TỔNG LIÊN ĐOÀN LAO ĐỘNG VIỆT NAM

**TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG**

**KHOA CÔNG NGHỆ THÔNG TIN**



**LÝ TUẤN AN – 52000620**

**NGUYỄN THÀNH AN – 52000621**

**PHÙNG PHÚC HẬU – 52000443**

**PHÂN TÍCH VÀ HỖ TRỢ RA QUYẾT ĐỊNH TRONG DỰ ĐOÁN BỆNH TIM**

**BÁO CÁO CUỐI KỲ**

**HỆ THỐNG**

**THƯƠNG MẠI THÔNG MINH**

**THÀNH PHỐ HỒ CHÍ MINH, NĂM 2023**

TỔNG LIÊN ĐOÀN LAO ĐỘNG VIỆT NAM

**TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG**

**KHOA CÔNG NGHỆ THÔNG TIN**



**LÝ TUẤN AN – 52000620**

**NGUYỄN THÀNH AN – 52000621**

**PHÙNG PHÚC HẬU – 52000443**

**PHÂN TÍCH VÀ HỖ TRỢ RA QUYẾT ĐỊNH TRONG DỰ ĐOÁN BỆNH TIM**

**BÁO CÁO CUỐI KỲ**

**HỆ THỐNG**

**THƯƠNG MẠI THÔNG MINH**

Người hướng dẫn

**ThS. Dương Hữu Phúc**

**THÀNH PHỐ HỒ CHÍ MINH, NĂM 2023**

LỜI CẢM ƠN

Trong suốt thời gian qua, nhờ sự giảng dạy tận tâm của quý Thầy Cô khoa Công Nghệ Thông Tin, trường Đại học Tôn Đức Thắng, chúng em đã học hỏi được rất nhiều điều bổ ích và tích lũy cho mình một số kiến thức để hoàn thành bài báo cáo này. Nhóm em xin chân thành cảm ơn.

Nhóm em xin cảm ơn thầy Dương Hữu Phúc đã tận tình chỉ bảo chúng em qua những buổi học tại lớp, thầy đã chỉ nhóm em cách thức làm bài, chỉ điểm những chỗ còn sai sót chưa phù hợp cũng như phải làm sao để trình bày bố cục đẹp. Nếu không có những lời hướng dẫn, dạy bảo của thầy thì bài thu hoạch của nhóm em cũng rất khó để hoàn thiện. Một lần nữa chúng em xin chân thành cảm ơn thầy.

Bước đầu đi vào thực tế với nền kiến thức mở rộng, kiến thức của nhóm em còn hạn chế và nhiều bỡ ngỡ. Vì thế, trong quá trình biên soạn khó tránh những sai sót, nhóm em rất mong nhận được những ý kiến đóng góp quý báu của thầy và các bạn để bài báo cáo hoàn thiện hơn.

ĐỒ ÁN ĐƯỢC HOÀN THÀNH

TẠI TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG

Chúng tôi xin cam đoan đây là sản phẩm đồ án của riêng chúng tôi và được sự hướng dẫn của ThS Dương Hữu Phúc. Các nội dung nghiên cứu, kết quả trong đề tài này là trung thực và chưa công bố dưới bất kỳ hình thức nào trước đây. Những số liệu trong các bảng biểu phục vụ cho việc phân tích, nhận xét, đánh giá được chính tác giả thu thập từ các nguồn khác nhau có ghi rõ trong phần tài liệu tham khảo.

Ngoài ra, trong đồ án còn sử dụng một số nhận xét, đánh giá cũng như số liệu của các tác giả khác, cơ quan tổ chức khác đều có trích dẫn và chú thích nguồn gốc.

**Nếu phát hiện có bất kỳ sự gian lận nào tôi xin hoàn toàn chịu trách nhiệm về nội dung đồ án của mình.** Trường đại học Tôn Đức Thắng không liên quan đến những vi phạm tác quyền, bản quyền do tôi gây ra trong quá trình thực hiện (nếu có).

*TP. Hồ Chí Minh, ngày 05 tháng 01 năm 2024*

*Tác giả*

*(ký tên và ghi rõ họ tên)*

*An*

*Lý Tuấn An*

*An*

*Nguyễn Thành An*

*Hậu*

*Phùng Phúc Hậu*

PHẦN XÁC NHẬN VÀ ĐÁNH GIÁ CỦA GIẢNG VIÊN

**Phần xác nhận của GV hướng dẫn**

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Tp. Hồ Chí Minh, ngày tháng năm

(kí và ghi họ tên)

**Phần đánh giá của GV chấm bài**

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Tp. Hồ Chí Minh, ngày tháng năm

(kí và ghi họ tên)

MỤC LỤC

[LỜI CẢM ƠN 1](#_Toc154146904)

[PHẦN XÁC NHẬN VÀ ĐÁNH GIÁ CỦA GIẢNG VIÊN 3](#_Toc154146905)

[MỤC LỤC 4](#_Toc154146906)

[DANH MỤC KÍ HIỆU VÀ CHỮ VIẾT TẮT 5](#_Toc154146907)

[DANH MỤC CÁC BẢNG BIỂU VÀ HÌNH ẢNH 6](#_Toc154146908)

[CHƯƠNG I: TỔNG QUAN 8](#_Toc154146909)

[1.1 Giới thiệu đề tài 8](#_Toc154146910)

[1.2 Phương pháp nghiên cứu 9](#_Toc154146911)

[1.3 Kết quả dự kiến 10](#_Toc154146912)

[CHƯƠNG II: THU THẬP VÀ XỬ LÝ DATASET 11](#_Toc154146913)

[2.1 Xây dựng dataset 11](#_Toc154146914)

[2.1.1 Mô tả dataset 11](#_Toc154146915)

[2.1.2 Đặc tả dữ liệu 11](#_Toc154146916)

[2.1.3 Làm sạch dữ liệu 14](#_Toc154146917)

[2.2 Trực quan hóa 18](#_Toc154146918)

[2.3 Ứng dụng Machine Learning 24](#_Toc154146919)

[2.3.1 Đặc tả mô hình Machine Learning 24](#_Toc154146920)

[2.3.2 Cách triển khai mô hình Machine Learning 29](#_Toc154146921)

[TÀI LIỆU THAM KHẢO 36](#_Toc154146922)

# DANH MỤC KÍ HIỆU VÀ CHỮ VIẾT TẮT

ML - Machine Learning

DT - Dataset

LG - Logistic Regression

KNN - K-Nearest Neighbors

DTC - Decision Tree Classical

SVM - Support Vector Machine

DANH MỤC CÁC BẢNG BIỂU VÀ HÌNH ẢNH

[Bảng 2. 2 Đặc tả dữ liệu 13](#_Toc154090937)

[Hình 1. 1 Dự đoán bệnh tim bằng ML 10](#_Toc154090947)

[Hình 2. 1 Phát hiện có phần tử rỗng 14](#_Toc154090954)

[Hình 2. 2 Phát hiện có phần tử rỗng ở cột num\_major\_vessels 15](#_Toc154090955)

[Hình 2. 3 Phát hiện có phần tử rỗng 15](#_Toc154090956)

[Hình 2. 4 Phát hiện có phần tử rỗng ở cột thalassemia 15](#_Toc154090957)

[Hình 2. 5 Tổng kết phát hiện số lỗi ở bộ dữ liệu 15](#_Toc154090958)

[Hình 2. 6 Lỗi sai kiểu dữ liệu 16](#_Toc154090959)

[Hình 2. 7 Loại bỏ phần tử bị null 16](#_Toc154090960)

[Hình 2. 8 Tổng kết không còn xuất hiện phần tử null 16](#_Toc154090961)

[Hình 2. 9 Code chuyển kiểu float thành int 17](#_Toc154090962)

[Hình 2. 10 Kiểu dữ liệu của num major vessels 17](#_Toc154090963)

[Hình 2. 11 Dữ liệu sau khi đã làm sạch 17](#_Toc154090964)

[Hình 2. 12 Chuyển toàn bộ dữ liệu khác sang int 18](#_Toc154090965)

[Hình 2. 13 Quy định giá trị cho từng thuộc tính ở kiểu int 18](#_Toc154090966)

[Hình 2. 14 Bộ dữ liệu hoàn chỉnh 19](#_Toc154090967)

[Hình 2. 15 Disease Distributions 19](#_Toc154090968)

[Hình 2. 16 Disease Ratio of Categorical Feature 20](#_Toc154090969)

[Hình 2. 17 Disease Status of Categorical Features 20](#_Toc154090970)

[Hình 2. 18 Distributions of Numerical Features 22](#_Toc154090971)

[Hình 2. 19 Corr Scatter Plot 22](#_Toc154090972)

[Hình 2. 20 Corr Heat Map 23](#_Toc154090973)

[Hình 2. 21 Khai báo hàm sử dụng 29](#_Toc154090974)

[Hình 2. 22 Tạo dữ liệu train và test 29](#_Toc154090975)

[Hình 2. 23 Code tính các chỉ số accuracy, precision, recall, f1-score cho các mô hình 29](#_Toc154090976)

[Hình 2. 24 Biểu đồ hiển thị 30](#_Toc154090977)

[Hình 2. 25 StandardScaler và SDA bộ dữ liệu để model SVM và KNN 30](#_Toc154090978)

[Hình 2. 26 Biểu đồ hiển thị sau khi cải thiện 30](#_Toc154090979)

[Hình 2. 27 Sử dụng GridSearchCV để tìm ra Hyperparameters tốt nhất. 31](#_Toc154090980)

[Hình 2. 28 Dùng Hyperparameters tốt nhất để tính các chỉ số trong model 31](#_Toc154090981)

[Hình 2. 29 Biểu diễn biểu đồ với mô hình LR 32](#_Toc154090982)

[Hình 2. 30 Tính các chỉ số trong model dựa theo hàm cross\_val\_score 32](#_Toc154090983)

[Hình 2. 31 Kết quả 33](#_Toc154090984)

[Hình 2. 32 Hiển thị bảng classification\_report 33](#_Toc154090985)

[Hình 2. 33 Hiển thị confusion\_matrix 33](#_Toc154090986)

[Hình 2. 34 Code dự đoán của model Decision Tree 34](#_Toc154090987)

[Hình 2. 35 Code dự đoán của model Logistic Regresssion 34](#_Toc154090988)

[Hình 2. 36 Code dự đoán của model KNN 34](#_Toc154090989)

[Hình 2. 37 Code dự đoán của model SVC 35](#_Toc154090990)

[Hình 2. 38 Code dự đoán của model Naïve Bayes 35](#_Toc154090991)

[Hình 2. 39 Test dự đoán đối với từng mô hình ML 35](#_Toc154090992)

CHƯƠNG I: TỔNG QUAN

* 1. **Giới thiệu đề tài**
* Trong số tất cả các cơ quan, trái tim là một bộ phận quan trọng của cơ thể chúng ta. Tim đập khoảng 2,5 tỷ lần trong suốt cuộc đời trung bình, đẩy hàng triệu gallon máu đến mọi bộ phận của cơ thể.
* Trong thời đại này, bệnh tim mạch ngày càng gia tăng do lối sống và thực phẩm hiện đại. Việc chẩn đoán bệnh tim là một nhiệm vụ đầy thách thức. Mô hình phân loại này sẽ dự đoán bệnh nhân có mắc bệnh tim hay không dựa trên các tình trạng/triệu chứng khác nhau của cơ thể họ.
* Đề tài sử dụng học máy để dự đoán bệnh tim mạch dựa trên các chỉ số cơ bản của bệnh nhân là một chủ đề quan trọng trong lĩnh vực y học và có thể đem lại những ứng dụng thực tiễn đáng kể. Các chỉ số cơ bản mà bạn có thể sử dụng để dự đoán bệnh tim mạch có thể bao gồm:
  + **Dữ liệu Lâm sàng cơ bản:** Như huyết áp, mức độ đường huyết, cholesterol, hàm lượng triglycerides, hành vi hút thuốc lá, BMI (Chỉ số khối cơ thể), và lịch sử bệnh lý.
  + **Chỉ số EKG (Electrocardiogram):** Các thông số từ EKG như nhịp tim, hình dạng sóng, và bất thường trong sóng điện có thể được sử dụng để dự đoán bệnh tim mạch.
  + **Thông tin về tuổi, giới tính và di truyền:** Các yếu tố này có thể đóng vai trò quan trọng trong việc dự đoán nguy cơ bệnh tim mạch.
  + **Lịch sử y tế:** Bao gồm lịch sử bệnh tim mạch trong gia đình, các vấn đề sức khỏe đã từng có và các phác đồ điều trị trước đó.
* Cách tiếp cận thông thường là thu thập dữ liệu từ các bệnh viện hoặc cơ sở y tế, sau đó sử dụng các thuật toán học máy như Logistic Regression, Random Forest, Support Vector Machines (SVM), hoặc Neural Networks để phân loại bệnh nhân thành các nhóm có nguy cơ bệnh tim mạch cao và thấp dựa trên thông tin này.
* Một số đặc điểm quan trọng sau cần lưu ý:
  + **Tiền xử lý dữ liệu:** Đôi khi, dữ liệu có thể không đầy đủ hoặc có nhiễu, việc xử lý dữ liệu để loại bỏ giá trị ngoại lai (outliers), điền các giá trị thiếu (missing values), và chuẩn hóa dữ liệu có thể là cần thiết.
  + **Đánh giá mô hình:** Để đảm bảo mô hình học máy là hiệu quả và có thể áp dụng trong thực tế, cần phải thực hiện các phương pháp đánh giá như cross-validation, ROC curves, và confusion matrices để đo lường hiệu suất của mô hình.
  + **Phân tích diễn giải:** Việc hiểu và diễn giải các biến quan trọng trong mô hình có thể giúp cải thiện hiểu biết về mối quan hệ giữa các yếu tố và bệnh tim mạch.
  + **Bảo vệ quyền riêng tư:** Trong việc sử dụng dữ liệu y tế, việc bảo vệ thông tin cá nhân và tuân thủ các quy định về quyền riêng tư là rất quan trọng.
  1. **Phương pháp nghiên cứu**
* Xác định mục tiêu nghiên cứu:
* Phân tích hành vi tiêu dùng: Thu thập dữ liệu về những chỉ số/tình trạng gặp phải trên cơ thể người, bao gồm giới tính, đường huyết, huyết áp, cholesterol, đau thắt ngực, …
* Sử dụng Tableau để trực quan hóa dữ liệu, xác định mối liên hệ giữa các chỉ số và khả năng chuẩn đoán bệnh tim.
* Thu thập dữ liệu:
* Lựa chọn mẫu nghiên cứu: Tập trung vào một nhóm người tham gia nghiên cứu có đủ thông tin về các chỉ số cơ thể và lịch sử bệnh lý về tim mạch.
* Thu thập chỉ số cơ thể: Thu thập dữ liệu về huyết áp, nhịp tim, cholesterol, BMI, glucose level, và các chỉ số cơ thể khác thông qua thiết bị y tế hoặc các phương pháp đo lường chuẩn.
* Xử lý và phân tích dữ liệu:
* Xử lý dữ liệu: Kiểm tra và xử lý dữ liệu thiếu sót, loại bỏ các giá trị ngoại lệ, chuẩn hóa dữ liệu nếu cần.
* Trực quan hóa:
* Import dữ liệu vào Tableau và tạo các biểu đồ trực quan hóa để khám phá mối quan hệ giữa các chỉ số cơ thể và bệnh tim.
* Tạo biểu đồ dạng scatterplot để thấy mối tương quan giữa các chỉ số.
* Áp dụng các kỹ thuật trực quan hóa như heatmap, biểu đồ đường, biểu đồ cột để so sánh sự biến đổi của các chỉ số theo thời gian hoặc theo nhóm người khác nhau.
* Phân tích và kết luận:
  + Phân tích kết quá: Phân tích biểu đồ và số liệu để nhận diện các mẫu, xu hướng, mối liên hệ giữa các chỉ số cơ thể và bệnh tim.
  + Kết luận: Đưa ra các kết luận hoặc dự đoán về khả năng chuẩn đoán bệnh tim dựa trên dữ liệu và mối liên hệ phát hiện được.
  + Lập đề xuất: Dựa trên kết quả, đề xuất các phương pháp chuẩn đoán hoặc quản lý bệnh tim dựa trên thông tin từ các chỉ số cơ thể.
  1. **Kết quả dự kiến**
* Cung cấp thông tin hữu ích và nhận thức sâu sắc về mối quan hệ giữa chỉ số cơ thể và bệnh tim, hỗ trợ việc chuẩn đoán và điều trị bệnh tim.
* Đóng góp vào lĩnh vực nghiên cứu y học về khả năng sử dụng dữ liệu chỉ số cơ thể để dự đoán bệnh tim.



Hình 1. Dự đoán bệnh tim bằng ML

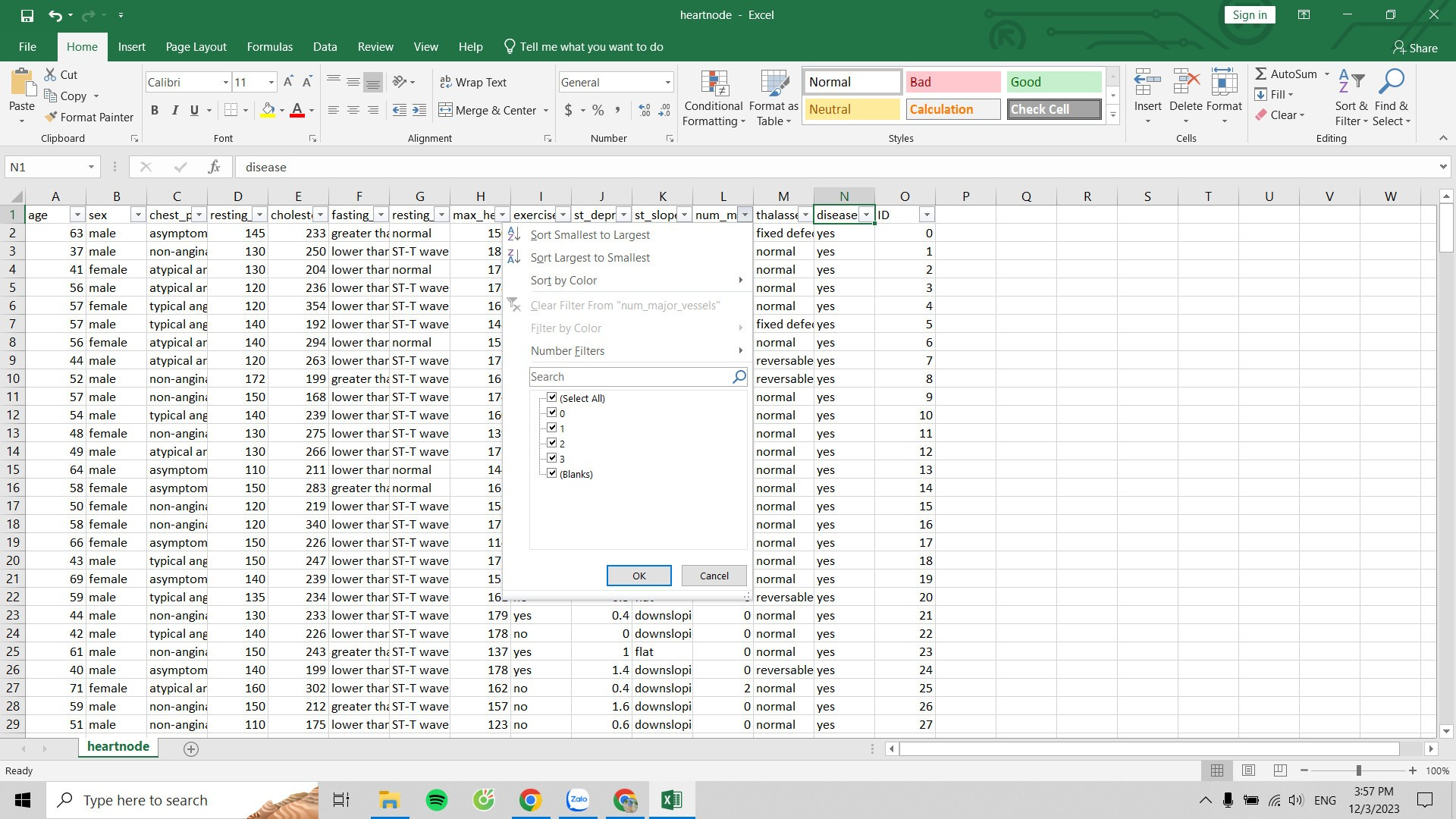
CHƯƠNG II: THU THẬP VÀ XỬ LÝ DATASET

* 1. Xây dựng dataset
     1. Mô tả dataset
* Bộ dữ liệu cung cấp thông tin đa dạng về chỉ số cơ thể và yếu tố rủi ro để đánh giá nguy cơ bệnh tim của một người.
* Việc sử dụng các công cụ như Tableau để trực quan hóa và phân tích dataset này có thể giúp xác định mối liên hệ giữa các chỉ số và khả năng dự đoán bệnh tim.
* Một thông tin cơ bản như: Giới tính, huyết áp, nhịp tim, mức cholesterol, chỉ số cơ thể, tiền sử bệnh án, thói quen, …Top of Form
  + 1. Đặc tả dữ liệu

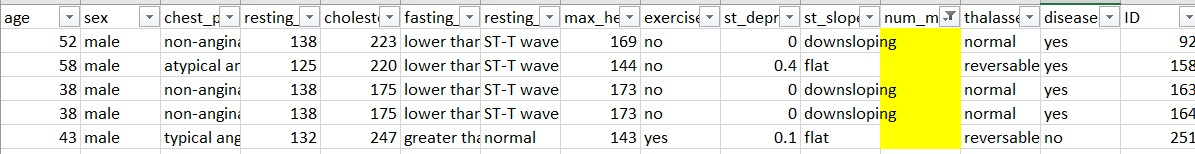
|  |  |  |
| --- | --- | --- |
| **STT** | **Tập dữ liệu** | **Mô tả** |
| 1 | age | Độ tuổi bệnh nhân (tính bằng năm) |
| 2 | gender | Giới tính bệnh nhân: nứ (0) - nam (1) |
| 3 | cp (chest pain type) | Loại đau ngực:  - typical angina – đau ngực điển hình (0): Đau ngực liên quán đến giảm cung cấp máu cho tim  - atypical angina – đau ngực không điển hình (1): Đau ngực không liên quan đến tim  - non-anginal pain – đau ngực không phải do tim: Thường là do co thắt dạ dày, không liên quan đến tim (2)  - asymptomatic – không có triệu trứng đau ngực (3): Không có dấu hiệu của bệnh |
| 4 | resting\_blood\_pressure | Huyết áp nghỉ:  - Là huyết áp của bệnh nhân, được đo bằng mm Hg khi nhập viện  - Giá trị trên 130 – 140 thường là nguyên nhân lo lắng |
| 5 | cholesterol | Cholesterol huyết thanh:  - Là lượng cholesterol trong huyết thanh, được tính bằng mg/dl  - Giá trị trên 200 thường là nguyên nhân lo lắng |
| 6 | fasting\_blood\_sugar | Đường huyết nhanh:  - Đường huyết nhanh > 120 ml/dl – Đúng (1)  - Đường huyết nhanh <= 120 ml/dl – Sai (0)  - Giá trị ‘> 126 m’ g/dl là dấu hiệu của bệnh tiểu đường |
| 7 | resting\_electrocardiogram | Điện tâm đồ nghỉ (Biến thiên của dòng điện do tim phát ra khi nghỉ ngơi):  - Bình thường (0)  - Biến động song ST-T – Dấu hiệu của nhịp tim không bình thường (1)  - Có thể hoặc chắc chắn có phì độ thất trái – tăng thể tích tim trái theo tiêu chí của Estes (2) |
| 8 | max\_heart\_rate\_achieved | Nhịp tim tối đa đạt được trong quá trình kiểm tra |
| 9 | exercise\_induced\_angina | Đau thắt ngực do tập thể dục:  - yes (1)  - no (0) |
| 10 | st\_depression | ST depression tạo ra bởi tập thể dục so với nghỉ ngơi:  - Đo lượng sự giảm độ lệch ST tạo ra bởi tập thể dục so với thời gian nghỉ.  - Nhìn vào sự căng trải qua thời gian thể dục, với tim không khỏe sẽ trải qua căng càng nhiều. |
| 11 | st\_slope | Độ dốc của đoạn ST khi tập luyện mạnh:  - upsloping – Tăng lên: Tốt hơn với nhiệp tim khi tập thể dục – Hiếm gặp (0)  - flatsloping – Bằng phẳng: Thay đổi tối thiểu – Thường là tim khỏe mạnh (1)  - downsloping – Giảm xuống: Dấu hiệu của tim không khỏe mạnh (2) |
| 12 | num\_major\_vessels | Số lượng mạch chính (0-3) được nhuộm màu bằng fluoroscopy:  - Số lượng mạch máu chính được nhuộm bằng fluoroscopy, từ 0 – 3  - Mạch máu nhuộm được xem như mạch máu đang di chuyển mà không có cục máu động |
| 13 | thalassemia | Chỉ số rối loạn máu di truyển (Kết quả thử Thallium Stress):  - Fixed defect – Khuyết tật cố định: Từng có khuyết tật nhưng hiện tại đã ổn định (0)  - Normal – Bình thường: Không có vấn đề với sự di chuyển của máu (1)  - Reversible defect – Khuyết tật có thể đảo ngược: Sự di chuyển của máu không bình thường khi tập thể dục (2) |
| 14 | disease | Chuẩn đoán:  - yes – Mắc bệnh tim mạch (1)  - no – Không mắc bệnh (0) |

Bảng 2. Đặc tả dữ liệu

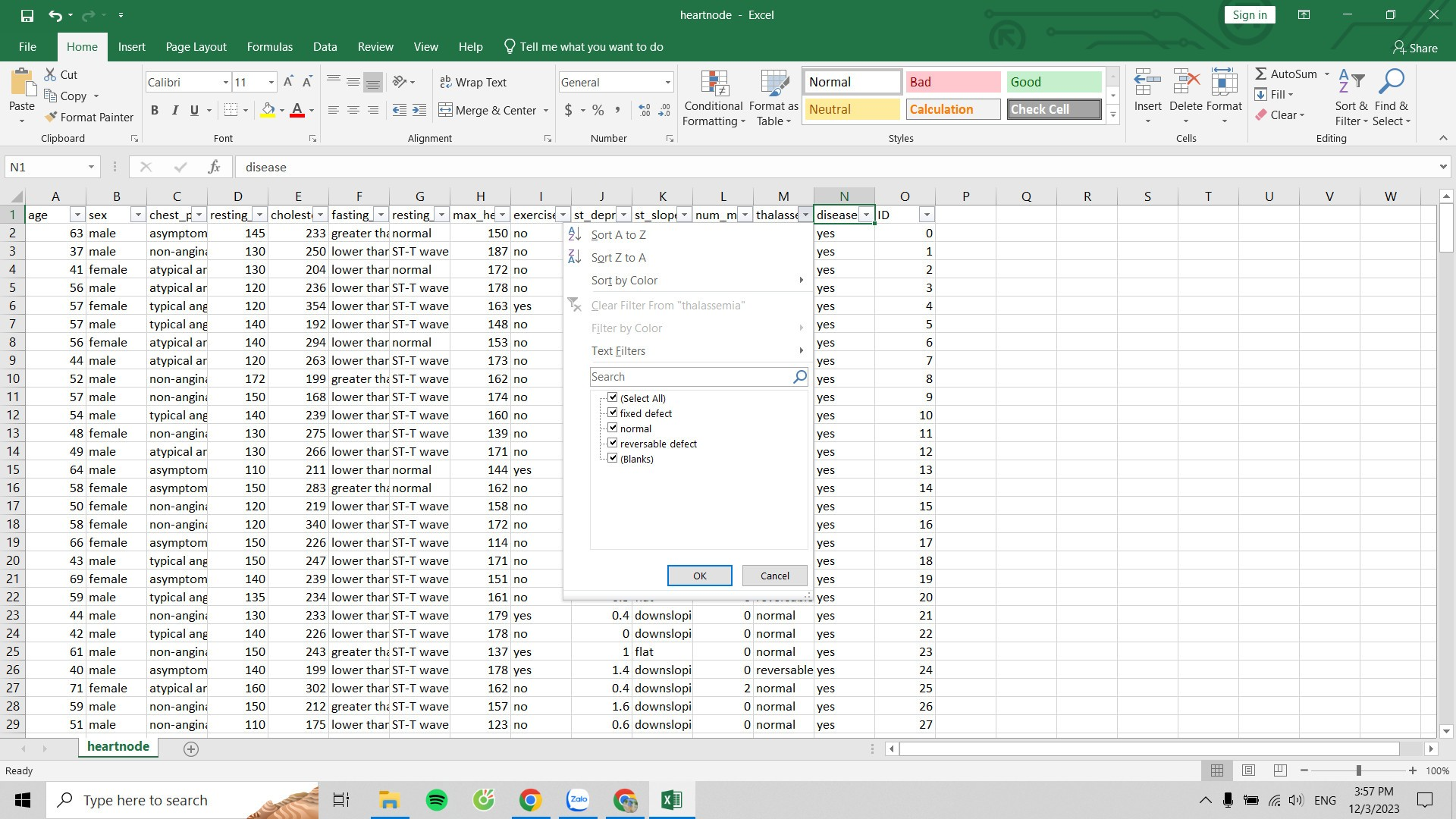
* Dưới đây là giải thích chi tiết hơn về các thuộc tính liên quan đến ST, ST-T, số mạch máu chính được nhuộm bằng fluoroscopy, và kết quả thử thallium stress:
  + ST và ST-T:
    - Trong điện tâm đồ (ECG), ST và ST-T là phần của chu kỳ điện tim được theo dõi để đánh giá hoạt động điện của tim.
    - ST Segment (Đoạn ST): Đoạn này bắt đầu từ kết thúc sóng Q và kéo dài đến đầu sóng T. Sự thay đổi của đoạn ST thường được quan sát để đánh giá sự cung cấp máu và khả năng hoạt động của tim.
    - ST-T Wave: Là sóng T kết hợp với đoạn ST. Sự thay đổi của ST-T thường được sử dụng để chẩn đoán các vấn đề về tim, đặc biệt là trong trường hợp có thể xuất hiện những biến động không bình thường.
  + Số mạch máu chính (0 - 3) được nhuộm màu bằng Fluoroscopy:
    - Đây là một ước lượng về số lượng mạch máu chính (các động mạch lớn cung cấp máu đến tim) được nhuộm bằng fluoroscopy.
    - Fluoroscopy là một phương pháp hình ảnh y khoa sử dụng tia X để theo dõi chuyển động và dòng chảy của chất nhuộm trong các mạch máu.
    - Số mạch máu nhuộm được thường được xem là một chỉ số của sự di chuyển của máu trong tim. Động mạch nhuộm được đánh giá lành mạnh nếu có nhiều mạch máu được nhuộm.
  + Kết quả thử Thallim Stress:
    - Là một phương pháp thử nghiệm thực hiện trong điện tâm đồ để đánh giá cách tim hoạt động khi được đặt trong tình trạng căng thẳng.
    - Fixed Defect (Khuyết Tật Cố Định): Kết quả 1 chỉ ra một khuyết tật đã cố định, nghĩa là có một khu vực trong tim mà không có sự di chuyển của máu và vẫn giữ nguyên tình trạng này sau khi tăng cường tập thể dục.
    - Normal (Bình Thường): Kết quả 2 cho thấy không có vấn đề đặc biệt với sự di chuyển của máu trong tim khi tập thể dục.
    - Reversible Defect (Khuyết Tật Có Thể Đảo Ngược): Kết quả 3 chỉ ra sự di chuyển của máu không bình thường khi tập thể dục, và tình trạng này có thể đảo ngược sau khi dừng tập thể dục.
* Những thông tin này cung cấp hiểu biết sâu sắc hơn về sự hoạt động của tim và các vấn đề tiềm ẩn mà có thể được đánh giá thông qua các kết quả điện tâm đồ và thử thallium stress.
  + 1. Làm sạch dữ liệu
* Những điểm bất thường ở bộ dữ liệu ban đầu:
  + Phát hiện có phần tử rỗng ở cột num\_major\_vessels và cột thalassemia:



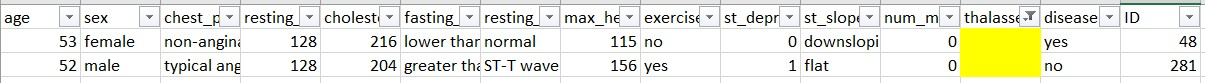
Hình 2. Phát hiện có phần tử rỗng



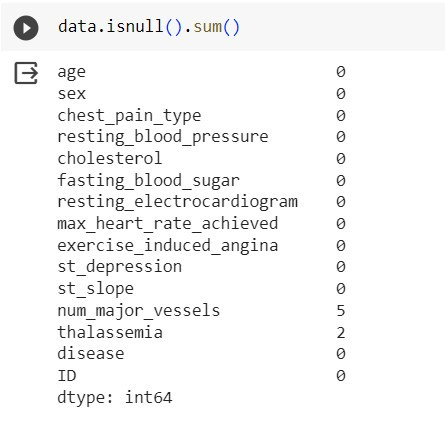
Hình 2. Phát hiện có phần tử rỗng ở cột num\_major\_vessels



Hình 2. Phát hiện có phần tử rỗng

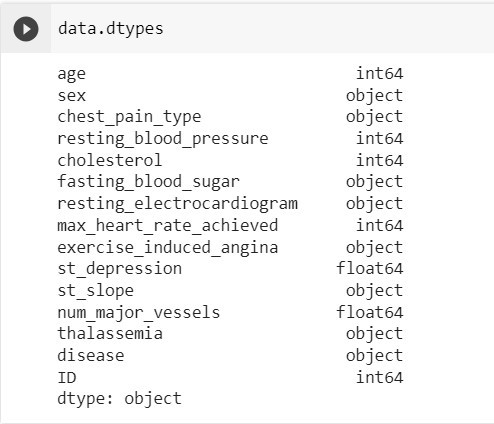


Hình 2. Phát hiện có phần tử rỗng ở cột thalassemia



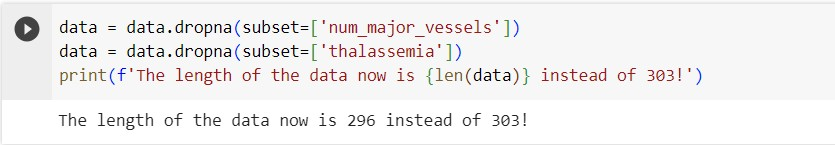
Hình 2. Tổng kết phát hiện số lỗi ở bộ dữ liệu

* + Kiểu dữ liệu ở cột num\_major\_vessels đang là kiểu float trong khi nó chỉ cần để kiểu int.

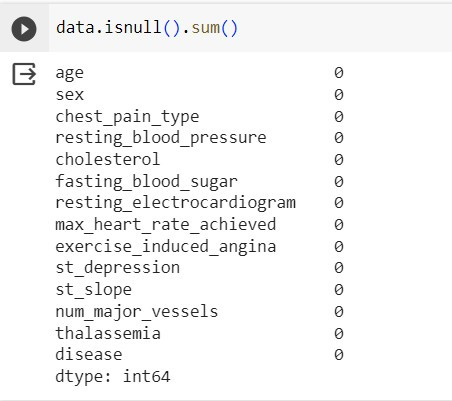


Hình 2. Lỗi sai kiểu dữ liệu

* Làm sạch dữ liệu:
  + Loại bỏ những phần tử bị null ở 2 cột.

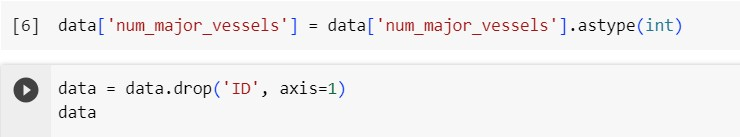


Hình 2. Loại bỏ phần tử bị null

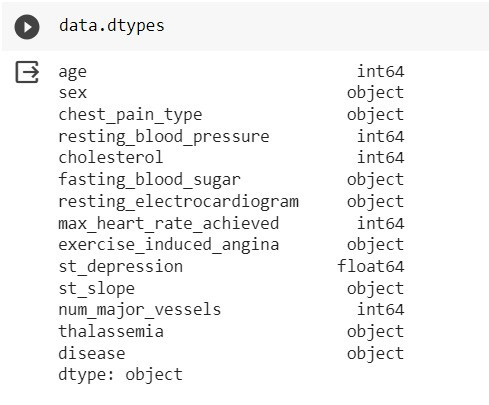


Hình 2. Tổng kết không còn xuất hiện phần tử null

* + Chuyển kiểu float ở cột num\_major\_vessels thành kiểu int.

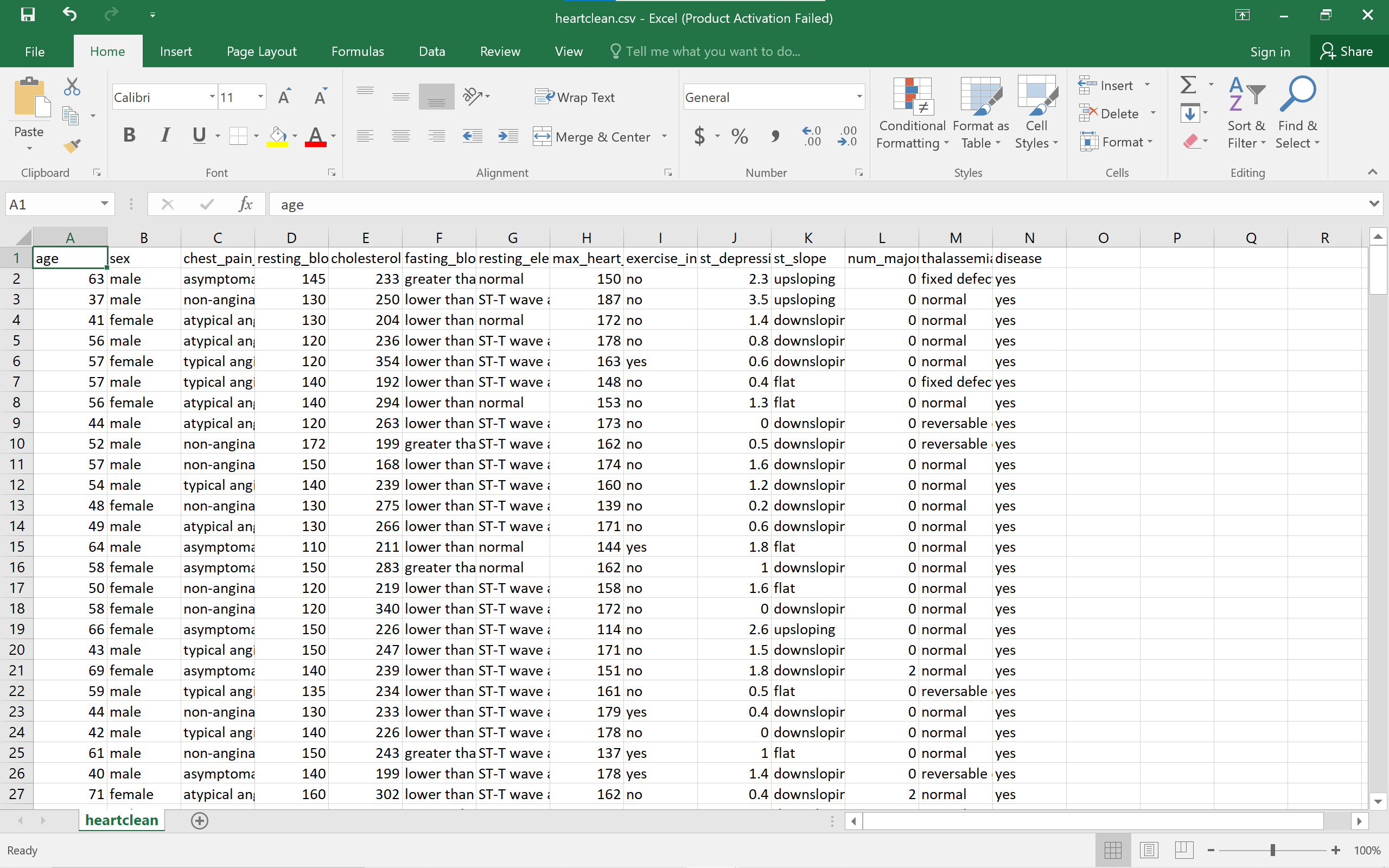


Hình 2. Code chuyển kiểu float thành int



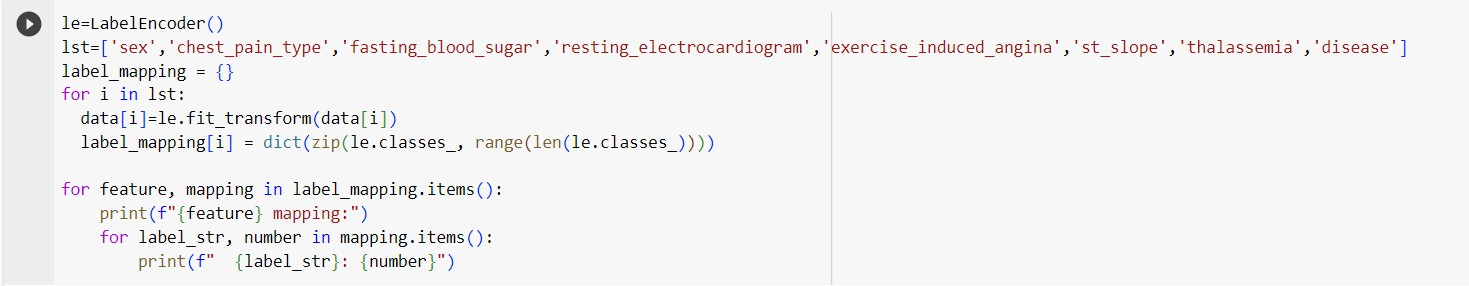
Hình 2. Kiểu dữ liệu của num major vessels

* Sau khi làm sạch dữ liệu thì ta đươc bộ dataset như sau:

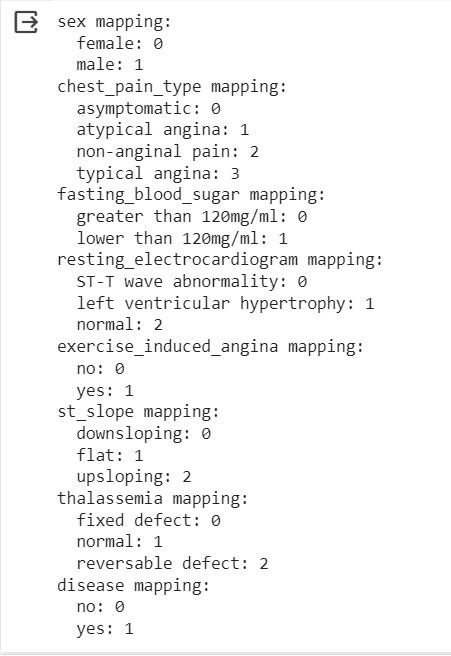


Hình 2. Dữ liệu sau khi đã làm sạch

* 1. Trực quan hóa
* Encoder label: Để thực hiện trực quan hóa thì tiến hành thay đổi các phần tử từ kiểu chuỗi sang kiểu int.

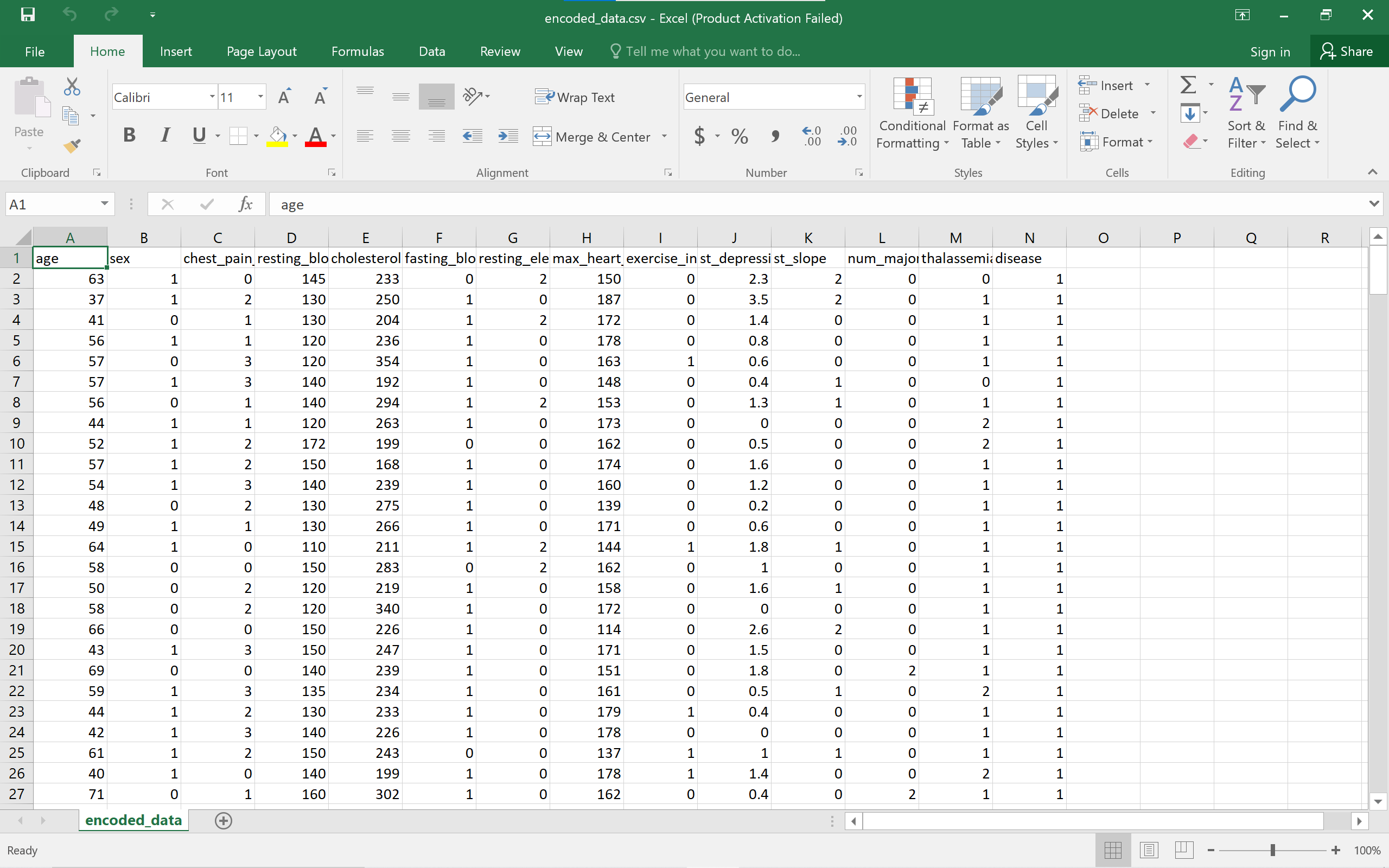


Hình 2. Chuyển toàn bộ dữ liệu khác sang int



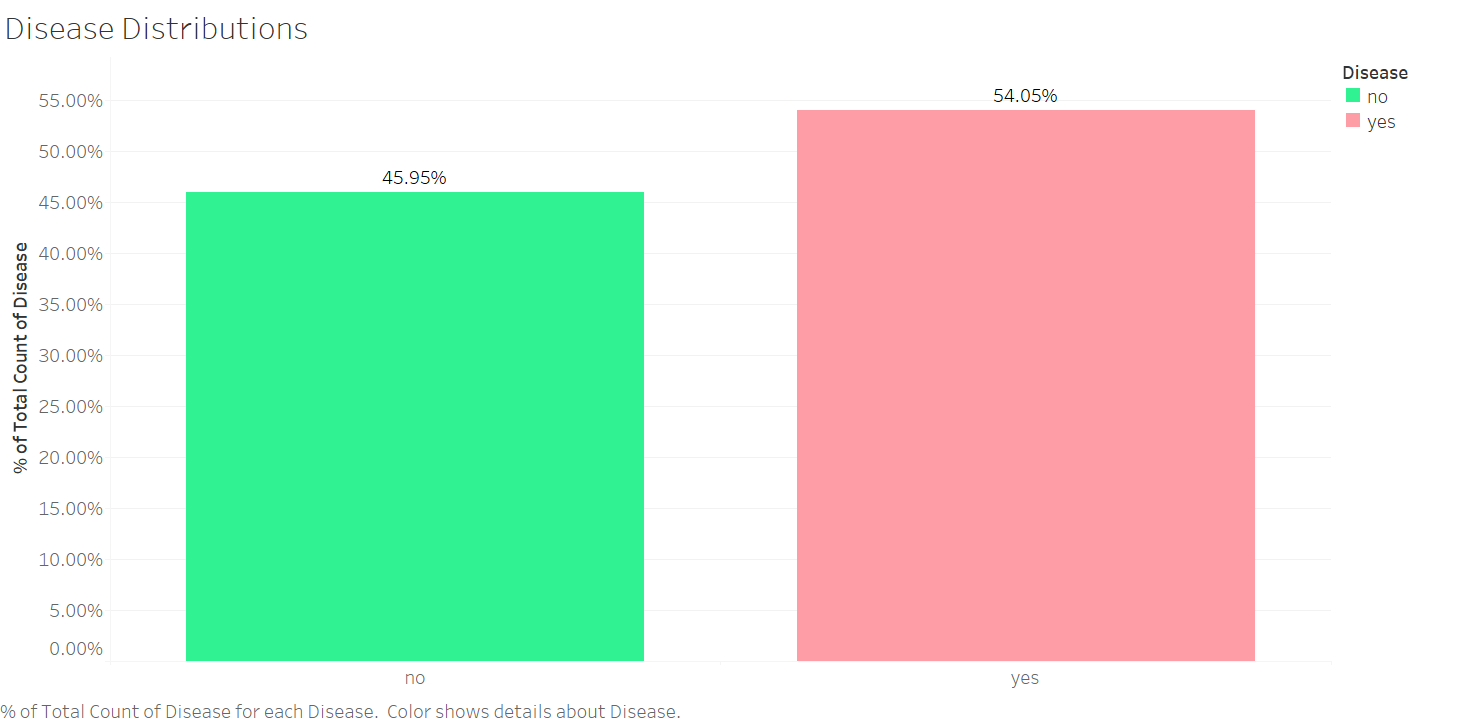
Hình 2. Quy định giá trị cho từng thuộc tính ở kiểu int

* Sau cùng thì ta được bộ dữ liệu như sau:



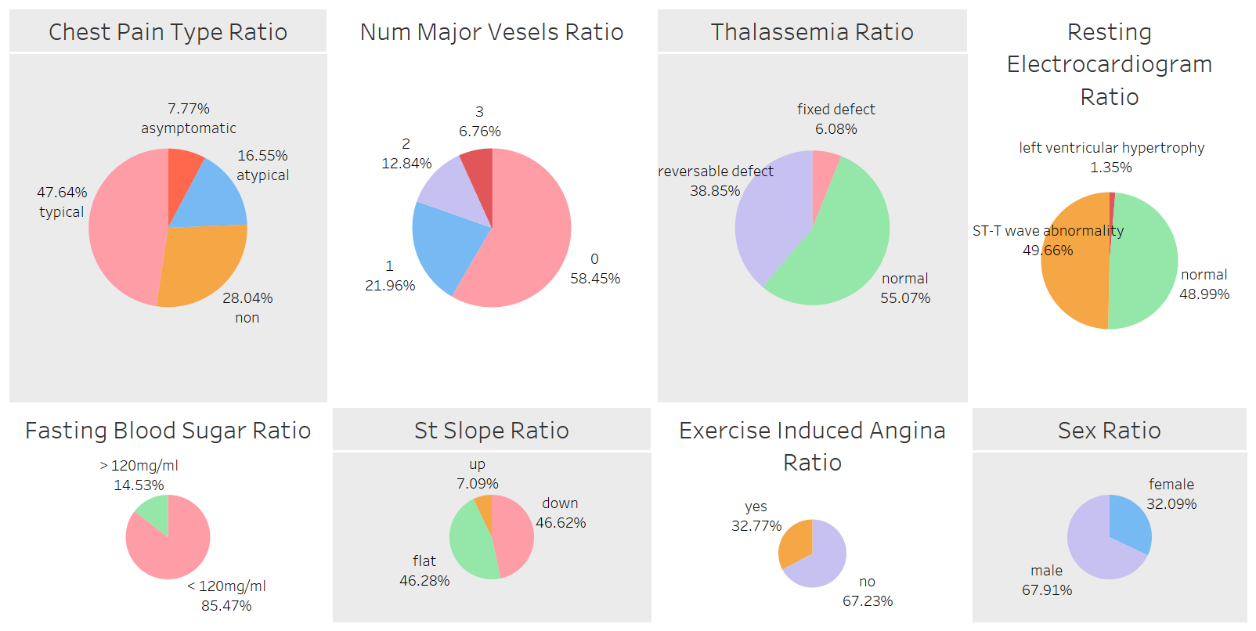
Hình 2. Bộ dữ liệu hoàn chỉnh

* Disease Distributions



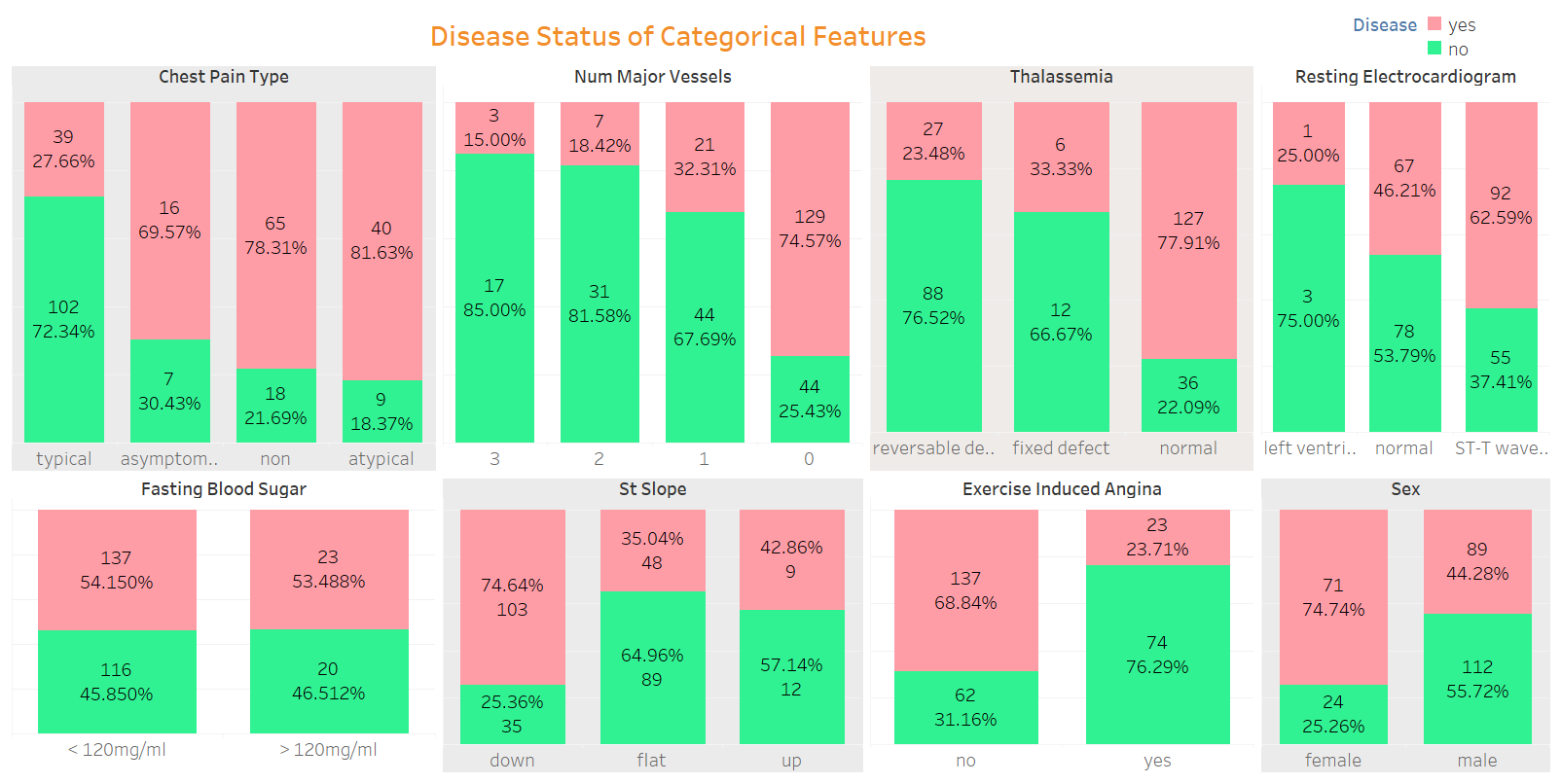
Hình 2. Disease Distributions

* Categorical Feature:
  + Disease Ratio of Categorical Feature



Hình 2. Disease Ratio of Categorical Feature

* + Disease Status of Categorical Features



Hình 2. Disease Status of Categorical Features

* + Tổng quan:
    - Chest Pain: Hơn 75% bệnh nhân bị đau ngực điển hình (typical angina) và đau ngực không phải do tim (non-angina pain) nên họ có nguy cơ mắc bênh tim.
    - Fasting Blood Sugar: Bệnh nhân có lượng đường trong máu lúc đói thấp hơn (dưới 120 mg/ml) chiếm đa số trong tập dữ liệu của chúng tôi bao gồm ~ 85% mẫu. Lượng đường trong máu khi nghỉ ngơi thấp hơn có xu hướng làm tăng nguy cơ mắc bệnh tim (~54%).
    - ST Slope: Hầu hết bệnh nhân đều có ST-Slope trong xét nghiệm REC của họ là dốc xuống (downsloping) hoặc bằng phẳng (flat). Độ dốc ST dốc xuống (downsloping) là dấu hiệu rõ ràng cho thấy bệnh nhân có thể mắc bệnh tim.
    - Thalassemia: Hầu hết bệnh nhân có khiếm khuyết bình thường (normal) hoặc có thể đảo ngược (reversible defect). Những bệnh nhân có khuyết tật bệnh thalassemia (có thể reversible + fixed) ít có khả năng mắc bệnh tim hơn. Trong khi đó, những người khiếm khuyết bình thường (normal) có nhiều khả năng mắc bệnh tim hơn.
    - Exercise Include Angina: Hai phần ba số bệnh nhân không có biểu hiện đau thắt ngực do tập thể dục. 76% bệnh nhân bị đau thắt ngực do tập thể dục không có bệnh tim. Trong khi đó ~ 69% bệnh nhân không bị đau thắt ngực do tập thể dục được chẩn đoán có nguy cơ mắc bệnh tim.
    - Resting Electrocardiogram: Bệnh nhân bị phì độ thất trái là ít nhất (~1,4%). Phần còn lại gần như là sự phân chia 50 - 50 giữa bệnh nhân có ST-T bất thường và những người có xét nghiệm REC bình thường. Phần lớn những bệnh nhân thực hiện loại xét nghiệm này đều có nguy cơ mắc bệnh tim.
    - Sex: Đa phần bệnh nhân trong dữ liệu là nam gới. Tuy nhiên, nữ giới lại có nguy cơ mắc bệnh tim cao hơn.
* Numerical Features:
  + Distributions of Numerical Features



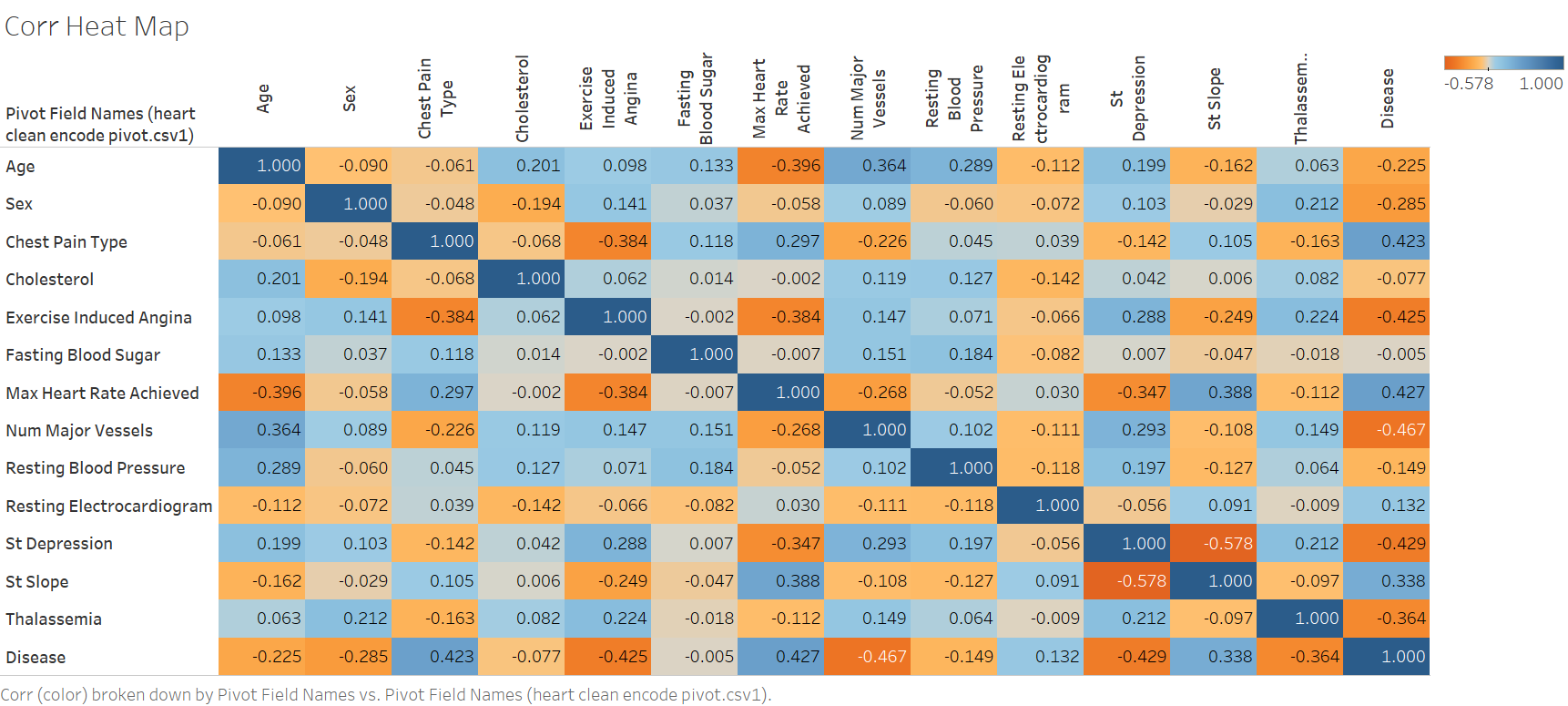
Hình 2. Distributions of Numerical Features

* + Corr Scatter Plot



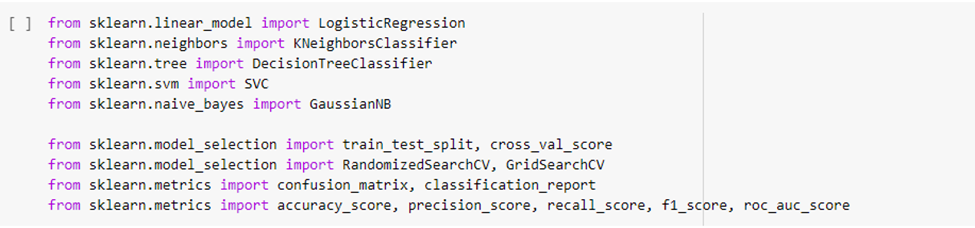
Hình 2. Corr Scatter Plot

* + Corr Heat Map



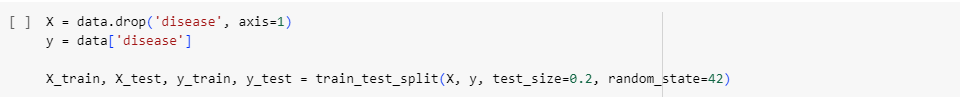
Hình 2. Corr Heat Map

* + Tổng quan:
    - Age: Độ tuổi trung bình ở trong tập dữ liệu sẽ là 54.5 tuổi, người lớn nhất là 77 tuổi và trẻ nhất là 29 tuổi.
    - Cholesterol: Mức cholesterol trung bình là 247.15, mức cao nhất là 564 và thấp nhất là 126 (Thường mức cholesterol khỏe mạnh là < 200mg/dl và ngược lại thì sẽ có liên quan đến bệnh tim)
    - Resting blood pressure: Mức huyết áp nghỉ trung bình là 132, mức cao nhất là 200 và thập nhất là 94.
    - Max heart rate achieved: Nhịp tim tối đa trung bình đạt 149.5 bpm, mức cao nhất đạt được là 200 bpm và thấp nhất đạt được là 71 bpm.
    - ST\_depression: Giá trị trung bình của ST\_depression là 1.06, mức tối đa đạt là 6.2 và tối thiểu là 0.
  1. Ứng dụng Machine Learning
     1. Đặc tả mô hình Machine Learning
* Bộ dữ liệu về bệnh tim là một bộ dữ liệu được sử dụng để giải quyết bài toán phân loại.
* Mục tiêu chính của bài toán này là dự đoán xem một người có rủi ro cao mắc bệnh tim hay không dựa trên các đặc trưng y tế và lâm sàng của họ. Vì nó là bài toán phân loại nên ta sẽ chọn các mô hình ML sau để triển khai:
  + LogisticRegression:
    - Lý do chọn: Logistic Regression là một lựa chọn phổ biến cho bài toán phân loại nhị phân. Nó dễ hiểu, có thể giải thích được và thường hiệu quả trong các trường hợp khi có mối quan hệ tuyến tính giữa đặc trưng và kết quả.
    - Ưu điểm:
      * Dễ hiểu và dễ triển khai: Logistic Regression là một mô hình đơn giản và dễ hiểu. Nó không yêu cầu nhiều tài nguyên tính toán so với các mô hình phức tạp hơn như Neural Networks.
      * Tính diễn giải cao: Kết quả của mô hình Logistic Regression có thể dễ dàng diễn giải. Bạn có thể hiểu được tác động của từng đặc trưng đến dự đoán xác suất.
      * Ít yêu cầu về dữ liệu: Không cần nhiều điều kiện cho dữ liệu. Logistic Regression hoạt động tốt trong những tình huống dữ liệu ít và các đặc trưng không phức tạp.
    - Nhược điểm:
      * Giả định tuyến tính: Logistic Regression giả định mối quan hệ giữa biến độc lập và xác suất là tuyến tính. Điều này có thể là hạn chế nếu mối quan hệ thực sự không phải là tuyến tính.
      * Dễ bị ảnh hưởng bởi nhiễu: Nếu dữ liệu chứa nhiều nhiễu hoặc outliers, mô hình Logistic Regression có thể bị ảnh hưởng và không đưa ra dự đoán chính xác.
      * Không xử lý được các tương tác phức tạp giữa đặc trứng: Logis Regression không thể mô hình hóa tương tác phức tạp giữa các đặc trưng.
  + KNeighborsClassifier:
    - Lý do chọn: KNN là một mô hình không parametric đơn giản và có thể hoạt động tốt với dữ liệu có cấu trúc đơn giản. việc sử dụng KNN có thể đem lại kết quả tốt nếu dữ liệu không quá lớn và không có quá nhiều nhiễu.
    - Ưu điểm:
      * Dễ hiểu và triển khai: KNN là một trong những mô hình đơn giản và dễ hiểu nhất trong Machine Learning. Nó không yêu cầu giả định về phân phối của dữ liệu và có thể được triển khai dễ dàng.
      * Khả năng áp dụng vào dữ liệu phức tạp: KNN hoạt động tốt trên dữ liệu có cấu trúc phức tạp và có thể xử lý các đặc trưng không tuyến tính.
      * Khả năng dự đoán linh hoạt: KNN không huấn luyện một mô hình cụ thể, mà chỉ lưu trữ dữ liệu huấn luyện. Điều này có nghĩa là khi có dữ liệu mới, việc dự đoán có thể được thực hiện một cách linh hoạt và không cần phải huấn luyện lại mô hình.
    - Nhược điểm:
      * Yêu cầu nhiều tài nguyên khi dữ liệu lớn: KNN cần lưu trữ toàn bộ tập dữ liệu huấn luyện trong bộ nhớ. Điều này có thể trở nên rất tốn kém về tài nguyên nếu tập dữ liệu lớn.
      * Nhảy cảm với dữ liệu nhiễu và không đồng nhất: KNN có thể không hiệu quả khi dữ liệu chứa nhiễu hoặc không đồng nhất. Các điểm dữ liệu "gần" nhau thực tế có thể không phản ánh chính xác về tính chất của dữ liệu.
      * Yêu cầu quá trình chuẩn hóa dữ liệu: Do KNN dựa vào khoảng cách giữa các điểm dữ liệu, việc chuẩn hóa dữ liệu (đảm bảo các đặc trưng có cùng phạm vi giá trị) có thể cần thiết để mô hình hoạt động tốt.
  + DecisionTreeClassifier:
    - Lý do chọn: Decision Tree có thể xử lý tốt cả với dữ liệu có mối quan hệ tuyến tính và không tuyến tính. Ngoài ra, chúng cũng có khả năng xử lý tương tác giữa các đặc trưng.
    - Ưu điểm:
      * Dễ hiểu và diễn giải: Decision Tree rất dễ diễn giải, có thể được trực quan hóa. Bạn có thể dễ dàng giải thích cho người không chuyên về machine learning về cách mô hình ra quyết định dựa trên các thuộc tính.
      * Xử lý dữ liệu không chuẩn: Mô hình này có khả năng xử lý dữ liệu không đồng nhất hoặc thiếu sót, không cần quá nhiều tiền xử lý dữ liệu trước khi huấn luyện.
      * Phát hiện mẫu: Decision Tree có khả năng phát hiện mẫu tự nhiên trong dữ liệu. Nó có thể dễ dàng xác định các mẫu tương đối phức tạp trong tập dữ liệu.
      * Áp dụng cho cả dữ liệu dạng số và phân loại: Mô hình này có thể xử lý cả dữ liệu dạng số và dạng phân loại, giúp bạn tính toán được các biến số liên tục cũng như các biến phân loại.
    - Nhược điểm:
      * Dễ bị overfitting: Decision Tree có thể dễ bị overfitting (quá khớp) trên tập dữ liệu huấn luyện, đặc biệt khi không có các biện pháp điều chuẩn như pruning hoặc sử dụng các biến thay thế.
      * Không ổn định với sự biến đổi nhỏ: Các thay đổi nhỏ trong dữ liệu đầu vào có thể dẫn đến sự thay đổi lớn trong cấu trúc của cây quyết định, làm cho mô hình không ổn định.
      * Có thể tạo ra cây quyết định phức tạp: Nếu không được kiểm soát, cây quyết định có thể trở nên rất phức tạp và khó diễn giải, dẫn đến hiện tượng quá phù hợp với dữ liệu huấn luyện.
  + SVM
    - Lý do chọn: SVM là một mô hình mạnh mẽ có thể tạo ra các biên quyết định phức tạp. Nó có thể hoạt động tốt với dữ liệu có cấu trúc phức tạp và có khả năng xử lý tốt với các đặc trưng không tuyến tính.
    - Ưu điểm:
      * Hiệu suất tốt trong không gian chiều cao: SVM hoạt động tốt khi số lượng đặc trưng lớn hơn số lượng mẫu, đặc biệt trong các không gian chiều cao. Trong trường hợp dự đoán bệnh tim, thông thường có nhiều đặc trưng y tế.
      * Hiệu suất tốt khi có biên phân chia rõ ràng: SVM tốt khi có một ranh giới phân chia rõ ràng giữa các lớp dữ liệu. Nếu có tính chất phân tách tốt giữa những người có rủi ro cao và thấp về bệnh tim, SVM có thể hoạt động hiệu quả.
      * Khả năng xử lý dữ liệu lớn: SVM có thể xử lý tốt dữ liệu lớn mà không gây overfitting.
      * Hỗ trợ việc sử dụng các hàm kernel: SVM có khả năng ánh xạ dữ liệu vào các không gian chiều cao hơn thông qua việc sử dụng các hàm kernel. Điều này giúp tăng khả năng phân loại khi dữ liệu không phân tách tuyến tính.
    - Nhược điểm:
      * Đòi hỏi lựa chọn kernel phù hợp: Việc chọn kernel phù hợp có thể ảnh hưởng đến hiệu suất của SVM. Đôi khi, việc lựa chọn kernel không phù hợp có thể dẫn đến hiện tượng overfitting hoặc underfitting.
      * Khó mở rộng cho dữ liệu lớn: Trong một số trường hợp, SVM có thể tốn nhiều tài nguyên tính toán và bộ nhớ, đặc biệt khi xử lý dữ liệu lớn.
      * Khả năng xử lý nhiễu: SVM có thể nhạy cảm với nhiễu trong dữ liệu, có thể dẫn đến việc không tốt trong việc dự đoán.
      * Đòi hỏi tối ưu tham số: Để đạt hiệu suất tốt nhất, SVM yêu cầu tinh chỉnh tham số cẩn thận, điều này có thể là một quá trình tốn thời gian.
  + NaiveBayes
    - Lý do chọn: Naive Bayes là một mô hình phân loại đơn giản và nhanh chóng. Nó thường là một lựa chọn tốt khi dữ liệu có nhiều đặc trưng và các đặc trưng này độc lập với nhau (giả định "ngây thơ" của Naive Bayes).
    - Ưu điểm:
      * Đơn giản và nhanh chóng: Naive Bayes (NB) là một mô hình đơn giản, dễ hiểu và dễ triển khai. Thuật toán này có thể được huấn luyện nhanh chóng trên các tập dữ liệu lớn.
      * Xử lý tốt với dữ liệu lớn: Dù đơn giản, NB vẫn có thể hoạt động tốt trên các tập dữ liệu lớn mà không gặp phải vấn đề về hiệu suất.
      * Hiệu suất tốt với các đặc trưng độc lập: Mô hình này hoạt động tốt khi các đặc trưng độc lập với nhau, điều này cũng phù hợp với một số trường hợp trong y học.
      * Độ chính xác cao ban đầu: Trong một số trường hợp, NB có thể đạt được độ chính xác tốt ngay cả khi giả định về sự độc lập không hoàn toàn chính xác.
    - Nhược điểm:
      * Giả định về độc lập: Giả định rằng các đặc trưng là độc lập với nhau có thể không phù hợp với thực tế trong nhiều trường hợp y tế, vì nhiều biến có thể ảnh hưởng lẫn nhau.
      * Đòi hỏi dữ liệu lớn cho độ chính xác cao: Mặc dù NB hoạt động tốt với dữ liệu lớn, để đạt được độ chính xác cao, nó vẫn cần một lượng dữ liệu đủ lớn và đủ đa dạng.
      * Không xử lý được mối quan hệ phức tạp: Mô hình Naive Bayes không thể hiện mối quan hệ phức tạp giữa các đặc trưng, điều này có thể làm giảm khả năng dự đoán chính xác trong một số trường hợp.
    1. Cách triển khai mô hình Machine Learning
* Khai báo các hàm cần thiết.



Hình 2. Khai báo hàm sử dụng

* Tạo bộ dữ liệu train, test và bộ dữ liệu dùng để dự đoán.



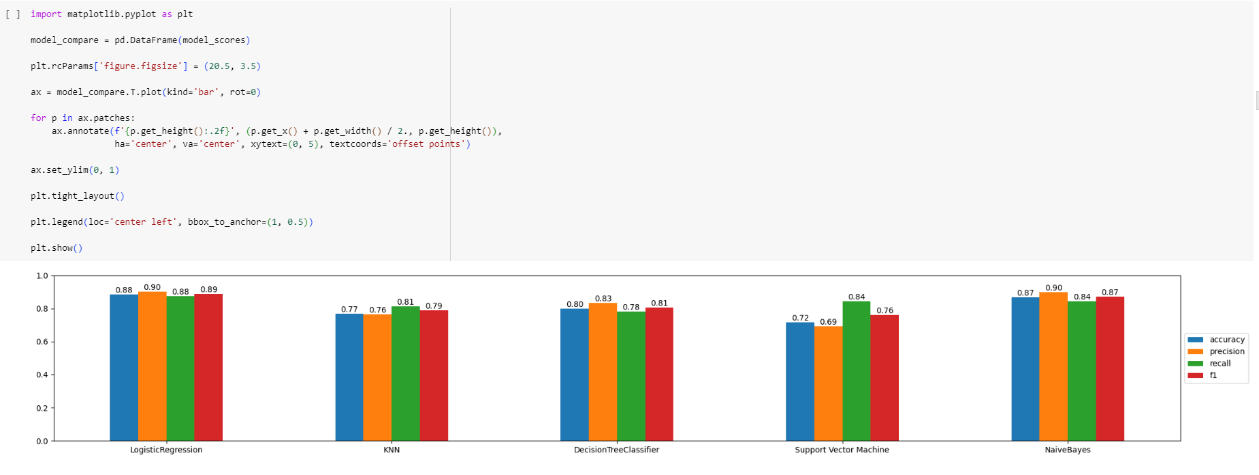
Hình 2. Tạo dữ liệu train và test

* Tính các chỉ số accuracy, precision, recall, f1-score cho các mô hình.



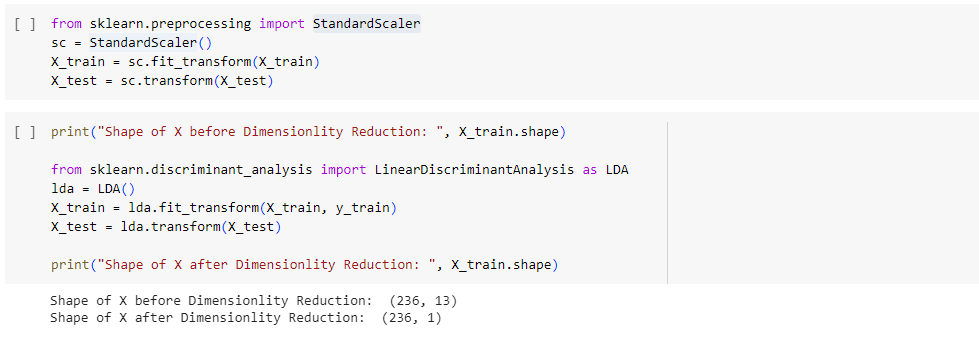
Hình 2. Code tính các chỉ số accuracy, precision, recall, f1-score cho các mô hình

* Vẽ biểu đồ hiển thị các chỉ số trên.



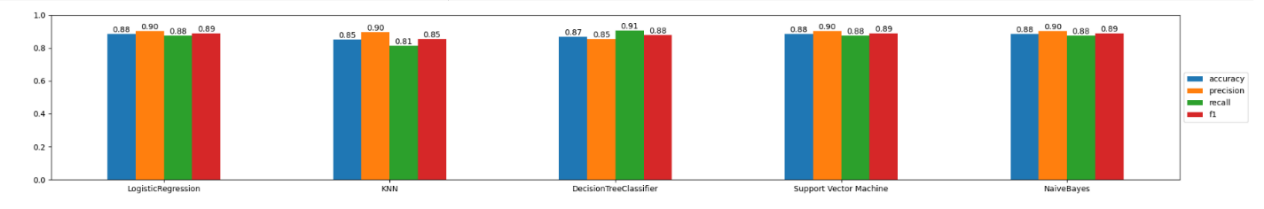
Hình 2. Biểu đồ hiển thị

* Nhìn vào biểu đồ cho thấy model Logistic Regression là mô hình chạy hiệu quả nhất trên bộ dữ liệu đang làm.
* Sau đó ta StandardScaler và SDA bộ dữ liệu để model SVM và KNN chạy tốt và hiệu quả hơn.



Hình 2. StandardScaler và SDA bộ dữ liệu để model SVM và KNN

* Tiếp đến, ta tính lại các chỉ số như bước ở trên và đo lại các chỉ số, hiển thị các chỉ số lên biểu đồ.



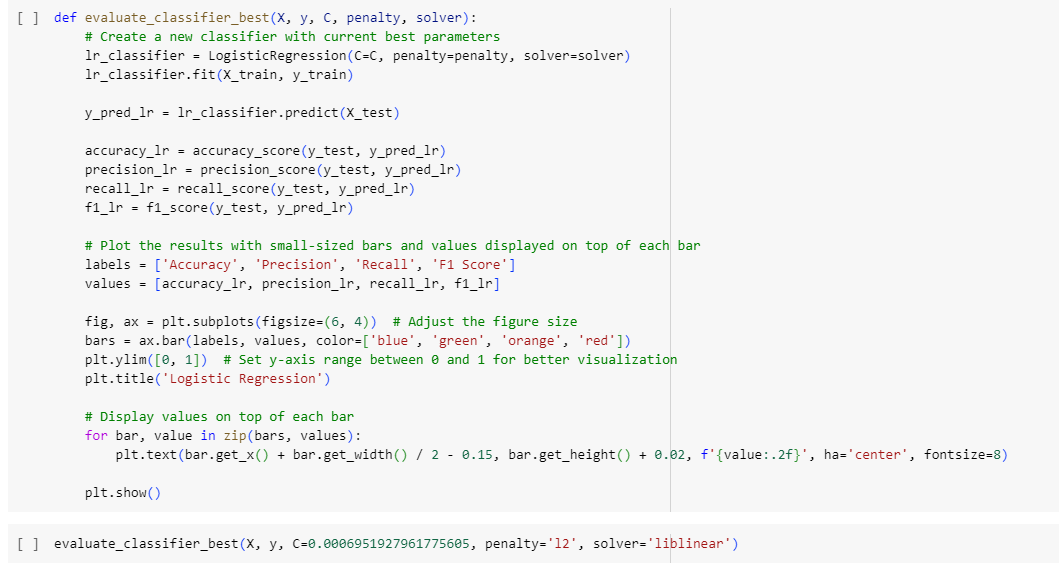
Hình 2. Biểu đồ hiển thị sau khi cải thiện

* Kết quả cho thấy, sau khi SDA và StandardScaler thì 2 mô hình KNN và SVM đã hoạt động tối ưu và hiệu quả hơn trước.
* Vì mô hình Logistic Regression là mô hình chạy hiệu quả nhất nên sẽ sử dụng mô hình này để mô tả cách triển khai của nó:
  + Sử dụng GridSearchCV để tìm ra Hyperparameters tốt nhất.



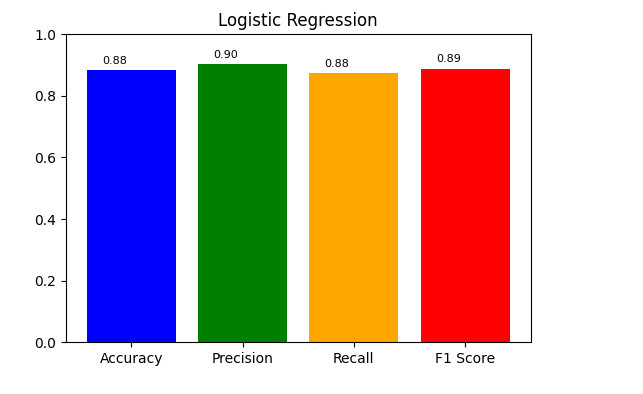
Hình 2. Sử dụng GridSearchCV để tìm ra Hyperparameters tốt nhất.

* + Dùng Hyperparameters tốt nhất để tính các chỉ số trong model.



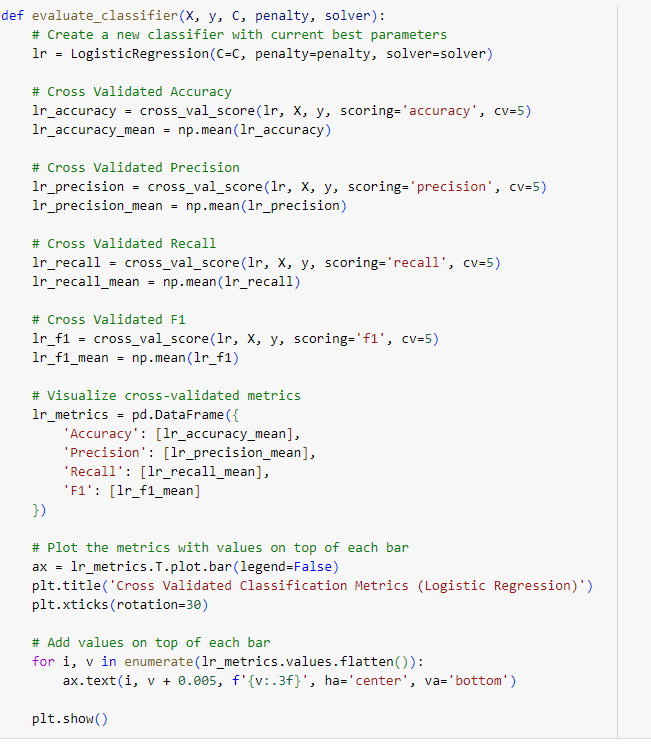
Hình 2. Dùng Hyperparameters tốt nhất để tính các chỉ số trong model

* + - Kết quả sau khi tính được các chỉ số và biểu diễn nó lên biểu đồ:



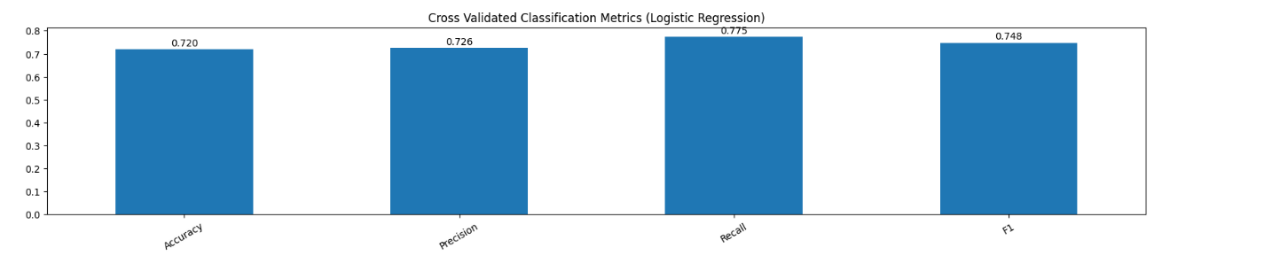
Hình 2. Biểu diễn biểu đồ với mô hình LR

* + - Tính các chỉ số trong model dựa theo hàm cross\_val\_score, hàm sẽ giúp chúng ta đánh giá hiệu quả của mô hình trên nhiều bộ train và test khác nhau:



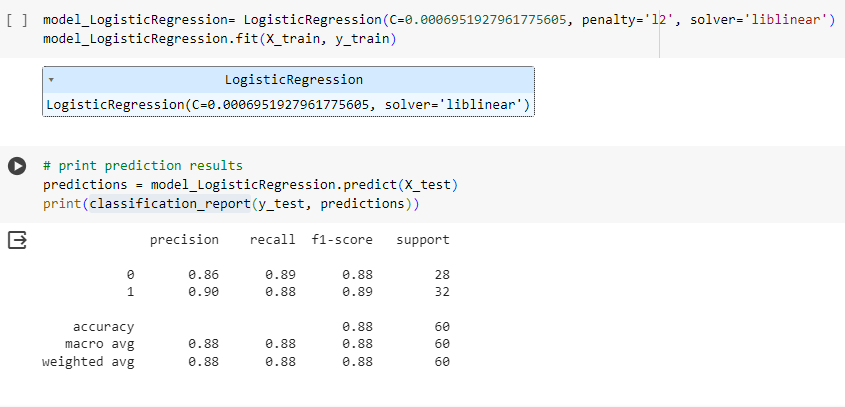
Hình 2. Tính các chỉ số trong model dựa theo hàm cross\_val\_score

* + - Kết quả:



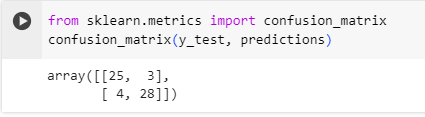
Hình 2. Kết quả

* + Hiển thị bảng classification\_report.



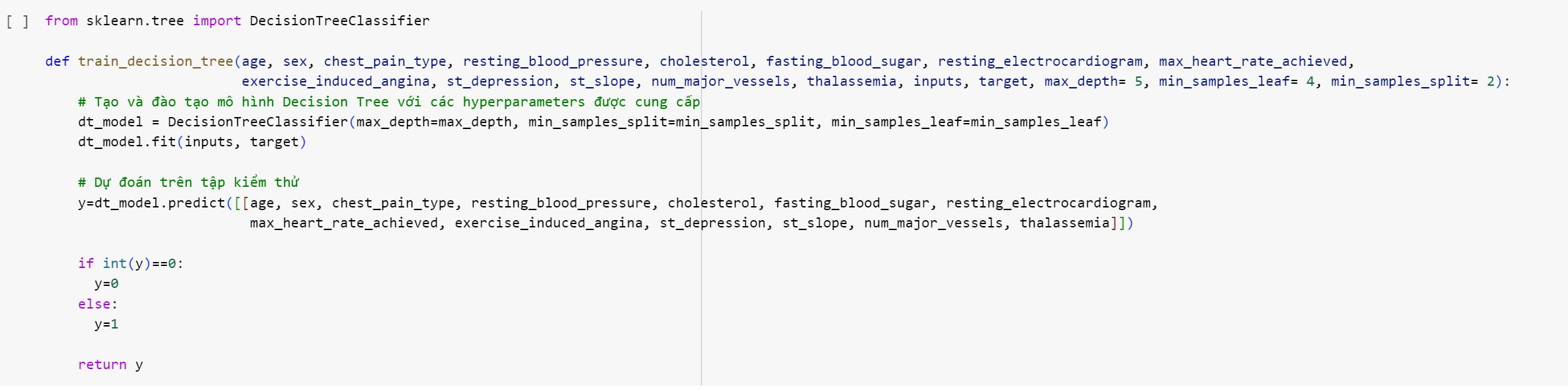
Hình 2. Hiển thị bảng classification\_report

* + Hiển thị confusion\_matrix



Hình 2. Hiển thị confusion\_matrix

* Cuối cùng thì ta sẽ có một vài hàm dự đoán của các model:
  + Decision Tree



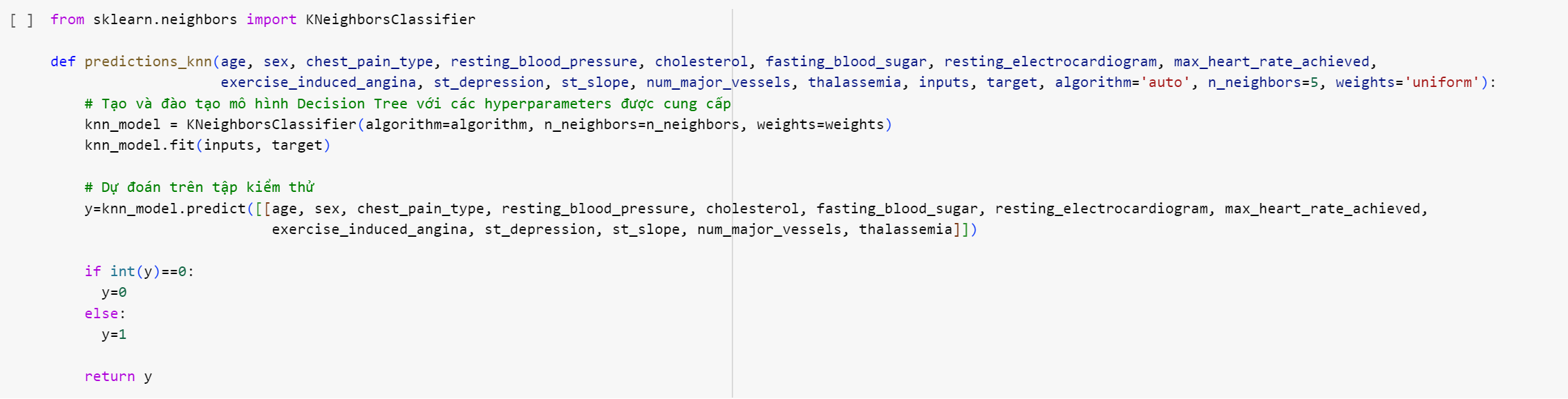
Hình 2. Code dự đoán của model Decision Tree

* + Logistic Regression



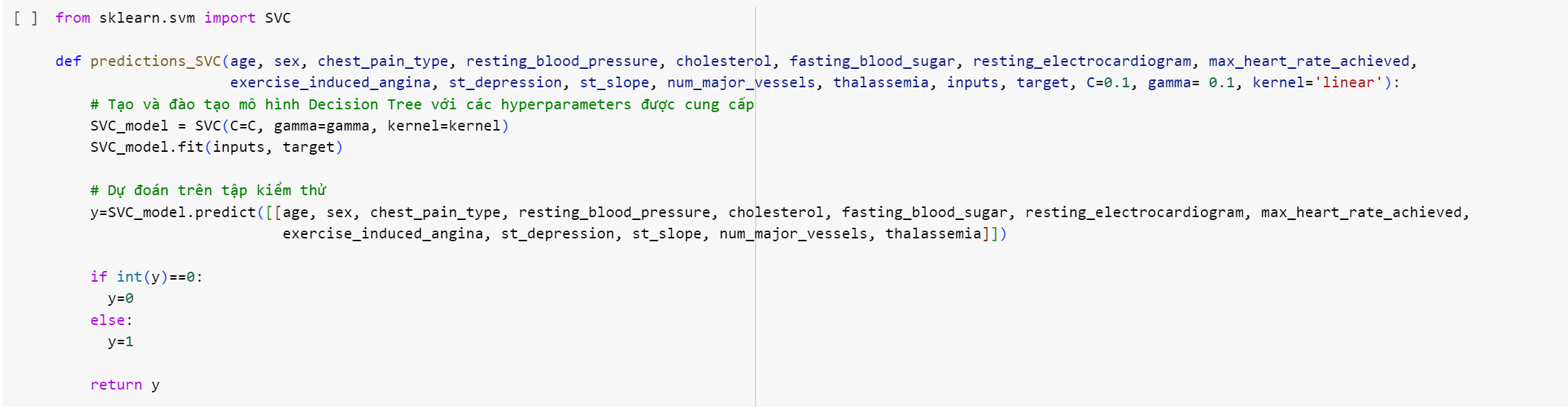
Hình 2. Code dự đoán của model Logistic Regresssion

* + KNN



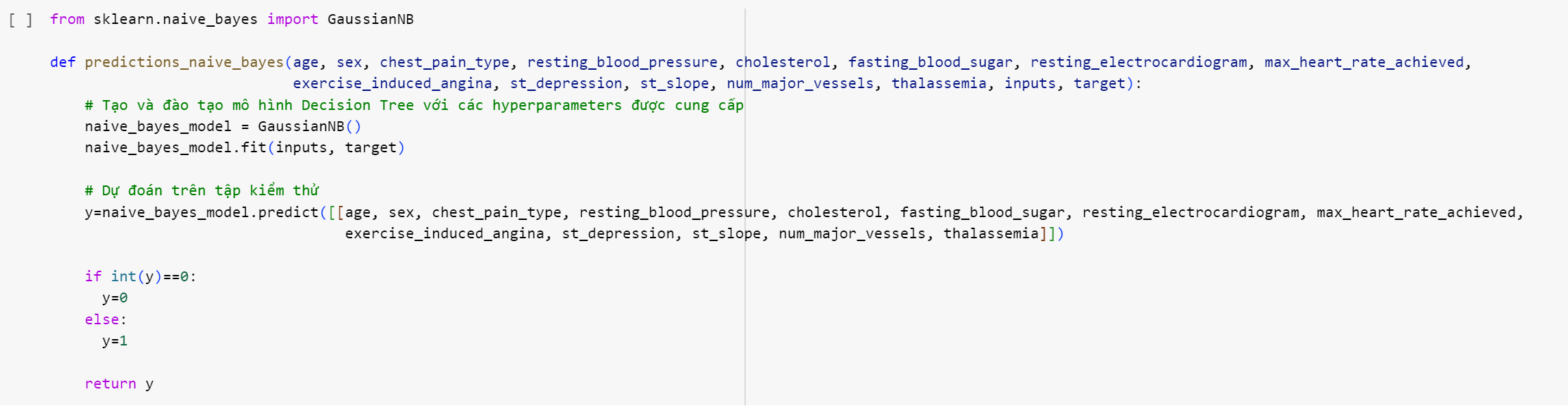
Hình 2. Code dự đoán của model KNN

* + SVC



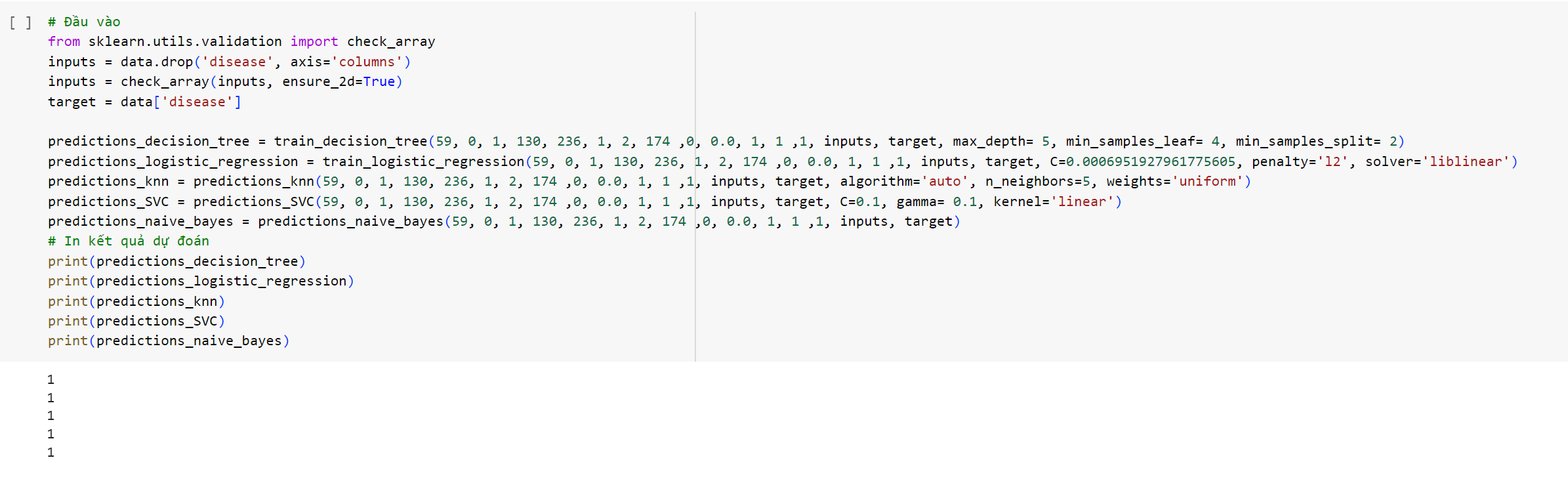
Hình 2. Code dự đoán của model SVC

* + Naive Bayes



Hình 2. Code dự đoán của model Naïve Bayes

* + Test dự đoán với 1 trường hợp bất kì chung cho tất cả các model:



Hình 2. Test dự đoán đối với từng mô hình ML

* Dự đoán trường hợp trên qua các model đều cho ra kết quả là 1, tức là người có các chỉ số điều kiện trên được dự đoán là bị mắc bệnh tim mạch.

# TÀI LIỆU THAM KHẢO

[1]: <https://drive.google.com/file/d/1fJyPy0mqglkQ_eSqaoRa8zJbZ9lFAP-b/view>

[2]: <https://drive.google.com/file/d/1xyJZKh6eX2KwSwDh3FpU4MOTMY8PNCBJ/view>

[3]: <https://drive.google.com/file/d/1-ZnbuiqPYihiDPk-_OG77iAfyVvDWOAp/view>

[4]: <https://drive.google.com/file/d/1f7QWTxF5KP1bRfr-4hp0ZkAUD-3aJ7Mv/view>

[6]: <https://machinelearningcoban.com/2017/01/27/logisticregression/>

[7]: <https://machinelearningcoban.com/2017/01/08/knn/>

[8]: <https://machinelearningcoban.com/tabml_book/ch_model/decision_tree.html>

[9]: <https://www.geeksforgeeks.org/classifying-data-using-support-vector-machinessvms-in-python/>

[10]: <https://www.geeksforgeeks.org/ml-naive-bayes-scratch-implementation-using-python/>