần thiết

Ta sử dụng hàm describe để tính toán các giá trị cần thiết như giá trị trung bình, phương sai, độ lệch chuẩn, tứ phân vị…

### Vẽ biểu đồ

#### Biểu đồ Histogram

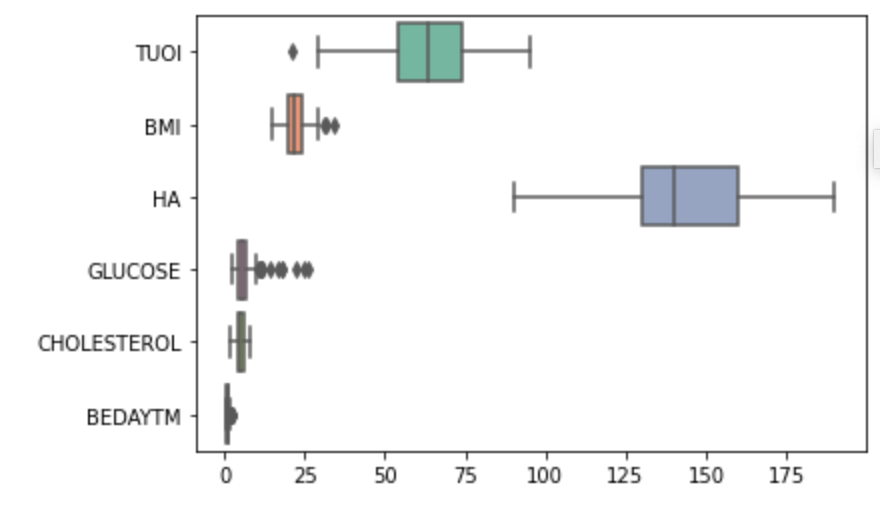
****

****

Hình . Biểu đồ histogram của các cột dữ liệu

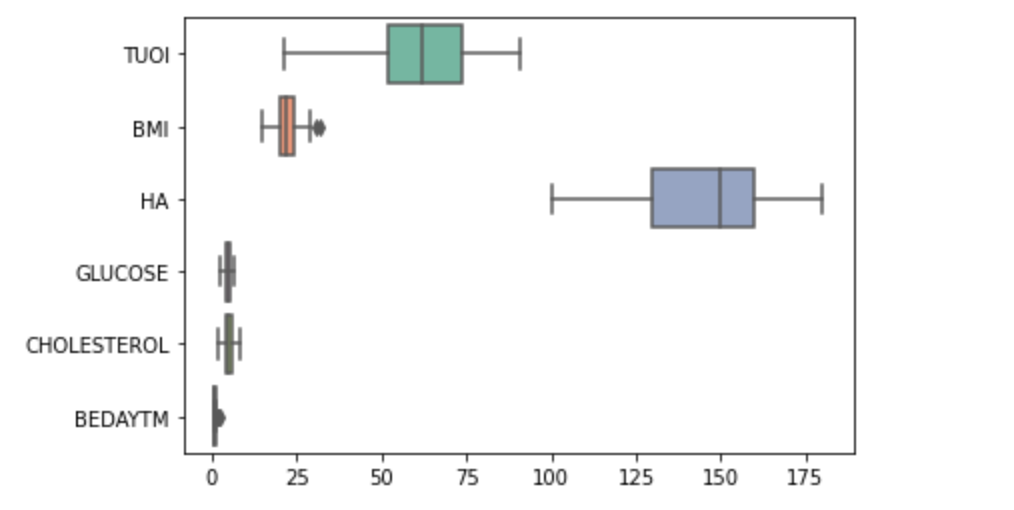
Nhìn vào biểu đồ ta có thể thấy dữ liệu k được đẹp lắm, không tuân theo phân phối chuẩn.

#### Biểu đồ Boxplot

****

Hình . Biểu đồ Boxplot của các cột dữ liệu

Nhìn vào biểu đồ ta thầy được ở cột GLUCOSE có khá nhiều giá trị ngoại lai làm cho quá trình huấn luyện mô hình trở nên khó khăn vì vậy ta cần loại bỏ các giá trị đó. Còn các cột như BMI, TUOI,... Cũng có giá trị ngoại lai nhưng ít nên ta có thể không cần loại bỏ chúng.



Hình . Biểu đồ sau khi đã loại bỏ giá trị ngoại lai

Nhìn vào hình ta cũng thấy được các giá trị ngoại lai ở cột GLUCOSO đã được loại bỏ. Như vậy quá trình huấn luyện mô hình sẽ trở nên dễ dàng hơn.

## Chia tập dữ liệu và xây dựng mô hình hồi quy

### Chia tập dữ liệu

Ta sẽ chia dữ liệu lần lượt là 70 30 tức là 70% để train 30% để test

Đây là phần code:

*#Phân loại dữ liệu train và test*X\_train, X\_test, Y\_train, Y\_test = train\_test\_split(X, Y, test\_size = 30)

### Huấn luyện mô hình

#### Giới thiệu thư viện scikit-learn

Scikit-learn (viết tắt là sklearn) là một thư viện mã nguồn mở dành cho học máy - một ngành trong trí tuệ nhân tạo, rất mạnh mẽ và thông dụng với cộng đồng Python, được thiết kế trên nền NumPy và SciPy. Scikit-learn chứa hầu hết các thuật toán machine learning hiện đại nhất, đi kèm với documentations, luôn được cập nhật. Trong sklearn có class sklearn.linear\_model.LinearRegression có các phương thức liên quan tới hồi quy tuyến tính.

#### Các phương thức và các biến của class LinearRegression

* **Sử dụng**

sklearn.linear\_model.LinearRegression(fit\_intercept=True, normalize=False, copy\_X=True, n\_jobs=1)

* **Các biến**
* **fit\_intercept:** Không tính toán những điểm cắt trục tung nếu cài đặt là false, mặc định là true
* **normalize:** Nếu cài dặt làt true thì chuẩn hoá biến giải thích trước, mặc định là false
* **copy\_X:** Lưu dữ liệu trong bộ nhớ rồi mới chạy, mặc định là true
* **n\_jobs:** Số job sử dụng khi tính toán, -1 sẽ dùng hết, mặc định là 1
* **Các thuộc tính của class LinearRegression**
* **coef\_:** Trả về hệ số hồi quy
* **intercept\_:** Trả về sai số
* **Các phương thức**
* **fit(X, y[, sample\_weight]):** Tiến hành tìm phương trình hồi quy tuyến tính
* **get\_params([deep]):** Lấy parameter đã sử dụng
* **predict(X):** Sử dụng model vừa tạo được tiến hành dự đoán
* **score(X, y[, sample\_weight]):** Đưa ra hệ số quyết định R2. Không phải lúc nào giá trị dự đoán cũng giống vs giá trị thực tế. Hệ số quyết định ~ 1 thì độ chính xác càng cao, ~ 0 thì sai lệch lớn

#### Cài đặt

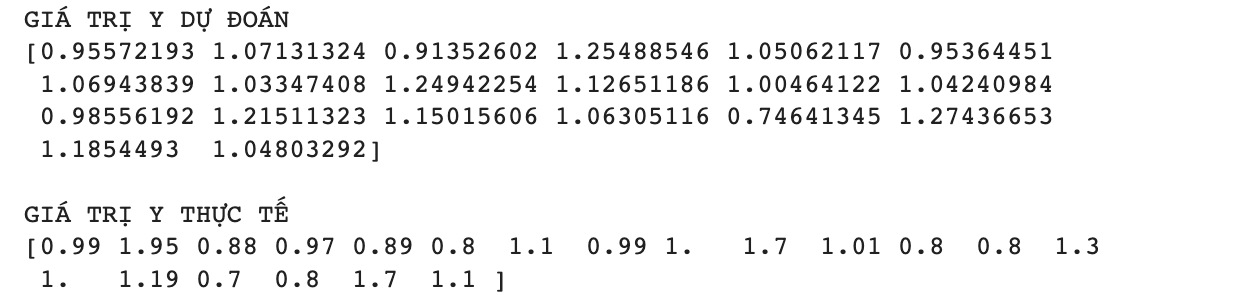
**from** sklearn.linear\_model **import** LinearRegression  
**from** sklearn.model\_selection **import** train\_test\_split  
**from** sklearn **import** metrics **as** sq  
**from** sklearn **import** linear\_model

#### Code huấn luyện mô hình

clf = linear\_model.LinearRegression()  
  
*#Đọc tệp*data = pd.read\_csv(**"DuLieuYKhoa.csv"**, sep=**";"**)  
data.head  
*#In dữ liệu*print(data)  
*#Lấy dataframe bề dày thành mạch làm biên mục tiêu*Y = data[**'BEDAYTM'**].to\_numpy()  
*#Lấy datafarame không chứa bề dày thành mạch làm biến giải thích*X = data.drop(**"BEDAYTM"**, axis = 1)  
*#Phân loại dữ liệu train và test*X\_train, X\_test, Y\_train, Y\_test = train\_test\_split(X, Y, test\_size = 20)  
*#Tạo model suy đoán*clf.fit(X\_train, Y\_train)  
*#In hệ số hồi quy của các biến giải thích xếp theo thứ tự tăng dần*print(**"\nHỆ SỐ HỒI QUY"**)  
print(pd.DataFrame({**"Tên"**: X\_train.columns, **"Hệ số"**: np.abs(clf.coef\_)}).sort\_values(by=**'Hệ số'**))  
*#In sai số*print(**"\nSAI SỐ"**)  
print(clf.intercept\_)  
*#print(clf.score())  
#Tiến hành dự đoán với bộ dữ liệu test*Y\_pred = clf.predict(X\_test)  
print(**"\nGIÁ TRỊ Y DỰ ĐOÁN"**)  
print(Y\_pred)  
*#In giá trị y test thực tế*print(**"\nGIÁ TRỊ Y THỰC TẾ"**) print(Y\_test)

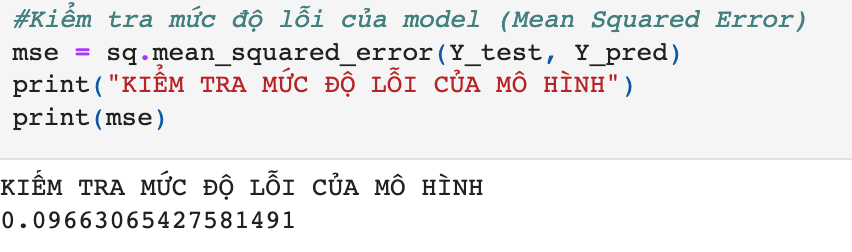
# CHƯƠNG 3. THÍ NGHIỆM VÀ KẾT LUẬN

## Kết quả

****

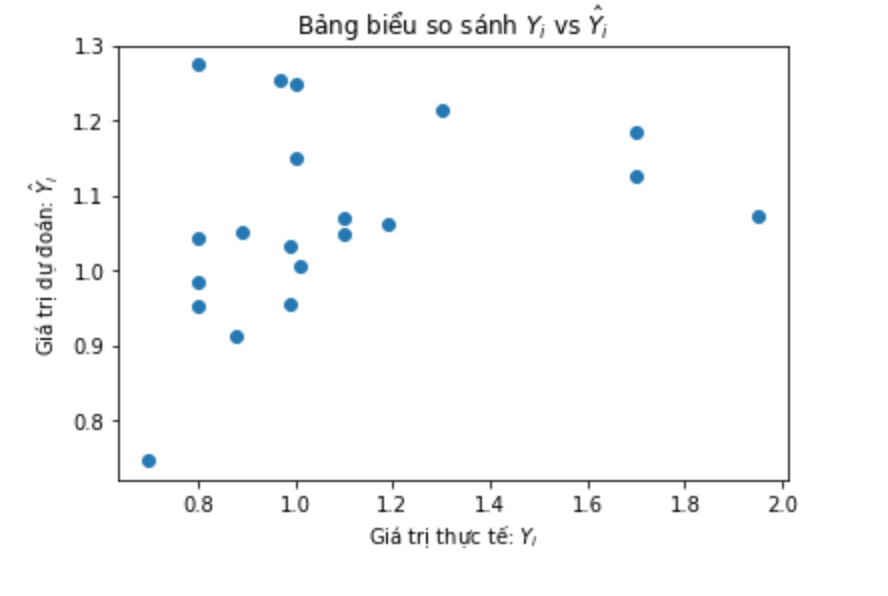
Hình . Kết quả dự đoán của mô hình hồi quy tuyến tính

Từ hình trên ta cũng thấy rõ được độ chính xác của của mô hình.



Hình . Mức độ lỗi của mô hình tuyến tính

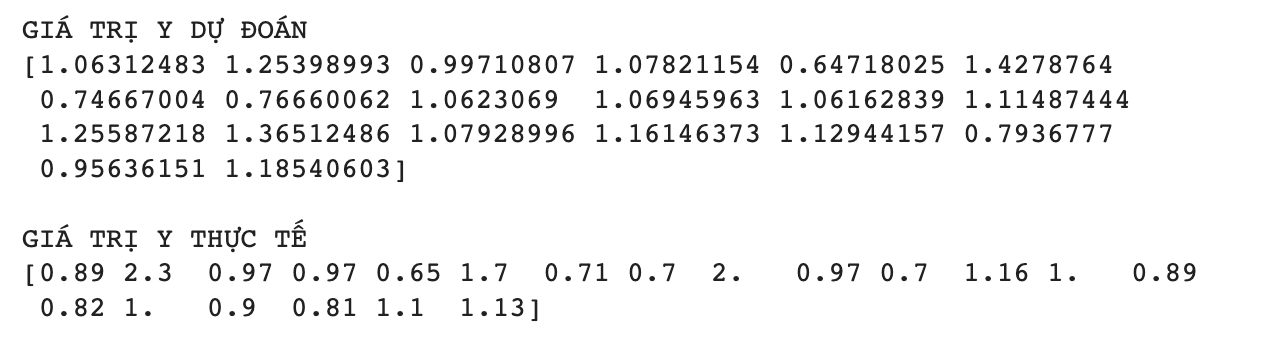
Để hình dung một cách rõ ràng hơn ta sẽ vẻ ra các điểm dự đoán.



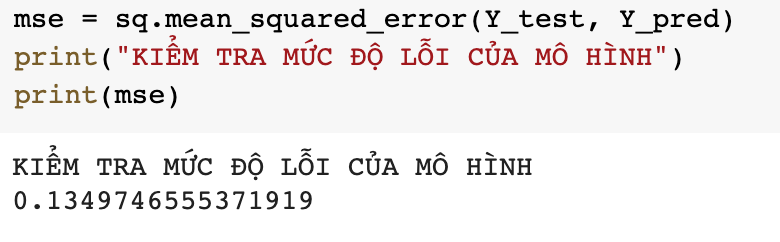
Hình . Bảng biểu so sánh giá trị y dự đoán và y thực tế

## So sánh các mô hình hồi quy

### Hồi quy ridge

****

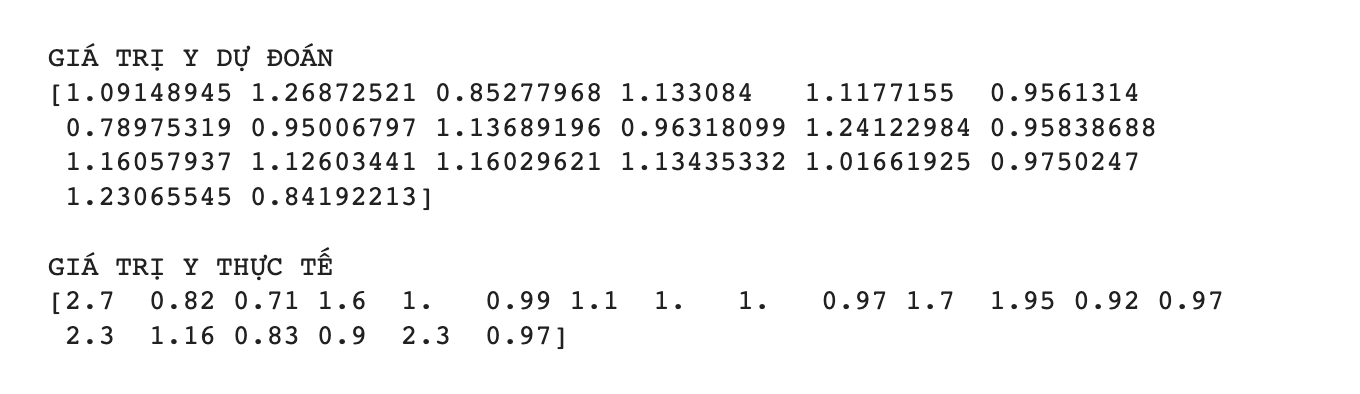
Hình . Kết quả dự đoán của mô hình ridge

****

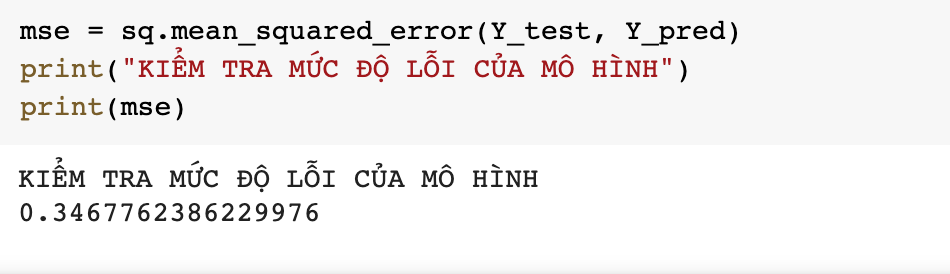
Hình . Mức độ lỗi của mô hình ridge

Từ 2 hình trên ta thấy được độ chính xác của mô hình này cũng tương đối cao.

### Hồi quy lasso

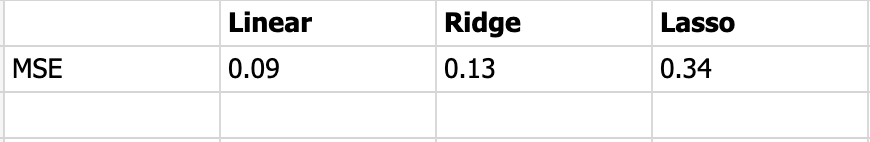
****

Hình . Kết quả dự đoán của mô hình lasso

****

Hình . Mức độ lỗi của mô hình lasso

### So sánh các kết quả

****

Hình . So sánh mức độ lỗi của các mô hình

Từ bảng trên ta thấy được mô hình hồi quy tuyến tính có độ chính xác cao nhất, tiếp đến là Ridge và cuối cùng là Lasso. Vậy nên trong bài toán này ta sử dụng mô hình hồi quy tuyến tính để huấn luyện là đúng đắng nhất.

## Kết luận

Kết quả đạt được : đã cài đặt được thuật toán và sử dụng dụng thư viện scikit-learn trong quá trình học tập. Nhưng bên cạnh đó thuật toán vẫn còn những ưu nhược điểm như:

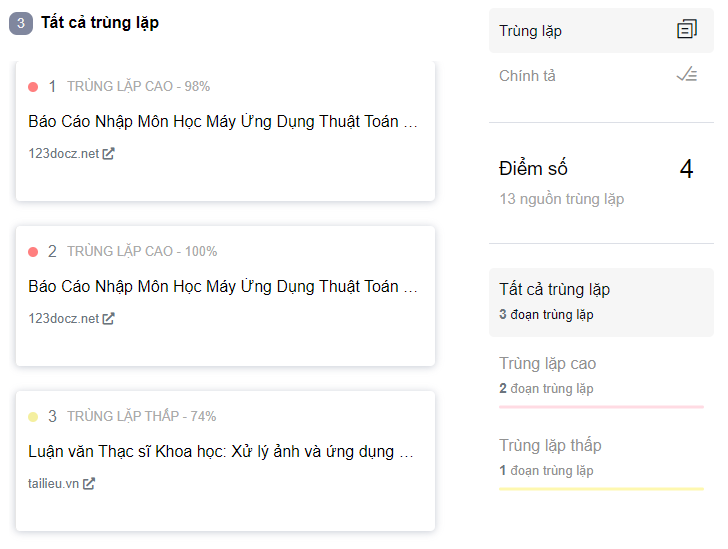
* Ưu điểm : Nhanh chóng để mô hình hóa và đặc biệt hữu ích khi mối quan hệ được mô hình hóa không quá phức tạp và nếu bạn không có nhiều dữ liệu. Hồi quy tuyến tính là đơn giản để hiểu, nó rất có giá trị cho các quyết định kinh doanh.
* Nhược điểm : Đối với dữ liệu phi tuyến tính, hồi quy đa thức có thể khá khó khăn để thiết kế, vì người ta phải có một số thông tin về cấu trúc của dữ liệu và mối quan hệ giữa các biến tính năng.

# TÀI LIỆU THAM KHẢO

1. [ML] Hồi quy tuyến tính : https://dominhhai.github.io/vi/2017/12/ml-linear-regression/
2. Machine Learning cơ bản: https://machinelearningcoban.com/2016/12/28/linearregression/
3. Phân tích hồi quy tuyến tính đa biến : <https://bit.ly/30xi3Vb>
4. 1. Schneider A, Hommel G, Blettner M. Linear regression analysis: part 14 of a series on evaluation of scientific publications. Dtsch Arztebl Int. 2010 Nov;107(44):776-82.
5. Tripepi G, Jager KJ, Stel VS, Dekker FW, Zoccali C. How to Deal with Continuous and Dichotomic Outcomes in Epidemiological Research: Linear and Logistic Regression Analyses. Nephron Clin Pract. 2011 Feb 23;118(4):c399- c406.
6. Hoàng Trọng và Chu Nguyễn Mộng Ngọc, Phân tích dữ liệu nghiên cứu với SPSS. Nhà xuất bản thống kê năm 2005

# KẾT QUẢ KIỂM TRA ĐẠO VĂN

<https://kiemtratailieu.vn/>



Hình . Kiểm tra đạo văn