**TRƯỜNG ĐẠI HỌC QUỐC TẾ HỒNG BÀNG**

**KHOA CÔNG NGHỆ THÔNG TIN**

**\_oOo\_**

A red and black logo

Description automatically generated

**BÁO CÁO CUỐI MÔN   
KHOA HỌC DỮ LIỆU**

**ĐỀ TÀI: Phân tích trên dữ liệu Cardiovascular Health Risk Assessment Dataset**

Giảng viên hướng dẫn: Lê Văn Hạnh

Sinh viên thực hiện: Phan Thành Đạt

Mã số sinh viên: 211110350

TP. Hồ Chí Minh, 2024

LỜI CẢM ƠN

Để hoàn thành đề tài cuối môn của mình, em muốn bày tỏ lòng biết ơn sâu sắc đến sự hướng dẫn, giúp đỡ và những góp ý nhiệt tình từ quý thầy cô tại Trường Đại Học Quốc Tế Hồng Bàng.

Em muốn bày tỏ lòng biết ơn đặc biệt đến thầy Lê Văn Hạnh, người đã dành rất nhiều thời gian và tâm huyết để hướng dẫn em trong quá trình nghiên cứu và hoàn thiện báo cáo cuối môn này.

Em xin chân thành cảm ơn đến tất cả thầy Lê Văn Hạnh tại Trường Đại Học Quốc Tế Hồng Bàng, những người đã tận tình dạy bảo và hỗ trợ em suốt thời gian học tập tại trường.

Mặc dù em đã cố gắng hết sức để hoàn thiện báo cáo cuối môn bằng tất cả năng lực của mình, nhưng không tránh khỏi những thiếu sót. Em rất mong nhận được những đóng góp quý báu từ quý thầy cô và các bạn để làm cho báo cáo này của em trở nên hoàn thiện hơn.

Chân thành cảm ơn và mong nhận được sự hỗ trợ và chỉ dẫn tiếp theo.

TP.HCM, ngày 01 tháng 6 năm 2024

Phan Thành Đạt

TRANG CAM KẾT

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

TP.HCM, ngày 01 tháng 6 năm 2024

Phan Thành Đạt

NHẬN XÉT CỦA GIẢNG VIÊN

TP.HCM, ngày 01 tháng 6 năm 2024

Chữ ký giảng viên

MỤC LỤC

[LỜI CẢM ƠN i](#_Toc169065511)

[TRANG CAM KẾT ii](#_Toc169065512)

[NHẬN XÉT CỦA GIẢNG VIÊN iii](#_Toc169065513)

[MỤC LỤC iv](#_Toc169065514)

[DOANH MỤC HÌNH ẢNH vi](#_Toc169065515)

[DOANH MỤC BẢNG BIỂU vii](#_Toc169065516)

[Lời nói đầu ix](#_Toc169065517)

[CHƯƠNG 1. TỔNG QUAN ĐỀ TÀI 1](#_Toc169065518)

[1.1. Tổng quan về cơ sở dữ liệu 1](#_Toc169065519)

[1.1.1. Cơ sở dữ liệu 1](#_Toc169065520)

[1.1.2. Sức khoẻ tim mạch và bệnh tim mạch 1](#_Toc169065521)

[1.2. Các thuộc tính của bộ dữ liệu 2](#_Toc169065522)

[1.2.1. Thuộc tính của bộ dữ liệu 2](#_Toc169065523)

[1.2.2. Kiểu dữ liệu của các thuộc tính (field) 3](#_Toc169065524)

[CHƯƠNG 2. PHÂN TÍCH - THỐNG KÊ THỦ CÔNG TRÊN CƠ SỞ DỮ LIỆU 8](#_Toc169065525)

[2.1. Tìm hiểu dữ liệu 8](#_Toc169065526)

[2.1.1. Vẽ biểu đồ 8](#_Toc169065527)

[2.1.2. Vẽ biểu đồ thuộc tính dạng danh nghĩa 12](#_Toc169065528)

[2.1.3. Sự tương đồng và khác biệt của dữ liệu 15](#_Toc169065529)

[2.2. Thực hiện khai thác dữ liệu 18](#_Toc169065530)

[2.2.1. Khai thác dữ liệu 18](#_Toc169065531)

[2.2.2. Đánh giá dữ liệu 24](#_Toc169065532)

[CHƯƠNG 3. PHÂN TÍCH – THỐNG KÊ CSDL BẰNG JUPYTER NOTEBOOK 26](#_Toc169065533)

[3.1. Giới thiệu công cụ 26](#_Toc169065534)

[3.1.1. Python 26](#_Toc169065535)

[3.1.2. Các thư viện của Python trong sự hỗ trợ phân tích dữ liệu 26](#_Toc169065536)

[3.1.3. Jupyter Notebook 27](#_Toc169065537)

[3.2. Phân tích – thống kê trên Jupyter Notebook 27](#_Toc169065538)

[3.2.1. Vẽ biểu đồ 27](#_Toc169065539)

[3.2.2. Vẽ Biểu đồ dạng danh nghĩa 33](#_Toc169065540)

[3.2.3. Sự tương đồng và khác biệt 33](#_Toc169065541)

[3.3. Thực hiện khai thác dữ liệu 37](#_Toc169065542)

[3.3.1. Khai thác dữ liệu 37](#_Toc169065543)

[3.3.2. Đánh giá dữ liệu 38](#_Toc169065544)

[CHƯƠNG 4. KẾT LUẬN 39](#_Toc169065545)

[CHƯƠNG 5. Tài liệu tham khảo 40](#_Toc169065546)

DOANH MỤC HÌNH ẢNH

[Ảnh 1. Biểu đồ Boxplot thuộc tính Age 8](#_Toc169065622)

[Ảnh 2. Biểu đồ Boxplot thuộc tính BP 9](#_Toc169065623)

[Ảnh 3. Biểu đồ Boxplot thuộc tính Cholesterol. 9](#_Toc169065624)

[Ảnh 4. Biểu đồ Q-Q Plot thuộc tính Age - BP 10](#_Toc169065625)

[Ảnh 5. Biểu đồ Histogram thuộc tính Age - BP 10](#_Toc169065626)

[Ảnh 6. Biểu đồ Histgram thuộc tính BP - Cholesterol 11](#_Toc169065627)

[Ảnh 7. Biểu đồ Scatter thuộc tính Age - BP 11](#_Toc169065628)

[Ảnh 8. Biểu đồ Scatter : Age - BP sau khi sắp xếp 12](#_Toc169065629)

[Ảnh 9. Biểu đồ Boxplot lọc theo Adult 13](#_Toc169065630)

[Ảnh 10. Biểu đồ Boxplot lọc theo Middle-aged 14](#_Toc169065631)

[Ảnh 11. Biểu đồ Boxplot lọc theo Elderly 14](#_Toc169065632)

[Ảnh 12. Ảnh biểu đồ Boxplot thuộc tính Age 28](#_Toc169065633)

[Ảnh 13. Ảnh biểu đồ Boxplot thuộc tính BP 28](#_Toc169065634)

[Ảnh 14. Ảnh biểu đồ thuộc tính Cholesterol 29](#_Toc169065635)

[Ảnh 15. Ảnh biểu đồ Q-Q Plot thuộc tính Age - BP 30](#_Toc169065636)

[Ảnh 16. Ảnh biểu đồ Histogram 31](#_Toc169065637)

[Ảnh 17. Ảnh biểu đồ thuộc tính Scatter thuộc tính Age - BP 32](#_Toc169065638)

[Ảnh 18. Ảnh biểu đồ Boxplot thuộc tính danh nghĩa 33](#_Toc169065639)

[Ảnh 19. Ảnh ma trận tương quan 35](#_Toc169065640)

[Ảnh 20. Ảnh độ đo Cosin 36](#_Toc169065641)

[Ảnh 21. Ảnh phân loại Bayes 37](#_Toc169065642)

[Ảnh 22. Ảnh đánh giá phân loại Bayes 38](#_Toc169065643)

DOANH MỤC BẢNG BIỂU

[Table 1. Bảng thuộc tính Age 3](#_Toc169065644)

[Table 2. Bảng thuộc tính Sex 3](#_Toc169065645)

[Table 3. Bảng thuộc tính Chest pain type 3](#_Toc169065646)

[Table 4. Bảng thuộc tính BP 4](#_Toc169065647)

[Table 5. Bảng thuộc tính Cholesterol 4](#_Toc169065648)

[Table 6. Bảng thuộc tính FBS over 120 4](#_Toc169065649)

[Table 7. Bảng thuộc tính EKG results 5](#_Toc169065650)

[Table 8. Bảng thuộc tính Max HR 5](#_Toc169065651)

[Table 9. Bảng thuộc tính Exercise angina 5](#_Toc169065652)

[Table 10. Bảng thuộc tính ST depression 6](#_Toc169065653)

[Table 11. Bảng thuộc tính Slope of ST 6](#_Toc169065654)

[Table 12. Bảng thuộc tính Number of vessels fluro 6](#_Toc169065655)

[Table 13. Bảng thuộc tính Thallium 7](#_Toc169065656)

[Table 14. Bảng thuộc tính Heart Disease 7](#_Toc169065657)

[Table 15. Bảng biểu đồ Boxplot Age 8](#_Toc169065658)

[Table 16. Bảng biểu đồ Boxplot BP 8](#_Toc169065659)

[Table 17. Bảng biểu đồ Boxplot thuộc tính Cholesterol 9](#_Toc169065660)

[Table 18. Bảng thuộc tính Age-Group 12](#_Toc169065661)

[Table 19. Bảng biểu đồ Boxplot lọc theo Adult 13](#_Toc169065662)

[Table 20. Bảng biểu đồ Boxplot lọc theo Middle-aged 13](#_Toc169065663)

[Table 21. Bảng biểu đồ Boxplot lọc theo Elderly 14](#_Toc169065664)

[Table 22. Bảng ma trận tương quan thuộc tính danh nghĩa 15](#_Toc169065665)

[Table 23. Bảng ma trận tương quan thuộc tính nhị phân 16](#_Toc169065666)

[Table 24. Bảng ma trận tương quan thuộc tính số 16](#_Toc169065667)

[Table 25. Bảng ma trận tương quan thuộc tính thứ tự 17](#_Toc169065668)

[Table 26. Bảng ma trận tương quan tổng hợp 17](#_Toc169065669)

[Table 27. Bảng độ đo Cosin 17](#_Toc169065670)

[Table 28. Bảng chuyển đổi giá trị BP 19](#_Toc169065671)

[Table 29. Bảng giá trị điểm centroid phân cụm K-means 23](#_Toc169065672)

[Table 30. Bảng đánh giá phân cụm K-means 25](#_Toc169065673)

Lời nói đầu

Khoa học dữ liệu (Data Science) là thuật ngữ đã bắt đầu quen thuộc ở Việt Nam trong giai đoạn hiện nay, nó đang dần trở thành xu hướng được các doanh nghiệp hướng tới để đẩy mạnh hoạt động kinh doanh nhờ việc sử dụng những thông tin được phân tích do công nghệ này mang lại.

Khoa học dữ liệu là ngành học dữ liệu, liên quan đến các công việc như tìm tòi, khai thác, thu thập, phân tích và xử lý dữ liệu, để từ đó tìm ra những thông tin có giá trị, sau đó chuyển hoá những thông tin này thành hành động.

Trước thời đại kỹ nguyên công nghệ số như hiện nay thì việc sử dụng dữ liệu đang ngày một gia tăng. Dữ liệu là một khía cạnh thiết yếu của bất kỳ doanh nghiệp nào, tất cả các công ty dù quy mô lớn nhỏ ra sao thì hiện tại vẫn đang phụ thuộc vào dữ liệu hơn bao giờ hết. Đây là một phần quan trọng của bất kỳ doanh nghiệp nào muốn có lợi thế cạnh tranh trên thị trường và đưa ra các quyết định sáng suốt. Điều quan trọng là phải tìm cách thu thập và xử lý, phân tích khối lượng dữ liệu lớn đó càng nhanh và chính xác càng tốt.

Một trong những ngành đi đầu về việc áp dụng kết quả của việc phân tích dữ liệu là sức khoẻ. Bệnh tim là một trong những căn bệnh rất phổ biến hiện nay, dù là người trẻ tuổi vẫn có thể bị do các lối sống không lành mạnh, với việc càng dễ mắc bệnh này khi tuổi tăng lên nên việc phân tích các triệu ứng, đặc điểm của căn bệnh tim là một điều cực kỳ quan trọng trong việc phòng bị, chữ trị. Điều này giúp phần nâng cao chất lượng đời sống và sức khoẻ của con người,

# TỔNG QUAN ĐỀ TÀI

## Tổng quan về cơ sở dữ liệu

### Cơ sở dữ liệu

Cơ sở dữ liệu về Đánh giá rủi ro sức khoẻ tim mạch (Cardiovascular Health Risk Assessment Dataset) là một tập hợp dữ liệu được sử dụng để phân tích và đánh giá các yếu tố liên quan đến sức khỏe tim mạch và nguy cơ mắc các bệnh tim mạch. Bộ dữ liệu này thường bao gồm thông tin về các đặc điểm cá nhân và các chỉ số y tế của các bệnh nhân, chẳng hạn như tuổi, giới tính, huyết áp, mức cholesterol, kết quả điện tâm đồ (EKG), và các yếu tố khác có thể ảnh hưởng đến sức khỏe tim mạch..

Bộ dữ liệu này được thu thập từ [Kaggle](https://www.kaggle.com/). Nguồn của cơ sở dữ liệu này [ở đây](https://www.kaggle.com/datasets/kapoorprakhar/cardio-health-risk-assessment-dataset).

### Sức khoẻ tim mạch và bệnh tim mạch

Sức khỏe tim mạch (Cardiovascular Health) là trạng thái hoạt động hiệu quả của hệ thống tim mạch, bao gồm tim và các mạch máu. Tim và hệ tuần hoàn chịu trách nhiệm cung cấp máu, oxy và các chất dinh dưỡng đến các mô và cơ quan trong cơ thể, đồng thời loại bỏ các chất thải từ quá trình trao đổi chất.

Heart disease, hay còn gọi là bệnh tim mạch, là một nhóm các rối loạn liên quan đến cấu trúc hoặc chức năng của tim. Đây là một trong những nguyên nhân hàng đầu gây tử vong trên toàn thể giới. Một số loại bệnh tim mạch phổ biến bao gồm:

* Bệnh động mạch vành: Gây ra do tích tụ cặn bám trong các động mạch chính nuôi dưỡng tim, làm hẹp và cứng lại các động mạch này, ảnh hưởng đến lưu thông máu đến cơ tim.
* Bệnh tim do thiếu máu cục bộ: Xảy ra khi một phần cơ tim không nhận đủ oxy và dưỡng chất do lưu thông máu bị giới hạn.
* Bệnh van tim: Các van tim không mở hoặc đóng hoàn toàn, ảnh hưởng đến khả năng bơm máu của tim.
* Rối loạn nhịp tim: Nhịp tim nhanh, chậm hoặc không đều, có thể do các vấn đề về hệ điều khiển nhịp tim.
* Bệnh cơ tim: Các bệnh làm cơ tim bị phì đại, cứng hoặc suy yếu, ảnh hưởng đến khả năng bơm máu của tim.

## Các thuộc tính của bộ dữ liệu

### Thuộc tính của bộ dữ liệu

Bộ dữ liệu này gồm 14 thuộc tính (field) và 270 bản ghi (records), dữ liệu không có giá trị null (missing data).

Tên và ý nghĩa của các thuộc tính:

1. **Age:** Tuổi của bệnh nhân (Độ tuổi từ 29 đến 80 tuổi).
2. **Sex:** Giới tính của bệnh nhân (Nữ = 0, Nam =1).
3. **Chest pain type:** Mức độ đau tức ngực (có giá trị từ 1 đến 4).
4. **BP (Blood Pressure):** Huyết áp, là lực mà máu tạo ra khi lưu thông trong các mạch máu.
5. **Cholesterol:** Mức Cholesterol trong máu.
6. **FBS over 120:** Đường huyết, Lượng đường trong máu lúc đói(không ăn uống gì trong khoảng 8 tiếng trước khi xét nghiệm) lớn hơn 120 mg/dl (False = 0, True = 1).
7. **EKG results (Electrocardiogram):** Kết quả điện tâm đồ, là một quy trình ghi lại các xung điện được tạo ra bởi sự co bóp của cơ tim (có giá trị từ 0 đến 2).
8. **Max HR (Heart Rate):** Nhịp tim tối đa đạt được trong vòng một phút.
9. **Exercise angina:** Đau thắt ngực khi tập thể dục (Không = 0, Có = 1).
10. **ST depression:** là sự giảm hoạt động co bóp của cơ tim trong giai đoạn ST (segment ST) giữa khi nghỉ ngơi và khi tập thể dục.
11. **Slope of ST:** Độ dốc của đoạn ST khi tập thể dục (có giá trị từ 1 đến 3).
12. **Number of vessels fluro :** Số lượng mạch máu chính được nhìn thấy qua phương pháp chụp X-quang huỳnh quang(fluoroscopy) (có giá trị từ 0 đến 3).
13. **Thallium :** Kết quả xét nghiệm stress với Thallium, là một kỹ thuật đánh giá chức năng tim, thường được sử dụng để phát hiện bệnh động mạch vành (có giá trị 3,6,7).
14. **Heart Disease :** Sự hiện diện của bệnh tim, có 2 giá trị là Presence (có) và Absence (không có).

### Kiểu dữ liệu của các thuộc tính (field)

1. **Age (Tuổi):** Thuộc tính số, có giá trị từ 29 đến 80.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Age** | | | | | | | | | |
| Min | Max | Mean  (Trung bình cộng) | Midrange | Mode | Five-number summary | | | | |
| Lower Outliners | Q1 | Median | Q3 | Upper  Outliners |
| 29 | 80 | 54 | 54.5 | 54 | 27.5 | 47 | 54 | 60 | 79.5 |

Table 1. Bảng thuộc tính Age

1. **Sex (Giới tính):** Thuộc tính nhị phân.

|  |  |  |
| --- | --- | --- |
| **Sex** | | |
| Giá trị | Số lượng | Tổng phần trăm |
| 0 (Nữ) | 87 | 32.22% |
| 1 (Nam) | 183 | 67.78% |
| Vậy giá trị Mode là 1 (Nam) | | |

Table 2. Bảng thuộc tính Sex

1. **Chest pain type (Mức độ đau tức ngực):** Thuộc tính thứ tự.

|  |  |  |
| --- | --- | --- |
| **Chest pain type** | | |
| Giá trị | Số lượng | Tổng phần trăm |
| 1 | 20 | 7.4% |
| 2 | 42 | 15.55% |
| 3 | 79 | 29.25% |
| 4 | 129 | 47.8% |
| Vậy giá trị mode là 4 | | |

Table 3. Bảng thuộc tính Chest pain type

1. **BP (Huyết áp):** Thuộc tính số, có giá trị từ 94 đến 200.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Blood Pressure** | | | | | | | | | |
| Min | Max | Mean  (Trung bình cộng) | Midrange | Mode | Five-number summary | | | | |
| Lower Outliners | Q1 | Median | Q3 | Upper  Outliners |
| 94 | 200 | 131.34 | 147 | 120 | 90 | 120 | 130 | 140 | 170 |

Table 4. Bảng thuộc tính BP

1. **Cholesterol:** Thuộc tính số, có giá trị từ 126 đến 564.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Cholesterol** | | | | | | | | | |
| Min | Max | Mean  (Trung bình cộng) | Midrange | Mode | Five-number summary | | | | |
| Lower Outliners | Q1 | Median | Q3 | Upper  Outliners |
| 126 | 564 | 249.66 | 345 | 234 | 114.5 | 212 | 244.5 | 277 | 374.5 |

Table 5. Bảng thuộc tính Cholesterol

1. **FBS over 120 (Đường huyết lớn hơn 120 mg/dl):** Thuộc tính nhị phân.

|  |  |  |
| --- | --- | --- |
| **FBS over 120** | | |
| Giá trị | Số lượng | Tổng phần trăm |
| 0 (Sai) | 230 | 85.18% |
| 1 (Đúng) | 40 | 14.82% |
| Vậy giá trị Mode là 0 (Sai) | | |

Table 6. Bảng thuộc tính FBS over 120

1. **EKG results (Kết quả điện tâm đồ):** Thuộc tính thứ tự.

|  |  |  |
| --- | --- | --- |
| **EKG results** | | |
| Giá trị | Số lượng | Tổng phần trăm |
| 0 | 131 | 48.51% |
| 1 | 2 | 0.74% |
| 2 | 137 | 50.75% |
| Vậy giá trị mode là 2 | | |

Table 7. Bảng thuộc tính EKG results

1. **Max HR (Nhịp tim tối đa trong một phút):** Thuộc tính số, có giá trị từ 71 đến 202.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Max HR** | | | | | | | | | |
| Min | Max | Mean  (Trung bình cộng) | Midrange | Mode | Five-number summary | | | | |
| Lower Outliners | Q1 | Median | Q3 | Upper  Outliners |
| 71 | 202 | 149.67 | 136.5 | 162 | 81 | 132 | 153 | 166 | 217 |

Table 8. Bảng thuộc tính Max HR

1. **Exercise angina (Đau ngực khi tập thể dục):** Thuộc tính nhị phân.

|  |  |  |
| --- | --- | --- |
| **Exercise angina** | | |
| Giá trị | Số lượng | Tổng phần trăm |
| 0 (Không có) | 181 | 67% |
| 1 (Có) | 89 | 33% |
| Vậy giá trị Mode là 0 (Không có) | | |

Table 9. Bảng thuộc tính Exercise angina

1. **ST depression:** Thuộc tính số, có giá trị từ 0 đến 6.2.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **ST depression** | | | | | | | | | |
| Min | Max | Mean  (Trung bình cộng) | Midrange | Mode | Five-number summary | | | | |
| Lower Outliners | Q1 | Median | Q3 | Upper  Outliners |
| 0 | 6.2 | 1.05 | 3.1 | 0 | -1.6 | 0 | 0.8 | 1.6 | 3.2 |

Table 10. Bảng thuộc tính ST depression

1. **Slope of ST:** Thuộc tính thứ tự.

|  |  |  |
| --- | --- | --- |
| **Slope of ST** | | |
| Giá trị | Số lượng | Tổng phần trăm |
| 1 | 130 | 48.15% |
| 2 | 122 | 45.18% |
| 3 | 18 | 6.67% |
| Vậy giá trị mode là 1 | | |

Table 11. Bảng thuộc tính Slope of ST

1. **Number of vessels fluro:** Thuộc tính thứ tự.

|  |  |  |
| --- | --- | --- |
| **Number of vessels fluro** | | |
| Giá trị | Số lượng | Tổng phần trăm |
| 0 | 160 | 59.26% |
| 1 | 58 | 21.48% |
| 2 | 33 | 12.22% |
| 3 | 19 | 7.04% |
| Vậy giá trị mode là 0 | | |

Table 12. Bảng thuộc tính Number of vessels fluro

1. **Thallium:** Thuộc tính thứ tự.

|  |  |  |
| --- | --- | --- |
| **Thallium** | | |
| Giá trị | Số lượng | Tổng phần trăm |
| 3 | 152 | 56.3% |
| 6 | 14 | 5.18% |
| 7 | 104 | 38.52% |
| Vậy giá trị mode là 3 | | |

Table 13. Bảng thuộc tính Thallium

1. **Heart Disease (Sự hiện diện của bệnh tim):** Thuộc tính nhị phân.

|  |  |  |
| --- | --- | --- |
| **Heart Disease** | | |
| Giá trị | Số lượng | Tổng phần trăm |
| Presence (Bị mắc bệnh tim) | 120 | 44.44% |
| Absence (Không có bệnh tim) | 150 | 55.56% |
| Vậy giá trị Mode là Absence (Không có bệnh tim) | | |

Table 14. Bảng thuộc tính Heart Disease

# PHÂN TÍCH - THỐNG KÊ THỦ CÔNG TRÊN CƠ SỞ DỮ LIỆU

## Tìm hiểu dữ liệu

### Vẽ biểu đồ

Chọn 3 thuộc tính là Age, BP và Cholesterol

#### Biểu đồ Boxplot

Biểu đồ Boxplot thuộc tính **Age** dựa trên five-number summary.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Age – five-number summary** | | | | |
| Min | Q1 | Median | Q3 | Max |
| 29 | 47 | 54 | 60 | 80 |

Table 15. Bảng biểu đồ Boxplot Age

A graph with blue lines

Description automatically generated

Ảnh 1. Biểu đồ Boxplot thuộc tính Age

Biểu đồ Boxplot thuộc tính **BP** dựa trên five-number summary.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **BP – five-number summary** | | | | |
| Min | Q1 | Median | Q3 | Max |
| 94 | 120 | 130 | 140 | 170 |

Table 16. Bảng biểu đồ Boxplot BP

A graph with blue lines and dots

Description automatically generated

Ảnh 2. Biểu đồ Boxplot thuộc tính BP

Biểu đồ Boxplot thuộc tính **Cholesterol** dựa trên five-number summary.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Cholesterol – five-number summary** | | | | |
| Min | Q1 | Median | Q3 | Max |
| 126 | 212 | 244.5 | 277 | 374.5 |

Table 17. Bảng biểu đồ Boxplot thuộc tính Cholesterol

A graph of a diagram

Description automatically generated with medium confidence

Ảnh 3. Biểu đồ Boxplot thuộc tính Cholesterol.

#### Biểu đồ Quantile-Quantile Plot

Biểu đồ Quantile-Quantile Plot dựa trên 2 thuộc tính **Age** và thuộc tính **BP**, 2 thuộc này có sự liên quan về ý nghĩa với nhau, vì theo độ tuổi càng cao, thì huyết áp càng dễ tăng lên**.**

A graph with blue dots and a black line

Description automatically generated

Ảnh 4. Biểu đồ Q-Q Plot thuộc tính Age - BP

#### Biểu đồ Histogram

Biểu đồ Histogram dựa trên 2 thuộc tính liên quan về ý nghĩa: **Age** và **BP**

A graph of a number of people

Description automatically generated

Ảnh 5. Biểu đồ Histogram thuộc tính Age - BP

Biểu đồ Histogram dựa trên 2 thuộc tính liên quan về ý nghĩa: **BP** và **Cholesterol**

A graph of a number of patients

Description automatically generated

Ảnh 6. Biểu đồ Histgram thuộc tính BP - Cholesterol

#### Biểu đồ Scatter

Biểu đồ Scatter dựa trên 2 thuộc tính liên quan về ý nghĩa: **Age** và **BP**

A graph with blue dots

Description automatically generated

Ảnh 7. Biểu đồ Scatter thuộc tính Age - BP

Biểu đồ Scatter dựa trên 2 thuộc tính liên quan về ý nghĩa: **Age** và **BP** sau khi 2 cột dữ liệu được sắp xếp theo thứ tự tăng dần

A graph with blue dots

Description automatically generated

Ảnh 8. Biểu đồ Scatter : Age - BP sau khi sắp xếp

### Vẽ biểu đồ thuộc tính dạng danh nghĩa

Vì bộ dữ liệu nguyên bản mà em đang sử dụng không có thuộc tính dạng danh nghĩa, nên em sẽ chia thuộc tính Age thành thuộc tính Age-Group có 3 nhóm tuổi: Adult (từ 18 đến 40 tuổi), Middle-aged (từ 40 đến 60 tuổi) và Elderly (trên 60 tuổi)

Thuộc tính Age-Group:

|  |  |  |
| --- | --- | --- |
| **Age-Group** | | |
| Giá trị | Số lượng | Tổng phần trăm |
| Adult | 16 | 5.92% |
| Middle-aged | 186 | 68.88% |
| Elderly | 68 | 25.2% |
| Giá trị mode là Middle-aged | | |

Table 18. Bảng thuộc tính Age-Group

Biểu đồ Boxplot của thuộc tính Age sau khi lọc theo Adult

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Age Group – Adult filter – five-number summary** | | | | |
| Min | Q1 | Median | Q3 | Max |
| 29 | 34.25 | 37 | 39 | 40 |

Table 19. Bảng biểu đồ Boxplot lọc theo Adult

A screenshot of a graph

Description automatically generated

Ảnh 9. Biểu đồ Boxplot lọc theo Adult

Biểu đồ Boxplot của thuộc tính Age sau khi lọc theo Middle-aged

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Age Group – Middle-aged – five-number summary** | | | | |
| Min | Q1 | Median | Q3 | Max |
| 41 | 46 | 52 | 57 | 60 |

Table 20. Bảng biểu đồ Boxplot lọc theo Middle-aged

A screenshot of a graph

Description automatically generated

Ảnh 10. Biểu đồ Boxplot lọc theo Middle-aged

Biểu đồ Boxplot của thuộc tính Age sau khi lọc theo Elderly

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Age Group – Middle-aged – five-number summary** | | | | |
| Min | Q1 | Median | Q3 | Max |
| 61 | 62.25 | 65 | 67 | 71 |

Table 21. Bảng biểu đồ Boxplot lọc theo Elderly

A screenshot of a graph

Description automatically generated

Ảnh 11. Biểu đồ Boxplot lọc theo Elderly

### Sự tương đồng và khác biệt của dữ liệu

4 thuộc tính thuộc các dạng dữ liệu:

* Thuộc tính danh nghĩa (Nominal Attributes): Age-Group
* Thuộc tính dạng nhị phân (Binary Attributes): Sex
* Thuộc tính dạng số (Numeric Attributes): Age
* Thuộc tính dạng thứ tự (Ordinal Attributes): Chest pain type

4 dòng dữ liệu được chọn:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Ký hiệu thứ tự** | **Age** | **Age Group** | **Sex** | **Chest pain type** |
| **A** | 70 | Elderly | 1 | 4 |
| **B** | 80 | Elderly | 0 | 3 |
| **C** | 55 | Middle-aged | 1 | 2 |
| **D** | 40 | Adult | 1 | 1 |

#### Ma trận tương quan:

##### Ma trận thuộc tính danh nghĩa (Age-Group)

Nếu giá trị của đối tượng khác nhau thì sẽ bằng 1, nếu giống nhau sẽ bằng 0.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **A** | **B** | **C** | **D** |
| **A** | 0 |  |  |  |
| **B** | 0 | 0 |  |  |
| **C** | 1 | 1 | 0 |  |
| **D** | 1 | 1 | 1 | 0 |

Table 22. Bảng ma trận tương quan thuộc tính danh nghĩa

##### Ma trận thuộc tính nhị phân (Sex)

Giá trị bằng 0 thì là Female (Nữ), còn bằng 1 là Male (Nam).

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **A** | **B** | **C** | **D** |
| **A** | 0 |  |  |  |
| **B** | 0 | 0 |  |  |
| **C** | 1 | 0 | 0 |  |
| **D** | 1 | 0 | 1 | 0 |

Table 23. Bảng ma trận tương quan thuộc tính nhị phân

##### Ma trận thuộc tinh số (Age)

Trong thuộc tính độ tuổi đã chọn ở trên, với giá trị lớn nhất (max) là 80 và giá trị nhỏ nhất (min) là 40. Tính ma trận với công thức:

A black text with a line

Description automatically generated with medium confidence

Ta có:

* d(A,B) = = 0.25
* d(A,C) = = 0.375
* d(A,D) = = 0.75
* d(B,C) = = 0.625
* d(B,D) = = 1
* d(C,D) = = 0.375

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **A** | **B** | **C** | **D** |
| **A** | 0 |  |  |  |
| **B** | 0.25 | 0 |  |  |
| **C** | 0.375 | 0.625 | 0 |  |
| **D** | 0.75 | 1 | 0.375 | 0 |

Table 24. Bảng ma trận tương quan thuộc tính số

##### Ma trận thuộc tính thứ tự (Chest pain type)

0 0.33 0.67 1

Rank = 1 Rank = 2 Rank = 3 Rank = 4

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **A** | **B** | **C** | **D** |
| **A** | 0 |  |  |  |
| **B** | 0.33 | 0 |  |  |
| **C** | 0.67 | 0.34 | 0 |  |
| **D** | 1 | 0.67 | 0.33 | 0 |

Table 25. Bảng ma trận tương quan thuộc tính thứ tự

##### Kết luận ma trận tương quan

Từ 4 ma trận thuộc tính trên, ta có ma trận tương quan sau:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **A** | **B** | **C** | **D** |
| **A** | 0 |  |  |  |
| **B** | 0.145 | 0 |  |  |
| **C** | 0.761 | 0.491 | 0 |  |
| **D** | 0.937 | 0.667 | 0.676 | 0 |

Table 26. Bảng ma trận tương quan tổng hợp

Kết luận: giữa 4 người trên thì người A và người B có sự tương đồng trên các thuộc tính đã xét cao nhất, người A và người D có sự tương đồng là thấp nhất.

#### Độ đo Cosin

##### Tính toán độ đo Cosin

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Ký hiệu thứ tự** | **Age** | **Adult** | **Middle-aged** | **Elderly** | **Sex** | **Chest pain type** |
| **A** | 70 | 0 | 0 | 1 | 1 | 4 |
| **B** | 80 | 0 | 0 | 1 | 0 | 3 |
| **C** | 55 | 0 | 1 | 0 | 1 | 2 |
| **D** | 40 | 1 | 0 | 0 | 1 | 1 |

Table 27. Bảng độ đo Cosin

Tính độ tương quan giữa 4 vector trên bằng công thức:

A black symbol with a white background

Description automatically generated

sim (A,B) = = 0.9997

sim (A,C) = = 0.99934

sim (A,D) = = 0.9923

sim (B,C) = = 0.9996

sim (B,D) = = 0.9992

sim (C,D) = = 0.9995

##### Kết luận độ đo Cosin

Giữa người A và người B có độ tương đồng cao nhất (0.9997) , người A và người D có độ tương đồng thấp nhất (0.9923)

#### Sự tương đồng và khác biệt giữa Ma trận tương quan và độ đo Cosin.

Từ kết quả giữa ma trận tương quan và độ đo Cosin, ta có thể thấy rằng cả 2 cách đều cho ra kết quả giống nhau rằng A và B có độ tương đồng cao nhất cũng như A và D có độ tương đồng thấp nhất.

## Thực hiện khai thác dữ liệu

### Khai thác dữ liệu

2 Phương pháp khai thác dữ liệu mà em chọn:

* Phương pháp phân loại Bayes (Bayes Classification Methoids).
* Phương pháp phân cụm K-means.

#### Phương pháp phân loại Bayes

Bốn thuộc tính được chọn có liên quan nhau với kết quả Bệnh tim (Heart Disease) có tồn tại hay không là : Nhóm tuổi (Age-Group) , Huyết áp (Blood Pressure), Đường huyết trên 120 (FBS over 120) và Số lượng mạch máu thấy qua phương pháp chụp X-quang (Number of vessels fluro)

Bời vì Huyết áp có nhiều giá trị khác nhau dạng số nên cần được chuyển đổi để có thể dễ phân loại hơn.

|  |  |
| --- | --- |
| **Blood Pressure (Huyết áp)** | |
| **Giá trị** | **Giá trị thay thế** |
| 90 - 120 | Normal |
| 120 – 160 | Pre-hypertension |
| Lớn hơn 160 | Hypertension |

Table 28. Bảng chuyển đổi giá trị BP

**Dữ liệu cần phân loại mà không bị trùng:**

**X** = (Age-Group = Middle-aged,

BP = Hypertension,

FBS over 120 = True,

Number of vessels fluro = 1).

**Xác xuất của kết quả Bệnh tim (Heart Disease):**

P (Heart Disease = Presence) = = 0.44

P (Heart Disease = Absence) = = 0.56

**Tính các xác suất có điều kiện:**

P(Age-Group = Middle-aged | Heart Disease = Presence) = = 0.641

P(Age-Group = Middle-aged | Heart Disease = Absence) = = 0.72

P(BP = Hypertension | Heart Disease = Presence) = = 0.075

P(BP = Hypertension | Heart Disease = Absence) = = 0.02

P(FBS over 120 = True | Heart Disease = Presence) = = 0.141

P(FBS over 120 = True | Heart Disease = Absence) = = 0.153

P(Number of vessels fluro = 1 | Heart Disease = Presence) = = 0.31

P(Number of vessels fluro = 1 | Heart Disease = Absence) = = 0.133

**Ta có được:**

**P(X | Heart Disease = Presence) =** P(Age-Group = Middle-aged | Heart Disease = Presence) \* P(BP = Hypertension | Heart Disease = Presence) \* P(FBS over 120 = True | Heart Disease = Presence) \* P(Number of vessels fluro = 1 | Heart Disease = Presence) = 0.641 \* 0.075 \* 0.141 \* 0.31 = 0.0013490

**P(X | Heart Disease = Absence) =** P(Age-Group = Middle-aged | Heart Disease = Absence) \* P(BP = Hypertension | Heart Disease = Absence) \* P(FBS over 120 = True | Heart Disease = Absence) \* P(Number of vessels fluro = 1 | Heart Disease = Absence) = 0.72 \* 0.02 \* 0.153 \* 0.133 = 0.0014555

P(X | Heart Disease = Presence) \* P (Heart Disease = Presence) = 0.44 \* 0.0013490 = 0.0005936

P(X | Heart Disease = Absence) \* P (Heart Disease = Absence) = 0.56 \* 0.0014555 = 0.0008151

Vậy **X** = (Age-Group = Middle-aged, BP = Hypertension, FBS over 120 = True, Number of vessels fluro = 1) sẽ có kết quả là Heart Disease = Absence.

#### Phương pháp phân cụm K-means

Vì việc biểu diễn cách phân cụm tất cả 270 dòng dữ liệu khó đưa lên word nên em xin ví dụ cách phân cụm K-means với 4 dòng dữ liệu và 2 đặc trưng. Cách em phân cụm đầy đủ với tất cả dữ liệu sẽ ở trên Excel.

##### Ví dụ về việc phân cụm K-means

**Ví dụ**: sử dụng 4 dòng dữ liệu với 2 đặc trưng X (Chest pain type) và Y (Number of vessels fluro) để phân cụm K-means với **K = 2**:

|  |  |  |
| --- | --- | --- |
| **Thứ tự** | **X (Chest pain type)** | **Y (Number of vessels fluro)** |
| **A** | 4 | 3 |
| **B** | 2 | 0 |
| **C** | 3 | 1 |
| **D** | 1 | 2 |

**Lặp lần 1:**

**Bước 1**: Khởi tạo tâm centroid

* Chọn A(4,3) là tâm của nhóm 1.
* Chọn B(2,0) là tâm của nhóm 2.

**Bước 2**: Tính khoảng cách từ các đối tượng đến tâm của các nhóm bằng khoảng cách Euclidean.

d(Nhóm 1,B) = = 3.6

d(Nhóm 1,C) = = 2.23

d(Nhóm 1,D) = = 3.16

d(Nhóm 2,A) = = 3.6

d(Nhóm 2,C) = = 1.41

d(Nhóm 2,D) = = 2.23

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | A | B | C | D |  |
| X | 4 | 2 | 3 | 1 |  |
| Y | 3 | 0 | 1 | 2 |  |
| D | 0 | 3.6 | 2.23 | 3.16 | Nhóm 1(4,3) |
| 3.6 | 0 | 1.41 | 2.23 | Nhóm 2(2,0) |

**Bước 3**: Nhóm các đối tượng vào nhóm có tâm gần nhất

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **A** | **B** | **C** | **D** |  |
| 1 | 0 | 0 | 0 | **Nhóm 1(4,3)** |
| 0 | 1 | 1 | 1 | **Nhóm 2(2,0)** |

* Nhóm 1 có 1 đối tượng là A.
* Nhóm 2 có 3 đối tượng là B,C,D.

Lặp lần 2:

Bước 1: Tính lại toạ độ tâm cho các nhóm mới.

* Nhóm 1 chỉ có 1 đối tượng A nên tâm không đổi.
* Nhóm 2 có toạ độ tâm mới: = ( , ) = (2,1)

Bước 2: Tính lại khoảng cách của các đối tượng đến tâm của các nhóm mới bằng khoảng cách Euclidean.

d(Nhóm 1,B) = = 3.6

d(Nhóm 1,C) = = 2.23

d(Nhóm 1,D) = = 3.16

d(Nhóm 2,A) = = 2.82

d(Nhóm 2,B) = = 1

d(Nhóm 2,C) = = 1

d(Nhóm 2,D) = = 1.41

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | A | B | C | D |  |
| X | 4 | 2 | 3 | 1 |  |
| Y | 3 | 0 | 1 | 2 |  |
| D | 0 | 3.6 | 2.23 | 3.16 | Nhóm 1(4,3) |
| 2.82 | 1 | 1 | 2.41 | Nhóm 2(2,1) |

Bước 3: Nhóm các đối tượng vào các tâm gần nhất.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **A** | **B** | **C** | **D** |  |
| 1 | 0 | 0 | 0 | **Nhóm 1(4,3)** |
| 0 | 1 | 1 | 1 | **Nhóm 2(2,1)** |

* Nhóm 1 có 1 đối tượng là A.
* Nhóm 2 có 3 đối tượng là B,C,D.
* Vì không có sự thay đổi thành viên của các nhóm sau lần lặp này nên thuật toán dừng.

##### Kết quả việc phân cụm K-means trên tất cả dữ liệu

Sử dụng phương pháp phân cụm K-means để phân cụm tất cả các điểm dữ liệu (270 records) trong dữ liệu với số cụm **K = 3**

Khởi tạo 3 điểm centroid của 3 cụm với 3 điểm dữ liệu ngẫu nhiên.

* Chọn dữ liệu dòng thứ 1 là tâm của nhóm 1 với giá trị (70, 1, 4, 130, 322, 0, 2, 109, 0, 2.4, 2, 3).
* Chọn dữ liệu dòng thứ 134 là tâm của nhóm 2 với giá trị (64, 1, 4, 120, 246, 0, 2, 96, 1, 2.2, 3, 1).
* Chọn dữ liệu dòng thứ 270 là tâm của nhóm 3 với giá trị (67, 1, 4, 160, 286, 0, 2, 108, 1, 1.5, 2, 3).

Sau 11 lần lặp việc phân cụm, em đã chọn kết quả của lần lặp thứ 8 vì lần lặp 8 có số điểm dữ liệu lệch so với kết quả phân cụm trước là 1 trong khi các lần lặp sau thứ 8, số điểm dữ liệu lệch lại tăng lên với lần lặp thứ 9 (2 điểm dữ liệu lệch so với lần lặp trước), lần lặp thứ 10 (2), lần lặp thứ 11 (3).

Giá trị điểm centroid của 3 cụm ở lần lặp thứ 8 là:

|  |  |  |  |
| --- | --- | --- | --- |
| **Giá trị điểm centroid của 3 cụm ở lần lặp thứ 8** | | | |
| **Thuộc tính (Field)** | **Cụm 1** | **Cụm 2** | **Cụm 3** |
| **Age** | 56.36734694 | 50.92857143 | 56.28440367 |
| **Sex** | 0.469387755 | 0.75 | 0.697247706 |
| **Chest pain type** | 3.346938776 | 3.089285714 | 3.183486239 |
| **BP** | 134.9183673 | 125.7678571 | 135.4678899 |
| **Cholesterol** | 326.9795918 | 205.6607143 | 260.1100917 |
| **FBS over 120** | 0.12244898 | 0.142857143 | 0.165137615 |
| **EKG results** | 1.081632653 | 0.794642857 | 1.229357798 |
| **Max HR** | 153.0408163 | 154.2053571 | 143.5137615 |
| **Exercise angina** | 0.326530612 | 0.276785714 | 0.385321101 |
| **ST depression** | 0.967346939 | 1.004464286 | 1.133944954 |
| **Slope of ST** | 1.448979592 | 1.535714286 | 1.697247706 |
| **Number of vessels fluro** | 0.734693878 | 0.482142857 | 0.834862385 |
| **Thallium** | 4.530612245 | 4.6875 | 4.779816514 |

Table 29. Bảng giá trị điểm centroid phân cụm K-means

Sau 8 lần lặp, số bản ghi (records) của mỗi cụm là:

Cụm 1: 50 bản ghi.

Cụm 2: 112 bản ghi.

Cụm 3: 108 bản ghi.

### Đánh giá dữ liệu

2 Phương pháp đánh giá dữ liệu em chọn là:

* Phương pháp phân loại: Số liệu để đánh giá hiệu suất phân loại
* Phương pháp phân cụm: Phương pháp không giám sát

#### Phương pháp không giám sát – Chỉ số Silhouette

Để đo chỉ số Silhouette, đầu tiên lấy 3 bản ghi từ mỗi cụm.

Cụm 1:

* Bản ghi A(51, 0, 3, 120, 295, 0, 2, 157, 0, 0.6, 1, 0, 3)
* Bản ghi B(59, 1, 4, 170, 326, 0, 2, 140, 1, 3.4, 3, 0,7)
* Bản ghi C(42, 1, 4, 136, 315, 0, 0, 125, 1, 1.8, 2, 0, 6)

Cụm 2:

* Bản ghi D(57, 1, 3, 128, 229, 0, 2, 150, 0, 0.4, 2, 1, 7)
* Bản ghi E(37, 0, 3, 120, 215, 0, 0, 170, 0, 0, 1, 0, 3)
* Bản ghi F(50,1, 4, 144, 200, 0, 2, 126, 1, 0.9, 2, 0,7)

Cụm 3:

* Bản ghi G(71, 0, 3, 110, 265, 1, 2, 130, 0, 0, 1, 1,3)
* Bản ghi H(66, 1, 4, 160, 228, 0, ,2, 138, 0, 2.3, 1, 0, 6)
* Bản ghi J(48, 1, 4, 130, 256, 1, 2, 150, 1, 0, 1, 2, 7)

Chỉ số Silhouette được tính bởi công thức:



Trong đó:

* a(i) là khoảng cách từ điểm i trong cụm A đến điểm centroid của cụm A.
* b(i) là khoảng cách từ điểm i trong cụm A đến điểm centroid của cụm B.

Điểm Silhouette có giá trị nằm trong khoảng từ -1 đến 1:

* Điểm dữ liệu có Silhouette từ 0 – 1 là nằm đúng cụm.
* Điểm dữ liệu có Silhouette bằng 0 là đang nằm giữa 2 cụm.
* Điểm dữ liệu có Silhouette giá trị âm thì khả năng nằm sai cụm.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Bản ghi** | **Thuộc cụm** | **Khoảng cách đến cụm của bản thân(1)** | **Khoảng cách đến cụm gần nhất(2)** | **Max(1,2)** | **Chỉ số Silhouette** |
| A | 1 | 61.489 | 75.176 | 75.176 | 0.182 |
| B | 1 | 61.820 | 115.945 | 115.945 | 0.466 |
| C | 1 | 62.118 | 94.372 | 94.372 | 0.341 |
| D | 2 | 41.711 | 51.009 | 51.009 | 0.182 |
| E | 2 | 50.593 | 113.455 | 113.455 | 0.554 |
| F | 2 | 59.622 | 98.701 | 98.701 | 0.395 |
| G | 3 | 65.234 | 126.308 | 126.308 | 0.483 |
| H | 3 | 78.230 | 94.259 | 94.259 | 0.170 |
| J | 3 | 32.904 | 71.022 | 71.022 | 0.536 |

Table 30. Bảng đánh giá phân cụm K-means

Giá trị trung bình của chỉ số Silhouette là = 0.367, vì chỉ số Silhouette nằm trong khoảng từ 0 đến 1 nên tất cả bản ghi đều đang nằm đúng cụm.

# PHÂN TÍCH – THỐNG KÊ CSDL BẰNG JUPYTER NOTEBOOK

## Giới thiệu công cụ

### Python

Python là một ngôn ngữ lập trình bậc cao, một ngôn ngữ lập trình hướng đối tượng, do Guido van Rossum cho ra mắt vào năm 1991. Mục đích ra đời của Python là cung cấp một ngôn ngữ lập trình có cấu trúc rõ ràng, sáng sủa , thuận tiện cho người mới học lập trình. Một trong những đặc điểm độc nhất của Python là ngôn ngữ này không dùng đến dấu chấm phẩy, dấu mở-đóng ngoặc để kết thúc câu lệnh hay khối lệnh, mà cách duy nhất để nó nhận biết là một dấy thụt đầu dòng.

Python đã trở thành ngôn ngữ lập trình dành cho khoa học dữ liệu và học máy (machine learning) trong các ứng dụng trí tuệ nhân tạo (AI). Sự phổ biến của nó không chỉ do tính đơn giản và dễ đọc mà còn vì hệ sinh thái rộng lớn gồm các thư viện và khung giúp các nhà phát triển dễ dàng triển khai các thuật toán phức tạp và các tác vụ xử lý dữ liệu. Khi AI tiếp tục cách mạng hóa các ngành công nghiệp trên toàn cầu, vai trò của Python trong khoa học dữ liệu và máy học ngày càng trở nên quan trọng.

A close-up of a logo

Description automatically generated

### Các thư viện của Python trong sự hỗ trợ phân tích dữ liệu

Các thư viện của Python như NumPy, pandas và SciPy cung cấp nền tảng vững chắc cho thao tác và phân tích dữ liệu. Ví dụ, NumPy cung cấp hỗ trợ cho các mảng và ma trận đa chiều, rất cần thiết để làm việc với các tập dữ liệu lớn. Mặt khác, Pandas là một thư viện phân tích dữ liệu mạnh mẽ cung cấp các cấu trúc dữ liệu như DataFrames và Series, giúp dễ dàng thao tác và phân tích dữ liệu. SciPy xây dựng dựa trên các khả năng của NumPy và thêm chức năng bổ sung cho điện toán khoa học, chẳng hạn như tối ưu hóa, xử lý tín hiệu và đại số tuyến tính.

### Jupyter Notebook

Jupyter là một thuật ngữ được ghép từ ba ngôn ngữ lập trình Julia, Python và R. Trước đây Jupyter Notebook có tên là IPython Notebook, đến năm 2014 tách ra khỏi IPython và đổi tên thành Jupyter Notebook.

Jupyter Notebook là một nền tảng tính toán khoa học mã nguồn mở, bạn có thể sử dụng để tạo và chia sẻ các tài liệu có chứa code trực tiếp, phương trình, trực quan hóa dữ liệu và văn bản tường thuật.

Jupyter cho phép người dùng xem kết quả của code in-line (mã inline) mà không cần phụ thuộc vào các phần khác của code. Trong Notebook mọi ô của code có thể được kiểm tra bất cứ lúc nào, điều này đã giúp Jupyter trở nên khác biệt so với các ID như Pycharm, VSCode. Việc Jupyter có thể xuất code in-line đã giúp ích rất nhiều trong quá trình phân tích khám phá dữ liệu.

## Phân tích – thống kê trên Jupyter Notebook

### Vẽ biểu đồ

Chọn 3 thuộc tính là Age, BP và Cholesterol

#### Biểu đồ Boxplot

Biểu đồ Boxplot thuộc tính **Age vẽ bằng Jupyter Notebook**

Để vẽ được biểu đồ, trước tiên cần phải import thư viện của của Python.

A screenshot of a computer screen

Description automatically generated

Sau đó đọc file csv bằng thư viện pandas.



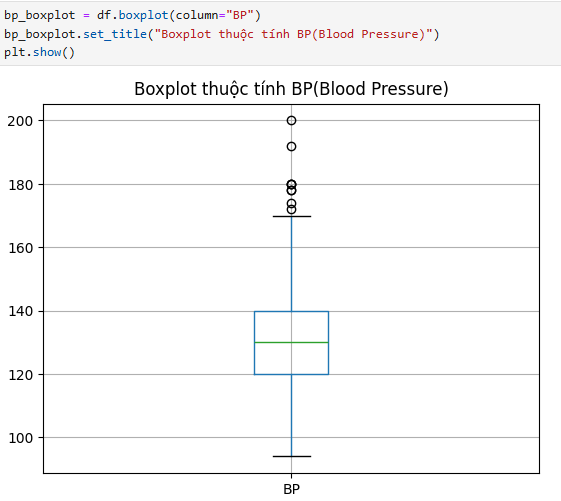
Vẽ biểu đồ Boxplot bằng thư viện matplotlib.

A screen shot of a graph

Description automatically generated

Ảnh 12. Ảnh biểu đồ Boxplot thuộc tính Age

Biểu đồ Boxplot thuộc tính BP



Ảnh 13. Ảnh biểu đồ Boxplot thuộc tính BP

Biểu đồ Boxplot thuộc tính Cholesterol

A screen shot of a graph

Description automatically generated

Ảnh 14. Ảnh biểu đồ thuộc tính Cholesterol

#### Biểu đồ Quantile-Quantile Plot

Biểu đồ Quantile-Quantile Plot dựa trên 2 thuộc tính **Age** và thuộc tính **BP**, 2 thuộc này có sự liên quan về ý nghĩa với nhau, vì theo độ tuổi càng cao, thì huyết áp càng dễ tăng lên**.**

**A graph with a red line and blue line

Description automatically generated**

Ảnh 15. Ảnh biểu đồ Q-Q Plot thuộc tính Age - BP

#### Biểu đồ Histogram

Biểu đồ Histogram dựa trên thuộc tính Age.

A graph with numbers and symbols

Description automatically generated with medium confidence

Ảnh 16. Ảnh biểu đồ Histogram

#### Biểu đồ Scatter

Biểu đồ Scatter dựa trên 2 thuộc tính liên quan về ý nghĩa: **Age** và **BP.**

A screen shot of a graph

Description automatically generated

Ảnh 17. Ảnh biểu đồ thuộc tính Scatter thuộc tính Age - BP

### Vẽ Biểu đồ dạng danh nghĩa

Biểu đồ Boxplot của thuộc tính Age sau khi lọc theo Age-Group

A screen shot of a graph

Description automatically generated

Ảnh 18. Ảnh biểu đồ Boxplot thuộc tính danh nghĩa

### Sự tương đồng và khác biệt

Chọn 4 thuộc tính thuộc các dạng dữ liệu:

* Thuộc tính danh nghĩa (Nominal Attributes): Age-Group
* Thuộc tính dạng nhị phân (Binary Attributes): Sex
* Thuộc tính dạng số (Numeric Attributes): Age
* Thuộc tính dạng thứ tự (Ordinal Attributes): Chest pain type

A screenshot of a computer code

Description automatically generated

Chọn 5 dòng đầu tiên của dữ liệu.

A screenshot of a data table

Description automatically generated

#### Ma trận tương quan

A screenshot of a computer

Description automatically generated

Ảnh 19. Ảnh ma trận tương quan

Vậy người có id 1 và 3 có độ tương đồng cao nhất, người có id 3 và 4 có độ tương đồng thấp nhất.

#### Ma trận độ đo Cosin

A screenshot of a computer

Description automatically generated

Ảnh 20. Ảnh độ đo Cosin

Vậy người có id 1 và 3 có độ tương đồng cao nhất, người có id 3 và 4 có độ tương đồng thấp nhất.

## Thực hiện khai thác dữ liệu

### Khai thác dữ liệu

2 Phương pháp khai thác dữ liệu mà em chọn:

* Phương pháp phân loại Bayes (Bayes Classification Methoids).
* Phương pháp phân cụm K-means.

#### Phương pháp phân loại Bayes

Bốn thuộc tính được chọn có liên quan nhau với kết quả Bệnh tim (Heart Disease) có tồn tại hay không là : Nhóm tuổi (Age-Group) , Nhóm huyết áp (BP-Group), Đường huyết trên 120 (FBS over 120) và Số lượng mạch máu thấy qua phương pháp chụp X-quang (Number of vessels fluro)

**Dữ liệu cần phân loại mà không bị trùng:**

**X** = (Age-Group = Middle-aged,

BP = Hypertension,

FBS over 120 = True,

Number of vessels fluro = 1).

A screenshot of a computer code

Description automatically generated

Ảnh 21. Ảnh phân loại Bayes

Với input trên thì chương trình cho ra kết quả là 1, ứng với **Heart Disease =** Presence. Kết quả này mâu thuẫn với kết quả đã tính được ở phần 2.

### Đánh giá dữ liệu

#### Đánh giá phân loại Bayes

Các chỉ số đánh phân lớp Bayes

Chỉ số chính xác (accuracy) là: 76%

Chỉ số chính xác (Precision) khi dự đoán kết quả Heart Disease là 0 (Absence): 83%

Chỉ số chính xác (Precision) khi dự đoán kết quả Heart Disease là 1 (Presence): 67%

Độ thu hồi (Recall) khi dự đoán kết quả Heart Disease là 0 (Absence): 76%

Độ thu hồi (Recall) khi dự đoán kết quả Heart Disease là 1 (Presence): 76%

Độ trung bình hài hoà của độ chính xác và thu hồi (f1-score) khi dự đoán kết quả Heart Disease là 0 (Absence): 79%

Độ trung bình hài hoà của độ chính xác và thu hồi (f1-score) khi dự đoán kết quả Heart Disease là 1 (Presence): 71%

A screenshot of a computer

Description automatically generated

Ảnh 22. Ảnh đánh giá phân loại Bayes

# KẾT LUẬN

Qua quá trình nghiên cứu và phân tích dữ liệu về bệnh tim trong đồ án này, em đã rút ra được kết luận sau:

* Khám phá dữ liệu và làm sạch dữ liệu: Việc khám phá và làm sạch dữ liệu ban đầu đóng vai trò rất quan trọng, giúp em hiểu rõ hơn về đặc điểm và chất lượng của tập dữ liệu. Điều này là tiền đề quan trọng để có thể tiến hành các bước phân tích và xây dựng mô hình hiệu quả.
* Phân tích các yếu tố ảnh hưởng: Thông qua việc phân tích các đặc trưng quan trọng, em đã xác định được các yếu tố như tuổi tác, huyết áp, đường huyết ảnh hưởng lớn nhất đến kết quả dự báo về việc có mắc bệnh tim hay không.

# Tài liệu tham khảo

1. Kaggle.
2. Các tài liệu của thầy Lê Văn Hạnh.