**TRƯỜNG ĐẠI HỌC ĐIỆN LỰC**

**KHOA CÔNG NGHỆ THÔNG TIN**

****

**BÁO CÁO MÔN HỌC  
 NHẬP MÔN HỌC MÁY**

**ĐỀ TÀI**

**ỨNG DỤNG THUẬT TOÁN HỒI QUY TUYẾN TÍNH ĐỂ DỰ ĐOÁN XƠ VỮA ĐỘNG MẠCH**

**Sinh viên thực hiện : PHẠM HỒNG THÁI**

**Giảng viên hướng dẫn : PHẠM THỊ KIM DUNG**

**Ngành : CÔNG NGHỆ THÔNG TIN**

**Chuyên ngành : CÔNG NGHỆ PHẦN MỀM**

**Lớp : D14CNPM1**

**Niên Khóa : 2019-2024**

**Hà Nội, tháng 06 năm 2022**

**PHIẾU CHẤM ĐIỂM**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **STT** | **Họ và Tên sinh viên** | **Nội dung thực hiện** | **Điểm** | **Chữ ký** |
| 1 | Phạm Hồng Thái |  |  |  |

|  |  |  |
| --- | --- | --- |
| **Giảng viên chấm thi** | **Chữ ký** | **Ghi chú** |
| Giảng viên chấm thi 1: |  |  |
| Giảng viên chấm thi 2: |  |  |

**MỤC LỤC**

[LỜI CẢM ƠN 6](#_Toc106891763)

[LỜI MỞ ĐẦU 7](#_Toc106891764)

[CHƯƠNG 1 : THUẬT TOÁN HỒI QUY TUYẾN TÍNH TRONG HỌC MÁY 8](#_Toc106891765)

[1.1. Giới thiệu 8](#_Toc106891766)

[1.2. Phân tích toán học 8](#_Toc106891767)

[1.2.1. Dạng của Linear Regression 8](#_Toc106891768)

[1.2.2. Hàm mất mát 9](#_Toc106891769)

[1.2.3. Tìm nghiệm của mô hình hồi quy tuyến tính 9](#_Toc106891770)

[1.2.4. Mức độ lỗi của mô hình hồi quy tuyến tính 10](#_Toc106891771)

[CHƯƠNG 2 : ÁP DỤNG THUẬT TOÁN 11](#_Toc106891772)

[2.1. Mô tả bài toán 11](#_Toc106891773)

[2.2. Xây dựng bộ dữ liệu 11](#_Toc106891774)

[2.3. Thực nghiệm với thư viện scikit-learn của python 13](#_Toc106891775)

[2.3.1. Giới thiệu thư viện scikit-learn 13](#_Toc106891776)

[2.3.2. Các phương thức và các biến của class LinearRegression 13](#_Toc106891777)

[2.3.3. Cài đặt 14](#_Toc106891778)

[2.3.4 Kiểm nghiệm 16](#_Toc106891779)

[KẾT LUẬN 18](#_Toc106891780)

**LIỆT KÊ CHỮ VIẾT TẮT**

|  |  |  |
| --- | --- | --- |
| BMI | Body Mass Index | Chỉ số khối cơ thể |
| BEDAYTM |  | Bề dày thành mạch |
| HA | SBP | Huyết áp tối đa |
| mmHg |  | Đơn vị milimet thủy ngân |
| mmol/l |  | Milimol trên lit |
| Mm |  | Milimet |

**DANH MỤC HÌNH ẢNH**

Trang

Hình 1: Các thuộc tính của bộ dữ liệu 12

Hình 3: Bảng biểu so sánh giá trị Y dự đoán và Y thực tế 16

Hình 4: Kết quả huấn luyện bằng thư viện sklearn 17

# LỜI CẢM ƠN

Trong lời đầu tiên của báo cáo môn học “ Nhập môn học máy ” , chúng em muốn gửi những lời cám ơn và biết ơn chân thành nhất của mình tới tất cả những người đã hỗ trợ, giúp đỡ chúng em về kiến thức và tinh thần trong quá trình thực hiện báo cáo.

Em xin chân thành gửi lời cảm ơn tới các thầy cô giáo trong Trường Đại học Điện Lực nói chung và các thầy cô giáo trong Khoa Công nghệ thông tin nói riêng đã tận tình giảng dạy, truyền đạt cho em những kiến thức cũng như kinh nghiệm quý báu trong suốt quá trình học tập.

Đặc biệt, em xin gửi lời cảm ơn đến giảng viên hướng dẫn Phạm Thị Kim Dung, giảng viên Khoa Công nghệ thông tin - Trường Đại học Điện Lực. Cô đã tận tình theo sát giúp đỡ, trực tiếp chỉ bảo, hướng dẫn trong suốt quá trình nghiên cứu và học tập của chúng em. Trong thời gian học tập với cô, nhóm em không những tiếp thu thêm nhiều kiến thức bổ ích mà còn học tập được tinh thần làm việc, thái độ nghiên cứu khoa học nghiêm túc, hiệu quả. Đây là những điều rất cần thiết cho chúng em trong quá trình học tập và công tác sau này.

Do thời gian thực hiện có hạn kiến thức còn nhiều hạn chế nên bài làm của em chắc chắn không tránh khỏi những thiếu sót nhất định. Em rất mong nhận được ý kiến đóng góp của thầy cô giáo và các bạn để em có thêm kinh nghiệm và tiếp tục hoàn thiện báo cáo của mình.

Em xin chân thành cảm ơn!

# LỜI MỞ ĐẦU

Công nghệ thông tin(CNTT) ngày càng có vai trò quan trọng trong cuộc sống hằng ngày của chúng ta. Việc ứng dụng CNTT vào các lĩnh vực trong đời sống giúp công việc được tiến hành nhanh chóng và hiệu quả hơn. Có rất nhiều công việc mới phát triển song song với sự phát triển của CNTT, một trong những số đó là phân tích chuẩn đoán dữ liệu – big data, hướng đi mang lại hiệu quả rất lớn.

Nhóm chúng em chọn đề tài “Ứng dụng thuật toán Linear Regresion để dự đoán xơ vữa động mạch (độ dày thành mạch)” nhằm tìm hiểu sâu hơn về dữ liệu cách khai thác và sử dụng hiệu quả, từ đó viết một ứng dụng cụ thể thử nghiệm làm cơ sở củng cố kiến thức và định hướng, kế hoạch xây dựng các ứng dụng trong tương lai.

# CHƯƠNG 1 : THUẬT TOÁN HỒI QUY TUYẾN TÍNH TRONG HỌC MÁY

## 1.1. Giới thiệu

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...

## 1.2. Phân tích toán học

### 1.2.1. Dạng của Linear Regression

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)

### 1.2.2. 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.

### 1.2.3. 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.

# 1.2.4. Mức độ lỗi của mô hình hồi quy tuyến tính

Ta có công thức tính mức độ lỗi của mô hình như sau :

MSE =

Với mức độ lỗi của mô hình cho ta biết mức độ học của mô hình.

# CHƯƠNG 2 : ÁP DỤNG THUẬT TOÁN

## 2.1. Mô tả bài toán

Trong y sinh học, thường có rất nhiều yếu tố khác nhau dẫn đến một hiện tượng. Ví dụ như hiện tượng xơ vữa động mạch không chỉ do lượng cholesterol máu mà còn do nhiều yếu tố khác như di truyền, chủng tộc, tiền sử mắc bệnh tim mạch, tuổi, giới, BMI, tăng huyết áp, đái tháo đường,…

Vì vậy cần phải có mô hình hồi quy tuyến tính đa biến. Trong đề tài tài này chúng em không đề cập các yếu tố di truyền, chủng tộc, giới, mắc bệnh tim mạch... mà chỉ lưu ý đến các biến số như: tuổi, cholesterol, glucose, huyết áp tâm thu và BMI. Khi đó mô hình hồi quy tuyến tính đa biến với 5 yếu tố ( x1=tuổi, x2 = cholesterol, x3 = glucose, x4 = huyết áp, x5 = BMI ) như sau:

Bề dày TM = w0 + w1(tuổi)+ w2(cholesterol) + w3(glucose) + w4(huyết áp) + w5(BMI)

## 2.2. Xây dựng bộ 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) làm biến giải thích.

|  |  |
| --- | --- |
| **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 1: Các thuộc tính của bộ dữ liệu

Hình trên là 40 tập dữ liệu đầu của bộ dự liệu huấn luyện mô hình

Đánh giá bộ dữ liệu này hoàn toàn có thể sử dụng cho mô hình hồi quy tuyến tính chẩn đoán xơ vữa động mạch vì những lý do sau:

* Mối quan hệ giữa biến mục tiêu (Y) và biến giải thích (X) là tham số. Ví dụ Y = WX thì mỗi lần X tăng lên 1 đơn vị thì Y thêm W đơn vị dù cho bất kể X, Y có giá trị là bao nhiêu.
* Các giá trị của biến mục tiêu (Y) phải độc lập với nhau
* Các giá trị của biến giải thích (X) không được có sai số ngẫu nhiên. Ví dụ: Giá trị huyết áp tối đa là 120 mmHg là 120 mmHg không được là 120 mmHg 2 mmHg

## 2.3. Thực nghiệm với thư viện scikit-learn của python

### 2.3.1. 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.

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

* **Sử dụng**

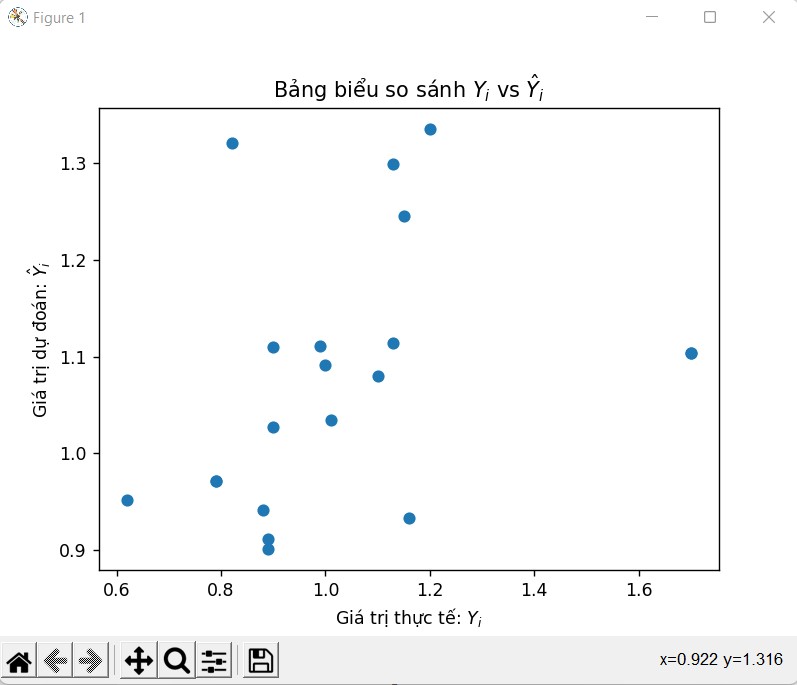
sklearn.linear\_model.LinearRegression(fit\_intercept=True, normalize=False, copy\_X=True, n\_jobs=1)

* **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

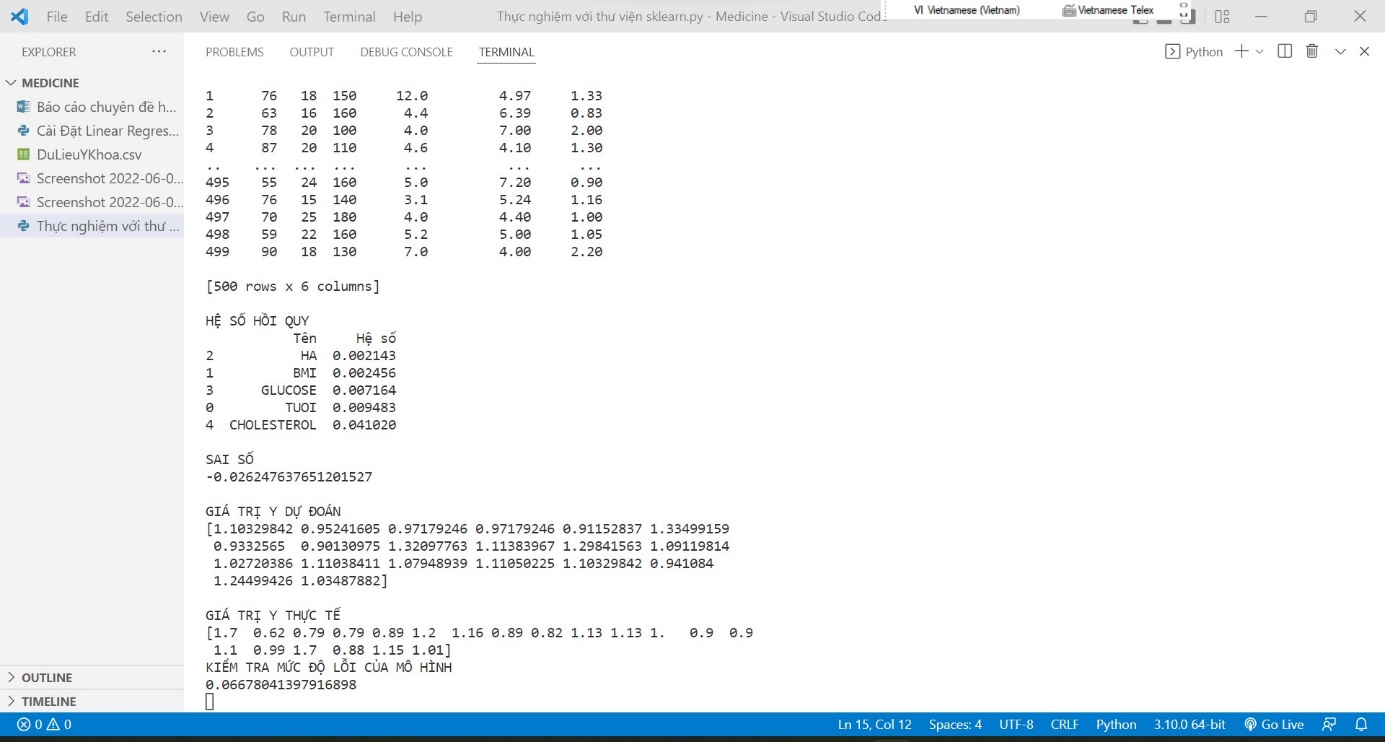
### 2.3.3. Cài đặt

**import** numpy **as** np  
**import** pandas **as** pd  
**import** matplotlib.pyplot **as** plt  
**import** sklearn  
**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  
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)  
*#Kiểm tra mức độ lỗi của model (Mean Squared Error)*mse = sq.mean\_squared\_error(Y\_test, Y\_pred)  
print(**"KIỂM TRA MỨC ĐỘ LỖI CỦA MÔ HÌNH"**)  
print(mse)  
*#Bảng biểu so sánh giá trị y dự đoán và y thực tế*plt.scatter(Y\_test, Y\_pred)  
plt.xlabel(**"Giá trị thực tế: $Y\_i$"**)  
plt.ylabel(**"Giá trị dự đoán: $\hat{Y}\_i$"**)  
plt.title(**"Bảng biểu so sánh $Y\_i$ vs $\hat{Y}\_i$"**)  
plt.show()

### 2.3.4 Kiểm nghiệm

****

Hình 2: Bảng biểu so sánh giá trị Y dự đoán và Y thực tế

****

Hình 3: Kết quả huấn luyện bằng thư viện sklearn

Từ kết quả trên ta có thể thấy kết quả huấn luyện bằng thư viện sklearn và phần cài đặt thuật toán khá tương đồng nhau, nếu có chênh lệch có thể là do sai số phần thập phân gây tích lũy nhiều sai số cho mô hình. Mô hình chạy trên thư viện sklearn cũng đã đưa ra yếu tố ảnh hưởng lớn tới giá trị dự đoán là nồng độ cholesterol và tuổi.

# KẾT LUẬN

Kết quả đạt được : chúng em 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.

Do thời gian và kiến thức có hạn nên báo cáo chúng em vẫn còn nhiều sai sót, rất mong các thầy cô góp ý giúp chúng em hoàn thiện báo hơn nữa.

**DANH MỤC 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