|  |
| --- |
| **BỘ GIÁO DỤC VÀ ĐÀO TẠO**  PHAN THANH HUYỀN CHUYÊN NGÀNH HTTT KHOÁ K63  **TRƯỜNG ĐẠI HỌC MỎ - ĐỊA CHẤT**  **---------------------------------------** |
|  |
| A picture containing graphical user interface  Description automatically generated |
| **PHAN THANH HUYỀN** |
|  |
| **ĐỐ ÁN TỐT NGHIỆP** |
|  |
|  |
|  |
| **NGÀNH CÔNG NGHỆ THÔNG TIN** |
|  |
|  |
| **TÊN ĐỀ TÀI: PHÂN TÍCH DỮ LIỆU VÀ MỘT SỐ KĨ THUẬT HỒI QUY TRONG DỰ ĐOÁN BỆNH TIM.** |
|  |
|  |
|  |

|  |
| --- |
| **Hà Nội - Năm 2022** |

**BỘ GIÁO DỤC VÀ ĐÀO TẠO**

**TRƯỜNG ĐẠI HỌC MỎ - ĐỊA CHẤT**

**ĐỒ ÁN TỐT NGHIỆP**

**CHUYÊN NGÀNH: HỆ THỐNG THÔNG TIN**

|  |
| --- |
| GIÁO VIÊN HƯỚNG DẪN SINH VIÊN THỰC HIỆN |
| **ThS. Vũ Lan Phương Phan Thanh Huyền** |

**Hà Nội - 2022**

**PHIẾU GIAO ĐỀ TÀI TỐT NGHIỆP**

1. **Thông tin chung**

**Tên đề tài**: Phân tích dữ liệu và một số kỹ thuật hồi quy trong dự đoán bệnh tim

|  |  |
| --- | --- |
| Họ và tên sinh viên: Phan Thanh Huyền |  |
| Điện thoại liên lạc: 0366521470 |  |
| Email: phanhuyen20x@gmail.com |  |
| Lớp: DCCTHT63B |  |
| Hệ đào tạo: | Đại học chính quy |
| Đồ án tốt nghiệp được thực hiện tại: | Hà Nội |
| Thời gian làm ĐATN: | 2022 |

1. **Mục tiêu của ĐATN**

* Khám phá dữ liệu, xử lý các giá trị nhiễu.
* Phân tích thống kê các biến định lượng, định tính.
* Phân tích song biến.
* Xây dựng mô hình dự đoán mắc bệnh tim từ thuật toán hồi quy.

1. **Các nhiệm vụ cụ thể của ĐATN**

Để đạt được mục tiêu đề ra của đề tài thì sẽ có những nội dung nghiên cứu sau:

* Sử dụng thành thạo các thư viện về phân tích dữ liệu và học máy.
* Xây dựng phương pháp xử lý dữ liệu nhiễu.
* Sử dụng thành thạo công cụ, ngôn ngữ lập trình đồ án:
* Python
* Jupyter notebook
* Xây dựng phương pháp phân tích phù hợp với bài toán:
  + Phân tích mô tả biến định lượng
  + Phân tích mô tả biến định tính
  + Phân tích song biến
  + Phân tích độ tương quan
* Kiểm tra các dữ liệu thiếu trong tập dữ liệu.
* Xử lý giá trị nhiễu (outliers) trước khi train dữ liệu.
* Thống kê mô tả biến định lượng.
* Trực quan hóa các phần phân tích mô tả các biến.
* Xây dựng biểu đồ hộp các biến định lượng
* Xây dựng biểu đồ phân phối tần suất các biến định lượng
* Xây dựng biểu đồ phân phối tần suất các biến định tính
* Xây dựng biểu đồ thống kê tỉ lệ người mắc bệnh trong tập dữ liệu
* Xây dựng biểu đồ phân tán – đường hồi quy giữa biến Chol và Target
* Xây dựng biểu đồ tỉ lệ mắc bệnh tim theo độ tuổi
* Xây dựng biểu đồ tương quan giữa các biến
* Xây dựng mô hình Học Máy phù hợp với tập dữ liệu:
  + Kiểm tra trước mô hình hồi quy Logistic với tập dữ liệu
  + Loại bỏ các biến không có sự liên quan tới biến kết quả
  + Diễn giải kết quả hồi quy Logistic
  + Xây dựng phương trình hồi quy Logistic với các thuộc tính trong tập dữ liệu
  + Xây dựng mô hình học máy với thuật toán hồi quy Logistic (Logistic regression)
  + Đánh giá kết quả của xây dựng mô hình Học Máy bằng confusion matrix

1. **Lời cam đoan của sinh viên:**

Em- Phan Thanh Huyền - cam kết ĐATN là công trình nghiên cứu của bản thân em, làm dưới sự hướng dẫn của cô Vũ Lan Phương.

Các kết quả nêu trên đồ án hoàn toàn là trung thực, không phải là sao chép toàn văn của bất kì công trình nào khác.

Hà Nội, ngày tháng năm

Tác giả ĐATN

1. **Xác nhận của giáo viên hướng dẫn về mức độ hoàn thành ĐATN và cho phép bảo vệ:**

Hà Nội, ngày tháng năm

Cán bộ hướng dẫn

|  |
| --- |
| ĐỀ CƯƠNG ĐỒ ÁN TỐT NGHIỆP |

1. **Thông tin người thực hiện**

|  |  |
| --- | --- |
| Họ và tên sinh viên: Phan Thanh Huyền  Ngành: Hệ thống thông tin  Điện thoại: 0366521470 | MSSV: 1821050957  Lớp: DCCTHT63B  Email: phanhuyen20x@gmail.com |
| Giảng viên hướng dẫn: ThS Vũ Lan Phương | |

1. **Tên đồ án: Phân tích dữ liệu và một số kĩ thuật hồi quy dự đoán bệnh tim**
2. **Mô tả**
   1. **Tổng quan đề tài**

Kể từ năm 2000, số ca tử vong do bệnh tim mạch gia tăng nhiều nhất, đặc biệt tăng tới hơn 2 triệu đến 8,9 triệu ca tử vong vào năm 2019. Tại Việt Nam, bệnh tim mạch chịu trách nhiệm cho 31% tổng số ca tử vong năm 2016, tương đương với hơn 170.000 ca tử vong. Dự đoán trong tương lai, con số này vẫn tiếp tục tăng lên.

Phân tích dữ liệu không chỉ ứng dụng trong doanh nghiệp hay kinh tế mà còn có ý nghĩa lớn trong lĩnh vực y tế. Đây là cách để quan sát các dữ liệu liên quan tới chủ đề liên quan và có hướng xử lý nhằm giảm thiểu thiệt hại.

**3.2. Mục tiêu dự kiến sẽ đạt được**

* Trình bày đầy đủ, chính xác, đúng trọng tâm nội dung lý thuyết.
* Trình bày code và trực quan hóa các nội dung phân tích, thuật toán một cách dễ hiểu.

**3.3. Kết quả đề tài**

* Phân tích chi tiết từng thuộc tính (chỉ số đo lường).
* Trực quan hóa để người đọc hiểu rõ được xu hướng của dữ liệu.
* Xử lý dữ liệu nhiễu để xây dựng mô hình dự đoán chính xác hơn.
* Xây dựng – đánh giá mô hình dự đoán mắc bệnh tim.

**4. Nội dung thực hiện**

**4.1. Tìm hiểu công cụ để xây dựng đồ án, trong đó bao gồm:**

* Ngôn ngữ: Python
* Công cụ: Jupyter Notebook

**4.2. Các bước thực hiện**

Bước 1: Tìm hiểu về đề tài

Bước 2: Tìm hiểu về các công cụ thực hiện, các phương pháp phân tích và thuật toán phù hợp.

Bước 3: Nghiên cứu lý thuyết về phân tích và học máy.

Bước 4: Phân tích và trực quan hóa dữ liệu.

Bước 5: Xử lý dữ liệu trước khi train dữ liệu.

Bước 6: Xây dựng mô hình học máy.

Bước 7: Đánh giá kết quả của thuật toán đã áp dụng.

Bước 9: Đưa ra kết luận sau khi hoàn thiện mô hình.

1. **Kế hoạch thực hiện**

|  |  |  |
| --- | --- | --- |
| **Thời gian** | **Nội dung thực hiện** | |
| Tuần thứ 1 (04 – 10/07/2022) | Nghiên cứu tài liệu và đọc các tài liệu liên quan đến phân tích dữ liệu. | |
| Tuần thứ 2 (11 – 17/07/2022) | Tìm hiểu về đề tài phân tích dữ liệu, các mô hình học máy đã có. | |
| Tuần thứ 3 (18 – 24/07/2022) | Liệt kê các công việc cần phải làm, các bước phải làm trong phân tích dữ liệu và dữ đoán. | |
| Tuần thứ 4  (25 – 31/07/2022) | Tiến hành phân tích dữ liệu. | |
| Tuần thứ 5 (01 – 07/08/2022) | Tiến hành trực quan hóa dữ liệu. | |
| Tuần thứ 6 (08 – 14/08/2022) | Xây dựng mô hình học máy. | |
| Tuần thứ 7 (15 – 21/08/2022) | Đánh giá kết quả của thuật toán. | |
| Tuần thứ 8 (22 – 28/08/2022) | Kết luận các nội dung đã phân tích và độ phù hợp của thuật toán. | |
| Ngày tháng năm 2022  **Giảng viên hướng dẫn**  *(ký và ghi rõ họ tên)* | | Ngày tháng năm 2022  **Sinh viên thực hiện**  *(ký và ghi rõ họ tên)* | |

# LỜI CẢM ƠN

Đồ án tốt nghiệp chuyên ngành *Hệ thống thông tin và Tri thức* với Đề tài “*Phân tích dữ liệu và một số kĩ thuật hồi quy trong dự đoán bệnh tim mạch*” là thành quả sau quá trình cố gắng không ngừng nghỉ của bản thân em dưới sự giúp đỡ tận tình, động viên khích lệ từ thầy cô, bạn bè và người thân. Qua đây, em xin gửi lời cảm ơn chân thành đến những người đã giúp đỡ em trong thời gian học tập – nghiên cứu đồ án vừa qua.

Trước hết, em xin chân thành cảm ơn quý thầy, cô khoa *Công nghệ thông tin*. Đặc biệt là các thầy cô trong bộ môn *Hệ thống thông tin và Tri thức* đã tận tình truyền dạy và trang bị cho em những kiến thức nòng cốt trong suốt hành trình ngồi trên ghế giảng đường, làm nền móng giúp em có thể hoàn thành đồ án này một cách trọn vẹn nhất.

Tiếp đó, em cũng xin trân trọng gửi lời tri ân đến *GV, Th.S Vũ Lan Phương* đã dành tâm huyết giúp đỡ, định hướng cho em cách tư duy và làm việc khoa học hơn. Đó là những lời góp ý hết sức quý báu không chỉ trong quá trình thực hiện đồ án này mà còn là hành trang tiếp bước cho em trong quá trình học tập và lập nghiệp sau này.

Em xin chân thành cảm ơn./

# LỜI CAM ĐOAN

Những nội dung trong đồ án tốt nghiệp được em phát triển dựa trên sự tích lũy kiến thức trong thời gian thực tập và đồ án môn học. Em xin cam đoan đồ án này không sao chép nội dung từ các đồ án khác và sản phẩm của đồ án là của chính tay bản thân em xây dựng lên.

MỤC LỤC

[ĐỀ CƯƠNG ĐỒ ÁN TỐT NGHIỆP 4](#_Toc117077268)

[LỜI CẢM ƠN 7](#_Toc117077269)

[LỜI CAM ĐOAN 8](#_Toc117077270)

[DANH MỤC HÌNH ẢNH 11](#_Toc117077271)

[CHƯƠNG 1: GIỚI THIỆU 12](#_Toc117077272)

[1.1 Tổng quan 12](#_Toc117077273)

[1.1.1 Lý do chọn đề tài 12](#_Toc117077274)

[1.1.2 Mục tiêu của đề tài 13](#_Toc117077275)

[1.1.3 Đối tượng và phạm vi nghiên cứu 13](#_Toc117077276)

[1.1.4 Nguồn tài liệu 13](#_Toc117077277)

[1.2 Phương pháp – Kết quả 13](#_Toc117077278)

[1.2.1 Phương pháp nghiên cứu, phân tích, thực nghiệm 13](#_Toc117077279)

[1.2.2 Kết quả mong muốn 14](#_Toc117077280)

[1.3 Cấu trúc đồ án 14](#_Toc117077281)

[CHƯƠNG 2: NGHIÊN CỨU VỀ LÝ THUYẾT 15](#_Toc117077282)

[2.1 Tổng quan về Phân tích dữ liệu 15](#_Toc117077283)

[2.1.1 Data Analytics là gì? 15](#_Toc117077284)

[2.1.2 Tại sao lại phải phân tích dữ liệu y tế (Healthcare Data Analytics)? 16](#_Toc117077285)

[2.1.3 Data Analytics vs. Data Analysis 16](#_Toc117077286)

[2.1.4 Ứng dụng và cơ hội nghề nghiệp của Data Analytics 17](#_Toc117077287)

[2.1.5 Các loại phân tích dữ liệu 18](#_Toc117077288)

[2.1.6 Quy trình phân tích dữ liệu 20](#_Toc117077289)

[2.2 Tổng quan về Machine Learning 23](#_Toc117077290)

[2.2.1 Machine learning là gì? 23](#_Toc117077291)

[2.2.2 Phân nhóm Machine Learning 24](#_Toc117077292)

[CHƯƠNG 3: PHÂN TÍCH DỮ LIỆU 26](#_Toc117077293)

[3.1 Mô tả bài toán 26](#_Toc117077294)

[3.1.1 Thu thập dữ liệu (Data collection) 26](#_Toc117077295)

[3.1.2 Mô tả tập dữ liệu (Dataset describe) 26](#_Toc117077296)

[3.2 Khám phá dữ liệu 28](#_Toc117077297)

[3.2.1 Đọc dữ liệu từ dataset 28](#_Toc117077298)

[3.2.2 Xác định kiểu dữ liệu trong dataset 30](#_Toc117077299)

[3.2.3 Kiểm tra dữ liệu thiếu trong tập dữ liệu 31](#_Toc117077300)

[3.3 Thống kê mô tả biến định lượng 32](#_Toc117077301)

[3.3.1 Boxplot 33](#_Toc117077302)

[3.3.2 Age 34](#_Toc117077303)

[3.3.3 Trestbps (huyết áp tâm trương) 35](#_Toc117077304)

[3.3.4 Chol (Cholesterol) 36](#_Toc117077305)

[3.3.5 Thalach (maximum heart rate achieved) 37](#_Toc117077306)

[3.4 Thống kê mô tả biến định tính 38](#_Toc117077307)

[3.4.1 Sex (giới tính) 38](#_Toc117077308)

[3.4.2 Cp (Chest pain type) 39](#_Toc117077309)

[3.4.3 Fbs (fasting blood sugar) 40](#_Toc117077310)

[3.4.4 Restecg (resting electrocardiographic results) 41](#_Toc117077311)

[3.4.5 Exang (exercise induced angina) 42](#_Toc117077312)

[3.4.6 Slope (độ dốc của đoạn ST) 43](#_Toc117077313)

[3.4.7 Thal (Thalassemia – bệnh tan máu bẩm sinh) 44](#_Toc117077314)

[3.5 Bivariate Analysis – Phân tích song biến 46](#_Toc117077315)

[3.5.1 Thống kê tỉ lệ người mắc bệnh và người không mắc bệnh tim 46](#_Toc117077316)

[3.5.2 Mối quan hệ giữa chỉ số Cholesterol và kết quả (target) 47](#_Toc117077317)

[3.5.3 Mối quan hệ giữa độ tuổi và kết quả mắc bệnh (target) 50](#_Toc117077318)

[CHƯƠNG 4: DỰ ĐOÁN VÀ ĐÁNH GIÁ KĨ THUẬT HỒI QUY LOGISTIC 51](#_Toc117077319)

[4.1 Độ tương quan 51](#_Toc117077320)

[4.2 Xử lý giá trị nhiễu 52](#_Toc117077321)

[4.3 Kĩ thuật hồi quy Logistic (Logistic Regression) 56](#_Toc117077322)

[4.3.1 Logistic Regression đa biến được sử dụng khi nào 56](#_Toc117077323)

[4.3.2 Bản chất hồi quy Logistic đa thức 57](#_Toc117077324)

[4.3.3 Xây dựng mô hình hồi quy Logistic 58](#_Toc117077325)

[4.3.4 So sánh với mô hình hồi quy tuyến tính (Linear regression) 64](#_Toc117077326)

[4.3.5 Đánh giá kết quả của thuật toán hồi quy Logistic 65](#_Toc117077327)

[KẾT LUẬN VÀ ĐỊNH HƯỚNG PHÁT TRIỂN 68](#_Toc117077328)

[1. Kết quả đạt được 68](#_Toc117077329)

[2. Định hướng phát triển 69](#_Toc117077330)

[TÀI LIỆU THAM KHẢO 70](#_Toc117077331)

# DANH MỤC HÌNH ẢNH

[Hình 2.1: Sự khác nhau giữa Analytics và Analysis 15](#_Toc115606032)

[Hình 3.1: Biểu đồ hộp của các biến định lượng 32](#_Toc115606033)

[Hình 3.2: Biểu đồ tần suất của biến Tuổi 33](#_Toc115606034)

[Hình 3.3: Biểu đồ tần suất của biến Trestbps. 34](#_Toc115606035)

[Hình 3.4: Biểu đồ tần suất của biến Chol. 35](#_Toc115606036)

[Hình 3.5: Biểu đồ tần suất của biến Thalach. 36](#_Toc115606037)

[Hình 3.6: Biểu đồ phân phối tần suất của Giới tính (sex). 37](#_Toc115606038)

[Hình 3.7: Biểu đồ phân phối tần suất của các loại đau tim (cp). 38](#_Toc115606039)

[Hình 3.8: Biểu đồ tần suất của chỉ số đường huyết lúc đói (fbs). 39](#_Toc115606040)

[Hình 3.9: Biểu đồ tần suất của kết quả điện tâm đồ lúc nghỉ (restecg). 40](#_Toc115606041)

[Hình 3.10: Biểu đồ tần suất của chỉ số đau thắt ngực do tập luyện (exang). 41](#_Toc115606042)

[Hình 3.11: Biểu đồ tần suất độ dốc ST (slope) 43](#_Toc115606043)

[Hình 3.12: Biểu đồ tần suất của bệnh tan máu bẩm sinh (Thal). 45](#_Toc115606044)

[Hình 3.13: Tỉ lệ số bệnh nhân mắc bệnh tim 46](#_Toc115606045)

[Hình 3.14: Biểu đồ phân tán Chol – Target 47](#_Toc115606046)

[Hình 3.15: Biểu đồ phân tán – đường hồi quy giữ Chol và Target. 48](#_Toc115606047)

[Hình 3.16: Tỉ lệ mắc bệnh tim mạch theo độ tuổi. 49](#_Toc115606048)

[Hình 4.1: Biểu đồ tương quan 50](#_Toc115606049)

[Hình 4.2: Biểu đồ cột nằm ngang của tương quan 51](#_Toc115606050)

[Hình 4.3: Xác định Outliers bằng biểu đồ hộp boxplot 52](#_Toc115606051)

[Hình 4.4: Kiểm tra các giá trị ngoại lai đã xóa 54](#_Toc115606052)

[Hình 4.5: Các chỉ số của confusion matrix. 63](#_Toc115606053)

[Hình 4.6: Confusion matrix của mô hình hồi quy logistic 65](#_Toc115606054)

# CHƯƠNG 1: GIỚI THIỆU

## **Tổng quan**

### **Lý do chọn đề tài**

Tháng 12/2020, Tổ chức Y tế Thế giới (WHO) công bố bệnh tim mạch là nguyên nhân gây tử vong hàng đầu thế giới theo số liệu thống kê của năm 2019. Kể từ năm 2000, số ca tử vong do bệnh tim mạch gia tăng nhiều nhất, đặc biệt tăng tới hơn 2 triệu đến 8,9 triệu ca tử vong vào năm 2019. Tại Việt Nam, bệnh tim mạch chiếm 31% tổng số ca tử vong năm 2016, tương đương với hơn 170.000 ca tử vong. Dự đoán trong tương lai, con số này vẫn tiếp tục tăng lên.

Bệnh tim mạch trước đây vốn được biết đến là bệnh của người già nhưng thực tế cho thấy số lượng người trẻ mắc bệnh ngày càng nhiều. Theo thống kê tại các bệnh viện lớn hiện nay, nhồi máu cơ tim ở người trẻ (dưới 45 tuổi) đã lên đến 10,5% và người rất trẻ (dưới 35 tuổi) là 1,8%. Đặc biệt ở trẻ em, tỉ lệ thừa cân béo phì tăng từ 8,5% năm 2010 lên thành 19% năm 2020 kèm theo đó là nguy cơ mắc bệnh tim mạch cũng tăng lên đáng kể. Đây là những con số rất đáng báo động về mức độ trẻ hóa của bệnh tim cũng như mức độ nguy hiểm tiềm tàng của nó đối với giới trẻ hiện nay.

Lý giải cho thực trạng này chính là do xã hội càng phát triển, đời sống vật chất tinh thần của con người ngày càng được nâng cao thì giới trẻ càng có nhiều lối sống thiếu lành mạnh như: ăn quá nhiều đồ ăn nhanh, ăn nhiều đường, thiếu hoạt động thể lực, stress và hút thuốc lá,...đây đều là những nguyên nhân chính dẫn đến bệnh tim mạch.

Đặc biệt trong tình hình dịch bệnh COVID-19 hiện nay, nguy cơ tử vong của người cao tuổi mắc bệnh lý nền như bệnh tim mạch càng đáng báo động. Theo số liệu thống kê từ Vũ Hán (Trung Quốc) cho thấy, tỉ lệ tử vong chiếm tới 10,9% ở những người mắc COVID-19 có bệnh tim. Mặc khác, việc lo lắng về tình hình dịch bệnh, hạn chế đi lại đã dẫn đến việc người dân mắc bệnh tim trì hoãn tái khám. Điều này thực sự nguy hiểm đối với bệnh tim, một bệnh mãn tính thường không có các triệu chứng hay dấu hiệu rõ ràng.

Trên đây chính là những lý do khiến em quyết định chọn phân tích và dự đoán bệnh tim làm đồ án tốt nghiệp.

### **Mục tiêu của đề tài**

Đề tài này đặt ra với mục đích tổng hợp, phân tích các yếu tố nguy cơ mắc bệnh tim mạch, dự đoán mắc mệnh hay không. Khi các thông số về nguy cơ đã được thu thập, sàng lọc và lượng giá từ các tổ chức và đại học uy tín về ngành tim mạch và y học, sẽ giúp ích cho việc đánh giá một người thuộc nhóm nguy cơ thấp hay cao và theo sau đó, một chương trình điều trị hướng tới việc điều chỉnh các yếu tố nguy cơ.

### **Đối tượng và phạm vi nghiên cứu**

Đối tượng nghiên cứu: là các đối tượng trong tập dữ liệu được tham gia khảo sát, được lấy số liệu từ các nhà chuyên môn về y tế - sức khỏe từ trước năm 1988.

Phương pháp nghiên cứu: Phân tích mô tả (Descriptive Analysis) và bài toán dự đoán (Machine Learning).

Công cụ được sử dụng là Jupyter Notebook.

### **1.1.4 Nguồn tài liệu**

Tập dữ liệu được lấy từ Kaggle – là một cộng đồng trực tuyến dành cho dân trí tuệ nhân tạo (AI) và khoa học dữ liệu (DS) để chia sẻ dữ liệu giữa người làm data.

Tập dữ liệu bệnh tim mạch (Heart Disease) có từ năm 1988, gồm 4 cơ sở dữ liệu là: Cleaveland, Hungary, Thụy Sĩ và Long Beach. Chứa tất cả 76 thuộc tính nhưng qua các cuộc thử nghiệm thì chỉ được sử dụng 14 thuộc tính bao gồm cả thuộc tính dự đoán.

## **Phương pháp – Kết quả**

### **Phương pháp nghiên cứu, phân tích, thực nghiệm**

* Phương pháp nghiên cứu, phân tích:
  + Thu thập dữ liệu.
  + Thu thập các tài liệu liên quan đến phân tích dữ liệu và học máy.
  + Tham khảo các dự án liên quan đến phân tích và dự đoán.
  + Áp dụng các môn học đã được học: Xác suất & Thống kê, Lập trình Python, Khoa học dữ liệu, Ngôn ngữ R trong phân tích dữ liệu, Học máy thống kê.
* Phương pháp thực nghiệm
  + Tìm hiểu, áp dụng Jupyter Notebook.
  + Tìm hiểu, áp dụng thư viện Numpy, Pandas, Matplotlib, Seaborn, Sklearn.

### **Kết quả mong muốn**

* Tìm hiểu các cơ sở lý thuyết, các kiến thức liên quan tới: phân tích dữ liệu, khoa học dữ liệu, công cụ code và phân tích, các thư viện hỗ trợ,…
* Phân tích dữ liệu và trực quan hóa chi tiết, chính xác các thuộc tính.
* Xây dựng các modul học máy với chức năng chính là dự đoán và tính độ chính xác của thuật toán trên tập dữ liệu.

## **Cấu trúc đồ án**

Với mục tiêu phân tích và dự đoán tỉ lệ mắc bệnh tim, giúp ngăn chặn khả năng cao và lan truyền thông điệp sống khỏe tới tất cả mọi người, giúp mọi người thấy rõ khả năng mắc bệnh của từng độ tuổi và giới tính, bên cạnh đó cũng dự đoán được khả năng mắc bệnh bằng thuật toán chính xác nhất mà em lựa chọn. Vậy nên em đã xây dựng đồ án của em gồm 4 chương, như sau:

* Chương 1: Giới thiệu
* Chương 2: Nghiên cứu về lý thuyết
* Chương 3: Phân tích dữ liệu
* Chương 4: Dự đoán và đánh giá kĩ thuật hồi quy Logistic
* Kết luận và hướng phát triển đề tài

# CHƯƠNG 2: NGHIÊN CỨU VỀ LÝ THUYẾT

## **Tổng quan về Phân tích dữ liệu**

Ngày nay, lĩnh vực data analytics dần trở nên quan trọng. Bất kỳ công ty nào đang muốn mở rộng quy mô dù muốn hay không cũng phải cần quan tâm đến nó. Một lượng thông tin khổng lồ vẫn đang được thu thập hàng năm và cần phải phân tích cẩn thận để biến tất cả dữ liệu đó thành thông tin chi tiết hữu ích.

Lĩnh vực này ngày càng hấp dẫn hơn nữa khi xuất hiện các chuyên viên data analyst những người kể chuyện bằng con số. Công việc của họ là biến những dữ liệu thô thành chiến lược cho công ty. Một công việc công nghệ đáng mơ ước với mức lương cao, một thị trường việc làm ổn định với đầy những cơ hội làm việc từ xa…

### **Data Analytics là gì?**

Data analytics là thực hành thu thập và sắp xếp thông tin, tìm kiếm các mẫu và trực quan hóa những phát hiện đó một cách rõ ràng, dễ hiểu.

Hay như **wikipedia** có định nghĩa: “*Phân tích dữ liệu là một quá trình kiểm tra, làm sạch, chuyển đổi và mô hình hóa dữ liệu với mục tiêu khám phá thông tin hữu ích, thông báo kết luận và hỗ trợ ra quyết định.*

*Phân tích dữ liệu có nhiều khía cạnh và cách tiếp cận, bao gồm các kỹ thuật đa dạng dưới nhiều tên khác nhau và được sử dụng trong các lĩnh vực kinh doanh, khoa học và khoa học xã hội khác nhau.*

*Trong thế giới kinh doanh ngày nay, phân tích dữ liệu đóng vai trò giúp đưa ra quyết định khoa học hơn và giúp doanh nghiệp hoạt động hiệu quả hơn*.”

### **Tại sao lại phải phân tích dữ liệu y tế (Healthcare Data Analytics)?**

Big Data và Data Analytics có tác động mạnh mẽ đến lĩnh vực chăm sóc sức khỏe. Điều đó tạo điều kiện để hình thành và phát triển lĩnh vực nghiên cứu mới - phân tích dữ liệu mảng y tế.

Phân tích dữ liệu mảng y tế và chăm sóc sức khỏe thu thập, phân tích dữ liệu giúp bác sĩ, nhân viên y tế, các tổ chức hoạt động trong lĩnh vực này nâng cao chất lượng khám chữa bệnh cho bệnh nhân, dự đoán có mang bệnh hay không dựa trên ứng dụng của Machine Learning.

### **Data Analytics vs. Data Analysis**



##### Hình 2.1: Sự khác nhau giữa Analytics và Analysis

Data Analytics và Data Analysis dịch ra đều có nghĩa là phân tích dữ liệu. Vì thế nhiều người lầm tưởng rằng nghĩa của 2 từ này hoàn toàn giống nhau.

Về cơ bản, sự khác biệt giữa 2 từ này chính là quy mô, phạm vi. Data Analytics có nghĩa rộng hơn Data Analysis và Data Analysis là tập hợp con trong Data Analytics.

Data analysis mô tả quá trình kiểm tra, chuyển đổi và sắp xếp dữ liệu theo trật tự nhất định để tìm hiểu và rút ra các thông tin hữu ích.

Còn Data Analytics mô tả một lĩnh vực bao gồm quá trình quản lý toàn bộ dữ liệu. Quá trình này không chỉ bao gồm việc phân tích (data analysis) mà còn thu thập, tổ chức, lưu trữ dữ liệu cũng như các công cụ và kỹ thuật liên quan.

### **Ứng dụng và cơ hội nghề nghiệp của Data Analytics**

* Vị trí Data Analyst: Đây là lĩnh vực đang được nhiều sự quan tâm trên thị trường Việt Nam và trên thế giới hiện nay. Vì Data Analytics giúp doanh nghiệp hiểu rõ được tình hình kinh doanh hiện tại. Từ đó họ có thể khắc phục được những sai lầm trong quá khứ và đưa ra những quyết định sáng suốt hơn trong tương lai dựa vào những số liệu cụ thể chứ không phải dựa trên cảm tính hay kinh nghiệm dự đoán.
* Vị trí Business Analyst: phân tích nhu cầu của khách hàng cùng các đối tác đối với việc kinh doanh của doanh nghiệp để tìm ra cách giải quyết cho những vấn đề phát sinh. BA góp phần giúp giảm thiểu chi phí hoạt động, cải thiện cách thức vận hành, sử dụng hiệu quả nguồn lực có hạn, phục vụ khách hàng tốt hơn.

BA đóng vai trò như cầu nối giữa hoạt động kinh doanh của doanh nghiệp và bộ phận CNTT. Là thành viên của nhóm dự án, họ đóng góp nhiều ý kiến và thông tin giá trị. Họ làm việc với các nhà quản lý và các nhà tư vấn, chịu trách nhiệm phát triển mô hình nghiệp vụ, thực hiện các công việc nghiên cứu và phân tích phức tạp.

* Vị trí Data Scientist: phân tích, sắp xếp và kể chuyện bằng dữ liệu. Công việc của họ có sự kết hợp giữa khoa học máy tính, thống kê và toán học. Họ là người phân tích, xử lý và “mô hình hóa” dữ liệu và diễn giải kết quả để rút ra những kế hoạch hoạt động cho cả team. Hay nói một cách đơn giản hơn, nghề data scientist làm việc với dữ liệu để cho ra những insight đắt giá. Họ sẽ truyền tải những insight này đến với các bên liên quan như lãnh đạo cấp cao, quản lý, khách hàng. Điều đó giúp công ty đưa ra những quyết định sáng suốt nhất, thúc đẩy tình hình kinh doanh và góp phần tăng lợi nhuận.
* Healthcare Data Analyst: phân tích dữ liệu mảng y tế và chăm sóc sức khỏe thu thập, phân tích dữ liệu giúp bác sĩ, nhân viên y tế, các tổ chức hoạt động trong lĩnh vực này nâng cao chất lượng dịch vụ cho bệnh nhân.

### **Các loại phân tích dữ liệu**

#### Text Analysis (Phân tích văn bản)

Text Analysis là một nhánh của Data Mining. Mục đích của nó là tìm kiếm, trích xuất thông tin trong văn bản, chuyển dữ liệu thô thành thông tin kinh doanh. Các công cụ Business Intelligence được sử dụng nhằm đưa ra những quyết định kinh doanh chiến lược. Nhìn chung, nó cung cấp cách để trích xuất và kiểm tra dữ liệu, bắt nguồn từ một mẫu và cuối cùng là giải thích dữ liệu. Hiện nay, dữ liệu văn bản ngày càng nhiều khiến text analysis ngày càng có nhiều ứng dụng trong thực tế. Ví dụ như đối chiếu lý lịch cá nhân, lọc thư rác, phân tích cảm nghĩ, phân loại tài liệu.

#### 2.1.5.2 Statistical Analysis (Phân tích thống kê)

Statistical Analysis thể hiện “Điều gì xảy ra?” bằng cách sử dụng dữ liệu trong quá khứ dưới dạng dashboards. Phân tích thống kê gồm việc thu thập, phân tích, giải thích, trình bày và mô hình hóa dữ liệu. Nó phân tích một tập hợp dữ liệu hoặc một mẫu dữ liệu. Có hai loại phân tích – Descriptive Analysis (phân tích mô tả) và Inferential Analysis (phân tích suy luận).

* Descriptive Analysis (phân tích mô tả): được sử dụng để phân tích và mô tả các dữ liệu đã có sẵn hay dữ liệu lịch sử của doanh nghiệp. Dữ liệu trong quá khứ thường được sử dụng để so sánh và rút ra kết luận. Hai kỹ thuật chính được sử dụng trong loại phân tích này là data aggregation (tổng hợp dữ liệu) và data mining (khai phá dữ liệu). Đầu tiên dữ liệu được thu thập, sau đó được sắp xếp thông qua quá trình tổng hợp dữ liệu. Tiếp theo, các mẫu và ý nghĩa được xác định thông qua quá trình khai phá dữ liệu.
* Inferential Analysis (phân tích suy luận): phân tích mẫu từ dữ liệu hoàn chỉnh. Trong loại phân tích này, bạn có thể tìm thấy các kết luận khác nhau từ cùng một dữ liệu nếu chọn các mẫu khác nhau.

#### Diagnostic Analysis: phân tích chẩn đoán

Đôi khi, có thể có những sai sót trong dữ liệu. Những bất thường này có thể ảnh hưởng tiêu cực đến hoạt động của doanh nghiệp. Mục đích của phân tích chẩn đoán là để chẩn đoán những vấn đề và những lỗi phát sinh này. Hiệu suất và chiến lược của một doanh nghiệp có thể được cải thiện nếu các vấn đề được chẩn đoán và khắc phục sớm.

#### Predictive Analysis: phân tích dự đoán

Như tên gọi của nó, loại phân tích này được sử dụng để dự đoán kết quả trong tương lai. Dựa trên dữ liệu quá khứ hoặc lịch sử, một tập hợp các sự kiện có khả năng xảy ra trong tương lai sẽ được xác định. Phân tích mô tả cũng là một phần của quá trình này, vì nó được sử dụng để tạo ra các dự đoán. Trong loại phân tích này, một lượng lớn dữ liệu được khai thác một cách khéo léo bằng cách sử dụng các mô hình dự đoán. Điều này đảm bảo rằng các dự đoán chính xác có thể được thực hiện.

Ví dụ đơn giản nhất về phân tích dự đoán là nếu năm ngoái bạn mua hai chiếc váy dựa trên khoản tiết kiệm của mình và năm nay lương của bạn tăng gấp đôi nên bạn dự đoán có thể mua được bốn chiếc váy. Nhưng tất nhiên không dễ như thế vì bạn phải suy nghĩ về các trường hợp có thể xảy ra như giá quần áo tăng trong năm nay hoặc có thể thay vì váy bạn muốn mua xe đạp mới, hoặc bạn cần mua nhà.

#### 2.1.5.4 Prescriptive Analysis: phân tích đề xuất

Đây là sự kết hợp của tất cả các loại phân tích khác. Thay vì tập trung vào giám sát dữ liệu, kỹ thuật này nhấn mạnh vào những hiểu biết sâu sắc về dữ liệu từ đó đề xuất những hành động phù hợp và dự đoán kết quả có thể xảy ra. Dữ liệu được thu thập bằng cách sử dụng cả mô hình mô tả và dự đoán, cũng như sự kết hợp giữa toán học và khoa học máy tính.

### **Quy trình phân tích dữ liệu**

#### 2.1.6.1 Business Problem Definition – Xác định vấn đề

Để chuyển đổi dữ liệu thành các quyết định kinh doanh, bạn nên bắt đầu từ những vấn đề cần đào sâu tìm hiểu của doanh nghiệp, trước khi thực hiện thu thập dữ liệu. Dựa vào chiến lược, mục tiêu, ngân sách và đối tượng khách hàng của doanh nghiệp, bạn sẽ chuẩn bị được danh sách câu hỏi của mình. Hãy đặt những câu hỏi có thể đo lường, rõ ràng và ngắn gọn.

#### 2.1.6.2 Data Preparation – Chuẩn bị dữ liệu

Xác định dữ liệu được lưu trữ ở vị trí nào, dữ liệu có sẵn hay không? Được lưu trữ ở định dạng nào?

Sau khi thu thập và tổng hợp dữ liệu từ nhiều nguồn khác nhau, bạn có được một núi thông tin cần phải xử lý. Lúc đó những sai sót trong dữ liệu là điều bạn sẽ gặp phải và chúng sẽ khiến bạn phân tích lệch hướng. Vì thế quá trình là sạch dữ liệu trở nên vô cùng quan trọng. Nó không chỉ tạo ra không gian để thêm dữ liệu mới mà còn giúp bạn rút ra những insight đảm bảo được sự chính xác và đáng tin cậy hơn.

Làm sạch dữ liệu là quy trình chuẩn bị dữ liệu thông qua việc loại bỏ những dữ liệu không đầy đủ, không chính xác, không phù hợp về định dạng, bị trùng lặp, không liên quan, không có giá trị... Bởi vì những dữ liệu này có thể ảnh hưởng đến kết quả cuối cùng.

#### 2.1.6.3 Descriptive Data – Mô tả dữ liệu

Ở bước này chúng ta cần tập trung vào 2 bước:

* **Data Visualization** (Trực quan hóa dữ liệu): là tạo ra những hình ảnh, biểu đồ, đồ thị nhằm diễn tả những thông tin đến người dùng. Trực quan hóa nghiên cứu trình bày một cách trực quan khối dữ liệu trừu tượng để tăng cường nhận thức của con người. Một bản tóm tắt thông tin trực quan giúp bạn hiểu mô hình và xu hướng dễ dàng và rõ ràng hơn so với việc xem hàng trăm cột số liệu khác nhau. Thậm chí đối với người có thể hiệu được những số liệu ấy thì việc không trực quan hóa dữ liệu cũng khiến họ gặp khó khăn trong việc diễn đạt ý của mình cho khách hàng và đồng nghiệp hiểu.
* **Descriptive Statistical** (Thống kê mô tả): là các hệ số mô tả ngắn gọn hay tóm tắt một tập dữ liệu nhất định, có thể là đại diện cho toàn bộ hoặc một mẫu của một tổng thể. Thống kê mô tả được chia thành đo lường *xu hướng tập trung* và *đo lường biến động*. Đo lường xu hướng tập trung có giá trị *trung bình* (mean), *trung vị* (median) và *yếu vị* (mode), trong khi các đo lường biến động gồm *độ lệch chuẩn* (standard deviation), *phương sai* (variance), *giá trị nhỏ nhất* (min) và *giá trị lớn nhất* (max), *độ nhọn* (kurtosis) và *độ lệch* (skewness).

#### Diagnosis Step – Chẩn đoán

Phân tích chuẩn đoán giúp trả lời câu hỏi tại sao tình trạng này lại diễn ra. Phân tích chuẩn đoán bao gồm các phân tích mô tả cơ bản, sử dụng những kết quả của phân tích mô tả để từ đó tìm ra nguyên nhân của các sự kiện. Từ đó, những chỉ số đo lường hiệu quả được phân tích kỹ hơn. Quá trình này thường bao gồm 3 bước:

* Tìm ra những điểm bất thường trong dữ liệu. Những điểm này thường là một biến đổi không lường trước được ở một chỉ số hoặc một thị trường nào đó.
* Thu thập các dữ liệu liên quan đến những điểm bất thường này.
* Sử dụng các kỹ thuật thống kê để tìm ra mối quan hệ và xu hướng giúp giải thích những điểm bất thường này.

Các phương pháp kĩ thuật được sử dụng ở bước này như: data discovery, drill-up, drill-down, data mining, và correlations (tương quan). Phân tích chẩn đoán giúp chúng ta nhận được giá trị từ dữ liệu của mình bằng cách đặt câu hỏi phù hợp và tìm hiểu sâu để tìm câu trả lời.

#### 2.1.6.5 Predictive – Dự đoán

Phân tích dự đoán giúp trả lời các câu hỏi về điều gì sẽ xảy ra trong tương lai. Phân tích dự đoán sử dụng các dữ liệu quá khứ để xác định các xu hướng cũng như khả năng xảy ra của xu hướng đó. Những công cụ phân tích dự đoán cung cấp những thông tin giá trị về tình hình tương lai. Các kỹ thuật phân tích dự đoán đa dạng bao gồm cả các kỹ thuật thống kê lẫn học máy (machine learning) như mạng nơ-ron nhân tạo(neural networks), cây quyết định (decision trees) và hồi quy (regression).

#### 2.1.6.7 Precriptive – Đề xuất

Phân tích đề xuất giúp trả lời các câu hỏi về hành động nào nên được thực hiện để đặt được một mục tiêu nhất định. Bằng cách sử dụng các kết quả từ phân tích đề xuất, các tổ chức có thể ra quyết định dựa trên dữ liệu. Phương pháp này giúp các tổ chức đưa ra các quyết định trong các hoàn cảnh không chắc chắn.

Phân tích đề xuất dựa trên học máy (machine learning) để tìm ra các hình thái (pattern) trong các dataset lớn. Bằng cách phân tích các quyết định & sự kiện trong quá khứ, các tổ chức có thể ước lượng xác suất xảy ra của các kết quả.

## **Tổng quan về Machine Learning**

### **2.2.1 Machine learning là gì?**

Machine learning gây nên cơn sốt công nghệ trên toàn thế giới trong vài năm nay. Trong giới học thuật, mỗi năm có hàng ngàn bài báo khoa học về đề tài này. Trong giới công nghiệp, từ các công ty lớn như Google, Facebook, Microsoft đến các công ty khởi nghiệp đều đầu tư vào machine learning. Hàng loạt các ứng dụng sử dụng machine learning ra đời trên mọi lĩnh vực của cuộc sống, từ khoa học máy tính đến những ngành ít liên quan hơn như vật lý, hóa học, y học, chính trị.

Machine learning là một lĩnh vực con của Trí tuệ nhân tạo (Artificial Intelligence) sử dụng các thuật toán cho phép máy tính có thể học từ dữ liệu để thực hiện các công việc thay vì được lập trình một cách rõ ràng, điều này đã được chứng minh qua câu *Machine learning is the subfield of computer science that “gives computers the ability to learn without being explicitly programmed”* – Wikipedia

Machine learning có mối quan hệ rất mật thiết đối với statistics (thống kê). Machine learning sử dụng các mô hình thống kê để "ghi nhớ" lại sự phân bố của dữ liệu. Tuy nhiên, không đơn thuần là ghi nhớ, machine learning phải có **khả năng tổng quát hóa** những gì đã được nhìn thấy và đưa ra dự đoán cho những trường hợp chưa được nhìn thấy. Chúng ta có thể hình dung một mô hình machine learning không có khả năng tổng quát như một đứa trẻ học vẹt (chỉ trả lời được những câu trả lời mà nó đã học thuộc lòng đáp án). Khả năng tổng quát là một khả năng tự nhiên và kì diệu của con người. Giả sử bạn không thể nhìn thấy tất cả các khuôn mặt người trên thế giới nhưng bạn có thể nhận biết được một thứ có phải là khuôn mặt người hay không với xác suất đúng gần như tuyệt đối. Đỉnh cao của machine learning sẽ là mô phỏng được khả năng tổng quát hóa và suy luận này của con người.

Như đã giới thiệu ở trên, nhắc đến machine learning là nói đến **dự đoán**, từ việc dự đoán nhãn phân loại đến dự đoán hành động cần thực hiện trong bước tiếp theo. Vậy machine learning có thể dự đoán tương lai hay không? Có thể có hoặc có thể không, machine learning có thể dự đoán được tương lai, nhưng chỉ khi tương lai có mối liên hệ mật thiết với hiện tại.

### **2.2.2 Phân nhóm Machine Learning**

Có hai cách phổ biến phân nhóm các thuật toán Machine learning. Một là dựa trên phương thức học (learning style), hai là dựa trên hàm (function).

#### Phân nhóm dựa trên phương thức học

* ***Supervised Learning (Học có giám sát)***
* *Classification (Phân loại)*
* *Regression (Hồi quy)*
* ***Unsupervised Learning (Học không giám sát)***
* *Clustering (phân nhóm)*
* *Association*
* ***Semi-Supervised Learning (Học bán giám sát)[1]***
* ***Reinforcement Learning (Học Củng Cố)[2]***

***Supervised learning*** là thuật toán dự đoán đầu ra (outcome) của một dữ liệu mới (new input) dựa trên các cặp (input, outcome) đã biết từ trước. Cặp dữ liệu này còn được gọi là (data, label), tức (dữ liệu, nhãn). Supervised learning là nhóm phổ biến nhất trong các thuật toán Machine Learning.

Với ***Unsupervised Learning***, chúng ta không biết được outcome hay nhãn mà chỉ có dữ liệu đầu vào. Thuật toán unsupervised learning sẽ dựa vào cấu trúc của dữ liệu để thực hiện một công việc nào đó, ví dụ như phân nhóm (clustering) hoặc giảm số chiều của dữ liệu (dimension reduction) để thuận tiện trong việc lưu trữ và tính toán.

***Semi-Supervised Learning*** là khi chúng ta có một lượng lớn dữ liệu ***X*** nhưng chỉ một phần trong chúng được gán nhãn. Những bài toán thuộc nhóm này nằm giữa hai nhóm được nêu bên trên (Supervised learning và Unsupervised learning)

***Reinforcement learning*** là các bài toán giúp cho một hệ thống tự động xác định hành vi dựa trên hoàn cảnh để đạt được lợi ích cao nhất (maximizing the performance). Hiện tại, Reinforcement learning chủ yếu được áp dụng vào Lý Thuyết Trò Chơi (Game Theory), các thuật toán cần xác định nước đi tiếp theo để đạt được điểm số cao nhất.

**[1],[2]** Có một số cách phân nhóm không có Semi-supervised learning hoặc Reinforcement learning.

#### Phân nhóm dựa trên chức năng

Có một cách phân nhóm thứ hai dựa trên chức năng của các thuật toán. Trong phần này, dưới đây em xin liệt kê ra một số thuật toán phổ biến:

* Regression Algorithms
* Classification Algorithms
* Instance-based Algorithms
* Regularization Algorithms
* Bayesian Algorithms
* Clustering Algorithms
* Artificial Neural Network Algorithms
* Dimensionality Reduction Algorithms
* Ensemble Algorithms
* ….

# CHƯƠNG 3: PHÂN TÍCH DỮ LIỆU

## **Mô tả bài toán**

Xây dựng mô hình học máy dự đoán khả năng mắc bệnh tim của những bệnh nhân bị chứng đau ngực với các thông số liên quan sẽ được mô tả chi tiết trong những phần dưới.

### **Thu thập dữ liệu (Data collection)**

Đây là bước đầu tiên trong vòng đời của một dự án học máy. Mục tiêu của bước này là thu thập tất cả các dữ liệu liên quan đến bài toán. Dữ liệu có thể được thu thập từ nhiều nguồn khác nhau như: files, database, internet, nghiên cứu...

Chất lượng và số lượng dữ liệu thu thập được sẽ quyết định đến chất lượng của kết quả đầu ra. Dữ liệu càng nhiều thì độ chính xác của mô hình càng cao.

Bước này bao gồm các công việc:

* Xác định các nguồn dữ liệu liên quan đến bài toán
* Thu thập các dữ liệu từ các nguồn này
* Tích hợp các dữ liệu thu thập được để tạo thành một tập dữ liệu nhất quán (Dataset) sử dụng cho các bước tiếp theo.

### **Mô tả tập dữ liệu (Dataset describe)**

Tên Dataset: **heart\_disease.csv**

File dữ liệu chứa thông tin hơn 1328 bệnh nhân bị bệnh đau tim

Mỗi dòng ứng với thông tin của một bệnh nhân, bao gồm 14 thuộc tính:

* **age**: Tuổi của bệnh nhân
* **sex**: Giới tính của bệnh nhân (1 = male; 0 = female)
* **cp**: Chest pain type (loại đau ngực được hiển thị cấp độ từ 0 đến 3)
  + Value 0: typical angina
  + Value 1: atypical angina
  + Value 2: non-anginal pain
  + Value 3: asymptomatic
* **trestbps**: resting blood pressure (huyết áp ở trạng thái nghỉ ngơi, được tính bằng *mmHg*)
* **chol**: Chỉ số cholesterol ở trong huyết thanh (đơn vị: *mg/dL*)
* **fbs**: fasting blood sugar (đường huyết lúc đói 120mg/dl; 1 = True, 0 = False)
* **restecg**: resting electrocardiographic results (điện tâm đồ lúc nghỉ, hiển thỉ cấp độ từ 0 đến 2)
  + Value 0: normal
  + Value 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV)
  + Value 2: showing probable or definite left ventricular hypertrophy by Estes' criteria
* **thalach**: maximum heart rate achieved (nhịp tim tối đa)
* **exang**: exercise induced angina (đau thắt ngực do tập thể dục; 1 = yes, 0 = no)
* **oldpeak:** ST depression induced by exercise relative to rest
* **slope**: Độ dốc của đỉnh đoạn ST lúc tập thể dục (0-1-2)
* **ca**: số lượng các mạch chính được tô màu bằng phương pháp soi huỳnh quang (từ 0 đến 4)
* **thal**: Thalassemia – bệnh tan máu bẩm sinh (0 = bình thường; 1 = nhẹ; 2 = nặng; 3 = rất nặng)
* **target**: Cho biết kết quả bệnh nhân có bị bệnh tim hay không? (0: Không bị bệnh tim mạch; 1: Bị bệnh tim mạch)

## **Khám phá dữ liệu**

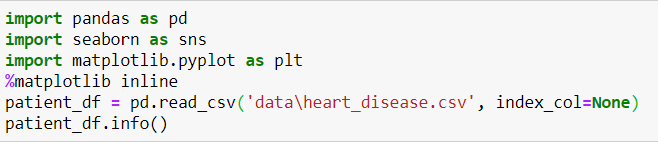
Thông thường dữ liệu thu thập được có rất nhiều phần bị nhiễu (noise), dữ liệu bị thiếu (missing value), dữ liệu ngoại lai (outliers)...dó đó dữ liệu cần phải được làm sạch và chuẩn hóa về dạng phù hợp. Đây là giai đoạn chiếm nhiều thời gian và nguồn lực nhất của một dự án Machine Learning.

Có rất nhiều nhiệm vụ phải thực hiện trong quá trình chuẩn bị dữ liệu. Một số vấn đề cơ bản cần giải quyết trong giai đoạn này bao gồm:

* Khám phá dữ liệu
* Làm sạch dữ liệu (xử lý giá trị thiếu, giá trị ngoại lai)
* Tích hợp dữ liệu
* Biến đổi, rời rạc hóa và chuẩn hóa dữ liệu
* Cân bằng dữ liệu.
* Rút gọn thuộc tính.

### **Đọc dữ liệu từ dataset**

Dữ liệu được đọc bằng thư viện pandas và import các thư viện cần thiết vào dự án phân tích này, với cú pháp như sau:

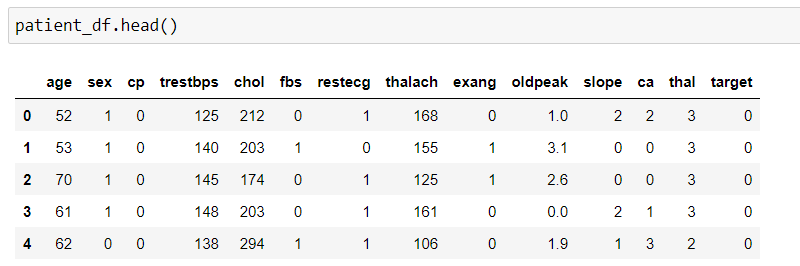


Dữ liệu đầu ra được hiển thị thông tin như sau:

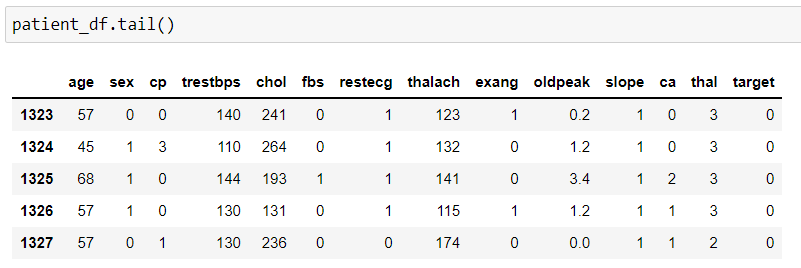
<class 'pandas.core.frame.DataFrame'>  
RangeIndex: 1328 entries, 0 to 1327  
Data columns (total 14 columns):  
 # Column Non-Null Count Dtype   
--- ------ -------------- -----   
 0 age 1328 non-null int64   
 1 sex 1328 non-null int64   
 2 cp 1328 non-null int64   
 3 trestbps 1328 non-null int64   
 4 chol 1328 non-null int64   
 5 fbs 1328 non-null int64   
 6 restecg 1328 non-null int64   
 7 thalach 1328 non-null int64   
 8 exang 1328 non-null int64   
 9 oldpeak 1328 non-null float64  
 10 slope 1328 non-null int64   
 11 ca 1328 non-null int64   
 12 thal 1328 non-null int64   
 13 target 1328 non-null int64   
dtypes: float64(1), int64(13)  
memory usage: 145.4 KB

Với 14 thuộc tính từ 0 đến 13, có 1328 dữ liệu trên mỗi thuộc tính, datatype của dataset chứa 13 thuộc tính kiểu int và 1 thuộc tính kiểu float. Có thể nhận định rằng, tệp dữ liệu này khá đầy đủ và không chứa chỉ số rỗng (null).

Chúng ta hiển thị 5 dòng dữ liệu đầu tiên với lệnh *patient\_df.head()* và nhận được kết quả đầu ra:



Tiếp tục dùng lệnh *patient\_df.tail()* để hiển thị 5 dòng dữ liệu cuối của tập dữ liệu:



### **Xác định kiểu dữ liệu trong dataset**

#### **Numerical** (Quantitative – định lượng)

Dữ liệu định lượng, như tên cho thấy là một dữ liệu liên quan đến số lượng hoặc các con số. Nó đề cập đến dữ liệu tính toán các giá trị và số lượng và có thể được biểu thị bằng số được gọi là dữ liệu định lượng. Trong thống kê, hầu hết các phân tích được thực hiện bằng cách sử dụng dữ liệu này.

Dữ liệu định lượng có thể được sử dụng trong tính toán và kiểm tra thống kê. Nó liên quan đến các phép đo như chiều cao, cân nặng, khối lượng, chiều dài, kích thước, độ ẩm, tốc độ, tuổi, v.v. Việc trình bày dữ liệu dạng bảng và dạng sơ đồ cũng có thể ở dạng biểu đồ, đồ thị, bảng, v.v. Hơn nữa, định lượng dữ liệu có thể được phân loại là dữ liệu rời rạc hoặc liên tục.

Ở trong tập dữ liệu này, các biến thuộc biến định lượng là: age, trestbps, chol, thalach, oldpeak.

#### **Categorical** (Qualitative – định tính)

Dữ liệu định tính đề cập đến dữ liệu cung cấp thông tin chi tiết và hiểu biết về một vấn đề cụ thể. Nó có thể được tính gần đúng nhưng không thể tính được. Do đó, nhà nghiên cứu cần có kiến ​​thức đầy đủ về loại đặc trưng, ​​trước khi thu thập dữ liệu.

Bản chất của dữ liệu là mô tả và vì vậy sẽ hơi khó phân tích nó. Loại dữ liệu này có thể được phân thành các loại, trên cơ sở các thuộc tính vật lý và thuộc tính của đối tượng. Dữ liệu được hiểu là những câu chuyện nói hoặc viết hơn là những con số. Nó liên quan đến dữ liệu có thể quan sát được về mùi, ngoại hình, mùi vị, cảm giác, kết cấu, giới tính, quốc tịch, v.v.

Trong tập dữ liệu, các biến định tính là: sex, cp, fbs, restecg, exang, slope, ca, thal, target.

### **Kiểm tra dữ liệu thiếu trong tập dữ liệu**

Nếu dữ liệu có chứa các giá trị thiếu (null/missing) nó sẽ gây ra vấn đề sai số cho các model học máy, gây ra hậu quả lớn cho các bước về sau. Vậy nên, quá trình chuẩn bị dữ liệu phải tìm ra các giá trị thiếu trong tập dữ liệu và xử lý triệt để.

Kiểm tra dữ liệu thiếu ta dùng lệnh *patient\_df.isnull().sum()*, trong đó: hàm *isnull()* để kiểm tra dữ liệu thiếu và *sum()* dùng để tính tổng các dữ liệu thiếu trong dataset.

patient\_df.isnull().sum()

Và kết quả nhận được là:

age 0

sex 0

cp 0

trestbps 0

chol 0

fbs 0

restecg 0

thalach 0

exang 0

slope 0

ca 0

thal 0

target 0

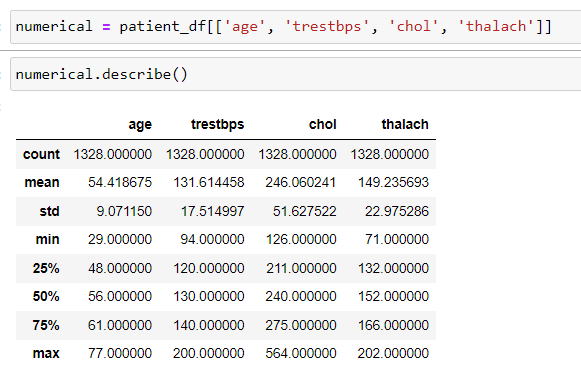
dtype: int64

Thật may mắn, tập dữ liệu này không chứa giá trị missing.

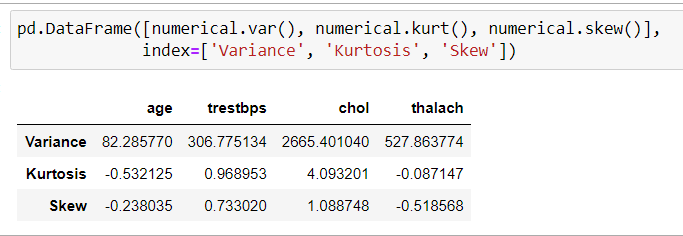
Tuy nhiên, để chắc chắn rằng tập data này không chứa giá trị bất thường, chúng ta sẽ dùng các phương pháp khác để kiểm tra giá trị outliers ở những phần sau.

## **3.3 Thống kê mô tả biến định lượng**

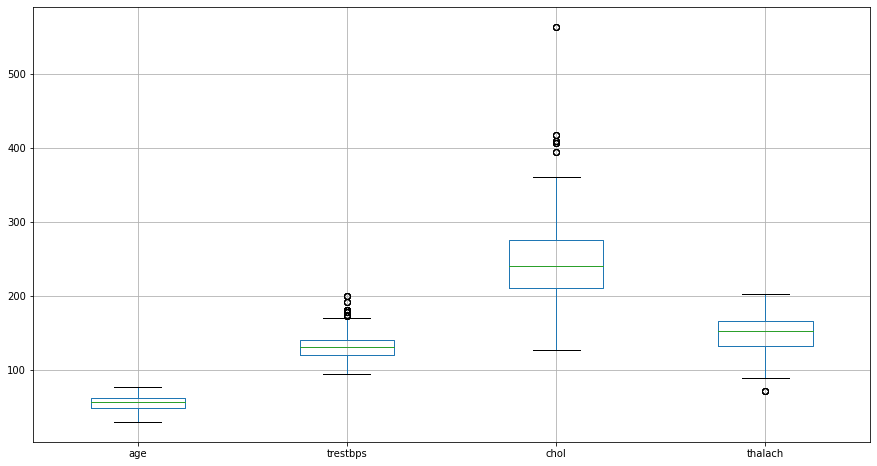
Em chia tệp dữ liệu gom thành định lượng và định tính, vì mỗi dữ kiểu dữ liệu sẽ có cách cách mô tả dữ liệu khác nhau.



Với các biến định lượng, chúng ta sẽ thống kê đo lường xu hướng tập trung và xu hướng biến động.



### **3.3.1 Boxplot**

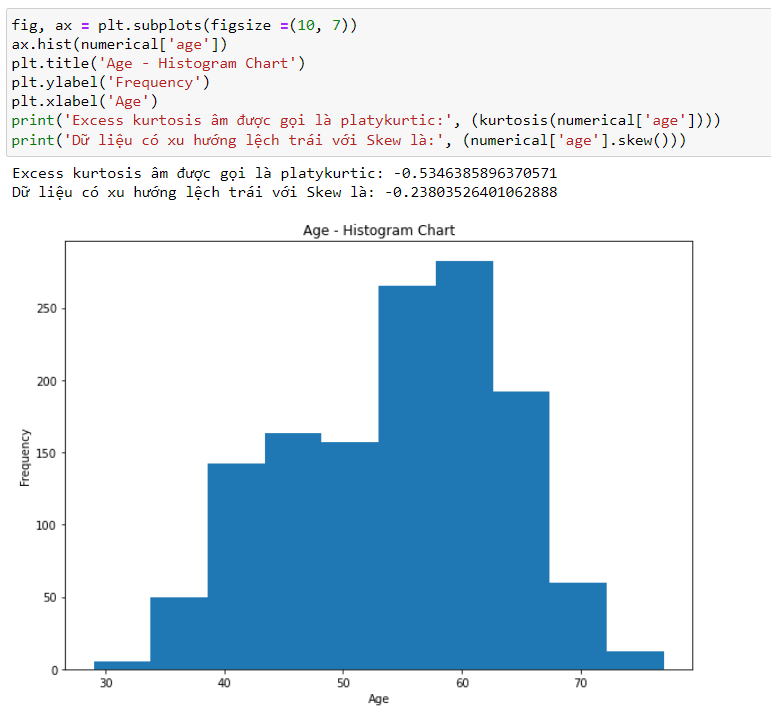
****

##### Hình 3.1: Biểu đồ hộp của các biến định lượng

Nhận xét từng biến về biểu đồ hộp bên trên:

* Age: có chỉ số trung vị (median) là 56, độ tuổi thấp nhất là 29 và độ tuổi già nhất là 77. Ở biểu đồ boxplot chúng ta không thấy sự xuất hiện của giá trị ngoại lai.
* Trestbps: trung vị là 130, phần tử lớn nhất không phải giả trị ngoại lai là 170, chỉ số huyết áp tâm trương thấp nhất là 94. Các giá trị ngoại lai nằm ở khoảng trên 170 đến 200.
* Chol: trung vị của chỉ số cholesterol là 240, tứ phân vị nằm ở khoảng từ 211 đến 275, thấp nhất là 126, các giá trị ngoại lai nằm ở vùng gần 400 đến hơn 500.
* Thalach: khoảng tứ phân vị của biến này nằm ở khoảng 132 đến 166, trung vị là 152, outlier được ghi nhận chỉ 1 giá trị duy nhất nằm ở dưới là 71, đây cũng chính là giá trị nhỏ nhất của biến này.

### **3.3.2 Age**

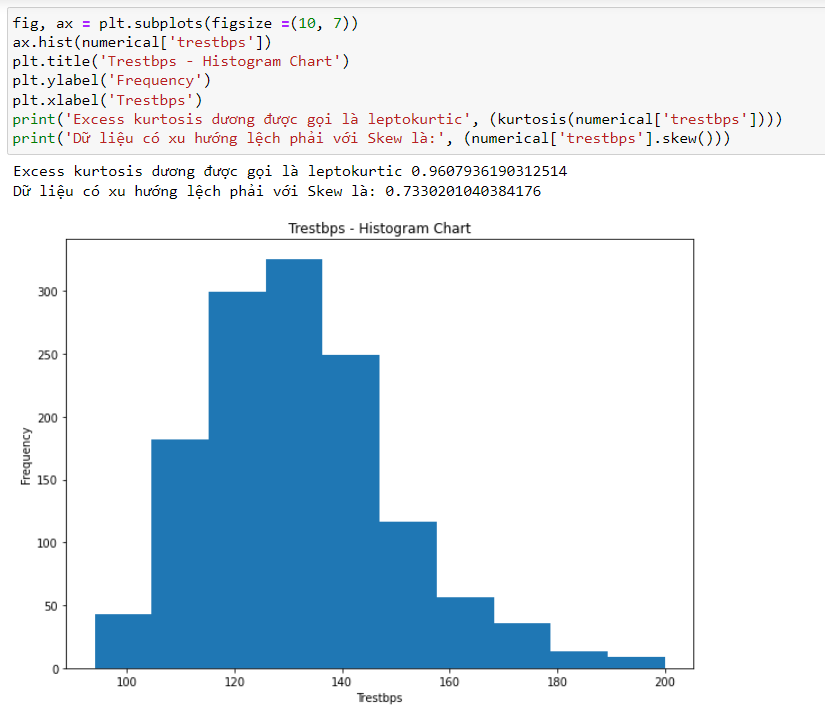
****

##### Hình 3.2: Biểu đồ tần suất của biến Tuổi

Từ histogram chart và kết quả của độ nhọn (Kurtosis), chúng ta có thể thấy đuôi biểu đồ “mỏng” hơn phân phối chuẩn, mang lại nguồn dữ liệu tích cực, ít outliers.

Với độ lệch (Skewness) là -0.23, quan sát biểu đồ có thể thấy dữ liệu có xu hướng lệch trái (negative skewness), đỉnh chuông lúc này nằm lệch thiên về bên phải.

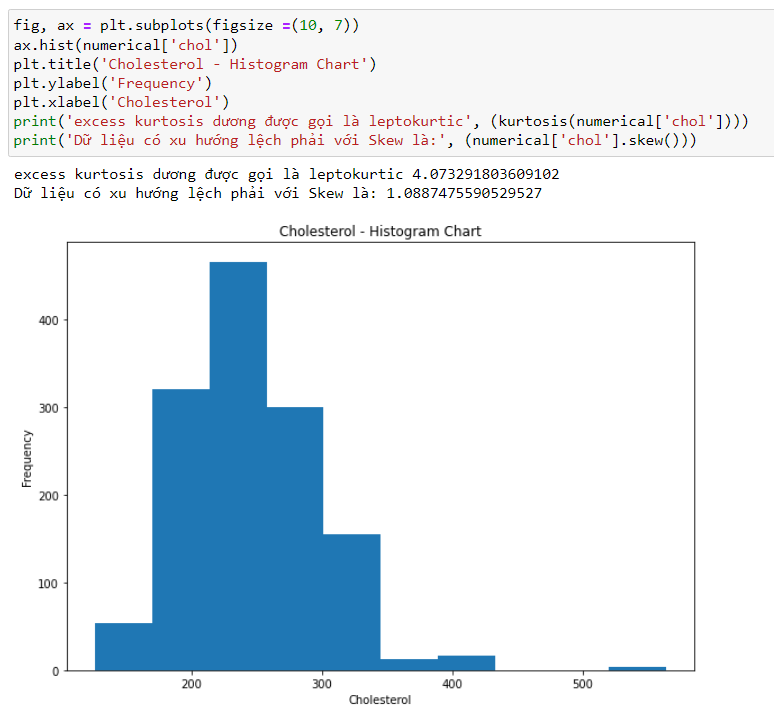
### **3.3.3 Trestbps (huyết áp tâm trương)**



##### Hình 3.3: Biểu đồ tần suất của biến Trestbps.

Chỉ số huyết áp tâm trương (trestbps) có độ nhọn với giá trị thặng dư gần bằng 1 cho thấy dữ liệu mang một phần nhỏ giá trị ngoại lai. Đỉnh nhọn tập trung ở phía bên trái, nơi tuần suất xuất hiện dữ liệu nhiều nhất ở khoảng 120-145, độ nhọn có thiên hướng lệch dương, phần đuôi bên phải không cân đối.

### **3.3.4 Chol (Cholesterol)**

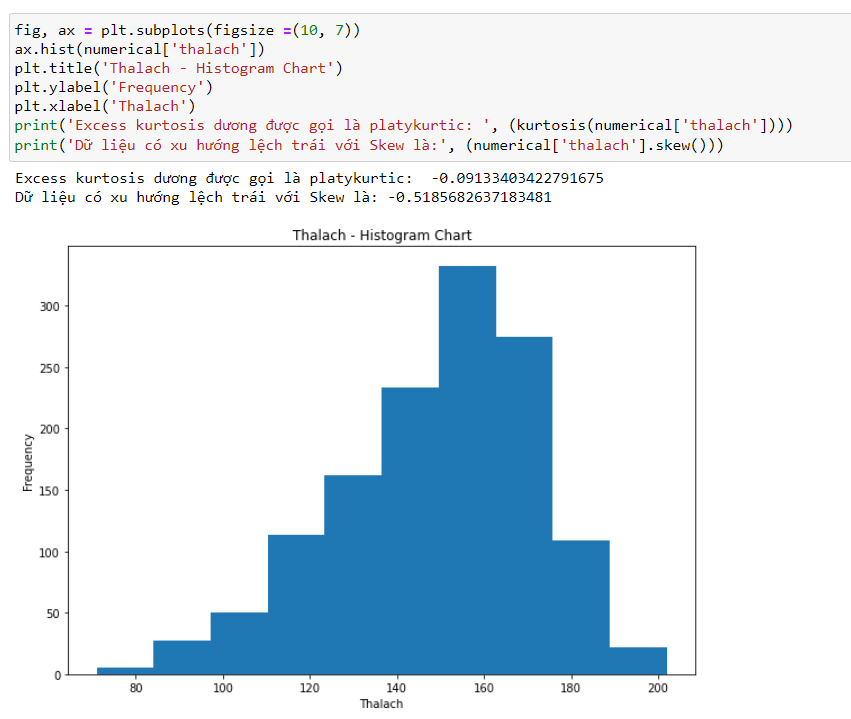


##### Hình 3.4: Biểu đồ tần suất của biến Chol.

Quan sát trên biểu đồ, chúng ta có thể thấy biểu đồ Cholesterol có đuôi hẹp, đỉnh chuông nhọn, với chỉ số Kurtosis là 4.07 (còn được gọi là Leptokurtic), chỉ số này cho ta thấy biến Chol có xác suất mang nhiều giá trị ngoại lai.

Độ lệch là dương 1.08, quan sát thêm biểu đồ, đỉnh dữ liệu nằm lệch sang trái vì phần đuôi bên phải của biểu đồ phân phối bị mất đối xứng do ít dữ liệu hơn.

### **3.3.5 Thalach (maximum heart rate achieved)**

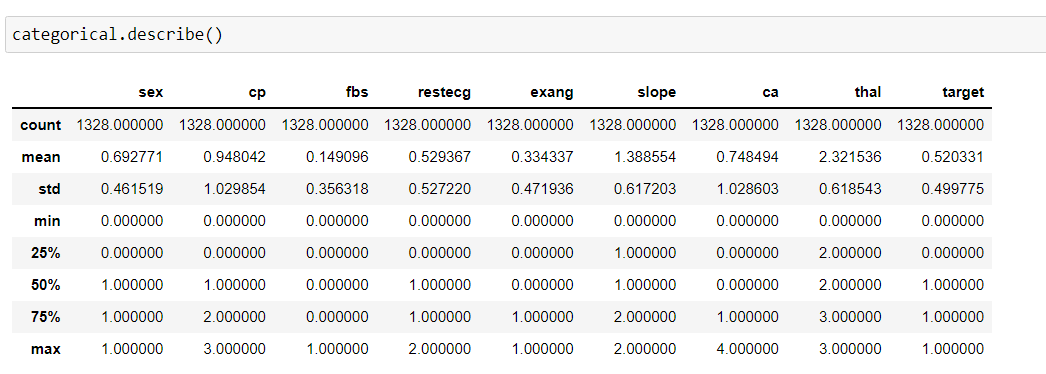


##### Hình 3.5: Biểu đồ tần suất của biến Thalach.

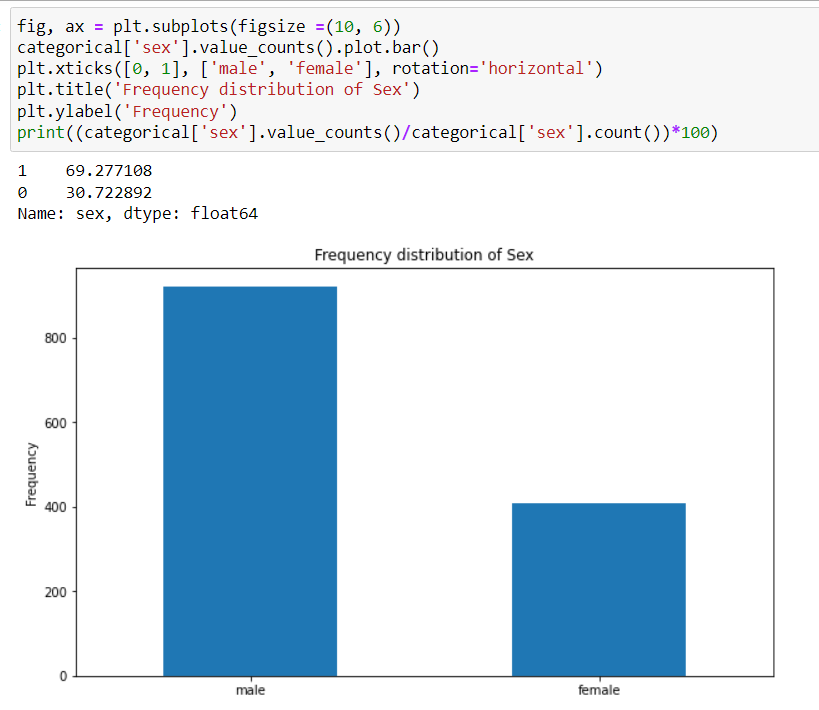
Nhịp tim tối đa trong tập dữ liệu có giá trị thặng dư của Kurtosis là -0.09, thuộc phân phối platykurtic, có ít giá trị ngoại lai.

Dữ liệu có xu hướng phần đuôi nằm lệch trái với đỉnh chuông nằm bên phải của biểu đồ, nhịp tim tối đa có giá trị 160 là khoảng giá trị mang xuất hiện nhiều nhất.

## **3.4 Thống kê mô tả biến định tính**



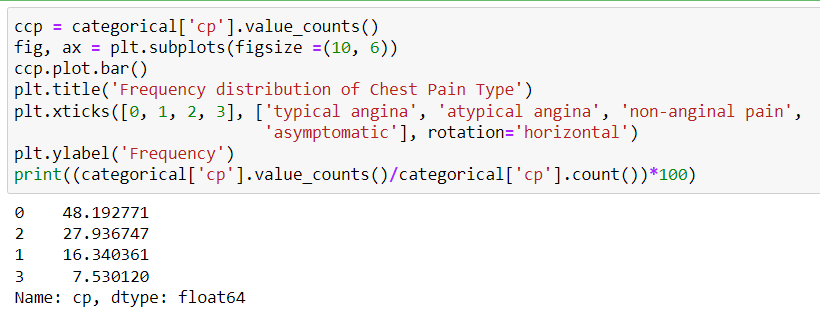
### **3.4.1 Sex (giới tính)**

****

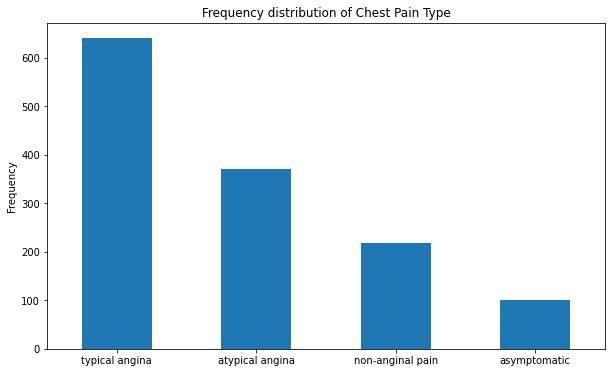
##### Hình 3.6: Biểu đồ phân phối tần suất của Giới tính (sex).

Tỉ lệ giới tính Nam được tham gia khảo sát chiếm gần 70% tổng số người được khảo sát, với hơn 920 lượt khảo sát. Giới tính Nữ chỉ chiếm hơn 30%, chênh lệch hơn 40% giữa Nam và Nữ.

### **3.4.2 Cp (Chest pain type)**

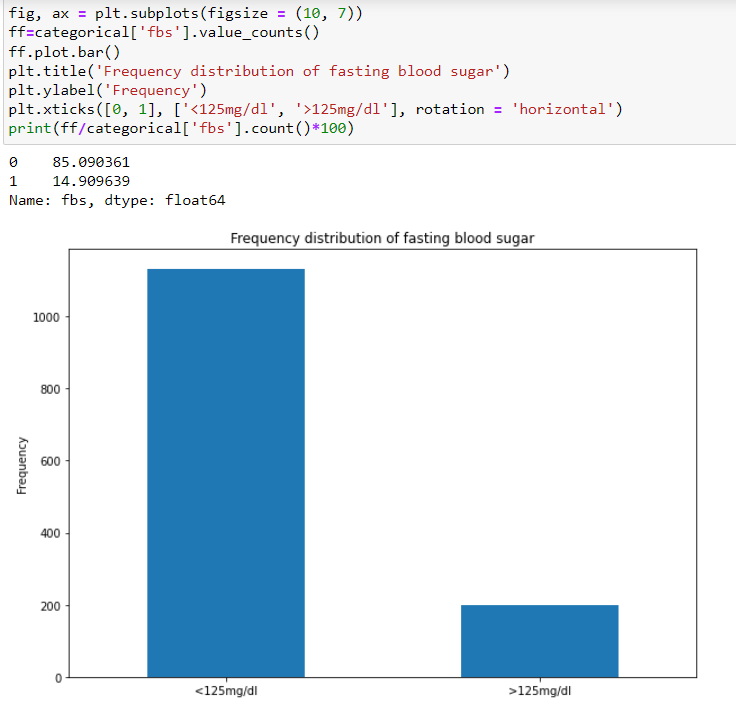
****

Tỉ lệ bệnh nhân được khảo sát bị đau thắt ngực điển hình (typical angina) chiếm gần ½ tổng số người tham gia khảo sát, với 640 người được ghi nhận; thấp dần ở loại atypical angina là gần 30% với 371 người; non-anginal pain là 16.3% gồm 217 người và tỉ lệ thấp nhất là asymptomatic với 7.5% với 100 người được ghi nhận.



##### Hình 3.7: Biểu đồ phân phối tần suất của các loại đau tim (cp).

### **3.4.3 Fbs (fasting blood sugar)**

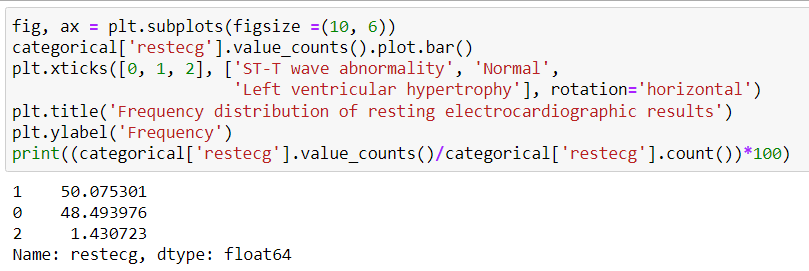
****

##### Hình 3.8: Biểu đồ tần suất của chỉ số đường huyết lúc đói (fbs).

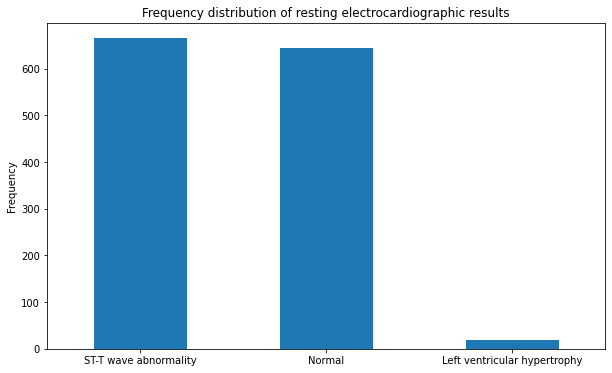
Theo cdc.gov, mức đường huyết lúc đói ở khoảng bình thường thấp hơn 99mg/dl, 100-125mg/dl thuộc nhóm tiền tiểu đường và trên 125mg/dl cho thấy bệnh nhân bị tiểu đường.

Thống kê trên cho thấy hơn 85% tỉ lệ người tham gia khảo sát có chỉ số đường huyết lúc đói trên 120mg/dl. Tỉ lệ bệnh nhân có chỉ số đường huyết lúc đói quá cao.

### **3.4.4 Restecg (resting electrocardiographic results)**

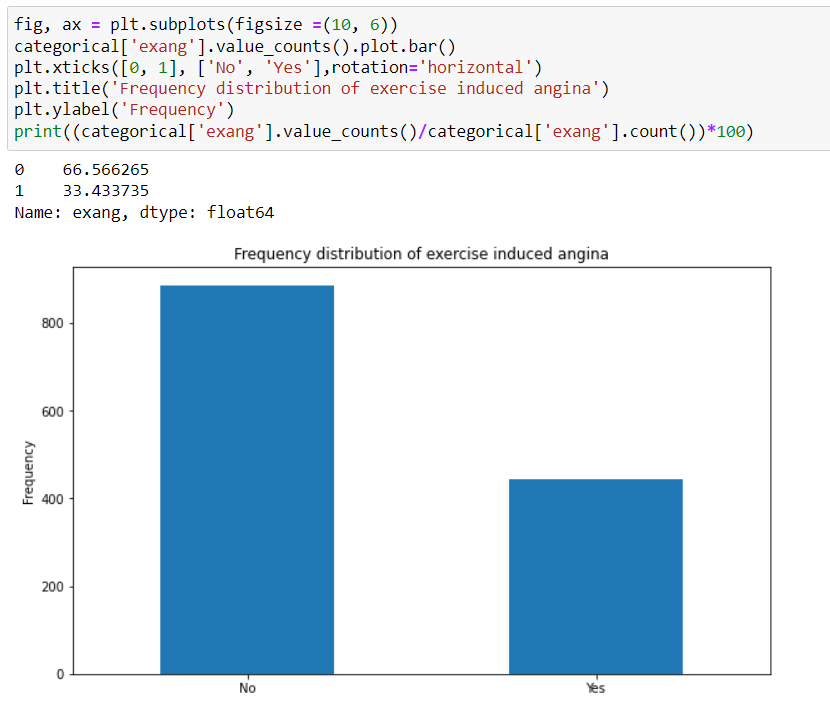
****

Tỉ lệ người có kết quả điện tâm đồ bình thường chiếm 48.5% tổng số người tham gia khảo sát; sóng ST-T có biểu hiện bất thường chiếm hơn 50% và phần còn lại là 1.4% số người tham gia có dấu hiệu phì đại thất trái có thể xảy ra hoặc xác định theo tiêu chí của Estes.



##### Hình 3.9: Biểu đồ tần suất của kết quả điện tâm đồ lúc nghỉ (restecg).

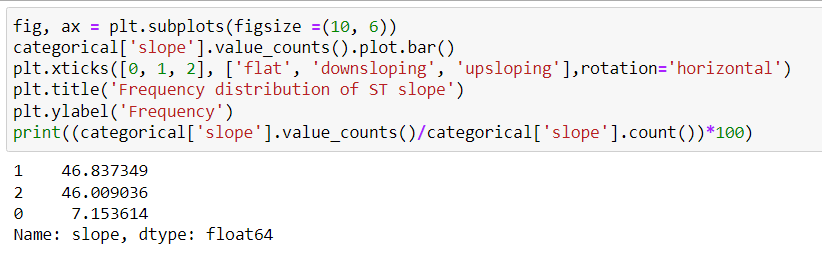
### **3.4.5 Exang (exercise induced angina)**

****

##### Hình 3.10: Biểu đồ tần suất của chỉ số đau thắt ngực do tập luyện (exang).

Tỉ lệ bệnh nhân có biểu hiện đau thắt ngực do tập luyện thể dục chiếm 33.4% tổng số người tham gia khảo sát; 66.5% bệnh nhân không có dấu hiệu đau sau khi tập luyện. Tần suất của bệnh nhân có biểu hiện đau thắt ngực thấp hơn các bệnh nhân không bị đau ngực là ½.

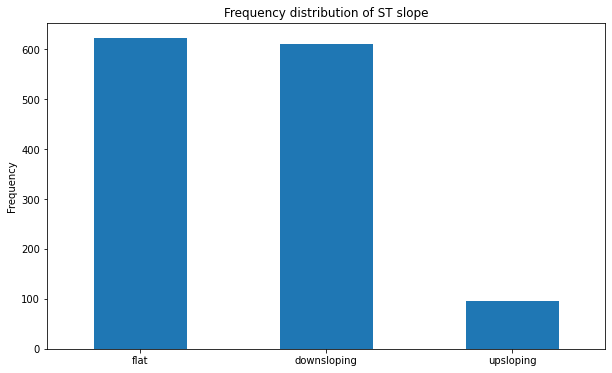
### **Slope (độ dốc của đoạn ST)**

****

Theo vinmec.com, độ dốc của đoạn ST mang sắc thái bình thường là không chênh lên cũng không đi xuống, mà đi ngang (flat). Nguyên nhân quan trọng nhất của đoạn ST bất thường (độ cao hoặc giảm xuống) là thiếu máu cục bộ cơ tim / nhồi máu.

Một số nguyên nhân của đoạn ST chênh lên (unsloping) là do: Nhồi máu cơ tim cấp tính, co thắt mạch vành, viêm màng ngoài tim, phì đại thất trái, phình vách thất, block nhánh trái,…

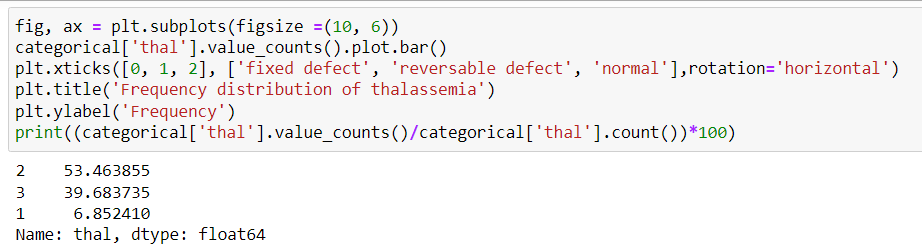
Một số nguyên nhân của đoạn ST chênh xuống (downsloping) là: Cơ tim thiếu máu cục bộ (NSTEMI), thay đổi đối ứng (STEMI), block nhánh phải, MI thành sau,…



##### Hình 3.11: Biểu đồ tần suất độ dốc ST (slope)

Quan sát biểu đồ có thể thấy, tỉ lệ người có đoạn ST đi ngang (flat) chiếm tỉ lệ cao nhất, với 46.8%; tỉ lệ người có đoạn ST đi xuống (downsloping) chiếm 46% tổng số người tham gia khảo sát. Tỉ lệ người có đoạn ST chênh lên (unsloping) là 7.15%.

### **3.4.7 Thal (Thalassemia – bệnh tan máu bẩm sinh)**

****

Thalassemia (còn được gọi là bệnh tan máu bẩm sinh), là một bệnh lý huyết học di truyền liên quan đến sự bất thường của hemoglobin (một cấu trúc protein trong hồng cầu có chức năng vận chuyển oxy). Ở bệnh nhân Thalassemia, các hồng cầu bị phá hủy quá mức dẫn đến tình trạng thiếu máu. Thalassemia là một bệnh di truyền lặn trên nhiễm sắc thể thường. Do đó, Thalassemia gây ra những hậu quả nghiêm trọng đến giống nòi, ảnh hưởng đến cuộc sống của bệnh nhân và cả cộng đồng.

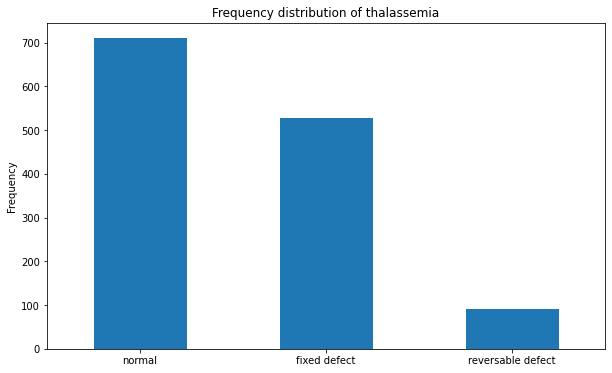
Có 2 loại bệnh Thalassemia chính là: **α-Thalassemia và β-Thalassemia.**

Bệnh nhân có thể xuất hiện các biến chứng như: Biến dạng xương ở mặt, mũi tẹt, răng vẩu; loãng xương, dễ gãy xương; da sạm, củng mạc mắt vàng,…và trong đó có cả các biến chứng tim mạch như: suy tim, rối loạn nhịp tim,…

* Khiếm khuyết có thể đảo ngược (reversable defect) trong bệnh thalassemia là gì?

Suy tim và rối loạn nhịp tim là nguyên nhân chính gây tử vong ở bệnh nhân b-thalassemia. Bệnh cơ tim do sắt có thể hồi phục, nếu bắt đầu thải sắt kịp thời, nhưng chẩn đoán thường bị trì hoãn do các triệu chứng xuất hiện muộn và bất thường trên siêu âm tim.

* Khiếm khuyết cố định (fixed defect) có thể hiểu là họ hoàn toàn không thấy máu chảy ở một số bộ phận của cơ thể.
* Normal là lưu lượng máu vẫn ở trạng thái bình thường.



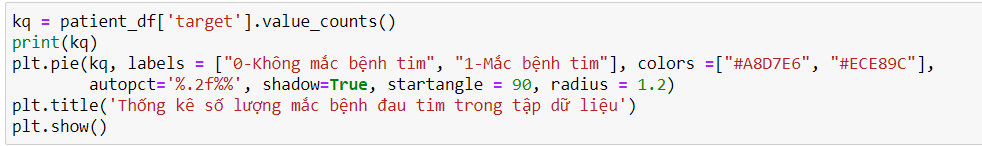
##### Hình 3.12: Biểu đồ tần suất của bệnh tan máu bẩm sinh (Thal).

Biểu đồ tần suất cho thấy 711 người mang dấu hiệu bình thường, chiếm 53% tổng số người tham gia khảo sát; hơn 500 người mang dấu hiệu của khiếm khuyết cố định với 39.6%; phần còn lại là khiếm khuyết có để đảo ngược chiếm 6.8% với 92 người khảo sát.

## **Bivariate Analysis – Phân tích song biến**

### **3.5.1 Thống kê tỉ lệ người mắc bệnh và người không mắc bệnh tim**

Để dễ hình dung tập dữ liệu, chúng ta tiến hành thống kê số lượng người mắc bệnh và không mắc bệnh, dưới đây là đoạn code:

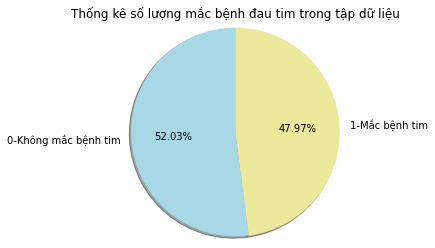


Kết quả nhận được:

1 691

0 637

Name: target, dtype: int64

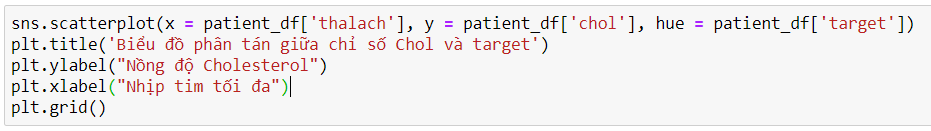


##### Hình 3.13: Tỉ lệ số bệnh nhân mắc bệnh tim

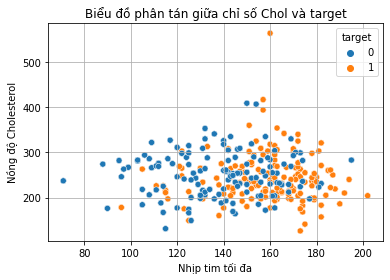
Chúng ta có thể thấy trên biểu đồ, tỉ lệ giữa người mắc bệnh và không mắc bệnh chênh lệch nhau khoảng ~4%, đây là con số đáng báo động bởi lẽ Tim mạch là bệnh lý nguy hiểm và chiếm tỉ lệ tử vong cao.

### **Mối quan hệ giữa chỉ số Cholesterol và kết quả (target)**

**3.5.2.1 Biểu đồ phân tán**



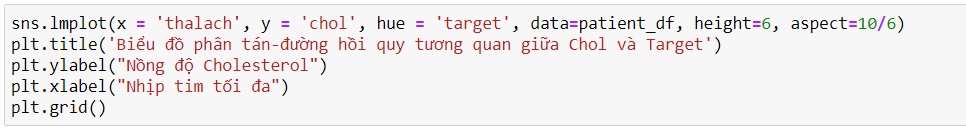
Chúng ta sử dụng hàm *sns.scatterplot()* và các đối số bên trong để tạo ra biểu đồ. Đầu tiên, đối số *x* thể hiện nhịp tim tối đa, *y* đại diện cho chỉ số *Cholesterol*; đặc biệt ở thông số *hue*, nhóm biến sẽ tạo ra các điểm có màu sắc khác nhau, ở đây chúng ta phân loại màu của chỉ số kết quả (thuộc tính *target*) nhằm áng chừng được độ phân tán của người mắc bệnh timm hay không mắc bệnh tim. Ngoài ra chúng ta sử dụng thêm *plt.title()* để thêm tên của biểu đồ, ở đây còn có hàm *plt.grid()* để biểu đồ có thêm những dòng lưới để dễ dàng quan sát cũng như ước lượng chỉ số của từng phân tử. Với hàm *plt.xlabel()* và *plt.ylabel()* tương ứng với nhãn trục tung và trục hoành của biểu đồ, và đây là biểu đồ mà chúng ta nhận được:



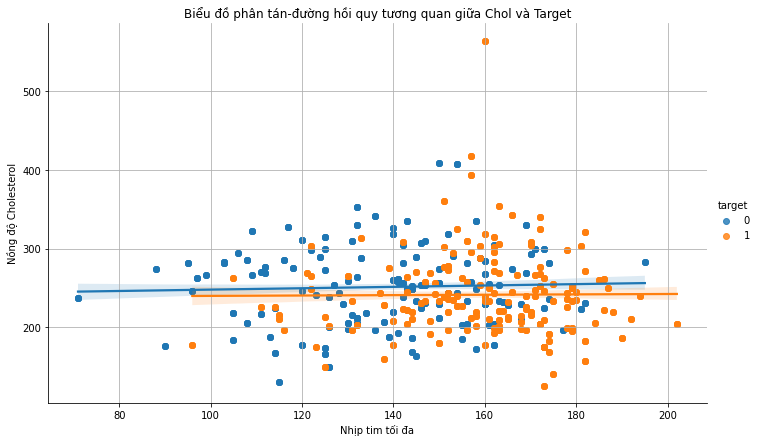
##### Hình 3.14: Biểu đồ phân tán Chol – Target

Với kết quả này, bằng mắt thường chúng ta có thể thấy điểm màu cam (tương ứng với người mắc bệnh) nằm hơi lệch sang phải và tịnh tiến lên trên. Để làm rõ vấn đề này, chúng ta nên tạo thêm 1 biểu đồ đường hồi quy.

* + - 1. **Biểu đồ phân tán-đường hồi quy**

****

sns.lmplot() là hàm kết hợp từ reglot() và FacetGrid(), nó được thiết kế như một giao diện thuận tiện để phù hợp với các mô hình hồi quy trên các tập con có điều kiện của tập dữ liệu. Các đối số và hàm còn lại của biểu đồ này thì cũng giống như biểu đồ 2.1

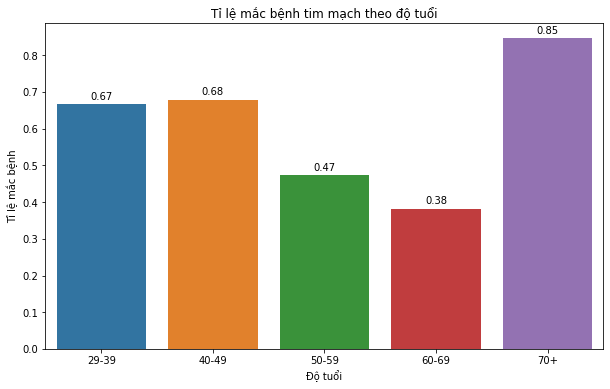


##### Hình 3.15: Biểu đồ phân tán – đường hồi quy giữ Chol và Target.

Nhìn từ biểu đồ có đường hồi quy, chúng ta có thể dễ xác định vùng người mắc bệnh nằm dày đặc ở khoảng x = [140; 190] và y = [200; 320] tương ứng với Nhịp tim và nồng độ Cholesterol.

Theo khuyến nghị của Hội Tim Mạch Học Việt Nam, chỉ số Cholesterol toàn phần <200 mg/dL cho biết tình trạng bình thường, nguy cơ mắc bệnh động mạch vành rất thấp; 200 - 239 mg/dL cho biết người bệnh đã hoặc đang có vấn đề về sức khỏe, cần chú trọng sinh hoạt điều độ và nên theo dõi sức khỏe định kỳ; với chỉ số Cholesterol toàn phần >= 240 mg/dL cho biết lượng Cholesterol trong máu tăng cao, người bệnh rất dễ có nguy cơ mắc bệnh xơ vữa động mạch. Từ khuyến nghị này áp dụng vào biểu đồ, có thể thấy mối tương quan giữa những bệnh nhân có Nhịp tim và Nồng độ Cholesterol cao phần lớn đều mắc bệnh tim. Phần lớn lời khuyên từ các tổ chức y tế đều nhắc đến Nhịp tim tăng cao đều do béo phì thừa cân gây nên nhiều vấn đề tiêu cực cho hệ tim mạch, hiệu quả và năng suất làm việc của mọi hoạt động trên cơ thể đều cần đến sự hỗ trợ từ hệ tim mạch. Vì vậy, nếu chúng ta đang có chỉ số Cholesterol>200mg/dL thì nên thay đổi nếp sống và ăn uống lành mạnh để duy trì một trái tim khỏe mạnh.

### **3.5.3 Mối quan hệ giữa độ tuổi và kết quả mắc bệnh (target)**

******

##### Hình 3.16: Tỉ lệ mắc bệnh tim mạch theo độ tuổi.

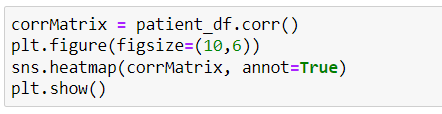
Từ biểu đồ trên, chúng ta có thể thấy bệnh tim ngày càng có xu hướng trẻ hóa, phát triển mạnh ở các độ tuổi trẻ, độ tuổi từ 29-39 chiếm 67% người mắc bệnh; tiếp sau đó là 40-49 chiếm 68%. Nguy cơ xảy ra các biến cố tim mạch cũng tăng lên theo độ tuổi già hóa, trong tập dữ liệu, các bệnh nhân có độ tuổi trên 70 tuổi mắc bệnh tim mạch chiếm 85%.Dấu hiệu mắc bệnh giảm xuống ở độ tuổi 60-69 với 38% và độ tuổi 50-59 chiếm 47%.

Đây là điều đáng báo động khi tỉ lệ mắc bệnh tim mạch ngày càng cao, ảnh hưởng tới sức khỏe, công việc và tuổi thọ. Chúng ta cần thay đổi cách ăn uống hợp lý, sinh hoạt, làm việc điều độ để giảm thiểu các tác nhân và hạn chế bệnh tim mạch, ngoài ra cần đi khám định kì để để biết được từng mức độ để sớm có biện pháp điều trị.

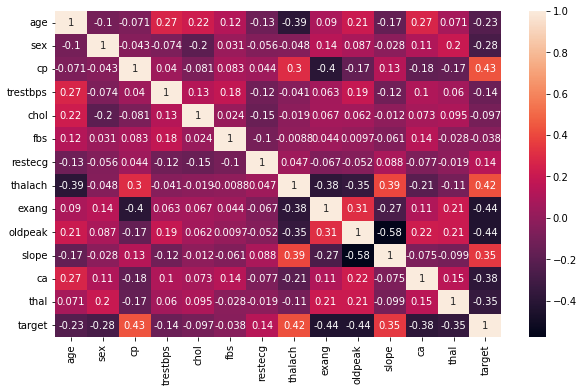
# CHƯƠNG 4: DỰ ĐOÁN VÀ ĐÁNH GIÁ KĨ THUẬT HỒI QUY LOGISTIC

## **Độ tương quan**

Ma trận tương quan là một cách để kiểm tra mối quan hệ tuyến tính giữa hai hoặc nhiều biến liên tục.

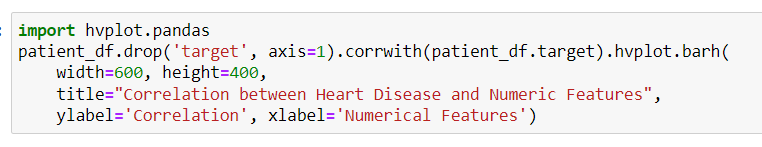


Kết quả hiển thị biểu đồ ma trận tương quan như bên dưới:

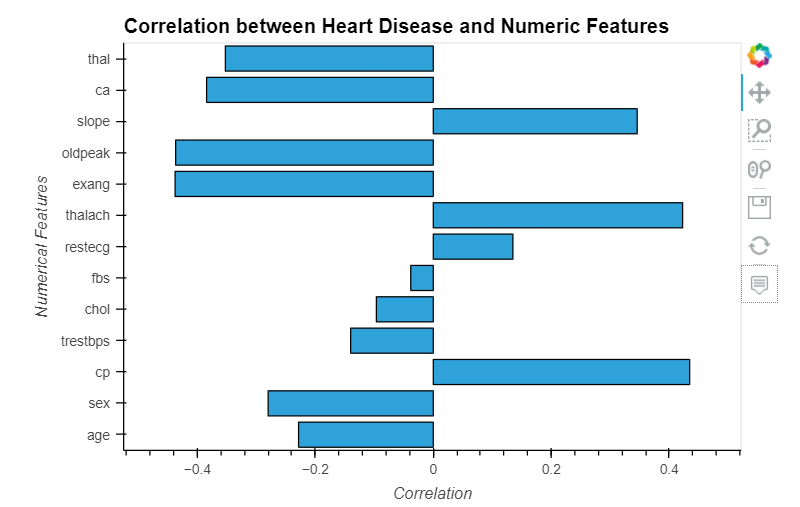


##### Hình 4.1: Biểu đồ tương quan

Để làm rõ biểu đồ này, chúng ta tạo thêm một biểu đồ cột nằm ngang, có thể thể hiện rõ được những chỉ số giúp ta phân tích chính xác hơn.



Có thể thấy, các chỉ số slope, thalach, cp, restecg đều có mối tương quan thuận (dương) đối với biến target. Các chỉ số còn lại đều có mối tương quan nghịch (âm) đối với biến target, trong đó chỉ số fbs và chol là có mỗi tương quan nhẹ (gần về 0)

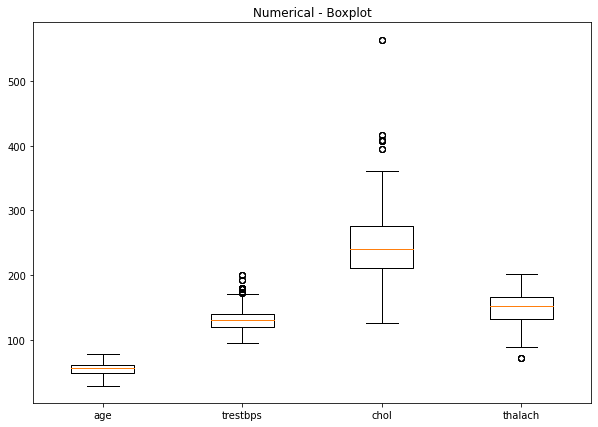


##### Hình 4.2: Biểu đồ cột nằm ngang của tương quan

## **Xử lý giá trị nhiễu**

Trước khi đưa dữ liệu vào mô hình Học Máy (Machine Learning) những dữ liệu xấu này cần được loại bỏ để tăng tính chính xác cho các model. Chúng ta sẽ quan sát các outlier của các biến định lượng (numeric) bằng biểu đồ hộp-boxplot để xác định các biến cần xóa outlier.

Dưới đây là biểu đồ diễn tả 5 vị trí phân bố của dữ liệu, đó là: giá trị nhỏ nhất (min), tứ phân vị thứ nhất (Q1), trung vị (median), tứ phân vị thứ 3 (Q3) và giá trị lớn nhất (max). Nếu dữ liệu ra khỏi vùng min-max thì đó là outlier



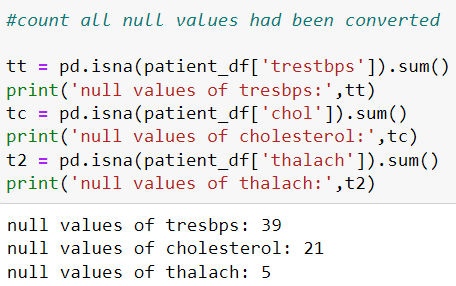
##### Hình 4.3: Xác định Outliers bằng biểu đồ hộp boxplot

Quan sát biểu đồ hộp có thể thấy các giá trị ngoại lai nằm ở các thuộc tính trestbps, chol và thalach. Chúng ta sẽ xóa các giá trị ngoại lai này bằng các bước như sau:

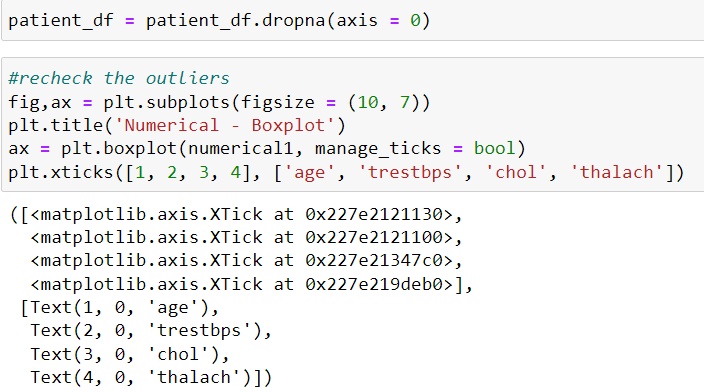
* Bước 1: Chuyển các outliers thành giá trị null



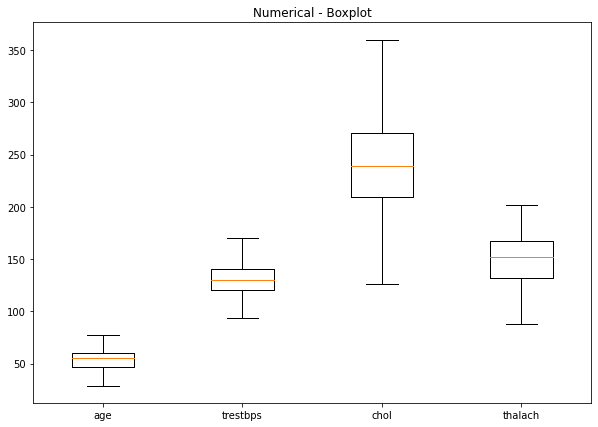
* Bước 2: Kiểm tra lại tổng số giá trị null đã được chuyển đổi



* Bước 3: Xóa các giá trị null (theo hàng) và kiểm tra lại bằng biểu đồ hộp



Kết quả sau khi xóa các giá trị ngoại lai ta nhận được biểu đồ hộp như sau:



##### Hình 4.4: Kiểm tra các giá trị ngoại lai đã xóa

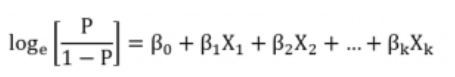
## **Kĩ thuật hồi quy Logistic (Logistic Regression)**

### **4.3.1 Logistic Regression đa biến được sử dụng khi nào**

Hồi quy logistic đa biến được sử dụng phổ biến nhất trong các nghiên cứu y học, bởi vì phần lớn các nghiên cứu có biến kết cục là biến nhị phân.

Mô hình hồi quy logistic nhị thức được sử dụng để dự đoán một biến phụ thuộc lưỡng phân (ví dụ: có/không, đạt/không đạt) bởi một hoặc nhiều biến độc lập liên tục hoặc phân loại.

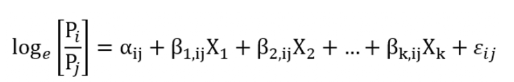
Phương trình liên hệ có dạng (logarit Odds):



Mô hình hồi quy đa thức (còn được gọi là hồi quy logistic đa thức) tương tự như mô hình hồi quy logistic nhị phân nhưng biến phụ thuộc là biến phân loại có nhiều hơn hai trạng thái.

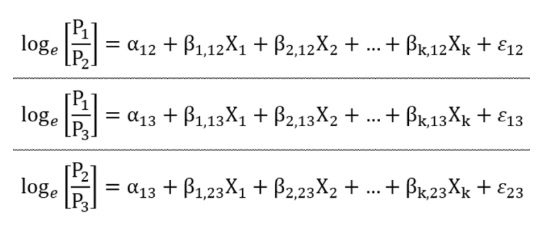
Ví dụ, mô hình hồi quy đa thức có thể được sử dụng để để dự đoán xem việc lựa chọn ngành học của các sinh viên có thể bị ảnh hưởng bởi nghề nghiệp của cha mẹ họ và thành tích học vấn phổ thông của họ. Biến phụ thuộc là lựa chọn ngành học như kĩ thuật, kinh tế, ngoại ngữ. Biến độc lập là thành tích học vấn phổ thông của học sinh và nghề nghiệp của cha. Ví dụ khác, học hết chương trình phổ thông trung học, lựa chọn tiếp theo của học sinh có thể là chương trình đại học, chương trình học nghề hoặc đi làm doanh nghiệp. Liệu sự lựa chọn của học sinh có phụ thuộc vào thành tích học vấn phổ thông và tình trạng kinh tế xã hội của họ.

Mô hình hồi quy logistic đa thức có dạng:



Trong đó, i và j là các phương án của phân loại trong biến độc lập. Gọi p1 là xác suất lựa chọn chương trình đại học, p2 là lựa chọn chương trình học nghề, p3 là xác suất lựa chọn đi làm doanh nghiệp.

Ta có 3 phương trình sau:



### **4.3.2 Bản chất hồi quy Logistic đa thức**

Sự phức tạp của hồi quy logistic đa thức khiến nó khó diễn giải cụ thể mà được giải thích theo ngữ cảnh khác nhau. Để diễn giả chi tiết hơn, mô hình hồi quy logistic đa thức được phân biệt trong 3 trường hợp:

* Trường hợp 1: Mô hình hồi quy logistic đa thức với dữ liệu đặc thù của người chọn.
* Trường hợp 2: Mô hình hồi quy logistic đa thức với dữ liệu đặc thù của lựa chọn.
* Trường hợp 3: Mô hình hồi quy logistic đa thức với dữ liệu kết hợp đặc thù của người chọn và đặc thù của lựa chọn.

Với trường hợp 1, các lựa chọn phụ thuộc vào các đặc điểm của người chọn, chẳng hạn như tuổi, thu nhập, giáo dục, tôn giáo… Các biến này là đặc thù đối với người chọn. Các loại mô hình này thường được ước lượng bằng các mô hình **logit (hoặc probit) đa thức**. Câu hỏi chủ yếu mà các mô hình này trả lời là: *Các đặc biểm của người chọn ảnh hưởng như thế nào đến việc lựa chọn của họ về một phương án cụ thể giữa một tập hợp nhiều phương án?*

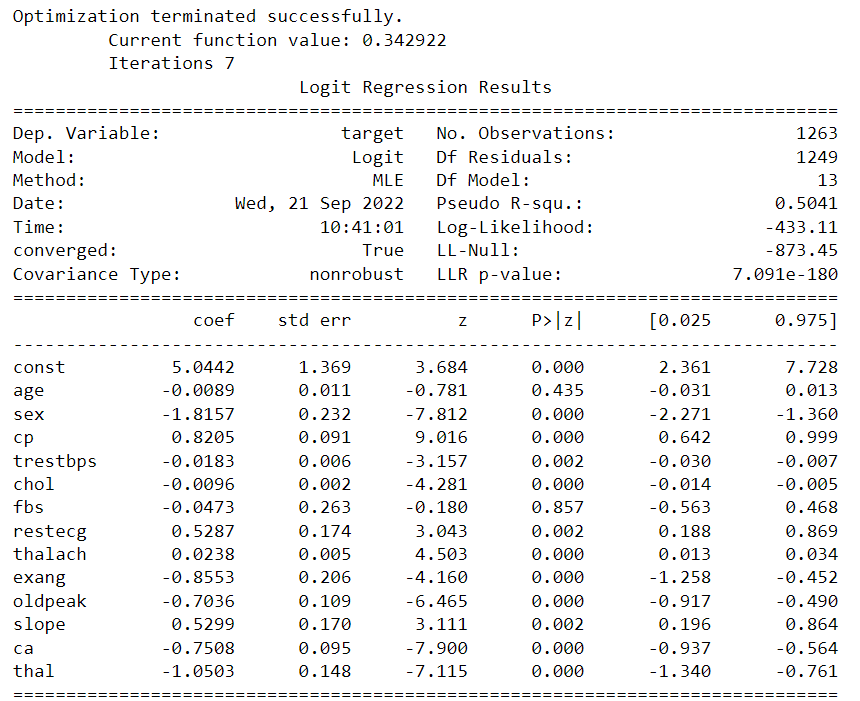
Với trường hợp 2, giả sử một học sinh sau trung học phải lựa chọn giữa ba loại ngành học: Kỹ thuật, Kinh tế, Ngoại ngữ. Như vậy, chúng có các thuộc tính đặc điểm đặc thù của lựa chọn không phụ thuộc vào cá nhân. Trong mô hình này cũng không có các đặc điểm đặc thù của cá nhân. Các mô hình như thể thường được ước lượng bằng các mô hình **logit (hoặc probit) có điều kiện**. Câu hỏi chính mà các mô hình như thế trả lời là: Các đặc điểm hoặc thuộc tính của các phương án khác nhau có ảnh hưởng đến sự lựa chọn của cá nhân giữa chúng hay không? Ví dụ, người ta có mua xe dựa trên các thuộc tính, như màu sắc, hình dáng, quảng cáo thương mại, và các đặc điểm khuyến mãi? Vì thế, **logit (hoặc probit) có điều kiện** là phù hợp khi các biến giải thích thay đổi khác nhau giữa các phương án.

Trong mô hình logit đa thức chuẩn, các biến giải thích là không đối với các phân loại của kết quả (outcome categories), nhưng các tham số thay đổi với kết quả. Trong mô hình logit có điều kiện, các biến giải thích thay đổi theo kết quả cũng như theo cá nhân, trong khi đó các tham số được giả định cố định qua tất cả các phân loại kết quả.

Với trường hợp 3: Ở đây chúng ta có dữ liệu về cả các đặc điểm đặc thù của người chọn và đặc thù của lựa chọn. Các mô hình như thế cũng có thể được ước lượng bằng mô hình logit có điều kiện bằng cách đưa thêm các biến giả vào mô hình. Ví dụ, khi chọn mua xe hơi, các thuộc tỉnh của các chiếc xe cũng như thu nhập và tuổi của cá nhân có thể ảnh hưởng đến lựa chọn của họ về chiếc xe. Hoặc một người tốt nghiệp phổ thông họ đang đứng trước ba lựa chọn: không đi học đại học/cao đẳng, học cao đẳng/trung cấp, và học đại học; chúng phụ thuộc vào tính hữu dụng kinh tế của tấm bằng sau tốt nghiệp, tổng thu nhập của gia đình, thành tích học tập phổ thông, giới tính, nghề nghiệp của cha mẹ...

### **4.3.3 Xây dựng mô hình hồi quy Logistic**

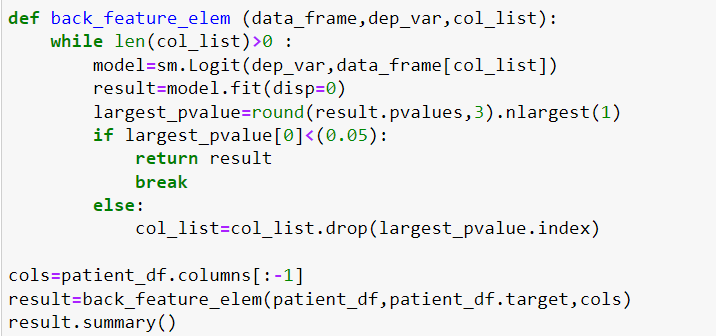
**Kiểm tra trước kết quả của hồi quy Logistic với tập dữ liệu:**

****

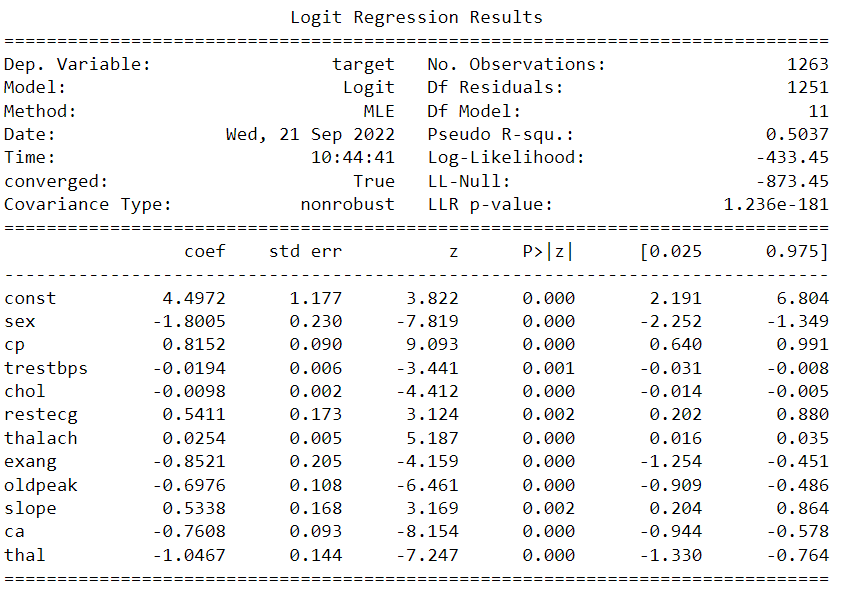
Trong bảng kết quả ta có các hệ số: coefficent, độ lệch chuẩn (standard errors), kiểm định z, p-values và khoảng tin cậy của coefficent.

Kết quả ở trên cho thấy một số thuộc tính có giá trị P cao hơn alpha được ưa thích (5%) và ở đó cho thấy mối quan hệ có ý nghĩa thống kê thấp với xác suất bệnh tim. Ở đây, phương pháp bầu cử ngược được sử dụng để loại bỏ từng thuộc tính có giá trị P cao nhất được tạo ra bằng cách chạy hồi quy lặp lại cho đến khi tất cả các thuộc tính có Giá trị P nhỏ hơn 0,05.

Ta dùng Backward elemination (P-value approach) để loại bỏ các feature không ảnh hưởng tới biến phụ thuộc (y) hoặc là biến dự đoán (prediction of value). Đây là phương pháp nhanh nhất nhằm thay đổi p-values với mục đính để giá trị đó có ý nghĩa thống kê. Nó sẽ tự động loại trừ các giá trị có p-value > 0.05 như age hoặc fbs.



Kết quả trả về như sau:



Quan sát kết quả bên trên, chúng ta có thể thấy các p-value có giá trị lớn hơn 0.05 là age và fbs đã được loại bỏ ra khỏi bảng kết quả hồi quy Logit.

Kết quả hồi quy được diễn giải như sau:

* sex = -1.8005, với hệ số hồi quy giới tính này có nghĩa giới tính là nam (đã được xét mặc định), nên tỉ lệ xác suất bị bệnh tim/không bị bệnh tim (trung bình giá trị logit) sẽ tăng 1.8005 đơn vị.
* cp = 0.8152 có nghĩa là khi chỉ số chest pain tăng lên 1 cấp độ thì tỉ lệ xác suất bị bệnh tim/không bị bệnh tim (trung bình giá trị logit) tăng lên 0.8152 đơn vị.
* trestbps = -0.0194 huyết áp ở trạng thái nghỉ ngơi giảm 1 mmHg thì tỉ lệ xác suất bị bệnh tim/không bị bệnh tim (trung bình giá trị logit) tăng lên 0.0194 đơn vị.
* chol = -0.0098 chỉ số Cholesterol giảm 1 mg/dL thì trung bình giá trị logit tăng lên 0.0914 đơn vị.
* restecg = 0.5411 chỉ số điện tâm đồ lúc nghỉ tăng 1 cấp độ thì trung bình giá trị logit tăng lên 0.5411 đơn vị.
* thalach = 0.0254, nhịp tim tối đa tăng lên 1 đơn vị thì trung bình giá trị logit tăng lên 0.0254 đơn vị.
* exang = -0.8521, cấp độ đau thắt ngực lúc tập thể dục giảm xuống 1 cấp độ thì trung bình giá trị logit tăng lên 0.8521 đơn vị.
* oldpeak = 0.6976 tăng lên 1 đơn vị thì trung bình giá trị logit tăng lên 0.6976 đơn vị.
* slope = 0.5338 tăng lên 1 đơn vị thì trung bình giá trị logit tăng lên 0.5338 đơn vị.
* ca = -0.7608 là số lượng mạch giảm xuống thì) trung bình giá trị logit tăng lên 0.7608 đơn vị.
* thal = -1.0467, cấp độ tan máu bẩm sinh giảm 1 đơn vị thì trung bình giá trị logit tăng lên 1.0467 đơn vị.

Vậy, ta có hàm logistic như sau:

= α + βx + ε

Từ kết quả của bảng hồi quy Logistic với 11 biến trong tập dữ liệu và áp dụng vào hàm logistic ta có phương trình sau:

= 4.4972 + 0.8152\*cp – 0.0194\*trestbps - 0.0098\*chol + 0.5411\*restecg + 0.0254\*thalach – 0.8521\*exang – 0.6976\*oldpeak + 0.5338\*slope – 0.7608\*ca – 1.0467\*thal

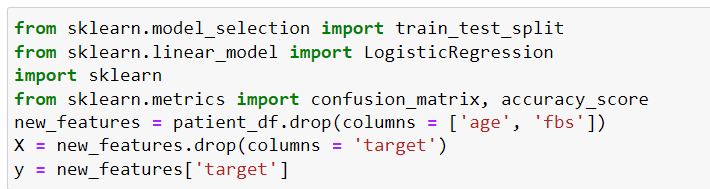
Ta biết chính là odd, vì vậy có thể viết:

= e(4.4972 + 0.8152\*cp – 0.0194\*trestbps – 0.0098\*chol + 0.5411\*restecg + 0.254\*thalach – 0.8521\*exang – 0.6976\*oldpeak + 0.5338\*slope – 0.7608\*ca – 1.0467\*thal)

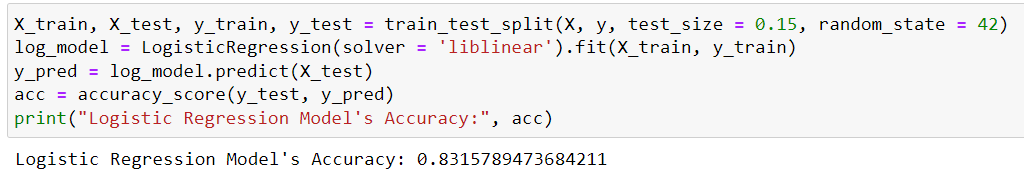
**Xây dựng mô hình Logistic Regression:**

Đầu tiên, import các thư viện cần thiết, xóa thuộc tính không có tính ảnh hưởng tới kết quả (age và fbs) và chia tập dữ liệu thành 2 phần, trong đó:

* X là biến độc lập (independent variables)
* y là biến phụ thuộc (dependent variable)



Chia dữ liệu thành 15% test và 85% train và áp dụng vào mô hình hồi quy tuyến tính như bên trên.



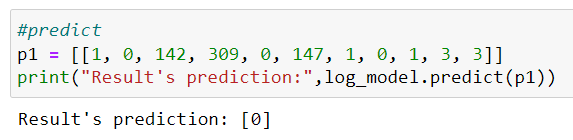
Xây dựng mô hình, train dữ liệu và tính độ chính xác của hồi quy Logistic dựa trên những hàm dựng sẵn trong python 3.

Có thể thấy, mô hình hồi quy logistic được đánh giá độ chính xác lên tới 83.1%, đây là 1 con số khá khả quan tới việc áp dụng mô hình vào tập dữ liệu.

**Dự đoán:**

Chúng ta tiếp tục tới mục dự đoán (predict) 1 bệnh nhân ngẫu nhiên, có các chỉ số sau đây:

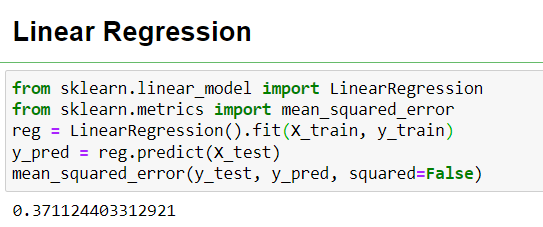
* Sex: 1
* Cp: 0
* Trestbps: 142
* Chol: 309
* Restecg: 0
* Thalach: 147
* Exang: 1
* Oldpeak: 0.0
* Slope: 1
* Ca: 3
* Thal: 3



Với các chỉ số trên thì kết quả được đưa ra là 0 (không mắc bệnh), trùng hợp với những chỉ số bình thường của những người không mắc bệnh.

### **4.3.4 So sánh với mô hình hồi quy tuyến tính (Linear regression)**

Em sẽ xây dựng một mô hình hồi quy để so sánh với hồi quy Logistic, đó là hồi quy tuyến tính. Dưới đây là mô hình bằng code python:



Chúng ta vẫn giữ phần train và test theo tỉ lệ 85 train và 15 test như lúc đầu, import thư viện và các hàm xây dựng sẵn vào, train dữ liệu và tính sai số toàn phương trung bình (mean squared error).

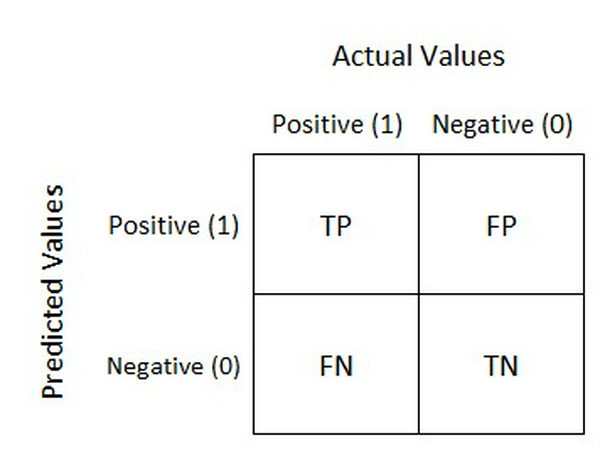
Sai số toàn phương trung bình (MSE) của một phép ước lượng là trung bình của bình phương các sai số, tức là sự khác biệt giữa các ước lượng và những gì được đánh giá. MSE là một hàm rủi ro, tương ứng với giá trị kỳ vọng của sự mất mát sai số bình phương hoặc mất mát bậc hai. Sự khác biệt xảy ra do ngẫu nhiên, hoặc vì các ước lượng không tính đến thông tin có thể cho ra một ước tính chính xác hơn.

Với MSE = 0.37 thì đây là con số không khả quan. Hơn nữa, tập dữ liệu nguyên trạng này cũng không phù hợp với mô hình hồi quy tuyến tính, bởi vì biến đầu ra (outcome) là dạng nhị phân, nếu dùng mô hình hồi quy tuyến tính thì cần xử lý dữ liệu để phù hợp với thuật toán hơn.

Vì vật, trong trường hợp này, thuật toán hồi quy Logistic phù hợp với tập dữ liệu nguyên thủy hơn so với thuật toán hồi quy tuyến tính.

### **4.3.5 Đánh giá kết quả của thuật toán hồi quy Logistic**

Để có cái nhìn khách quan hơn về các kết quả thực tế và kết quả dự đoán. Đây là một phương pháp đánh giá kết quả của những bài toán phân loại với việc xem xét cả những chỉ số về độ chính xác và độ bao quát của các dự đoán cho từng lớp. Một confusion matrix gồm 4 chỉ số sau đối với mỗi lớp phân loại:



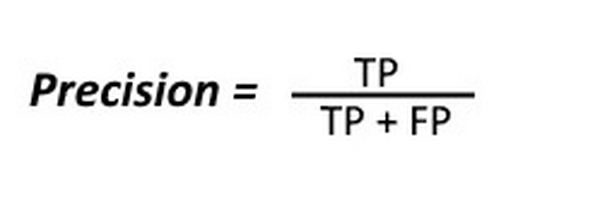
##### Hình 4.5: Các chỉ số của confusion matrix.

Trong bài toán chuẩn đoán tim mạch ta có 2 lớp: Lớp bị tim mạch được chuẩn đoán Positive và lớp không bị tim mạch được chuẩn đoán là Negative:

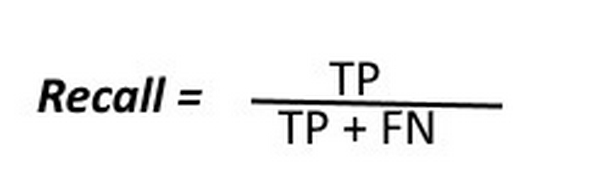
* TP (True Positive): Số lượng dự đoán chính xác. Là khi mô hình dự đoán đúng một người bị tim mạch.
* TN (True Negative): Số lượng dự đoán chính xác một cách gián tiếp. Là khi mô hình dự đoán đúng một người không bị tim mạch, tức là việc không chọn trường hợp bị tim mạch là chính xác.
* FP (False Positive - Type 1 Error): Số lượng các dự đoán sai lệch. Là khi mô hình dự đoán một người bị tim mạch và người đó hoàn toàn khỏe mạnh.
* FN (False Negative - Type 2 Error): Số lượng các dự đoán sai lệch một cách gián tiếp. Là khi mô hình dự đoán một người không bị tim mạch nhưng người đó bị tim mạch, tức là việc không chọn trường hợp bị tim mạch là sai.

Từ 4 chỉ số này, ta có 2 con số để đánh giá mức độ tin cậy của một mô hình:

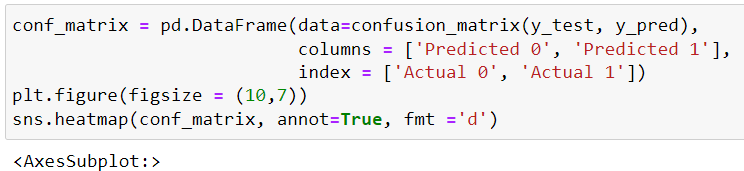
* Precision: Trong tất cả các dự đoán Positive được đưa ra, bao nhiêu dự đoán là chính xác? Chỉ số này được tính theo công thức:



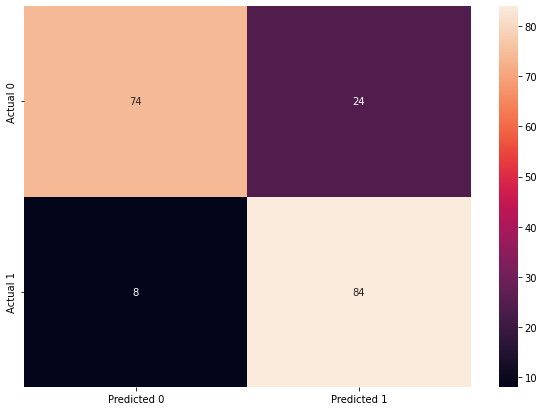
* Recall: Trong tất cả các trường hợp Positive, bao nhiêu trường hợp đã được dự đoán chính xác? Chỉ số này được tính theo công thức:



Confusion Matrix Code:



Kết quả hiển thị:



##### Hình 4.6: Confusion matrix của mô hình hồi quy logistic

Từ diễn họa bên trên có thể dễ hình dung những con số được dự đoán đúng sai. Với kết quả thực tế là không bị bệnh tim (0) nhưng lại dự đoán có bị (1) là 25, và ngược lại có bị bệnh tim (1) nhưng lại dự đoán là không bị (0) là 8.

Từ những kết quả bên trên có thể thấy, mặc dù độ chính xác (accuracy) ở mức khá cao, với tỉ lệ hơn 83%. Đây là một kết quả khá khả quan với các tham số mặc định, chưa được thay đổi.

# KẾT LUẬN VÀ ĐỊNH HƯỚNG PHÁT TRIỂN

## **Kết quả đạt được**

Qua quá trình làm đồ án, mang lại rất nhiều kinh nghiệm, kiến thức bổ ích cũng như có cơ hội áp dụng các kiến thức đã tích lũy được sau những năm tháng học tập tại trường Đại học Mỏ - Địa chất, trong đó phải kể đến các kiến thức nổi bật như: Xác suất & thống kê, lập trình Python, khoa học dữ liệu, ngôn ngữ R trong phân tích dữ liệu, học máy thống kê để hoàn thành đề tài tốt nghiệp ‘phân tích dữ liệu và một số kĩ thuật hồi quy trong bài toán dự đoán bệnh tim’ sử dụng python đã đạt được những kết quả như sau:

* Phân tích mô tả và phân tích song biến chi tiết, chính xác các thuộc tính.
* Nghiên cứu các thuật toán Machine Learning phù hợp với tập dữ liệu.
* Cải thiện kĩ năng phân tích dữ liệu bằng python
* Xây dựng các modul học máy với chức năng chính là dự đoán và tính độ chính xác của các thuật toán trên tập dữ liệu.

Từ những phân tích có thể thấy, ở độ tuổi nào cũng có thể mắc bệnh tim, và đặc biệt bệnh tim càng ngày càng trẻ hóa. Bên cạnh đó chỉ số Cholesterol cũng tương quan với tỉ lệ người mắc bệnh. Chúng ta nên có chế độ ăn uống và lối sống lành mạnh để hạn chế nguy cơ mắc bệnh tim mạch, bởi vì tim mạch không trừ độ tuổi hay bất kì giới tính nào.

Mặc dù đã rất cố gắng nhưng do trình độ cũng như hiểu biết của em còn hạn chế nên đồ án này không tránh khỏi những sai sót và hạn chế nhất định. Bên cạnh những mặt ưu cần phát huy thì còn một số hạn chế nhất định cần được khắc phục và hoàn thiện, em rất mong được sự đóng góp, chỉ bảo của thầy cô, bạn bè để em có thể biết được những thiếu sót và có thể hoàn thiện hơn.

## **Định hướng phát triển**

Trong tương lai gần, em sẽ tiếp tục hoàn thiện độ chính xác của mô hình đã xây dựng, và tìm những thuật toán, tham số tối ưu hơn cho tập dữ liệu bệnh tim. Em mong nhận được sự giúp đỡ quý báu của thầy cô và bạn bè để em có thể hoàn thiện và bổ sung đồ án tốt hơn để xây dựng hoàn thiện các chức năng cụ thể như:

* Xây dựng app/web để dự đoán.
* Áp dụng thuật toán phù hợp hơn.
* Tích lũy và áp dụng những kiến thức đã học được để trở thành nhà phân tích dữ liệu.
* Tiếp tục học tập/nghiên cứu sau đại học.

Cuối cùng, một lần nữa em xin gửi lời cảm ơn sâu sắc đến quý thầy cô trường đại học Mỏ - Địa Chất, bạn bè, các anh chị đi trước đã chỉ dạy cho em rất nhiều kiến thức bổ ích và đặc biệt là Thạc sĩ Vũ Lan Phương đã rất nhiệt tình hướng dẫn em hoàn thành đề tài này.

# TÀI LIỆU THAM KHẢO

***Tài liệu tiếng Việt***

[1] Vũ Hữu Tiệp, (2020), *Machine Learning cơ bản*

***Tài liệu tiếng Anh***

[1] Wes McKinney, (2017)*, Python for Data Analysis Data Wrangling with Pandas, NumPy, and IPython*

[2]Sebastian Raschka, Vahid Mirjalili, (2020), *Python Machine Learning - Second Edition: Machine Learning and Deep Learning with Python, scikit-learn, and TensorFlow 2nd Edition*

[3]Nussbaumer Knaflic, (2015), *Storytelling with Data: A Data Visualization Guide for Business Professionals*

[4] Charles Wheelan, (2013), *Naked Statistics: Stripping the Dread from the Data by Charles Wheelan*

[5]Bernard Marr, (2017), *Data Strategy: How to Profit from a World of Big Data, Analytics and the Internet of Things*

[6] [*https://realpython.com/logistic-regression-python/#logistic-regression-overview*](https://realpython.com/logistic-regression-python/#logistic-regression-overview)

[7][*https://en.wikipedia.org/wiki/Data\_analysis*](https://en.wikipedia.org/wiki/Data_analysis)