**TRƯỜNG ĐẠI HỌC QUỐC TẾ HỒNG BÀNG**

**KHOA CÔNG NGHỆ - KỸ THUẬT**

**BỘ MÔN CÔNG NGHỆ THÔNG TIN**

-----🙟🙝🕮🙟🙝-----



**ĐỒ ÁN KHOA HỌC DỮ LIỆU**

**ĐỀ TÀI :**

**PHÂN TÍCH DỮ LIỆU SỨC KHOẺ   
TIM MẠCH**

**Cố vấn học tập : Lê Văn Hạnh**

**Sinh viên thực hiện : Đường Hoàng Hải Yến**

**Mã số sinh viên : 2111111681**

**Lớp : TH21DH-CN2**

**TP. Hồ Chí Minh, tháng 06 năm 2024**

LỜI CẢM ƠN

Để hoàn thành đề tài này, em đã nhận được sự hướng dẫn, giúp đỡ và góp ý nhiệt tình của thầy Lê Văn Hạnh.

Em xin gửi lời biết ơn sâu sắc đến thầy Lê Văn Hạnh đã dành nhiều thời gian và tâm huyết hướng dẫn nghiên cứu và giúp em hoàn thành luận văn tốt nghiệp.

Em cũng xin chân thành cảm ơn đến quý thầy cô trường Đại học Quốc Tế Hồng Bàng, đặc biệt là những thầy cô đã tận tình dạy bảo cho em suốt thời gian học tập tại trường.

Em xin chân thành cảm ơn Ban Giám Hiệu trường Đại Học Quốc Tế Hồng Bàng cùng quý thầy cô trong Khoa Công Nghệ Thông Tin đã tạo rất nhiều điều kiện để em học tập và hoàn thành tốt khóa học.

Em đã có nhiều cố gắng hoàn thiện đề tài bằng tất cả năng lực của mình, tuy nhiên không thể tránh khỏi nhiều thiếu sót, rất mong nhận được những đóng góp quý báu của thầy và các bạn.

TRANG CAM KẾT

Em xin cam kết báo cáo này được hoàn thành dựa trên các kết quả nghiên cứu của em và các kết quả nghiên cứu này chưa được dùng cho bất cứ báo cáo cùng cấp nào khác.

TP.HCM, ngày ….. tháng ….. năm …..

Người thực hiện

Đường Hoàng Hải Yến

MỤC LỤC

[LỜI CẢM ƠN i](#_Toc169088411)

[TRANG CAM KẾT ii](#_Toc169088412)

[MỤC LỤC iii](#_Toc169088413)

[NHẬN XÉT CỦA GIẢNG VIÊN vi](#_Toc169088414)

[DANH MỤC BIỂU ĐỒ HÌNH VẼ vii](#_Toc169088415)

[DANH MỤC BẢNG x](#_Toc169088416)

[CHƯƠNG 1: GIỚI THIỆU VỀ CƠ SỞ DỮ LIỆU SỬ DỤNG CHO ĐỀ TÀI 1](#_Toc169088417)

[1.1. Tổng quan về cơ sở dữ liệu 1](#_Toc169088418)

[1.2. Giới thiệu các thuộc tính (fields) 1](#_Toc169088419)

[1.2.1. Tên của các fields 1](#_Toc169088420)

[1.2.2. Ý nghĩa chi tiết của các fields trong tập dữ liệu 1](#_Toc169088421)

[1.2.3. Số giá trị null của các fields 2](#_Toc169088422)

[1.2.4. Số giá trị unique của các fields 3](#_Toc169088423)

[1.2.5. Kiểu dữ liệu của các fields 3](#_Toc169088424)

[1.2.5.1. Field age 5](#_Toc169088425)

[1.2.5.2. Field trestbps 6](#_Toc169088426)

[1.2.5.3. Field cholesterol 6](#_Toc169088427)

[1.2.5.4. Field thalch 6](#_Toc169088428)

[1.2.5.5. Field oldpeak 6](#_Toc169088429)

[1.2.5.6. Field sex 7](#_Toc169088430)

[1.2.5.7. Field fbs 7](#_Toc169088431)

[1.2.5.8. Field exang 7](#_Toc169088432)

[CHƯƠNG 2: PHÂN TÍCH – THỐNG KÊ THỦ CÔNG TRÊN CSDL ĐÃ CHỌN 8](#_Toc169088433)

[2.1. Tìm hiểu dữ liệu 8](#_Toc169088434)

[2.1.1. Chọn 3 thuộc tính để vẽ các đồ thị 8](#_Toc169088435)

[2.1.1.1. Boxplot dựa trên five-number summary 8](#_Toc169088436)

[2.1.1.2. Quantile–Quantile Plot trên 2 thuộc tính bất kỳ nhưng có liên quan về ý nghĩa 9](#_Toc169088437)

[2.1.1.3. Histogram trên 2 thuộc tính bất kỳ nhưng có liên quan về ý nghĩa 10](#_Toc169088438)

[2.1.1.4. Scatter trên 2 thuộc tính bất kỳ nhưng có liên quan về ý nghĩa 11](#_Toc169088439)

[2.1.2. Nhóm dữ liệu đang có theo một thuộc tính dạng danh nghĩa 12](#_Toc169088440)

[2.1.2.1. Boxplot dựa trên five-number summary 12](#_Toc169088441)

[2.1.2.2. Histogram trên 2 thuộc tính bất kỳ nhưng có liên quan về ý nghĩa 13](#_Toc169088442)

[2.1.3. Đo lường sự tương đồng và khác biệt của dữ liệu bằng 2 cách: ma trận tương quan và độ đo Cosin 14](#_Toc169088443)

[2.1.3.1. Ma trận tương quan 14](#_Toc169088444)

[2.1.3.2. Độ đo Cosin 16](#_Toc169088445)

[2.2. Thực hiện khai thác dữ liệu 21](#_Toc169088446)

[2.2.1. Sử dụng các phương pháp khai phá dữ liệu 21](#_Toc169088447)

[2.2.1.1. Phương pháp phân loại Navie Bayesian 21](#_Toc169088448)

[2.2.1.2. Phương pháp phân cụm K-mean 22](#_Toc169088449)

[2.2.2. Thực hiện đánh giá các mẫu thu được bằng các phương pháp 26](#_Toc169088450)

[2.2.2.1. Phương pháp phân loại 26](#_Toc169088451)

[CHƯƠNG 3: PHÂN TÍCH – THỐNG KÊ BẰNG CÔNG CỤ PYTHON TRÊN CƠ SỞ DỮ LIỆU ĐÃ CHỌN 34](#_Toc169088452)

[3.1. Thực hiện lại phần 1 bằng Python 34](#_Toc169088453)

[3.1.1. Giá trị null 34](#_Toc169088454)

[3.1.2. Kiểu dữ liệu của các fields 35](#_Toc169088455)

[3.1.3. Kiểu dữ liệu thuộc tính số 39](#_Toc169088456)

[3.2. Phân tích – Thống kê 42](#_Toc169088457)

[3.2.1. Chọn 3 thuộc tính để vẽ các đồ thị 42](#_Toc169088458)

[3.2.1.1. Boxplot dựa trên five-number summary 43](#_Toc169088459)

[3.2.1.2. Quantile–Quantile Plot trên 2 thuộc tính bất kỳ nhưng có liên quan về ý nghĩa 46](#_Toc169088460)

[3.2.1.3. Histogram trên 2 thuộc tính bất kỳ nhưng có liên quan về ý nghĩa 48](#_Toc169088461)

[3.2.1.4. Scatter trên 2 thuộc tính bất kỳ nhưng có liên quan về ý nghĩa 49](#_Toc169088462)

[3.2.2. Nhóm dữ liệu đang có theo một thuộc tính dạng danh nghĩa 50](#_Toc169088463)

[3.2.2.1. Boxplot dựa trên five-number summary 51](#_Toc169088464)

[3.2.2.2. Histogram trên 2 thuộc tính bất kỳ nhưng có liên quan về ý nghĩa 51](#_Toc169088465)

[3.2.3. Đo lường sự tương đồng và khác biệt của dữ liệu bằng 2 cách: ma trận tương quan và độ đo Cosin 53](#_Toc169088466)

[3.2.3.1. Ma trận tương quan 54](#_Toc169088467)

[3.2.3.2. Độ đo Cosin 56](#_Toc169088468)

[3.3. Thực hiện khai thác dữ liệu 58](#_Toc169088469)

[3.3.1. Sử dụng các phương pháp khai phá dữ liệu 58](#_Toc169088470)

[3.3.1.1. Phương pháp phân loại Navie Bayesian 58](#_Toc169088471)

[3.3.1.2. Phương pháp phân cụm K-mean 61](#_Toc169088472)

[3.4. So sánh thực hiện của việc sử dụng công cụ Python và thủ công 62](#_Toc169088473)

[CHƯƠNG 4: KẾT LUẬN 63](#_Toc169088474)

[TÀI LIỆU THAM KHẢO 65](#_Toc169088475)

NHẬN XÉT CỦA GIẢNG VIÊN

TP.HCM, Ngày…Tháng… Năm 2024

Chữ ký giảng viên

DANH MỤC BIỂU ĐỒ HÌNH VẼ

[Hình 2.1. Boxplot age 8](#_Toc169088476)

[Hình 2.2.Boxplot trestbps 8](#_Toc169088477)

[Hình 2.3. Boxplot Cholesterol 9](#_Toc169088478)

[Hình 2.4.Boxplot Thalch 9](#_Toc169088479)

[Hình 2.5.Q-Q Plot giữa Thalch và Cholesterol 10](#_Toc169088480)

[Hình 2.6.Histogram của cholesterol 10](#_Toc169088481)

[Hình 2.7.Histogram của thalch 11](#_Toc169088482)

[Hình 2.8.Scatter thalch-cholesterol 11](#_Toc169088483)

[Hình 2.9. Boxplot Age lv hypertrophy 12](#_Toc169088484)

[Hình 2.10. Boxplot Age normal 12](#_Toc169088485)

[Hình 2.11.Boxplot Age st-t abnormality 13](#_Toc169088486)

[Hình 2.12.Historgram Normal của Cholesterol 13](#_Toc169088487)

[Hình 2.13.Historgram Normal của Thalch 14](#_Toc169088488)

[Hình 2.14.Thuộc tính các dạng 14](#_Toc169088489)

[Hình 2.15.Ma trận phân sai restecg 15](#_Toc169088490)

[Hình 2.16.Ma trận phân sai thal 15](#_Toc169088491)

[Hình 2.17.Ma trận phân sai fbs 15](#_Toc169088492)

[Hình 2.18.Ma trận phân sai cholesterol 15](#_Toc169088493)

[Hình 2.19.Chuyển sang dữ liệu số 23](#_Toc169088494)

[Hình 2.20.Cụm ngẫu nhiên 23](#_Toc169088495)

[Hình 2.21.Cụm K3 24](#_Toc169088496)

[Hình 2.22.Các cụm lần lượt 25](#_Toc169088497)

[Hình 2.23.So sánh các cụm 25](#_Toc169088498)

[Hình 2.24.Các cột dùng đánh giá 26](#_Toc169088499)

[Hình 2.25.Đã chuyển đổi dữ liệu 28](#_Toc169088500)

[Hình 2.26.Đếm thuộc tính (1) 29](#_Toc169088501)

[Hình 2.27.Đếm thuộc tính (2) 30](#_Toc169088502)

[Hình 2.28. Tính (P|Yes) và (P|No) 30](#_Toc169088503)

[Hình 2.29.Xác xuất dữ liệu (1) 31](#_Toc169088504)

[Hình 2.30.Xác xuất dữ liệu (1) 31](#_Toc169088505)

[Hình 2.31.Test Set (1) 32](#_Toc169088506)

[Hình 2.32.Xác định giá trị 33](#_Toc169088507)

[Hình 2.33.Vẽ Confusion Matrix 33](#_Toc169088508)

[Hình 3.1.Đọc dữ liệu 34](#_Toc169088509)

[Hình 3.2.Thuộc tính các cột 35](#_Toc169088510)

[Hình 3.3.Code kiểu dữ liệu 36](#_Toc169088511)

[Hình 3.4.Kiểu dữ liệu sex, origin 36](#_Toc169088512)

[Hình 3.5.Kiểu dữ liệu chestpain và fbs 37](#_Toc169088513)

[Hình 3.6.Kiểu dữ liệu restecg và exang 37](#_Toc169088514)

[Hình 3.7.Kiểu dữ liệu slope và ca 38](#_Toc169088515)

[Hình 3.8.Kiểu dữ liệu thal và num 38](#_Toc169088516)

[Hình 3.9.Tính giá trị theo yêu cầu 39](#_Toc169088517)

[Hình 3.10.Hiện thị kết quả 40](#_Toc169088518)

[Hình 3.11.Five-number summary(1) 40](#_Toc169088519)

[Hình 3.12.Five-number summary(2) 41](#_Toc169088520)

[Hình 3.13.Five-number summary(3) 42](#_Toc169088521)

[Hình 3.14. Boxplot age 43](#_Toc169088522)

[Hình 3.15.Boxplot trestbps 44](#_Toc169088523)

[Hình 3.16. Boxplot Cholesterol 45](#_Toc169088524)

[Hình 3.17.Boxplot Thalch 46](#_Toc169088525)

[Hình 3.18.Code Q-Q Plot giữa Thalch và Cholesterol 47](#_Toc169088526)

[Hình 3.19.Q-Q Plot giữa Thalch và Cholesterol 47](#_Toc169088527)

[Hình 3.20.Histogram của cholesterol 48](#_Toc169088528)

[Hình 3.21.Histogram của thalch 49](#_Toc169088529)

[Hình 3.22.Scatter thalch-cholesterol 50](#_Toc169088530)

[Hình 3.23.Code vẽ boxplot thuộc tính danh nghĩa 51](#_Toc169088531)

[Hình 3.24.Boxplot danh nghĩa 51](#_Toc169088532)

[Hình 3.25.Historgram Normal của Cholesterol 52](#_Toc169088533)

[Hình 3.26.Historgram Normal của Thalch 53](#_Toc169088534)

[Hình 3.27.Thuộc tính các dạng 54](#_Toc169088535)

[Hình 3.28.Code ma trận tương quan 54](#_Toc169088536)

[Hình 3.29.In kêt quả ma trận tương quan 55](#_Toc169088537)

[Hình 3.30.Biểu đồ heatmap của ma trận tương quan 55](#_Toc169088538)

[Hình 3.31.Code độ đo cosine 57](#_Toc169088539)

[Hình 3.32.Dữ liệu ban đầu 58](#_Toc169088540)

[Hình 3.33.Dữ liệu dã chuyển đổi 59](#_Toc169088541)

[Hình 3.34.Thêm cột suy tim 59](#_Toc169088542)

[Hình 3.35.Đã thêm cột suy tim 59](#_Toc169088543)

[Hình 3.36.Xây dựng mô hình NavieBayes 60](#_Toc169088544)

[Hình 3.37.Xây dựng mô hình NavieBayes(2) 60](#_Toc169088545)

[Hình 3.38.Độ chính xác NavieBayes 61](#_Toc169088546)

[Hình 3.39.Vẽ biểu đồ K-mean 61](#_Toc169088547)

[Hình 3.40.Biểu đồ k-mean 62](#_Toc169088548)

DANH MỤC BẢNG

[Bảng 1. Field age 5](#_Toc169001876)

[Bảng 2. Field trestbps 6](#_Toc169001877)

[Bảng 3. Field cholesterol 6](#_Toc169001878)

[Bảng 4 Field thalch 6](#_Toc169001879)

[Bảng 6 Field sex 7](#_Toc169001880)

[Bảng 7. Field fbs 7](#_Toc169001881)

[Bảng 8. Field exang 7](#_Toc169001882)

# GIỚI THIỆU VỀ CƠ SỞ DỮ LIỆU SỬ DỤNG CHO ĐỀ TÀI

## Tổng quan về cơ sở dữ liệu

Tập dataset mà em thu thập và phân tích có tên là “***UCI Heart Disease Data***”, có ý nghĩa là ***tâp dữ liệu dự đoán khả năng mắc bệnh tim mạch*** . Tập dataset này được em thu tập từ nguồn [Kaggle: UCI Heart Disease Data](https://www.kaggle.com/) .

## Giới thiệu các thuộc tính (fields)

Tập dataset này có 299 records và 16 fields (column).

### Tên của các fields

* id
* age
* origin
* sex
* chest pain
* trestbps
* cholesterol
* fbs
* restecg
* thalach
* exang
* oldpeak
* slope
* ca
* thal
* num

### Ý nghĩa chi tiết của các fields trong tập dữ liệu

|  |  |  |
| --- | --- | --- |
| **STT** | **Tên Field** | **Ý nghĩa** |
| 1 | Id | Id duy nhất cho mỗi bệnh nhân |
| 2 | Age | Tuổi của bệnh nhân tính bằng năm |
| 3 | Origin | Nguồn gốc |
| 4 | Sex | Giới tính nam / nữ |
| 5 | Chest pain | Đau thắt ngực |
| 6 | Trestbps | Huyết áp |
| 7 | Cholesterol | Cholesterol huyết thanh |
| 8 | Fbs | Đường huyết lúc đói |
| 9 | Restecg | Kết quả điện tâm đồ lúc nghỉ |
| 10 | thalch | Nhịp tim tối đa đạt được |
| 11 | Exang | Đau thắt ngực do tập thể dục |
| 12 | Oldpeak | Mức độ ST depression trong một bài kiểm tra căng thẳng tim |
| 13 | Slope | Độ dốc của đoạn ST vận động đỉnh |
| 14 | Ca | Sơ mạch chính được nhuộm màu bằng soi huỳnh quang |
| 15 | Thal | Thalassemia là bệnh thiếu máu huyết tán bẩm sinh |
| 16 | Num | Giá trị dự đoán suy tim theo cấp độ |

### Số giá trị null của các fields

* Id : không giá trị null .
* age : không giá trị null .
* origin : không giá trị null .
* sex : không giá trị null .
* chest pain : không giá trị null
* trestbps : không giá trị null
* cholesterol : không giá trị null
* fbs : không giá trị null
* restecg : không giá trị null
* thalach : không giá trị null
* exang : không giá trị null
* oldpeak : không giá trị null
* slope: không giá trị null
* ca : không giá trị null
* thal : không giá trị null
* num : không giá trị null

### Số giá trị unique của các fields

* Id : không giá trị unique .
* Age : không giá trị unique.
* Origin : 3 giá trị (Cleveland , Hungary, VA Long Beach)
* Sex: 2 giá trị (Male và Female).
* chest pain : 4 giá trị (Typical angina, atypical angina , non-anginal, asymptomatic).
* Trestbps : không giá trị unique
* Cholesterol : không giá trị unique.
* fbs : 2 giá trị (TRUE = 1 và FALSE = 0).
* Restecg : 3 giá trị ( lv hypertrophy, normal, st-t abnormality).
* Thalach : không giá trị unique.
* Exang : 2 giá trị (TRUE =1 và FALSE = 0)
* Oldpeak : không giá trị unique .
* Slope : 2 giá trị (downsloping , flat, upsloping ).
* Ca : không giá trị unique .
* Thal : 3 giá trị (fixed defect, normal, reversable defect).
* num : không giá trị unique

### Kiểu dữ liệu của các fields

* Id :
* Kiểu dữ liệu: Số nguyên (numeric).
* Không phải kiểu nhị phân hoặc rời rạc.
* Từ đó suy ra mode: Không có mode rõ ràng, vì mỗi giá trị id là duy nhất.
* Sex :
* Kiểu dữ liệu: Nhị phân (binary).
* Gồm các giá trị: 0 (Female) hoặc 1 (Male).
* Tỷ lệ phần trăm: Female : 32,1% và Male : 67,89%.
* Mode: Giá trị có tỷ lệ phần trăm cao nhất là Male.
* Age :
  + Kiểu dữ liệu: Số nguyên (numeric).
  + Không phải kiểu nhị phân hoặc rời rạc.
  + Từ đó suy ra mode: Không có mode rõ ràng, vì mỗi giá trị age có thể là duy nhất.
* Chest pain :
* Kiểu dữ liệu: Rời rạc (categorical).
* Gồm các giá trị: 0 (asymptomatic) ,1 (atypical angina), 2(non- anginal) ,3(typical angina).
* Tỷ lệ phần trăm: Asymptomatic: 48,18%, atypical angina: 16,38% , non- anginal : 27.75%, typical angina: 7,69%.
* Mode: Giá trị có tỷ lệ phần trăm cao nhất là Asymptomatic.
* Trestbps
  + Kiểu dữ liệu: Số nguyên (numeric).
  + Không phải kiểu nhị phân hoặc rời rạc.
  + Từ đó suy ra mode: Không có mode rõ ràng, vì mỗi giá trị huyết áp có thể là duy nhất.
* Cholesterol
  + Kiểu dữ liệu: Số nguyên (numeric).
  + Không phải kiểu nhị phân hoặc rời rạc.
  + Từ đó suy ra mode: Không có mode rõ ràng, vì mỗi giá trị cholesterol có thể là duy nhất.
* fbs :
* Kiểu dữ liệu: Nhị phân (binary).
* Gồm các giá trị: 0 (fbs ≤ 120 mg/dl giá trị False) hoặc 1 (fasting blood sugar > 120 mg/dl giá trị True).
* Tỷ lệ phần trăm: False :85,62% và True : 14.38% .
* Mode: Giá trị có tỷ lệ phần trăm cao nhất là False.
* Restecg:
* Kiểu dữ liệu: Rời rạc (categorical).
* Gồm các giá trị: 0 (lv hypertrophy) ,1(normal), 2(st-t abnormality).
* Tỷ lệ phần trăm: 0: 48,82%, 1: 49,83% , 2 : 1,33% .
* Mode: Giá trị có tỷ lệ phần trăm cao nhất là normal.
* Thalach
  + Kiểu dữ liệu: Số nguyên (numeric).
  + Không phải kiểu nhị phân hoặc rời rạc.
  + Từ đó suy ra mode: Không có mode rõ ràng, vì mỗi giá trị tần suất nhịp tim có thể là duy nhất.
* Exang
* Kiểu dữ liệu: Nhị phân (binary).
* Gồm các giá trị: 0 (False) hoặc 1 (True).
* Tỷ lệ phần trăm: False : 66,88% và True : 33,11% .
* Mode: Giá trị có tỷ lệ phần trăm cao nhất là False.
* Oldpeak :
* Kiểu dữ liệu: Số thực (numeric).
* Không phải kiểu nhị phân hoặc rời rạc.
* Từ đó suy ra mode: Không có mode rõ ràng, vì mỗi giá trị có thể là duy nhất.
* Slope
* Kiểu dữ liệu: Rời rạc (categorical).
* Gồm các giá trị: 0(downsloping) ,1 (upsloping), 2(flat).
* Tỷ lệ phần trăm: 0: 58,86%, 1: 21,73% , 2: 46,48%.
* Mode: Giá trị có tỷ lệ phần trăm cao nhất là upsloping và flat.
* Ca :
* Kiểu dữ liệu: Rời rạc (categorical).
* Gồm các giá trị: 0,1,2,3
* Tỷ lệ phần trăm: 0: 7,02%, 1: 46,48% , 2: 12,71%, 3 : 6,68%.
* Mode: Giá trị có tỷ lệ phần trăm cao nhất là 1.
* Thal :
* Kiểu dữ liệu: Rời rạc (categorical).
* Gồm các giá trị: 0(fixed defect) ,1 (normal), 2(reversable defect).
* Tỷ lệ phần trăm: 0: 6,02%, 1: 54,84% , 2: 39,13%.
* Mode: Giá trị có tỷ lệ phần trăm cao nhất là normal.
* num :
* Kiểu dữ liệu: Rời rạc (categorical).
* Gồm các giá trị: 0, 1, 2, 3 .
* Tỷ lệ phần trăm:0: 53,51%, 1: 18,73%,2: 11,71%,3: 11,71%,4: 4,34 %.
* Mode: Giá trị có tỷ lệ phần trăm cao nhất là 0.

#### Field age

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **AGE** | | | | | | | |
| Mean | Midrange | Mode | Five-number summary | | | | |
| 54.5 | 53 | 58 | Min | Q1 | Median | Q3 | Max |
|  |  | Xuất hiện 18 lần | 29 | 48 | 56 | 61 | 77 |

Bảng 1. Field age

#### Field trestbps

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **TRESTBPS** | | | | | | | |
| Mean | Midrange | Mode | Five-number summary | | | | |
| 131.7 | 147 | 120 | Min | Q1 | Median | Q3 | Max |
|  |  | Xuất hiện 38 lần | 94 | 120 | 130 | 140 | 200 |

Bảng 2. Field trestbps

#### Field cholesterol

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **CHOLESTEROL** | | | | | | | |
| Mean | Midrange | Mode | Five-number summary | | | | |
| 246.8 | 332 | 234 | Min | Q1 | Median | Q3 | Max |
|  |  | Xuất hiện 6 lần | 100 | 211 | 242 | 275.5 | 564 |

Bảng 3. Field cholesterol

#### Field thalch

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **THALCH** | | | | | | | |
| Mean | Midrange | Mode | Five-number summary | | | | |
| 149.3 | 136.5 | 162 | Min | Q1 | Median | Q3 | Max |
|  |  | Xuất hiện 11 lần | 71 | 132.5 | 152 | 165.5 | 202 |

Bảng 4 Field thalch

#### Field oldpeak

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **OLDPEAK** | | | | | | | |
| Mean | Midrange | Mode | Five-number summary | | | | |
| 1.06 | 3.1 | 0 | Min | Q1 | Median | Q3 | Max |
|  |  | Xuất hiện 99 lần | 0 | 0 | 0.8 | 1.6 | 6.2 |

Bảng 5. Field oldpeak

#### Field sex

|  |  |  |  |
| --- | --- | --- | --- |
| **SEX** | | | |
| Giá trị | Số lượng | Chiếm (%) | Mode |
| Male | 203 | 67.9 | ✓ |
| Female | 96 | 32.1 |  |
| Tổng | 299 | 100 |  |

Bảng 6 Field sex

#### Field fbs

|  |  |  |  |
| --- | --- | --- | --- |
| **FBS** | | | |
| Giá trị | Số lượng | Chiếm (%) | Mode |
| FALSE | 43 | 14.38 |  |
| TRUE | 256 | 85.62 | ✓ |
| Tổng | 299 | 100 |  |

Bảng 7. Field fbs

#### Field exang

|  |  |  |  |
| --- | --- | --- | --- |
| **EXANG** | | | |
| Giá trị | Số lượng | Chiếm (%) | Mode |
| FALSE | 200 | 66.9 | ✓ |
| TRUE | 99 | 33.1 |  |
| Tổng | 299 | 100 |  |

Bảng 8. Field exang

# PHÂN TÍCH – THỐNG KÊ THỦ CÔNG TRÊN CSDL ĐÃ CHỌN

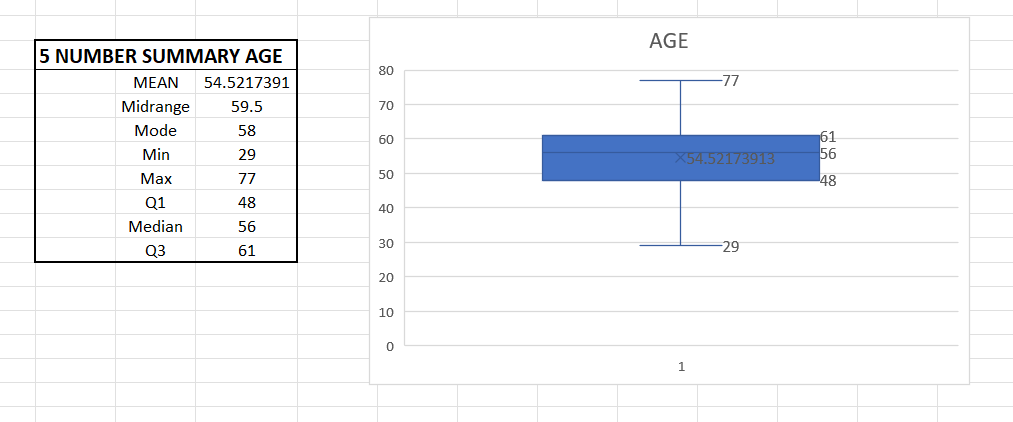
## Tìm hiểu dữ liệu

### Chọn 3 thuộc tính để vẽ các đồ thị

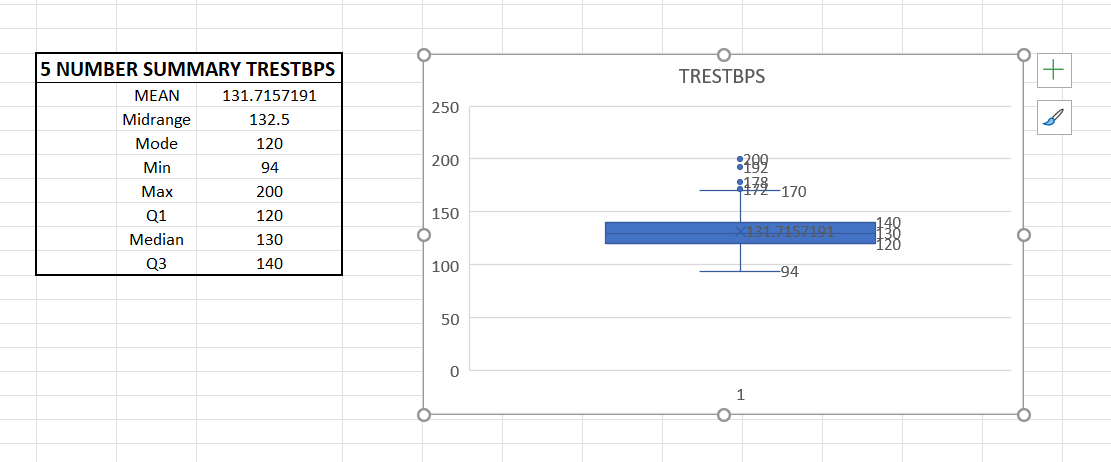
- 4 thuộc tính em chọn để vẽ các đồ thị là: age, trestbps , cholesterol và thalch

#### Boxplot dựa trên five-number summary

###### **Age**

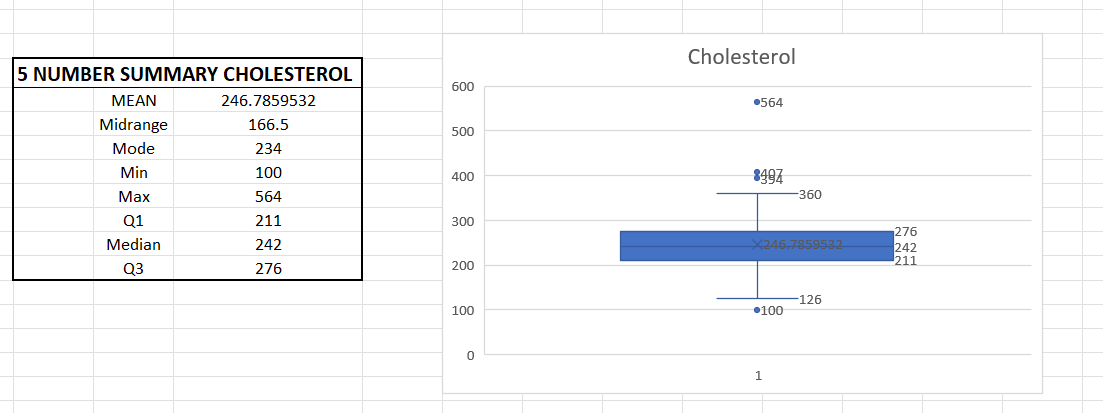


Hình .. Boxplot age

***Trestbps*** 

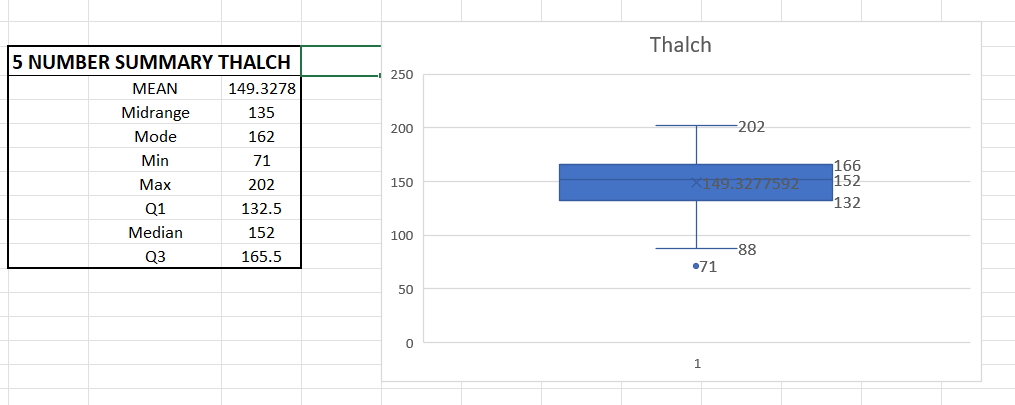
Hình ..Boxplot trestbps

###### **Cholesterol**



Hình .. Boxplot Cholesterol

###### **Thalch**

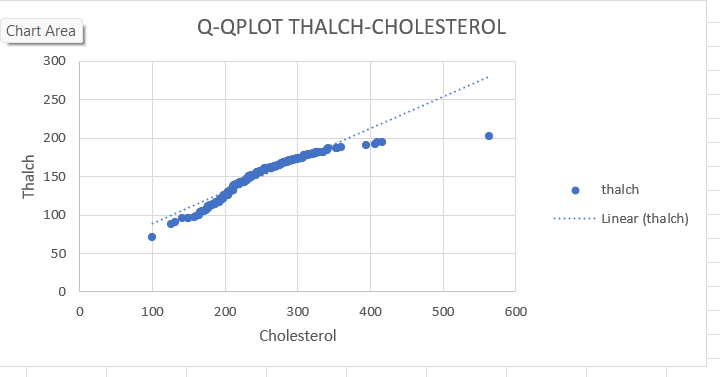


Hình ..Boxplot Thalch

#### Quantile–Quantile Plot trên 2 thuộc tính bất kỳ nhưng có liên quan về ý nghĩa

Để có thể vẽ được biểu đồ Q-Q Plot thì ta cần sắp xếp các datapoint theo thứ tự từ bé đến lớn và xếp hạng các datapoint bằng hàm Rank() trong excel, sau đó ta cần tính giá trị Percentile : (chỉ số xếp hạng của dòng datapoint - 0,5)/ tổng các datapoint.

Ở đây 2 thuộc tính liên quan về ý nghĩa em chọn là: cholesterol và thalch. Vì cholesterol là một trong những chỉ số chính liên quan đến sức khỏe tim mạch. Nó có thể ảnh hưởng đến tốc độ mà trái tim phải đập để cung cấp đủ oxy cho cơ thể, mà được thể hiện qua tần suất nhịp tim cao nhất đạt được (thalach). Mối quan hệ giữa cholesterol và thalach có thể được sử dụng để đánh giá tình trạng sức khỏe tim mạch của một người.

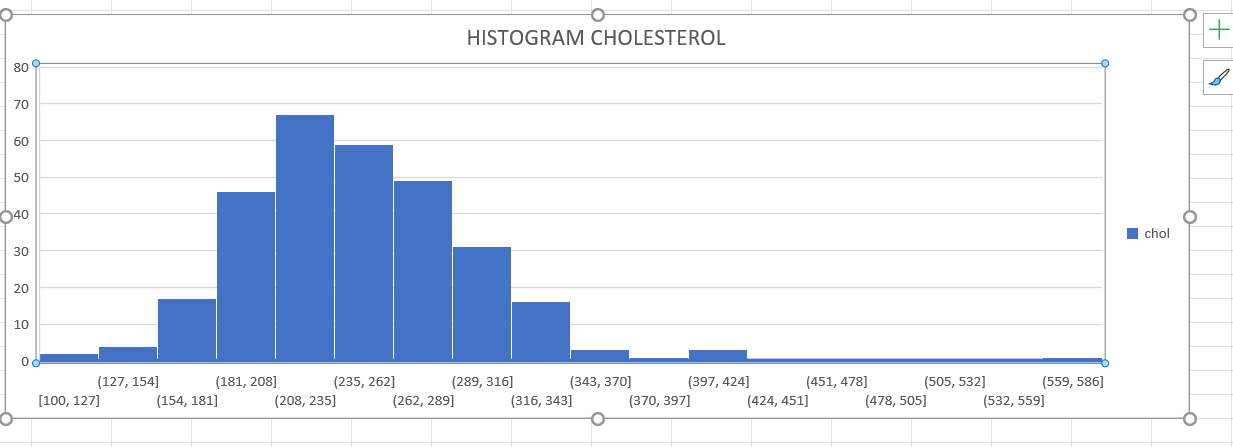


Hình ..Q-Q Plot giữa Thalch và Cholesterol

#### Histogram trên 2 thuộc tính bất kỳ nhưng có liên quan về ý nghĩa

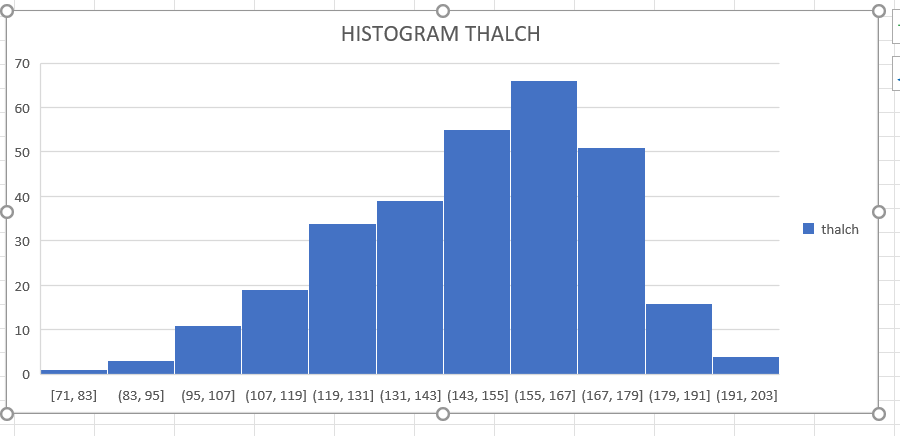
Ở đây 2 thuộc tính liên quan về ý nghĩa em chọn tiếp tục là: cholesterol và thalch.

###### Histogram của Cholesterol



Hình ..Histogram của cholesterol

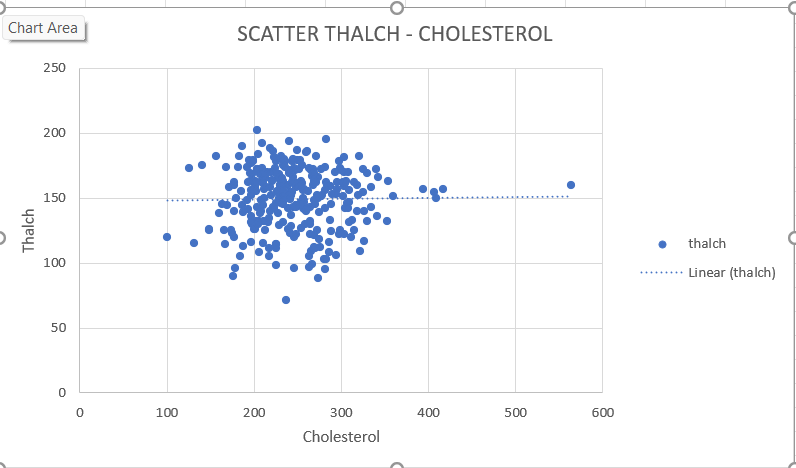
###### Histogram của Thalch



Hình ..Histogram của thalch

#### Scatter trên 2 thuộc tính bất kỳ nhưng có liên quan về ý nghĩa

Ở đây 2 thuộc tính liên quan về ý nghĩa em chọn tiếp tục là: cholesterol và thalch.



Hình ..Scatter thalch-cholesterol

### Nhóm dữ liệu đang có theo một thuộc tính dạng danh nghĩa

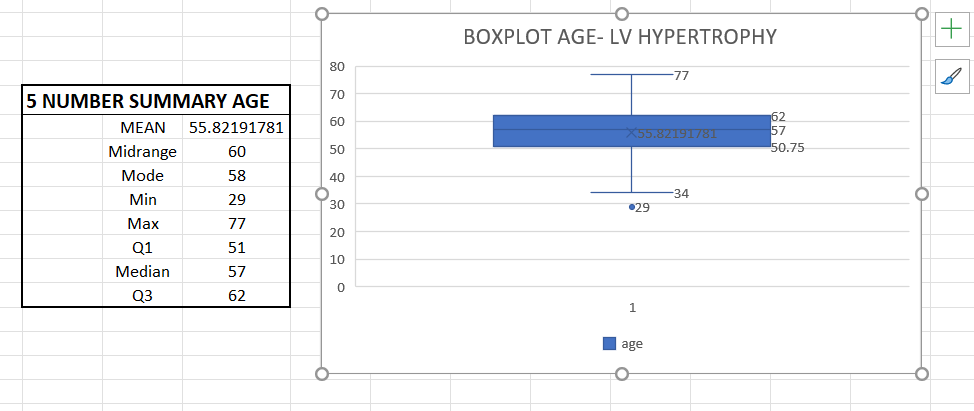
Ở đây em chọn thuộc tính restecg gồm 3 giá trị là:lv hypertrophy, normal, st-t abnormality .

#### Boxplot dựa trên five-number summary

Filter theo field restecg gồm 3 giá trị là: lv hypertrophy, normal, st-t abnormality.

**Filter theo lv hypertrophy**

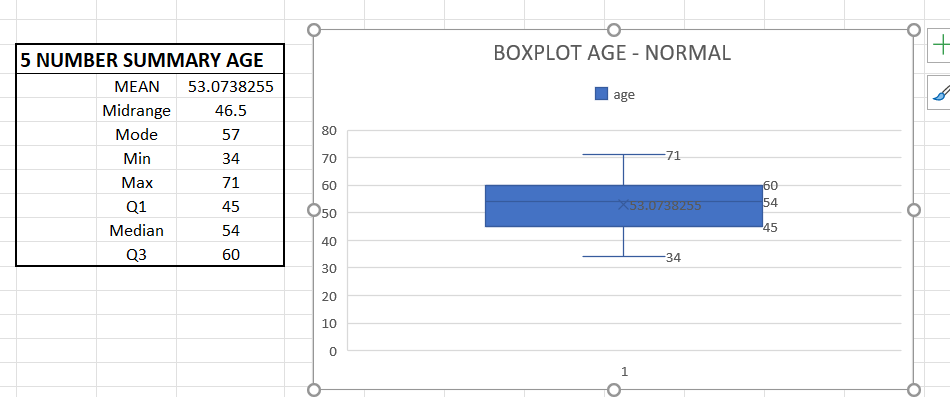
Boxplot của field age sau khi đã filter theo lv hypertrophy



Hình .. Boxplot Age lv hypertrophy

**Filter theo normal**

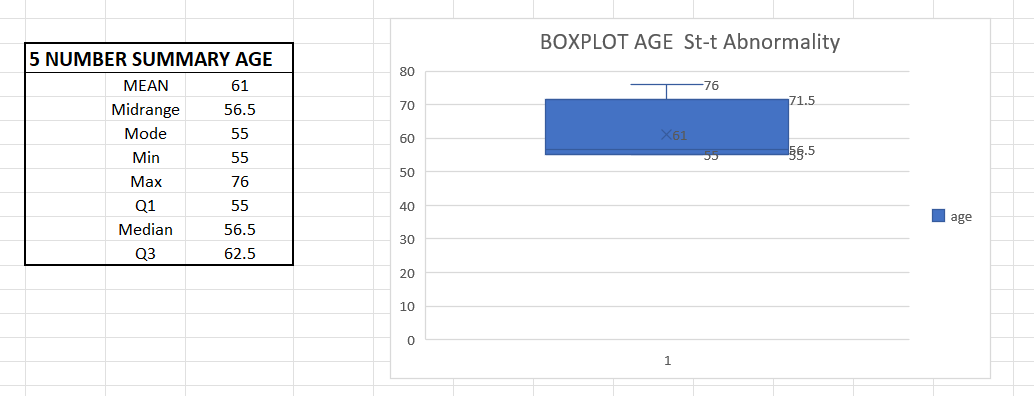
Boxplot của field age sau khi đã filter theo normal



Hình .. Boxplot Age normal

**Filter theo st-t abnormality**

Boxplot của field age sau khi đã filter theo st-t abnormality

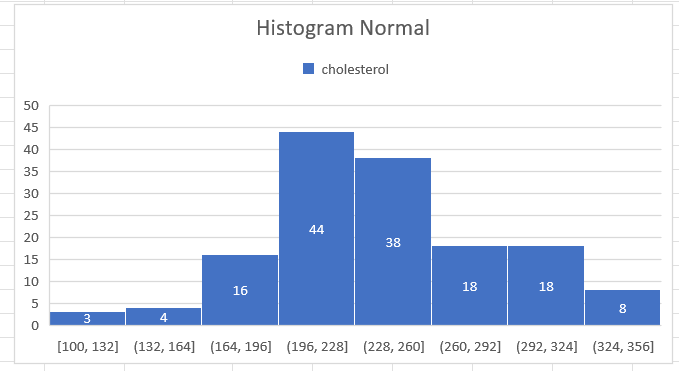


Hình ..Boxplot Age st-t abnormality

#### Histogram trên 2 thuộc tính bất kỳ nhưng có liên quan về ý nghĩa

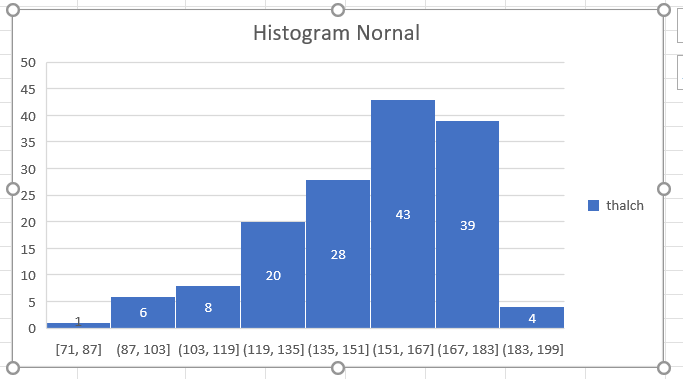
Histogram trên 2 thuộc tính có liên quan về ý nghĩa là cholesterol và thalch đã filter theo normal.

**Cholesterol**



Hình ..Historgram Normal của Cholesterol

**Thalch**



Hình ..Historgram Normal của Thalch

### Đo lường sự tương đồng và khác biệt của dữ liệu bằng 2 cách: ma trận tương quan và độ đo Cosin

• Chọn tối thiểu 4 thuộc tính lần lượt thuộc các dạng dữ liệu sau:

Thuộc tính dạng danh nghĩa (Nominal Attributes) : Restecg

Thuộc tính dạng nhị phân (Binary Attributes):Fbs

Thuộc tính dạng số (Numeric Attributes): Cholesterol

Thuộc tính dạng thứ tự (Ordinal Attributes) : Thal

• Chọn tối thiểu 4 dòng dữ liệu trong CSDL :

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| STT | Restecg | Thal | Fbs | Cholesterol |
| 1 | lv hypertrophy | fixed defect | TRUE | 233 |
| 2 | normal | normal | FALSE | 250 |
| 3 | normal | reversable defect | TRUE | 199 |
| 4 | st-t abnormality | normal | FALSE | 327 |

Hình ..Thuộc tính các dạng

#### Ma trận tương quan

Trước tiên tính ma trận phân sai từng thuộc tính :

*Ma trận phân sai thuộc tính Restecg*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| STT | 1 | 2 | 3 | 4 |
| 1 | 0 |  |  |  |
| 2 | 1 | 0 |  |  |
| 3 | 1 | 0 | 0 |  |
| 4 | 1 | 1 | 1 | 0 |

Hình 2.15.Ma trận phân sai restecg

*Ma trận phân sai thuộc tính Thal*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| STT | 1 | 2 | 3 | 4 |
| 1 | 0 |  |  |  |
| 2 | 1 | 0 |  |  |
| 3 | 0.5 | 0.5 | 0 |  |
| 4 | 1.0 | 0 | 0.5 | 0 |

Hình ..Ma trận phân sai thal

*Ma trận phân sai thuộc tính Fbs*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| STT | 1 | 2 | 3 | 4 |
| 1 | 0 |  |  |  |
| 2 | 1 | 0 |  |  |
| 3 | 0 | 1 | 0 |  |
| 4 | 1 | 0 | 1 | 0 |

Hình 2.17.Ma trận phân sai fbs

*Ma trận phân sai thuộc tính Cholesterol*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| STT | 1 | 2 | 3 | 4 |
| 1 | 0 |  |  |  |
| 2 | 0.17 | 0 |  |  |
| 3 | 0.31 | 0.48 | 0 |  |
| 4 | 0.69 | 0.52 | 1 | 0 |

Hình 2.18.Ma trận phân sai cholesterol

Ma trận sai phân thu được dữ liệu được mô tả bởi bốn thuộc tính của các loại hỗn hợp

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| STT | 1 | 2 | 3 | 4 |
| 1 | 0 |  |  |  |
| 2 | 0.85 | 0 |  |  |
| 3 | 0.65 | 0.83 | 0 |  |
| 4 | 0.13 | 0.71 | 0.79 | 0 |

Kết luận :

Ma trận tương quan cột d(1,4) là giống nhau nhất khoảng 0.85.

Ma trận tương quan cột d(1,2) là khác nhau nhất khoảng 0.13.

#### Độ đo Cosin

Trước tiên chuyển đổi giá trị của cột “restecg”,”thal”,”fbs” thành số:

**Restecg:**

* Lv hypertrophy = 2
* Normal = 0
* St-t abnormality =1

**Thal:**

* Fixed defect =1
* Normal =3
* Reversable defect =2

**Fbs :**

* TRUE = 1
* FALSE = 0

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| STT | Restecg | Thal | Fbs | Cholesterol |
| 1 | 2 | 1 | 1 | 233 |
| 2 | 0 | 3 | 0 | 250 |
| 3 | 0 | 2 | 1 | 199 |
| 4 | 1 | 3 | 0 | 327 |

***Tính độ đo cosine hai cột restecg và thal***

A = [2, 0, 0, 1]

B = [1, 3, 2 ,3]

**Tính tích vô hướng :**

A.B = (2 \* 1) + (0 \* 3) + (0\*2) + (1 \* 3)

= 2 + 0 + 0 + 3 = 5

|A|= =

= 2.236

|B| = =

= 4.8

0.46

**Độ đo cosine giữa hai cột restecg và thal là khoảng 0.46**

**=>** khá khác nhau

***Tính độ đo cosine hai cột restecg và fbs***

A = [2, 0, 0, 1]

B = [1, 0, 1 ,0]

**Tính tích vô hướng** :

A.B = (2 \* 1) + (0 \* 0) + (0 \* 1) + (1 \* 0)

= 2 + 0 + 0 + 0 = 2

|A|= =

= 2.236

|B| = =

= 1.41

0.63

**Độ đo cosine giữa hai cột restecg và fbs là khoảng 0.63**

=> khá giống nhau

***Tính độ đo cosine hai cột restecg và cholesterol***

A = [2, 0, 0, 1]

B = [233, 250, 199 ,327]

**Tính tích vô hướng :**

A.B = (2 \* 233)+(0 \* 250)+(0 \* 199)+(1 \* 327)

=466 + 0 + 0 + 327 = 793

|A|= =

= 2.236

|B| =

=

= 513.14

0.691

**Độ đo cosine giữa hai cột restecg và cholesterol là khoảng 0.691**

=> khá giống nhau

***Tính độ đo cosine hai cột thal và fbs***

A = [1, 3, 2, 3]

B = [1, 0, 1 ,0]

**Tính tích vô hướng :**

A.B = (1 \* 1)+(3 \* 0)+(2 \* 1)+(3 \* 0)

=1 + 0 + 2 + 0 = 3

|A|= =

= 4.8

||B| = =

= 1.41

0.44

**Độ đo cosine giữa hai cột thal và fbs là khoảng 0.44**

=> khá khác nhau

***Tính độ đo cosine hai cột thal và cholesterol***

A = [1, 3, 2, 3]

B = [233, 250, 199 ,327]

**Tính tích vô hướng :**

A.B = (1 \* 233)+(3 \* 250)+(2 \* 199)+(3 \* 327)

=233 + 750 + 398 + 981 = 2362

|A|= =

= 4.8

|B| =

=

= 513.14

0.96

**Độ đo cosine giữa hai cột thal và cholesterol là khoảng 0.96**

=> rất giống nhau

***Tính độ đo cosine hai cột fbs và cholesterol***

A = [1, 0, 1, 0]

B = [233, 250, 199 ,327]

**Tính tích vô hướng :**

A.B = (1 \* 233)+(0 \* 250)+(1 \* 199)+(0 \* 327)

=233 + 0 + 199 + 0 = 432

|A|= =

= 1.41

|B| =

=

= 513.14

0.58

**Độ đo cosine giữa hai cột fbs và cholesterol là khoảng 0.58**

=> khá giống nhau

**Kết luận :**

Độ đo cosine giữa cột thal và cholesterol là giống nhau nhất khoảng 0.96

Độ đo cosine giữa cột restecg và thal là khác nhau nhất là 0.44

**Thực hiện so sánh kết quả của 2 cách đo lường sự tương đồng và khác biệtcủa dữ liệu**

**Phương pháp ma trận tương quan**

* Cho biết mức độ liên quan tuyến tính giữa các cột trong dữ liệu . Số 0 chỉ ra rằng không có mối tương quan tuyến tính giữa các biến . Ví dụ : Restecg và Cholesterol có mối quan hệ giống nhau nhất .

**Phương pháp ma trận cosine**

* Ma trận cosine đo lường sự tương tự về hướng giữa các vecto biểu diễn các cột . Ví dụ Restecg vs thal có độ giống nhau nhất .

So Sánh

Mối tương quan giữa Restecg

+ Tương quan với Thal là 0.85 , fbs là 0.31, Cholesterol là 0.69 .

+ Cosine thal 0.46 ,fbs 0.63 , cholesterol 0.691

Mối tương quan giữa thal

+ Tương quan với fbs là 0.83 , cholesterol là 0.71

+ Cosine fbs là 0.44 , cholesterol là 0.96

Mối tương quan giữa fbs

+ Tương quan với cholesterol 0.79

+ Cosine với cholesterol là 0.58

## Thực hiện khai thác dữ liệu

### Sử dụng các phương pháp khai phá dữ liệu

#### Phương pháp phân loại Navie Bayesian

Em đã tạo thêm cột suy tim với 2 thuộc tính yes và no dựa trên cột num . Dưới đây là những cột liên quan đến có suy tim hay không .

X= (age=63, sex=male, chest pain= aptypical, trestbps=150, cholesterol=234 , fbs=TRUE, restecg = st-t-abnormality , thalch =120, exang=FALSE , oldpeak=2.6, slope = upsloping , ca=2,thal= reversable )

**Tính toán dựa trên các bộ dữ liệu huấn luyện:**

P(suytim=No) = 216/299 = 0.722

P(suytim=Yes) = 83/299 =0.28

**Ta tính các xác suất có điều kiện sau :**

P(age=63) | (suytim = no) = 5/216

P(age=63) | (suytim =Yes) =5/83

P(sex=male) | (suytim = no)=136/216

P(sex=male) | (suytim = yes) =67/83

P(cp=aptypical) | (suytim = no) = 46/216

P(cp=aptypical) | (suytim = yes) = 3/83

P(cp=trestbps) | (suytim = yes) = 5/83

P(cp=trestbps) | (suytim = no) = 11/216

P(cp=cholesterol) | (suytim = no) = 1/216

P(cp=cholesterol) | (suytim = yes) =4/83

P(fbs=true) | (suytim = no) = 26/216

P(fbs=true) | (suytim = yes) = 17/216

P(restecg=st-t) | (suytim = no) = 1/216

P(restecg=st-t) | (suytim = yes) = 3/83

P(thalch=120) | (suytim = no) =1/216

P(thalch=120) | (suytim = yes)=3/83

P(exang=false) | (suytim = no) = 167/216

P(exang=false) | (suytim = yes) = 33/83

P(oldpeak=2.6) | (suytim = no) = 2/216

P(oldpeak=2.6) | (suytim = yes) = 4/83

P(slope=upsloping) | (suytim = no) = 125/216

P(slope=upsloping) | (suytim = yes) =14/83

P(ca=2) | (suytim = no) = 13/216

P(ca=2) | (suytim = yes) = 26/83

P(thal=reverable defect) | (suytim = no) =58/216

P(thal=reverable defect) | (suytim = yes) =59/83

Sử dụng những xác suất này, ta có được :

P(X| Suytim=no) = P(age=63) | (suytim = no) \* P(sex=male) | (suytim = no) \*

P(cp=aptypical) | (suytim = no) \* P(cp=trestbps) | (suytim = no) \* P(cp=cholesterol) | (suytim = no) \* P(fbs=true) | (suytim = no) \* P(restecg=st-t) | (suytim = no) \* P(thalch=120) | (suytim = no) \* P(exang=false) | (suytim = no) \* P(oldpeak=2.6) | (suytim = no) \* P(slope=upsloping) | (suytim = no) \* P(ca=2) | (suytim = no) \* P(thal=reverable defect) | (suytim = no) = 7.5483x10^-17

P(X| Suytim=yes) = P(age=63) | (suytim = yes) \* P(sex=male) | (suytim = yes) \*

P(cp=aptypical) | (suytim = yes) \* P(cp=trestbps) | (suytim = yes) \* P(cp=cholesterol) | (suytim = yes) \* P(fbs=true) | (suytim = yes) \* P(restecg=st-t) | (suytim = yes) \* P(thalch=120) | (suytim = yes) \* P(exang=false) | (suytim = yes) \* P(oldpeak=2.6) | (suytim = yes) \* P(slope=upsloping) | (suytim = yes) \* P(ca=2) | (suytim = yes) \* P(thal=reverable defect) | (suytim = yes) = 1.05757x10^-12

P(X) = P(X| Suytim=No) \* P(suytim=No) + P(X| Suytim=yes) \* P(suytim=Yes) =

2.93628x10^-13

P(Yes | X) = P(X| Suytim=Yes) \* P(suytim=Yes) / P(X) = 0.9998

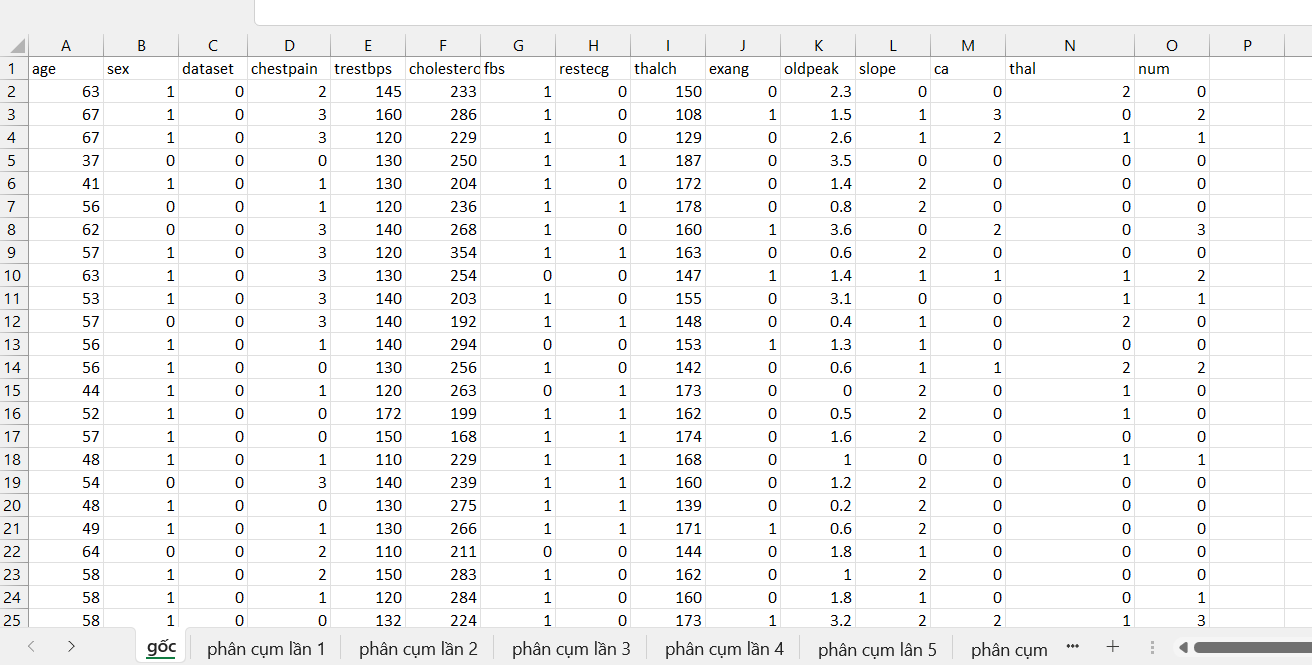
P(No | X) = P(X| Suytim=No) \* P(suytim=No) / P(X) = 0.0002

=> Do đó , trình phân loại Navie Bayesian dự đoán suytim= 1

=> Yes cho bộ X.

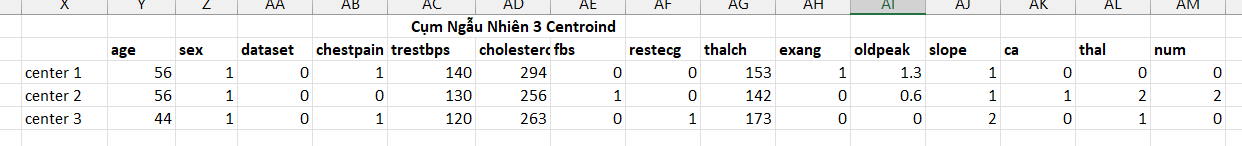
#### Phương pháp phân cụm K-mean

Phương pháp phân cụm K-mean bước đầu chuyển sang thành dữ liệu số



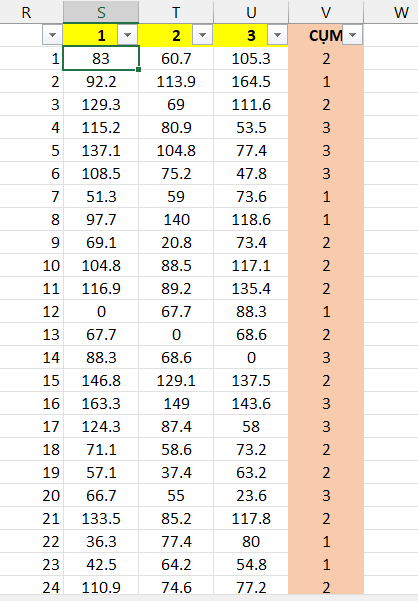
Hình ..Chuyển sang dữ liệu số

Chọn cụm ngẫu nhiên 3 centroid :

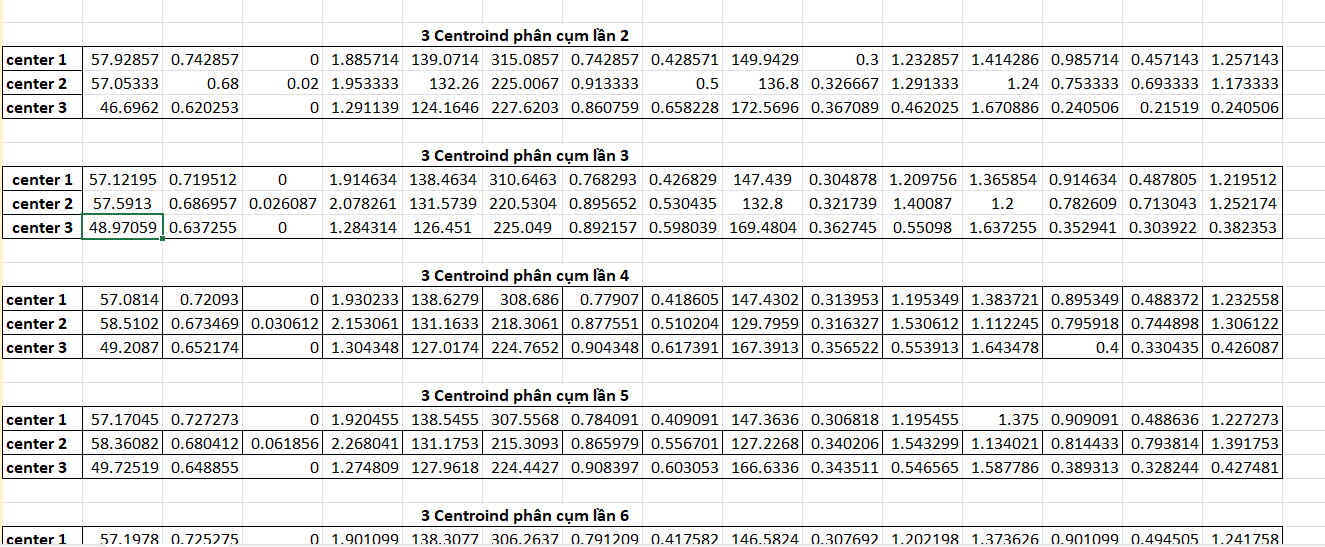


Hình ..Cụm ngẫu nhiên

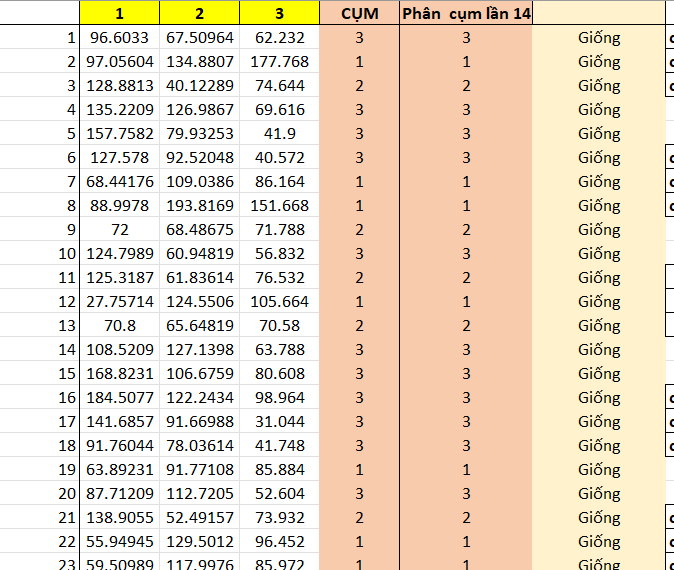
Tính từng cụm , K=3



Hình 2.21.Cụm K3



Hình ..Các cụm lần lượt



Hình ..So sánh các cụm

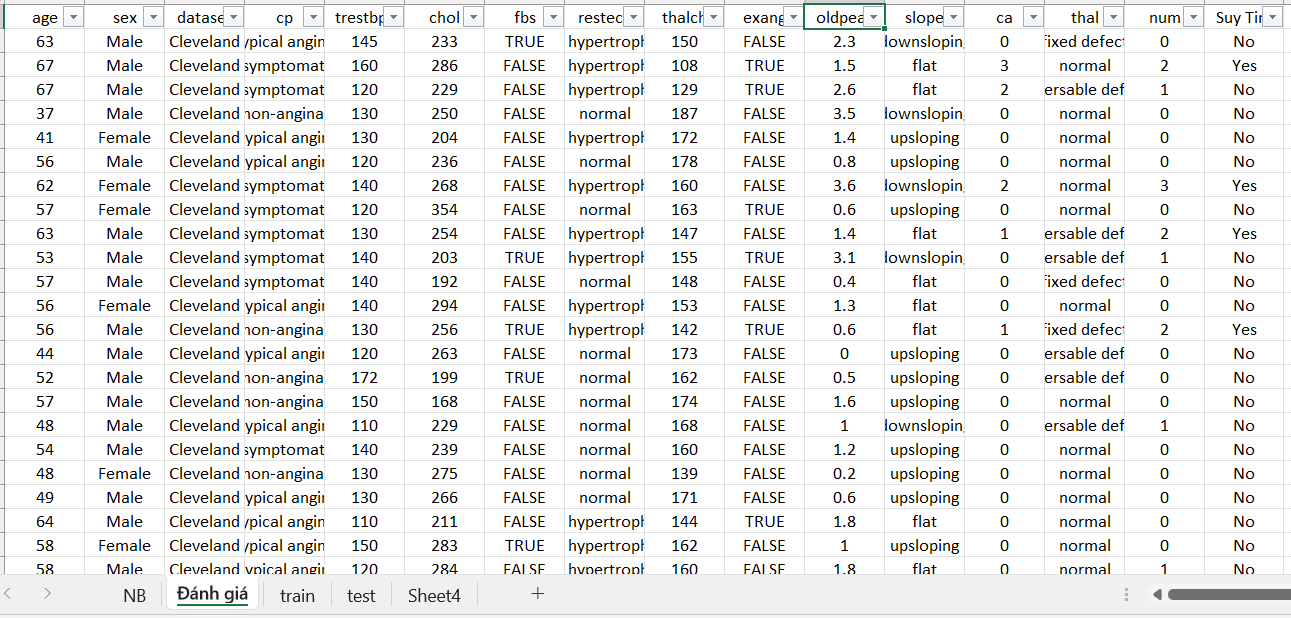
Ta lần lượt phân từng cụm cho tới phân cụm lần 15 rồi so sánh sự khác nhau.Sau khi so sánh thì không có giá trị nào khác nhau ở cụm lần 15

### Thực hiện đánh giá các mẫu thu được bằng các phương pháp

#### Phương pháp phân loại

Số liệu để đánh giá hiệu suất phân loại (Metrics for Evaluating Classifier Performance)

Chọn những cột age, trestbps, cholesterol, thalch, exang, oldpeak, sex, ca, thal ,suy tim.



Hình ..Các cột dùng đánh giá

Chuyển đổi các cột số thành danh nghĩa

Age :

+ Elderly từ 60 đến 80 tuổi

+ Middle-aged từ 35 đến 50 tuổi

+ Còn lại là young

Trestbps:

+ Huyết áp thấp (low) : Trestbps < 90 mmHg

+ Huyết áp bình thường (normal) : 90 mmHg ≤ Trestbps < 120 mmHg

+ Huyết áp cao : 120 mmHg trở lên

Cholesterol:

+ Mức cholesterol thấp (low): < 125 mg/dL

+ Mức cholesterol bình thường (normal) : 125 mg/dL ≤ Cholesterol < 200 mg/dL

+ Mức cholesterol cao (high): Total Cholesterol ≥ 200 mg/dL

Thalch :

Công thức tính nhịp tim tối đa :

Nhịp tim tối đa : 220- tuổi

+ Low : Nhịp tim tối đa thấp hơn 50% của nhịp tim tối đa dự kiến.

+ High : Nhịp tim cao hơn 85% của nhịp tim tối đa dự kiến.

+ Normal : Nhịp tim nằm trong khoảng 50-85% của nhịp tim tối đa dự kiến.

Oldpeak :

+ Low : ST depression < 1 mm

+ Normal : ST depression 1 đến 2 mm

+ High : ST depression > 2 mm

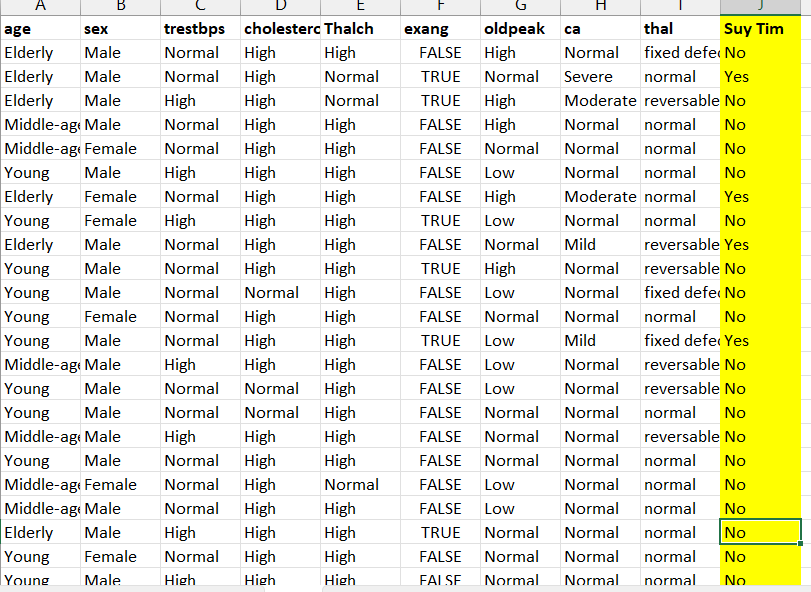
Ca:

+ Normal (bình thường) : 0

+ Mild (nhẹ): 1

+ Moderate (Vừa): 2

+ Severe (Nặng) : 3



Hình ..Đã chuyển đổi dữ liệu

**Chia tập dữ liệu :**

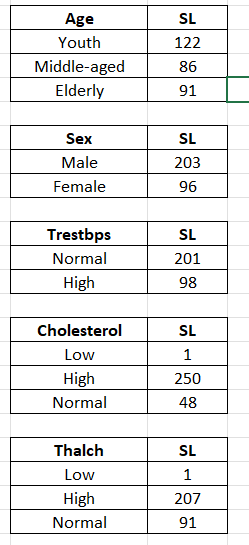
Tập dữ liệu ban đầu được chia thành hai phần:

+ 80% dữ liệu đầu tiên được sử dụng làm tập huấn luyện (train set) gồm 239 dòng

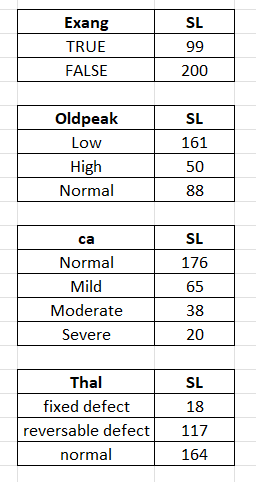
+ 20% dữ liệu còn lại được sử dụng làm tập kiểm tra (test set) gồm 60 dòng

**Tính toán xác suất có điều kiện P(X|Dead Event):**

Tiếp theo đếm từng thuộc tính có trong cột

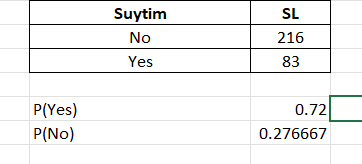


Hình ..Đếm thuộc tính (1)



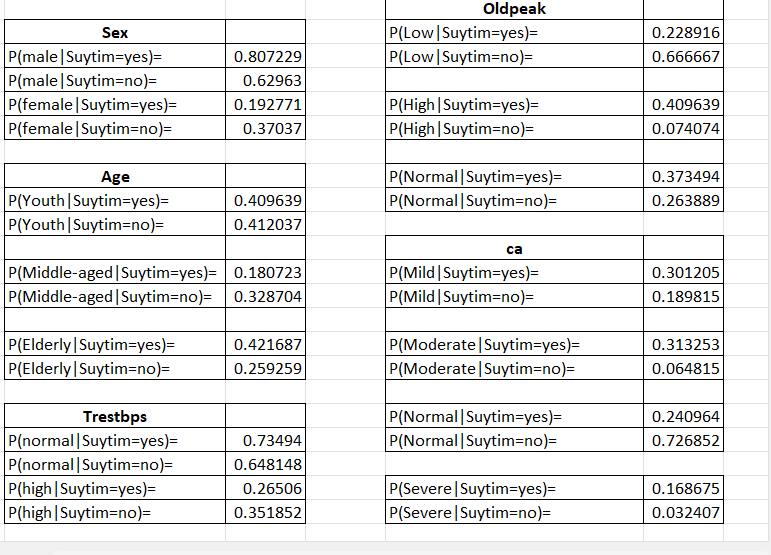
Hình ..Đếm thuộc tính (2)

Tính (P|Yes) và (P|No) của cột suy tim

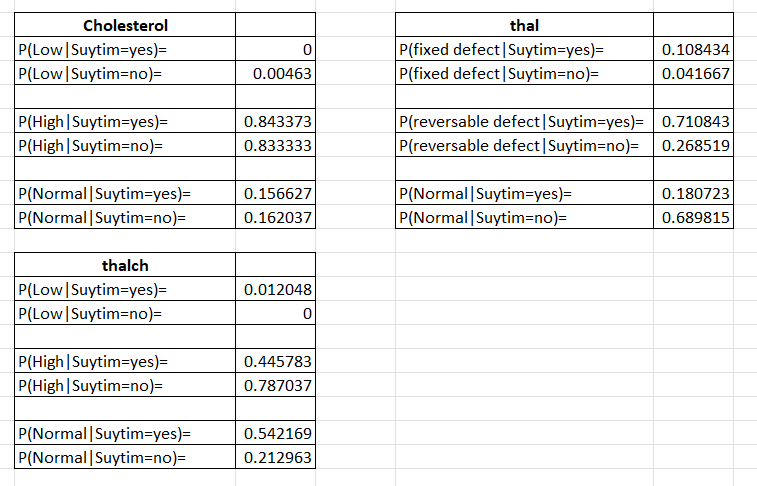


Hình .. Tính (P|Yes) và (P|No)

Tính tập huấn luyện (train set), đã tính toán xác suất có điều kiện P(X|SuyTim) cho từng đặc trưng đã chọn.



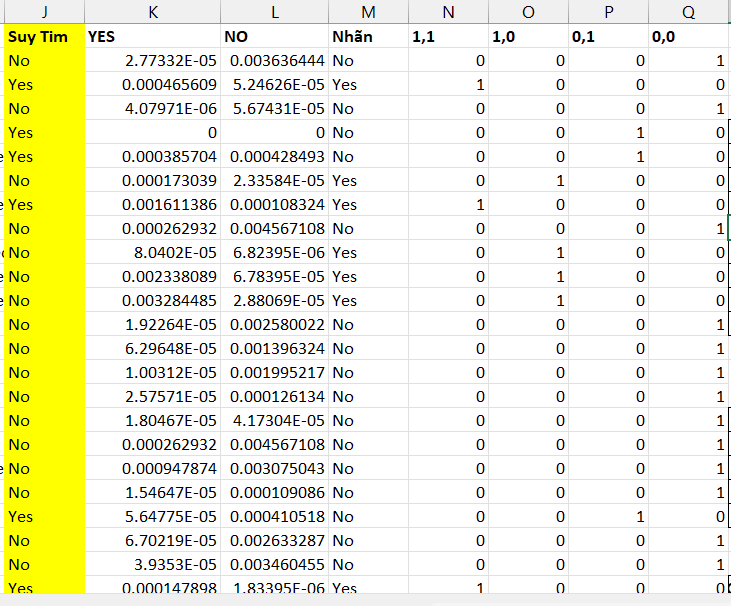
Hình ..Xác xuất dữ liệu (1)



Hình ..Xác xuất dữ liệu (1)

**Dự đoán trên tập kiểm tra:**

Dựa vào các xác suất có điều kiện đã tính toán từ tập huấn luyện (train set),tiếp tục thực hiện dự đoán SuyTim trên tập kiểm tra (test set).

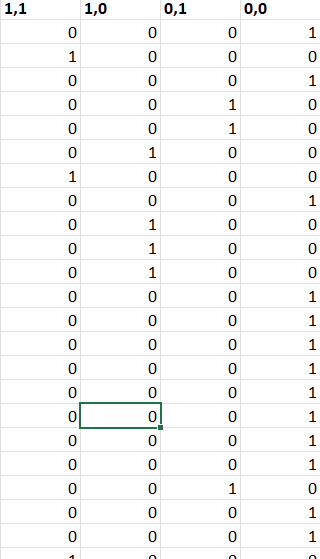


Hình ..Test Set (1)

**So sánh với tập dữ liệu ban đầu :**

Sau khi dự đoán SuyTim cho các dòng trong tập kiểm tra, so sánh các dự đoán này với giá trị thực tế của SuyTim trong tập dữ liệu ban đầu. Từ đó, xác định các giá trị:

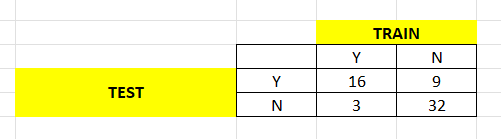
* True Positives (TP)
* False Negatives (FN)
* False Positives (FP)
* True Negatives (TN)



Hình 2.32.Xác định giá trị

Vẽ Confusion Matrix:

Sau khi có được các giá trị True Positives (TP), False Negatives (FN), False Positives (FP), và True Negatives (TN), đã vẽ được Confusion Matrix như sau

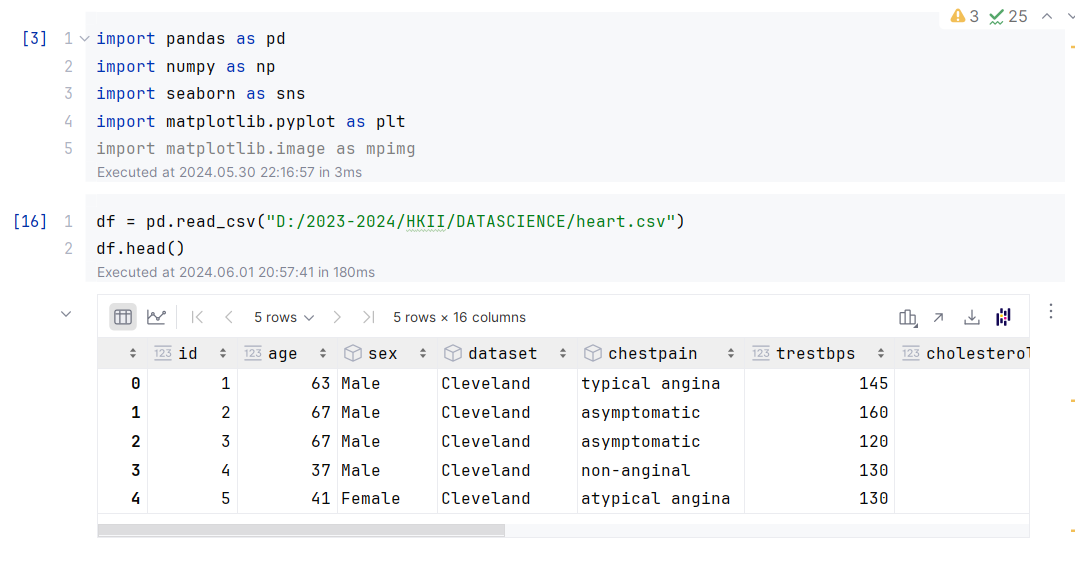


Hình ..Vẽ Confusion Matrix

# PHÂN TÍCH – THỐNG KÊ BẰNG CÔNG CỤ PYTHON TRÊN CƠ SỞ DỮ LIỆU ĐÃ CHỌN

## Thực hiện lại phần 1 bằng Python

Đọc dữ liệu từ file csv và import thư viện cần dùng



Hình ..Đọc dữ liệu

### Giá trị null

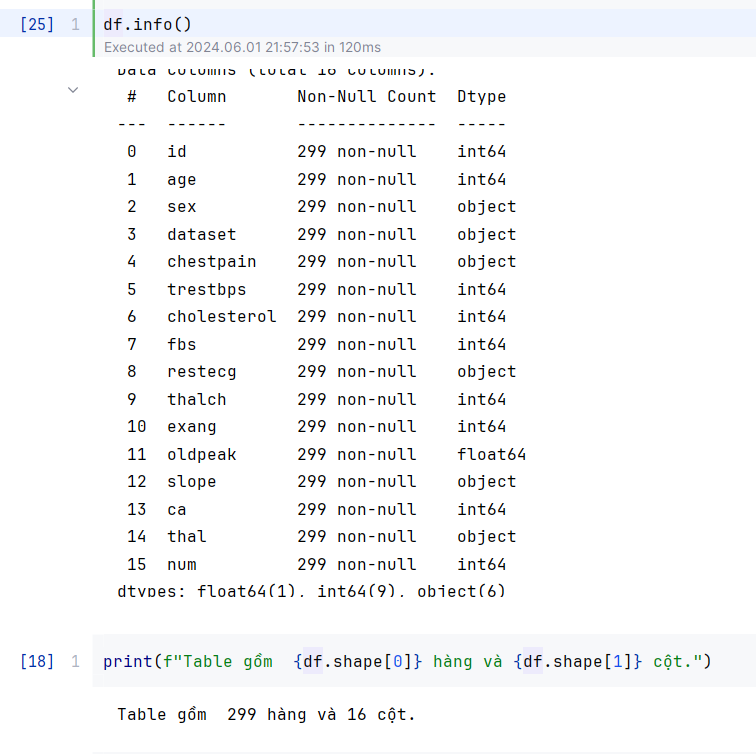
Table gồm 299 hàng và 16 cột .

- Các thuộc tính với 299 giá trị không null (non-null) tức data set này không bị thiếu .

+ float64 có 1 cột : kiểu dữ liệu cột là số thực 64-bit . Điều này có nghĩa là các giá trị trong cột oldpeak có thể là số thực và bao gồm số nguyên.

+ int64 có 9 cột : kiểu dữ liệu cột là số nguyên 64-bit gồm cột id ,age , trestbps, cholesterol , fbs, thalch, exang, ca, num .

+ object có 6 cột : kiểu dữ liệu chứa các giá trị đối tượng. Thường chứa các chuỗi ký tự (string) gồm các cột sex, origin , chestpain , restecg , slope, thal.



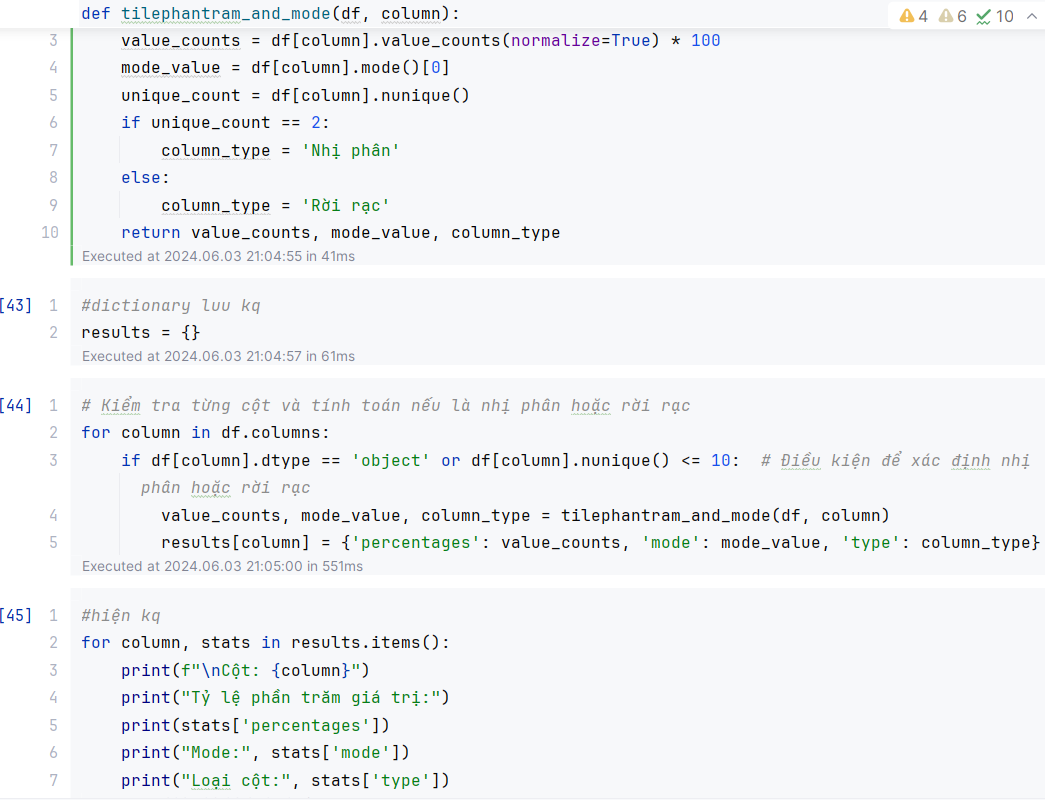
Hình ..Thuộc tính các cột

### Kiểu dữ liệu của các fields

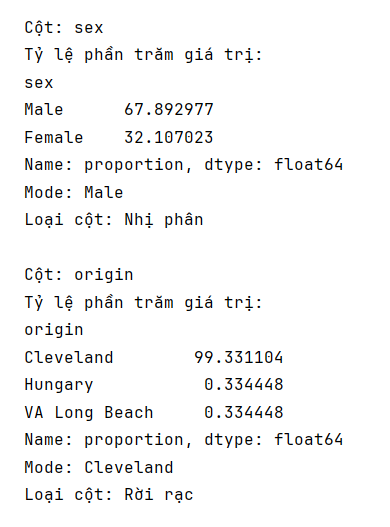
- Kiểu nhị phân hay rời rạc

- Tỷ lệ phần tram của từng giá trị

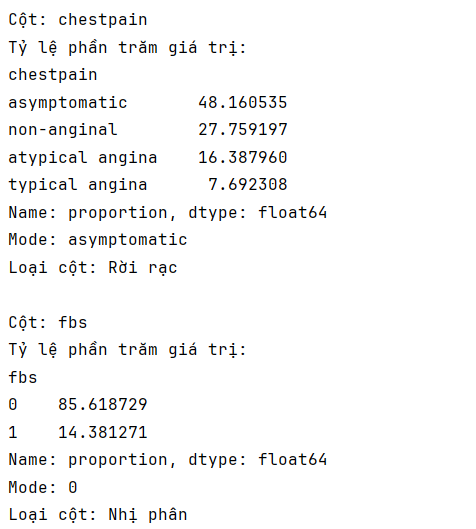
- Giá trị của mode.



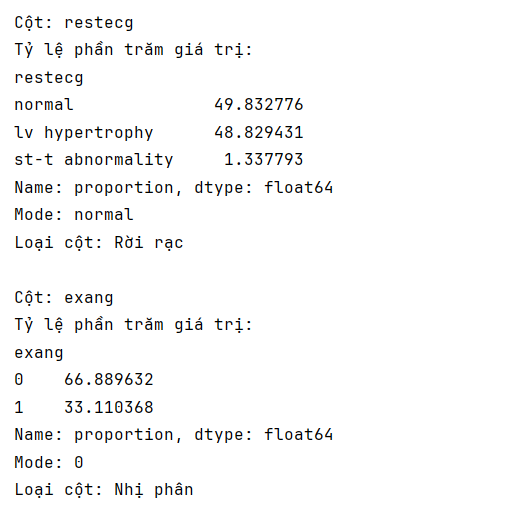
Hình ..Code kiểu dữ liệu



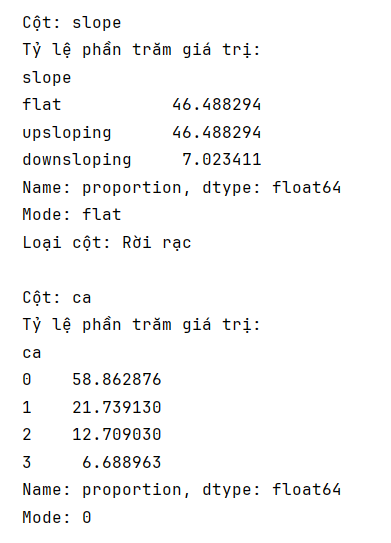
Hình ..Kiểu dữ liệu sex, origin



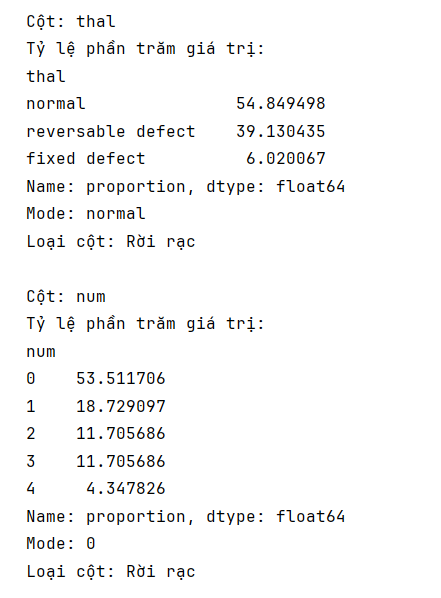
Hình ..Kiểu dữ liệu chestpain và fbs



Hình ..Kiểu dữ liệu restecg và exang



Hình ..Kiểu dữ liệu slope và ca



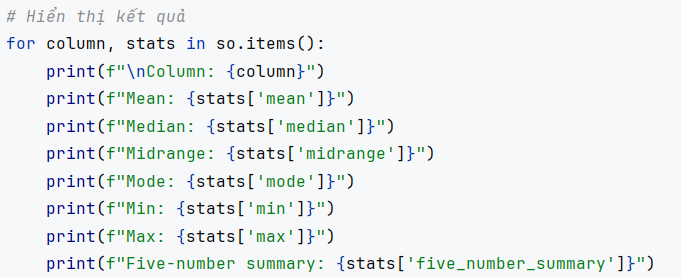
Hình ..Kiểu dữ liệu thal và num

### Kiểu dữ liệu thuộc tính số

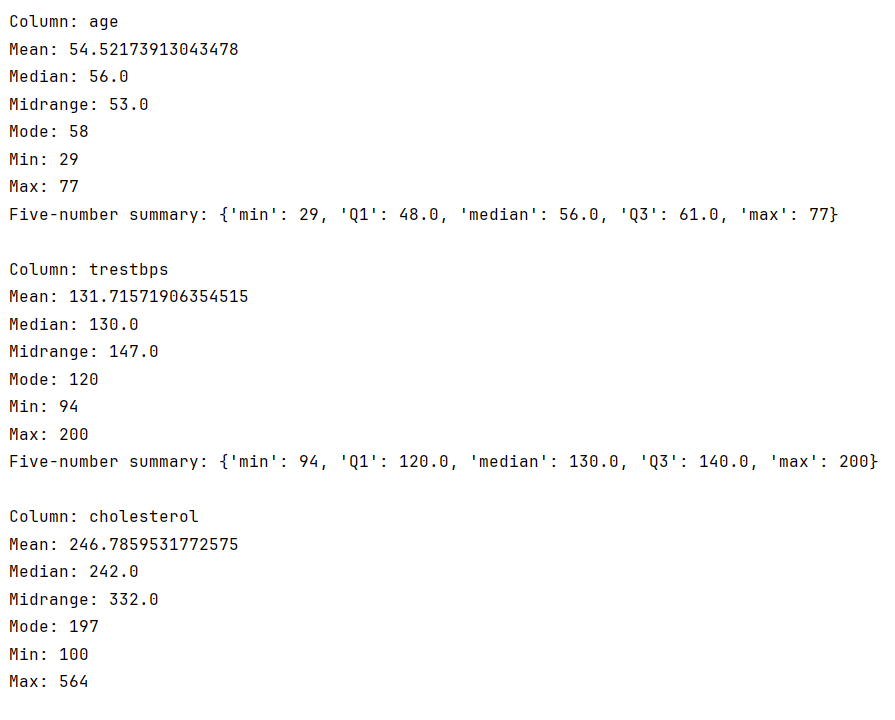
Tính giá trị mean , median, midrange, mode, min, max, five-number summary



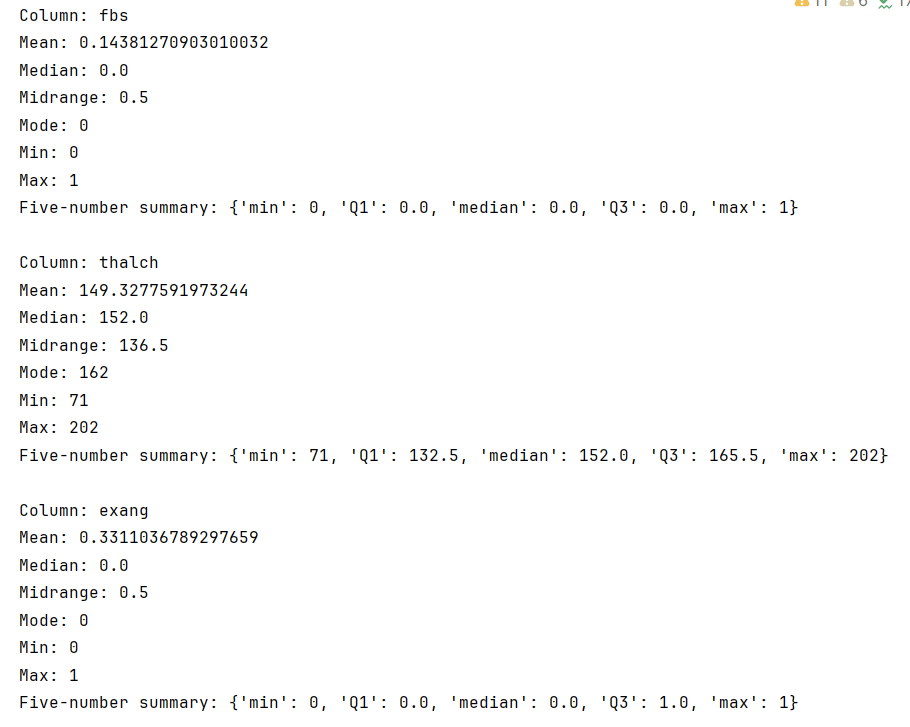
Hình ..Tính giá trị theo yêu cầu



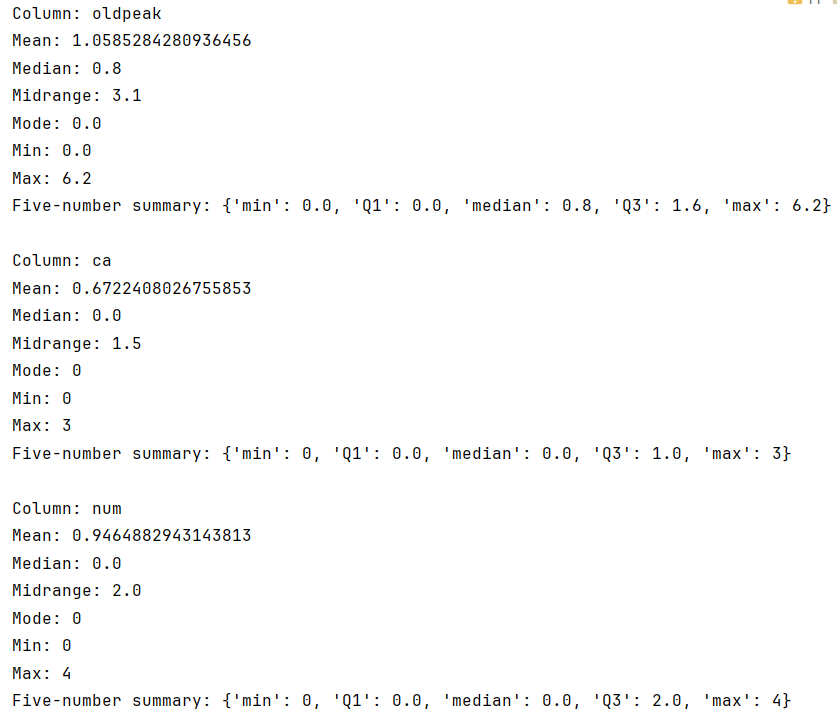
Hình ..Hiện thị kết quả



Hình ..Five-number summary(1)



Hình ..Five-number summary(2)



Hình ..Five-number summary(3)

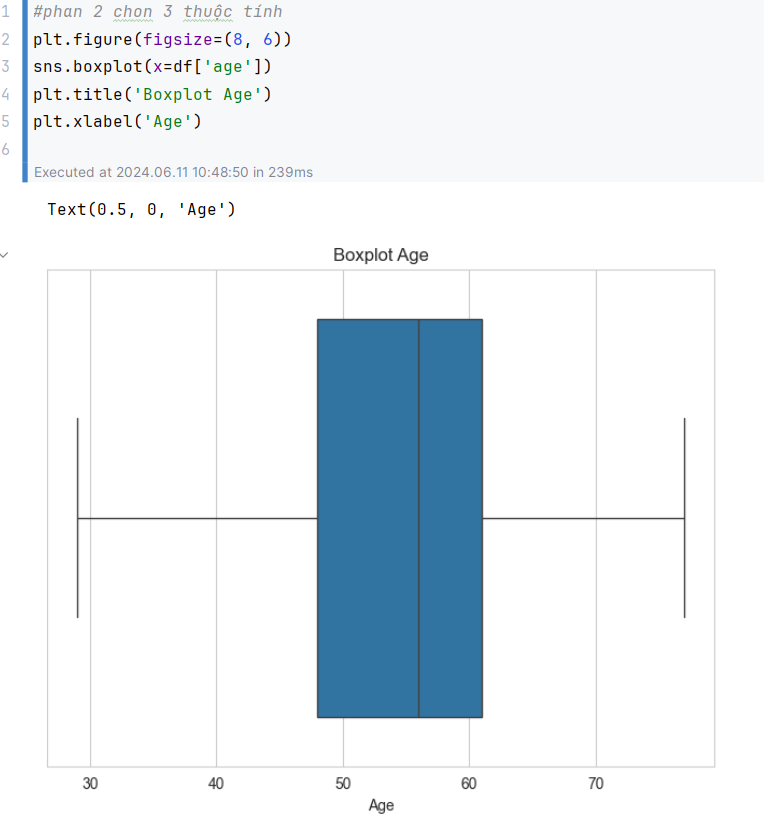
## Phân tích – Thống kê

### Chọn 3 thuộc tính để vẽ các đồ thị

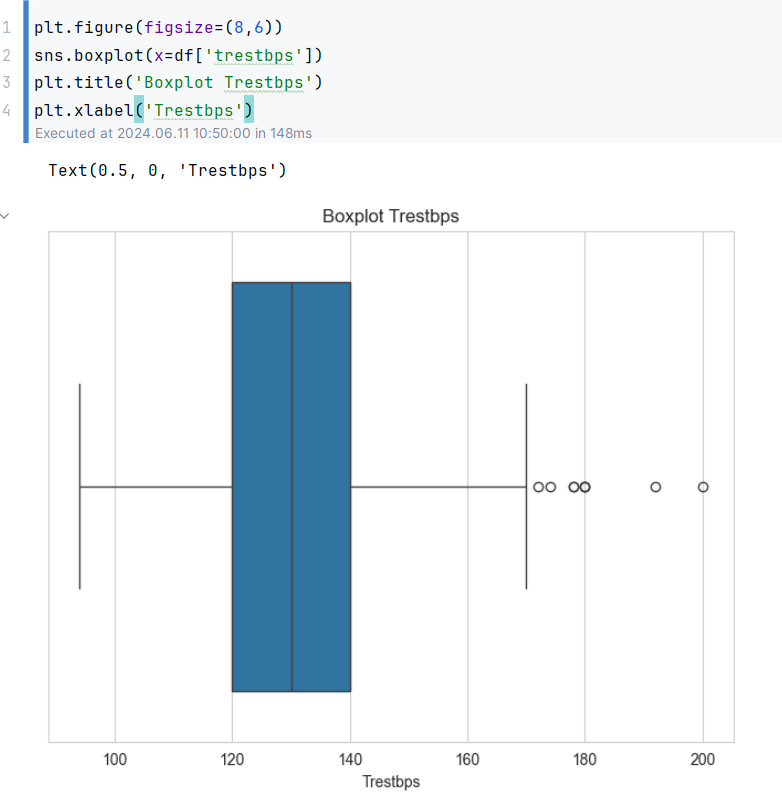
- 4 thuộc tính em chọn để vẽ các đồ thị là: age, trestbps , cholesterol và thalch

#### Boxplot dựa trên five-number summary

###### **Age**

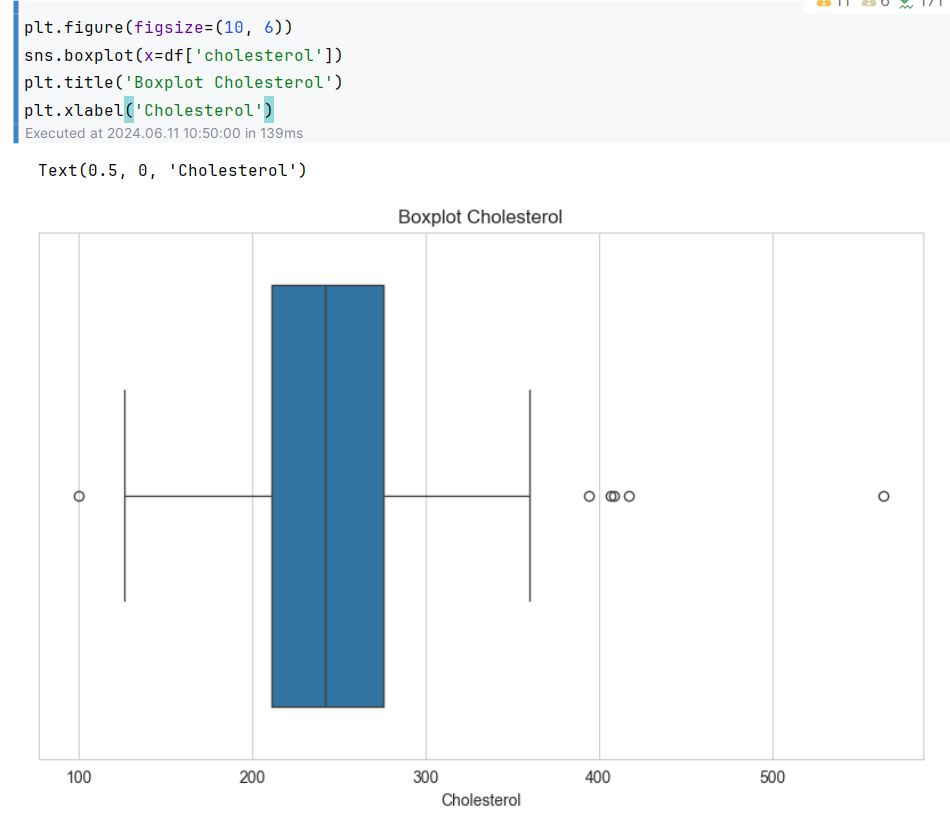


Hình .. Boxplot age

***Trestbps ***

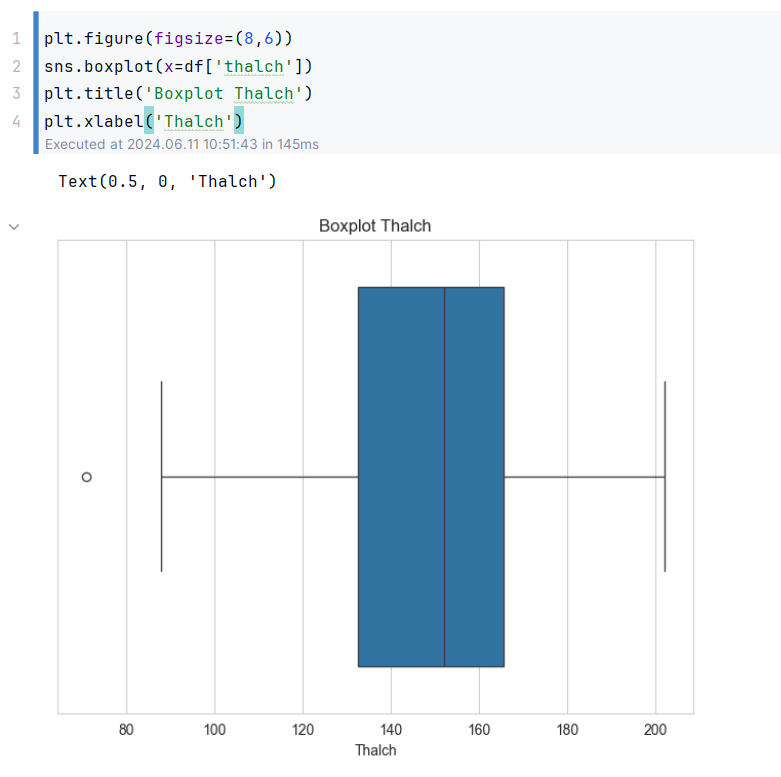
Hình ..Boxplot trestbps

###### **Cholesterol**



Hình .. Boxplot Cholesterol

###### **Thalch**



Hình ..Boxplot Thalch

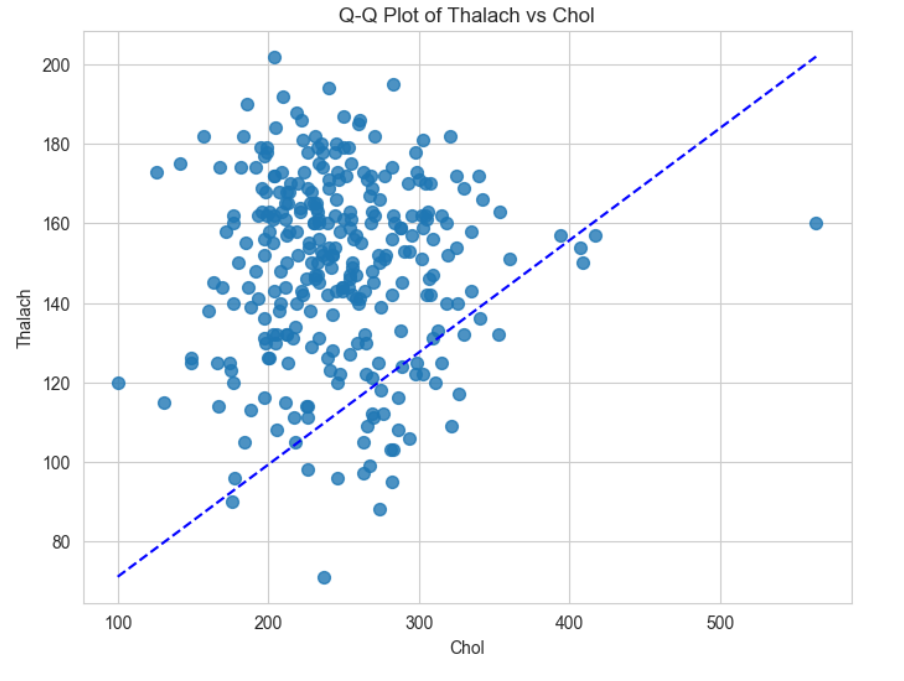
#### Quantile–Quantile Plot trên 2 thuộc tính bất kỳ nhưng có liên quan về ý nghĩa

Để có thể vẽ được biểu đồ Q-Q Plot thì ta cần sắp xếp các datapoint theo thứ tự từ bé đến lớn và xếp hạng các datapoint bằng hàm Rank() trong excel, sau đó ta cần tính giá trị Percentile : (chỉ số xếp hạng của dòng datapoint - 0,5)/ tổng các datapoint.

Ở đây 2 thuộc tính liên quan về ý nghĩa em chọn là: cholesterol và thalch. Vì cholesterol là một trong những chỉ số chính liên quan đến sức khỏe tim mạch. Nó có thể ảnh hưởng đến tốc độ mà trái tim phải đập để cung cấp đủ oxy cho cơ thể, mà được thể hiện qua tần suất nhịp tim cao nhất đạt được (thalach). Mối quan hệ giữa cholesterol và thalach có thể được sử dụng để đánh giá tình trạng sức khỏe tim mạch của một người.



Hình ..Code Q-Q Plot giữa Thalch và Cholesterol

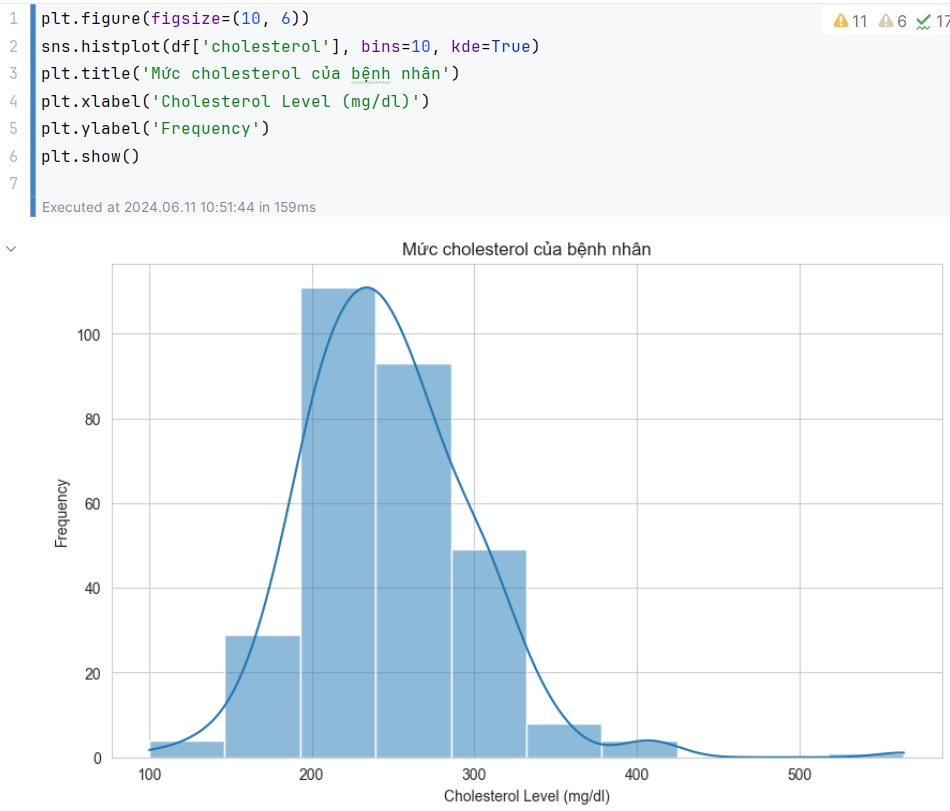


Hình ..Q-Q Plot giữa Thalch và Cholesterol

#### Histogram trên 2 thuộc tính bất kỳ nhưng có liên quan về ý nghĩa

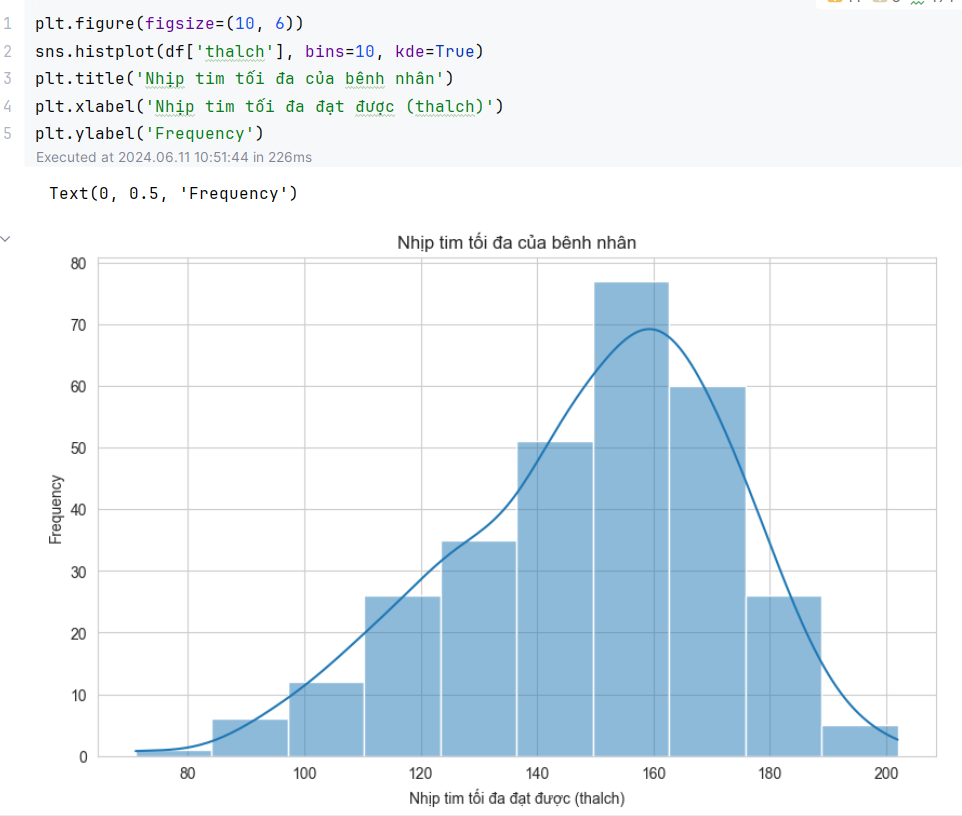
Ở đây 2 thuộc tính liên quan về ý nghĩa em chọn tiếp tục là: cholesterol và thalch.

###### Histogram của Cholesterol



Hình ..Histogram của cholesterol

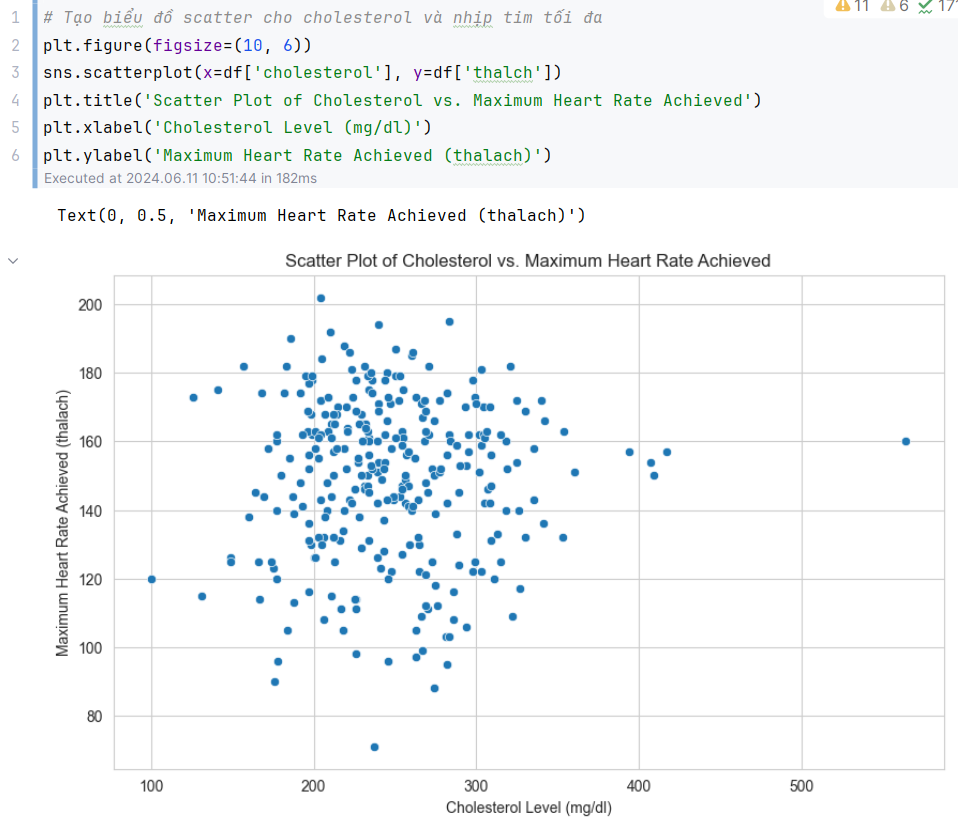
###### Histogram của Thalch



Hình ..Histogram của thalch

#### Scatter trên 2 thuộc tính bất kỳ nhưng có liên quan về ý nghĩa

Ở đây 2 thuộc tính liên quan về ý nghĩa em chọn tiếp tục là: cholesterol và thalch.

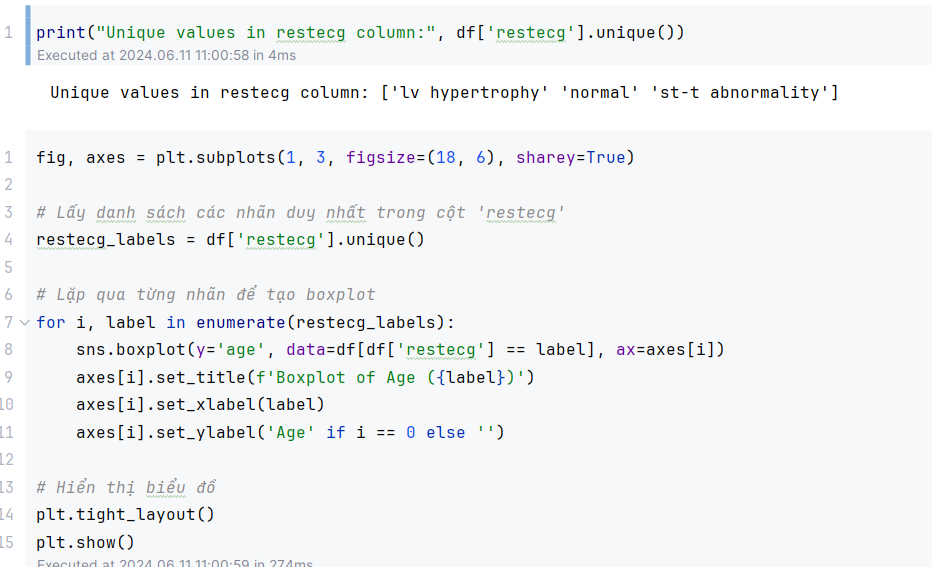


Hình ..Scatter thalch-cholesterol

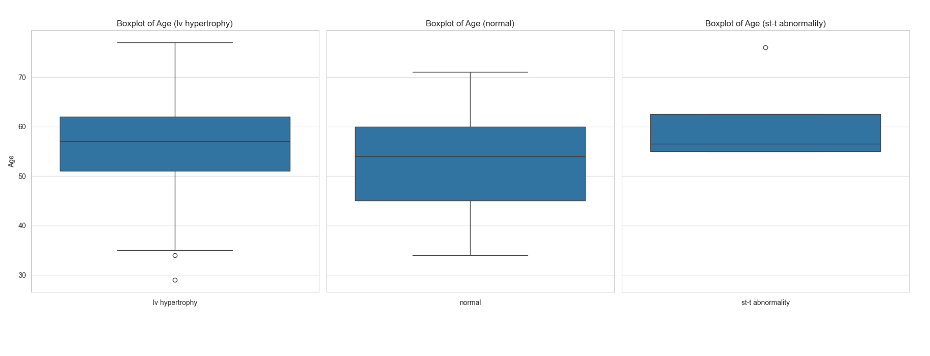
### Nhóm dữ liệu đang có theo một thuộc tính dạng danh nghĩa

Ở đây em chọn thuộc tính restecg gồm 3 giá trị là:lv hypertrophy, normal, st-t abnormality .

#### Boxplot dựa trên five-number summary



Hình ..Code vẽ boxplot thuộc tính danh nghĩa

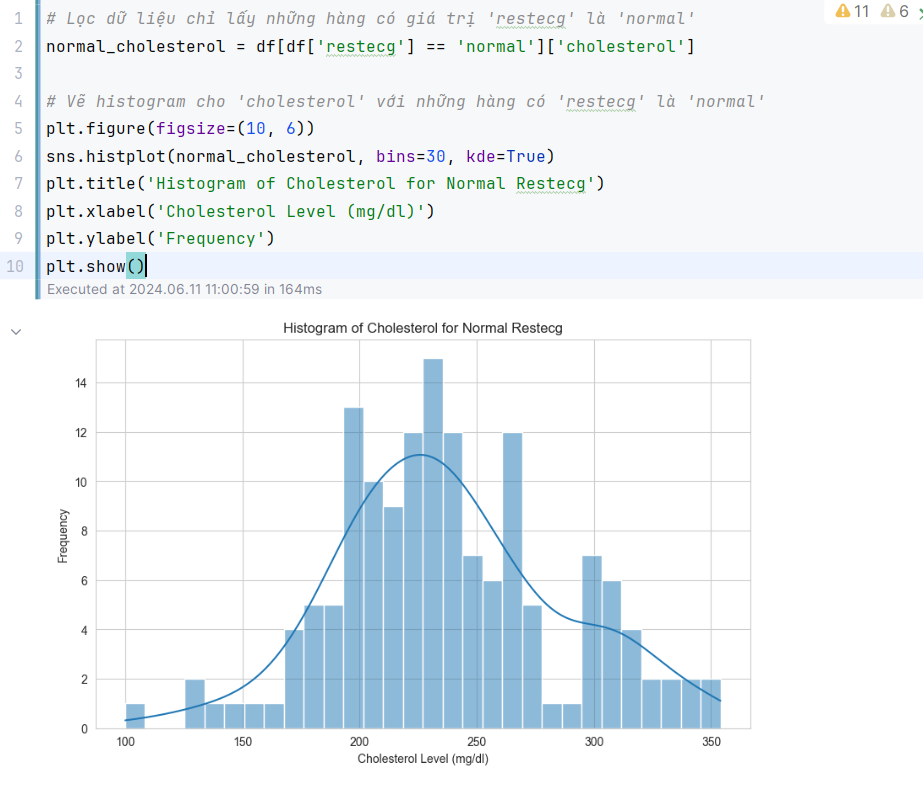


Hình ..Boxplot danh nghĩa

#### Histogram trên 2 thuộc tính bất kỳ nhưng có liên quan về ý nghĩa

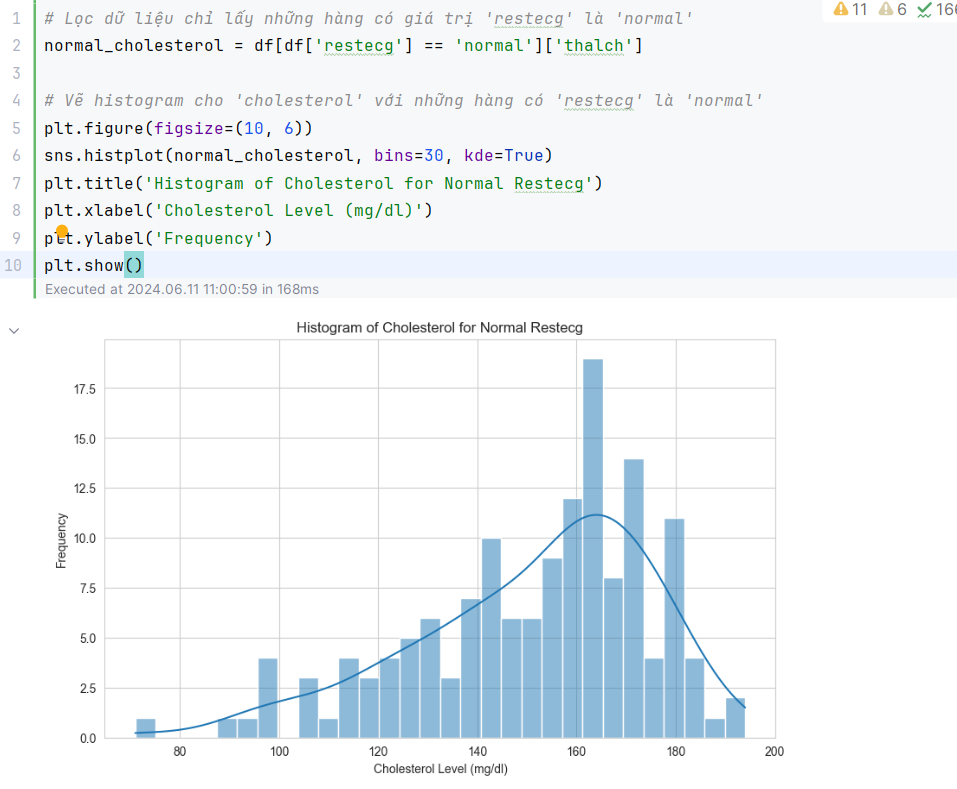
Histogram trên 2 thuộc tính có liên quan về ý nghĩa là cholesterol và thalch đã filter theo normal.

**Cholesterol**



Hình ..Historgram Normal của Cholesterol

**Thalch**



Hình ..Historgram Normal của Thalch

### Đo lường sự tương đồng và khác biệt của dữ liệu bằng 2 cách: ma trận tương quan và độ đo Cosin

• Chọn tối thiểu 4 thuộc tính lần lượt thuộc các dạng dữ liệu sau:

Thuộc tính dạng danh nghĩa (Nominal Attributes) : Restecg

Thuộc tính dạng nhị phân (Binary Attributes):Fbs

Thuộc tính dạng số (Numeric Attributes): Cholesterol

Thuộc tính dạng thứ tự (Ordinal Attributes) : Thal

• Chọn tối thiểu 4 dòng dữ liệu trong CSDL :

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| STT | Restecg | Thal | Fbs | Cholesterol |
| 1 | lv hypertrophy | fixed defect | TRUE | 233 |
| 2 | normal | normal | FALSE | 250 |
| 3 | normal | reversable defect | TRUE | 199 |
| 4 | st-t abnormality | normal | FALSE | 327 |

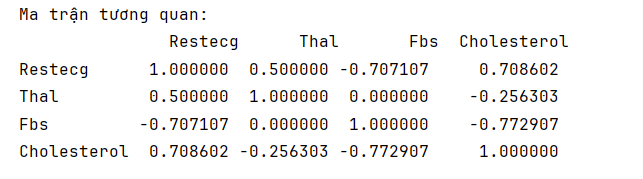
Hình ..Thuộc tính các dạng

#### Ma trận tương quan

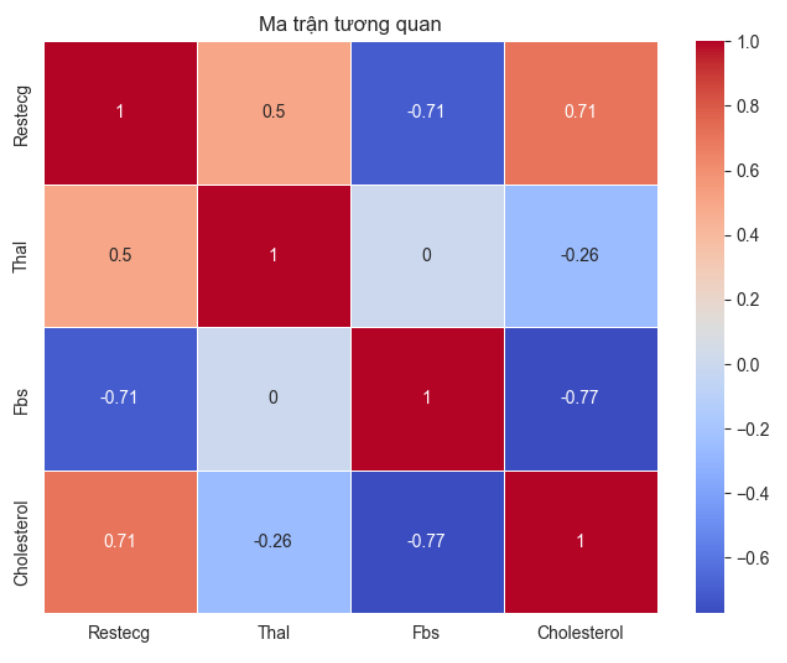
Trước tiên tính ma trận phân sai từng thuộc tính :



Hình 3.28.Code ma trận tương quan



Hình ..In kêt quả ma trận tương quan



Hình ..Biểu đồ heatmap của ma trận tương quan

Kết luận :

* Tương quan dương ở mức trung bình (0.5) cho thấy khi giá trị của Restecg tăng thì giá trị của Thal cũng có xu hướng tăng theo và ngược lại.Restecg và Cholesterol có tương quan dương nhẹ (0.306).
* Tương quan âm mạnh (-0.707) cho thấy khi giá trị của Restecg tăng thì giá trị của Fbs có xu hướng giảm và ngược lại.
* Tương quan dương mạnh (0.709) cho thấy khi giá trị của Restecg tăng thì giá trị của Cholesterol cũng có xu hướng tăng và ngược lại.
* Tương quan bằng 0 cho thấy không có mối quan hệ tuyến tính giữa Thal và Fbs.
* Tương quan âm nhẹ (-0.256) cho thấy khi giá trị của Thal tăng thì giá trị của Cholesterol có xu hướng giảm, nhưng mối quan hệ này không mạnh.
* Tương quan âm mạnh (-0.773) cho thấy khi giá trị của Fbs tăng thì giá trị của Cholesterol có xu hướng giảm và ngược lại.
* Các ô vuông màu càng đỏ đậm biểu thị mối tương quan dương càng mạnh, còn các ô vuông màu xanh đậm biểu thị mối tương quan âm càng mạnh.

#### Độ đo Cosin

Trước tiên chuyển đổi giá trị của cột “restecg”,”thal”,”fbs” thành số:

**Restecg:**

* Lv hypertrophy = 2
* Normal = 0
* St-t abnormality =1

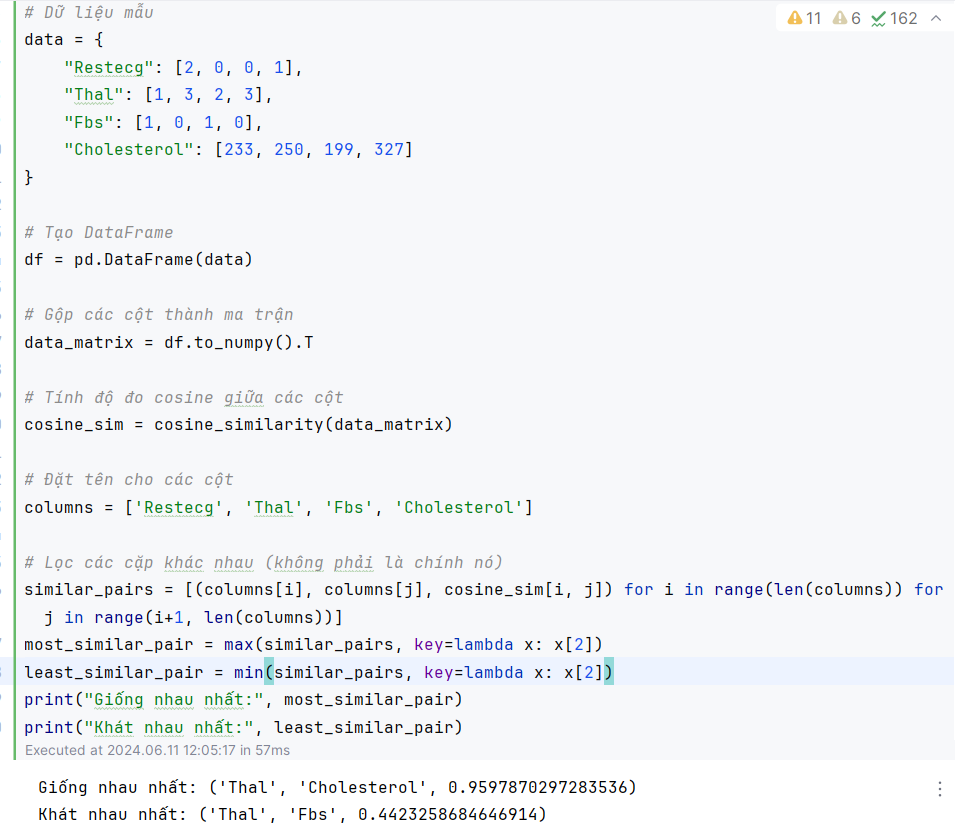
**Thal:**

* Fixed defect =1
* Normal =3
* Reversable defect =2

**Fbs :**

* TRUE = 1
* FALSE = 0

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| STT | Restecg | Thal | Fbs | Cholesterol |
| 1 | 2 | 1 | 1 | 233 |
| 2 | 0 | 3 | 0 | 250 |
| 3 | 0 | 2 | 1 | 199 |
| 4 | 1 | 3 | 0 | 327 |



Hình ..Code độ đo cosine

**Kết luận :**

Độ đo cosine giữa cột thal và cholesterol là giống nhau nhất khoảng 0.96

Độ đo cosine giữa cột restecg và thal là khác nhau nhất là 0.44

**Thực hiện so sánh kết quả của 2 cách đo lường sự tương đồng và khác biệtcủa dữ liệu**

**Phương pháp ma trận tương quan**

* Cho biết mức độ liên quan tuyến tính giữa các cột trong dữ liệu . Số 0 chỉ ra rằng không có mối tương quan tuyến tính giữa các biến . Ví dụ : Restecg và Cholesterol có mối quan hệ giống nhau nhất .

**Phương pháp ma trận cosine**

* Ma trận cosine đo lường sự tương tự về hướng giữa các vecto biểu diễn các cột . Ví dụ Restecg vs thal có độ giống nhau nhất .

So Sánh

Mối tương quan giữa Restecg

+ Tương quan với Thal là 0.85 , fbs là 0.31, Cholesterol là 0.69 .

+ Cosine thal 0.46 ,fbs 0.63 , cholesterol 0.691

Mối tương quan giữa thal

+ Tương quan với fbs là 0.83 , cholesterol là 0.71

+ Cosine fbs là 0.44 , cholesterol là 0.96

Mối tương quan giữa fbs

+ Tương quan với cholesterol 0.79

+ Cosine với cholesterol là 0.58

## Thực hiện khai thác dữ liệu

### Sử dụng các phương pháp khai phá dữ liệu

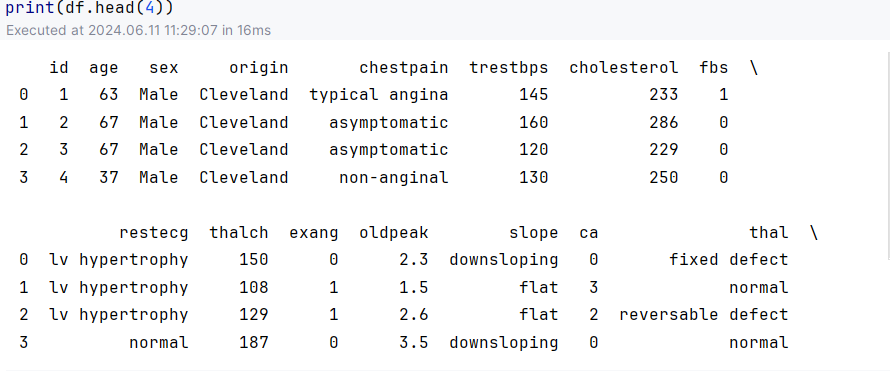
#### Phương pháp phân loại Navie Bayesian

Em đã tạo thêm cột suy tim với 2 thuộc tính yes và no dựa trên cột num . Dưới đây là những cột liên quan đến có suy tim hay không .

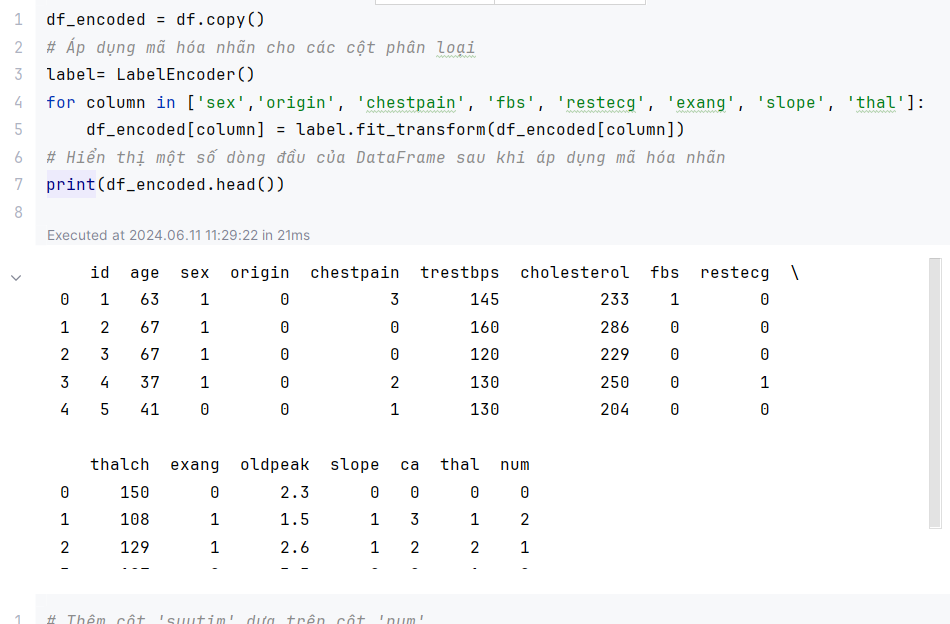
X= (age=63, sex=male, chest pain= aptypical, trestbps=150, cholesterol=234 , fbs=TRUE, restecg = st-t-abnormality , thalch =120, exang=FALSE , oldpeak=2.6, slope = upsloping , ca=2,thal= reversable )

**Chuyển đổi dữ liệu**

Dữ liệu ban đầu chưa xử lý

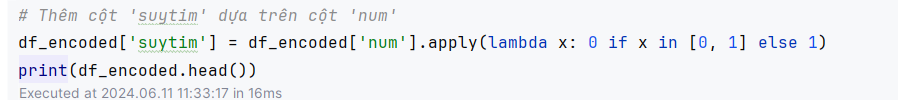


Hình ..Dữ liệu ban đầu

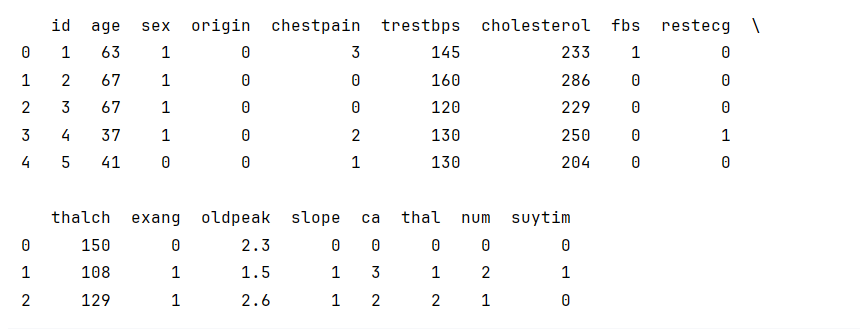
Mã hoá nhãn bằng LabelEncoder() 

Hình 3.33.Dữ liệu dã chuyển đổi

Thêm cột suy tim dựa trên cột num với điều kiện thuộc tính 0,1 sẽ là 0 trong cột suy tim



Hình ..Thêm cột suy tim

****

Hình ..Đã thêm cột suy tim

****

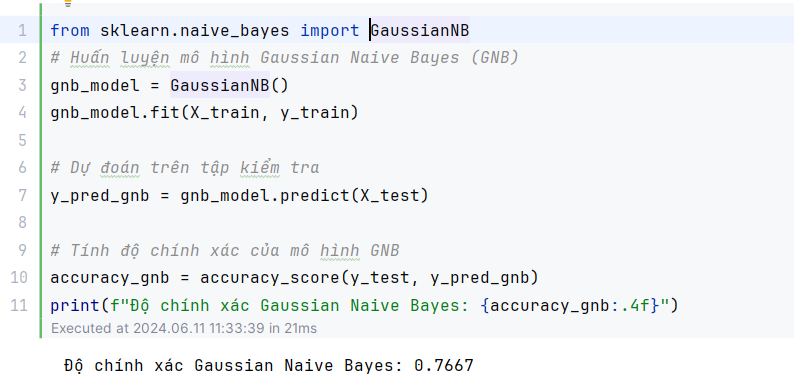
Hình ..Xây dựng mô hình NavieBayes



Hình ..Xây dựng mô hình NavieBayes(2)

=> Do đó , trình phân loại Navie Bayesian dự đoán suytim= 1

=> Yes cho bộ X.



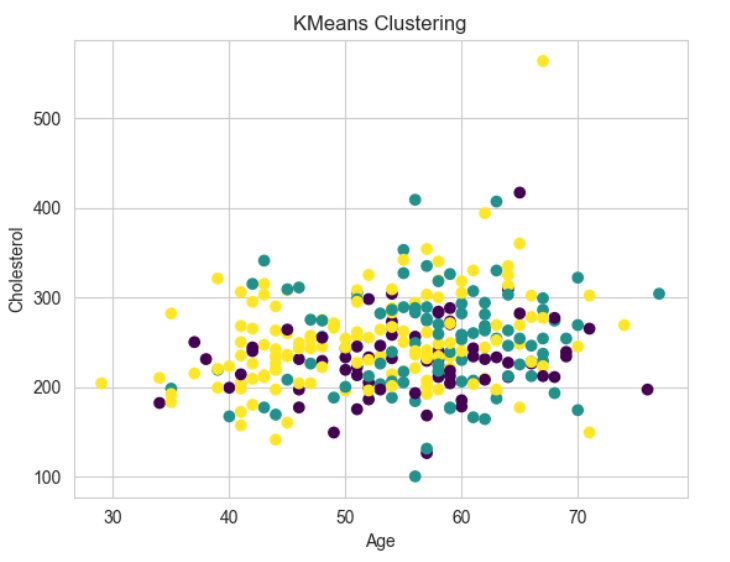
Hình ..Độ chính xác NavieBayes

#### Phương pháp phân cụm K-mean

Vẽ biểu đồ phân cụm Kmean



Hình ..Vẽ biểu đồ K-mean



Hình ..Biểu đồ k-mean

## So sánh thực hiện của việc sử dụng công cụ Python và thủ công

Hai công cụ đều có thể mang lại kết quả tương tự về mặt thống kê và phân tích dữ liệu. Excel tốt cho các phân tích đơn giản, nhanh chóng, và trực quan hóa dữ liệu trực tiếp nên việc phân tích dataset có phần đơn giản hơn vì giao diện dễ dùng và các hàm tính toán đơn giản .

Python thích hợp cho các phân tích phức tạp, tự động hóa quy trình, và làm việc với các tập dữ liệu lớn.Sử dụng các dòng code ngắn gọn dễ xửa lý nguyên tập data

Khác nhau về kết quả là ở ma trận tương quan đã cho kết quả khác nhau .

# KẾT LUẬN

Những phần đã đạt được trong đồ án: Phân tích dữ liệu bệnh tim .

- Giới thiệu về dataset sử dụng cho đề tài , giới thiệu chi tiết từng thuộc tính gồm table có bao nhiêu records, tên field , ý nghĩa , giá trị null, giá trị unique, kiểu dữ của fields, nhị phân , rời rạc , tính 5 giá trị , xử lý dữ liệu .

- Phân tích – thống kê thủ công gồm vẽ các đồ thị cho 3 thuộc tính để vẽ như đồ thị boxplot, quantile-quantile Plot , histogram, Scatter .

- Đo lường sự tương đồng và khác biệt dữ liệu bằng 2 cách : ma trận tương quan và độ đo cosine

- Khai thác dữ liệu đã chọn phân lớp (Navie Bayes), Phân cụm (K-means)

- Thực hiện đánh giá các mẫu bằng phương pháp phân loại số liệu để đánh giá hiệu suất phân loại (Metrics for Evaluating Classifier performance)

Tương tự phân tích và thống kê cơ sở dữ liệu bằng Python những phần đạt được ở trên.

Những phần chưa đạt được trong đồ án :

- Chỉ đánh giá được một phương pháp phân loại của thủ công, còn công cụ Python thì chưa đánh giá được phương pháp .

- Công cụ Python chưa khai thác dữ liệu bằng tập phân cụm K-mean.

Qua quá trình thực hiện giữa hai công cụ thủ công và công cụ Python nhận xét là:

**Python**

Ưu điểm :

- Python cung cấp các thư viện mạnh mẽ như Pandas, NumPy, SciPy, Matplotlib, Seaborn, và Scikit-learn.

- Có khả năng xử lý và phân tích các tập dữ liệu lớn .

Nhược điểm :

- Cần thời gian để học và làm quen với cú pháp và các thư viện.

**Excel**

Ưu điểm ;

- Giao diện đồ họa thân thiện, dễ dùng .

- Hỗ trợ trực quan hóa dữ liệu nhanh chóng với các biểu đồ, đồ thị .

- Tích hợp sẵn bao gồm các công cụ tính toán và thống kê cơ bản, không cần cài đặt thêm phần mềm.

Nhược điểm :

- Giới hạn về số lượng dữ liệu ,excel có giới hạn về số lượng hàng và cột (ví dụ: Excel 2016 có giới hạn 1,048,576 hàng và 16,384 cột).

- Không phù hợp với các tập dữ liệu lớn và phân tích phức tạp.

TÀI LIỆU THAM KHẢO

[1]. *Tài liệu môn học Khoa Học Dữ Liệu*, thầy Lê Văn Hạnh.

[2].https://scikitlearn.org/stable/modules/generated/sklearn.metrics.pairwise.cosine\_similarity.html

[3].https://www.youtube.com/watch?si=L3S1b1N0Mw-eRoz0&v=wE8H-MEHSKs&feature=youtu.be