**TRƯỜNG ĐẠI HỌC QUỐC TẾ HỒNG BÀNG**

**KHOA CÔNG NGHỆ THÔNG TIN**

**🙢🕮🙠**

A red and white logo with wings

Description automatically generated

**BÁO CÁO CUỐI MÔN**

**MÔN: KHOA HỌC DỮ LIỆU**

**ĐỀ TÀI:**

**PHÂN TÍCH TRÊN DỮ LIỆU**

**HEART FAILURE PREDICTION**

**Giảng viên hướng dẫn:** Lê Văn Hạnh

**Sinh viên thực hiện:** Nguyễn Quý Trọng

**Mã số sinh viên:** 2111111548

**TP. Hồ Chí Minh, 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 đồ án môn học.

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.

TP.HCM, ngày 11 tháng 6 năm 2024

Người thực hiện

Nguyễn Quý Trọng

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 11 tháng 6 năm 2024

Người thực hiện

Nguyễn Quý Trọng

MỤC LỤC

[LỜI CẢM ƠN i](#_Toc169055617)

[TRANG CAM KẾT ii](#_Toc169055618)

[MỤC LỤC iii](#_Toc169055619)

[NHẬN XÉT CỦA GIẢNG VIÊN v](#_Toc169055620)

[DANH MỤC BIỂU ĐỒ HÌNH VẼ vi](#_Toc169055621)

[DANH MỤC CÁC BẢNG viii](#_Toc169055622)

[CHƯƠNG 1: GIỚI THIỆU VỀ CƠ SỞ DỮ LIỆU SỬ DỤNG CHO ĐỀ TÀI 1](#_Toc169055623)

[1.1. Tổng quan về cơ sở dữ liệu 1](#_Toc169055624)

[1.2. Giới thiệu về Dataset: 1](#_Toc169055625)

[1.3. Giới thiệu các thuộc tính (fields) 1](#_Toc169055626)

[1.3.1. Tên của các fields 1](#_Toc169055627)

[1.3.2. Ý nghĩa chi tiết của các fields trong tập dữ liệu 2](#_Toc169055628)

[1.3.3. Số giá trị null của các fields 3](#_Toc169055629)

[1.3.4. Số giá trị unique của các fields 3](#_Toc169055630)

[1.3.5. Đặc điểm của Boolean : 4](#_Toc169055631)

[1.3.6. Kiểu dữ liệu của các fields 4](#_Toc169055632)

[1.3.6.1. Field Age 5](#_Toc169055633)

[1.3.6.2. Field Age Group 5](#_Toc169055634)

[1.3.6.3. Field Anaemia 5](#_Toc169055635)

[1.3.6.4. Field creatinine\_phosphokinase(CPK ) 6](#_Toc169055636)

[1.3.6.5. Field Diabetes 6](#_Toc169055637)

[1.3.6.6. Field Ejection Fraction 6](#_Toc169055638)

[1.3.6.7. Field High Blood Pressure 6](#_Toc169055639)

[1.3.6.8. Field Platelets 7](#_Toc169055640)

[1.3.6.9. Field Serum Creatinine 7](#_Toc169055641)

[1.3.6.10. Field Serum Sodium 7](#_Toc169055642)

[1.3.6.11. Field Sex 7](#_Toc169055643)

[1.3.6.12. Field Smoking 8](#_Toc169055644)

[1.3.6.13. Field Time 8](#_Toc169055645)

[1.3.6.14. Field Death 8](#_Toc169055646)

[CHƯƠNG 2: PHÂN TÍCH – THỐNG KÊ TRÊN CƠ SỞ DỮ LIỆU ĐÃ CHỌN 9](#_Toc169055647)

[2.1. Tìm hiểu dữ liệu 9](#_Toc169055648)

[2.1.1. Chọn 3 thuộc tính để vẽ các đồ thị 9](#_Toc169055649)

[2.1.1.2. Quantile–Quantile Plot trên 2 thuộc tính bất kỳ nhưng có liên quan về ý nghĩa 10](#_Toc169055650)

[2.1.1.3. Histogram trên 2 thuộc tính bất kỳ nhưng có liên quan về ý nghĩa 11](#_Toc169055651)

[2.1.1.4. Scatter trên 2 thuộc tính bất kỳ nhưng có liên quan về ý nghĩa 12](#_Toc169055652)

[2.1.2. Nhóm dữ liệu đang có theo một thuộc tính dạng danh nghĩa 13](#_Toc169055653)

[2.1.2.1. Boxplot dựa trên five-number summary 13](#_Toc169055654)

[2.1.2.2. Histogram trên 2 thuộc tính bất kỳ nhưng có liên quan về ý nghĩa : 15](#_Toc169055655)

[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 18](#_Toc169055656)

[2.1.3.1. Ma trận tương quan : 19](#_Toc169055657)

[2.1.3.2. Độ đo Cosin 22](#_Toc169055658)

[2.2. Thực hiện khai thác dữ liệu 24](#_Toc169055659)

[2.2.1. Sử dụng các phương pháp khai phá dữ liệu đã biết để khai thác dữ liệu đã chọn trong phần 1 (tập phổ biến, phân lớp, phân cụm) với yêu cầu thực hiện tối thiểu 2 phương pháp bất kỳ do SV tự chọn (ví dụ sử dụng Apriori và FP-growth) 24](#_Toc169055660)

[2.2.1.1. Thực hiện phân loại bằng Naive Bayes 24](#_Toc169055661)

[2.2.1.1. Thực hiện phân cụm với K-Means 26](#_Toc169055662)

[2.2.2. Thực hiện đánh giá các mẫu thu được bằng các phương pháp đã biết bằng cách chọn 2 trong số các phương pháp đánh giá để đánh giá kết quả của việc thực hiện ở phần 2.2.1 27](#_Toc169055663)

[2.2.2.1. Đánh giá phương pháp phân loại : 27](#_Toc169055664)

[CHƯƠNG 3: PHÂN TÍCH – THỐNG KÊ BẰNG PYTHON 29](#_Toc169055665)

[3.1. Chọn 3 thuộc tính để vẽ đồ thị trên python : 29](#_Toc169055666)

[3.1.1. Vẽ đồ thị Boxplot 29](#_Toc169055667)

[3.1.1.1. Boxplot của Age : 29](#_Toc169055668)

[3.1.1.2. Boxplot của Ejection Fraction : 30](#_Toc169055669)

[3.1.1.3. Boxplot của Platelets : 30](#_Toc169055670)

[3.1.2. Quantile–Quantile Plot trên 2 thuộc tính bất kỳ nhưng có liên quan về ý nghĩa 31](#_Toc169055671)

[3.1.3. Histogram trên 2 thuộc tính bất kỳ nhưng có liên quan về ý nghĩa 31](#_Toc169055672)

[3.1.3.1. Histogram của Age : 32](#_Toc169055673)

[3.1.3.2. Histogram của Serum Creatine : 32](#_Toc169055674)

[3.1.3.1. Scatter trên 2 thuộc tính bất kỳ nhưng có liên quan về ý nghĩa 33](#_Toc169055675)

[3.1.4. Nhóm dữ liệu đang có theo một thuộc tính dạng danh nghĩa 33](#_Toc169055676)

[3.1.4.1. Filter theo Middle-aged : 34](#_Toc169055677)

[3.1.4.2. Filter theo Elderly : 36](#_Toc169055678)

[3.1.4.3. Filter theo Senior citizens: 38](#_Toc169055679)

[3.1.5. Đo lường sự tương đồng và khác biệt của dữ liệu bằng 2 cách: ma trận tương quan và độ đo Cosin 40](#_Toc169055680)

[3.1.5.1. Ma trận tương quan : 40](#_Toc169055681)

[3.1.5.2. Độ đo Cosin : 41](#_Toc169055682)

[3.2. Thực hiện khai thác dữ liệu : 41](#_Toc169055683)

[3.2.1. Thực hiện phân loại bằng Naive Bayes 41](#_Toc169055684)

[3.2.2. Thực hiện phân cụm với K-Means : 42](#_Toc169055685)

[3.3. Đánh giá phương pháp phân loại : 43](#_Toc169055686)

[3.4. So sánh kết quả thực hiện của việc sử dụng công cụ với việc thực hiện thủ công : 43](#_Toc169055687)

[PHỤ LỤC 44](#_Toc169055688)

[TÀI LIỆU THAM KHẢO 45](#_Toc169055689)

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 9](#_Toc169055298)

[Hình 2.2 Boxplot Ejection Fraction 9](#_Toc169055299)

[Hình 2.3 Boxplot Platelets 10](#_Toc169055300)

[Hình 2.4 Q-Q Plot giữa Age và Serum Creatine 11](#_Toc169055301)

[Hình 2.5 Histogram của Age 11](#_Toc169055302)

[Hình 2.6 Histogram của Serum Creatinine 12](#_Toc169055303)

[Hình 2.7 Scatter của Serum Creatinine và Age 12](#_Toc169055304)

[Hình 2.8 Boxplot của Age filter theo Middle-aged 13](#_Toc169055305)

[Hình 2.9 Boxplot của Serum Creatine filter theo Middle-aged 13](#_Toc169055306)

[Hình 2.10 Boxplot của Age filter theo Elderly 14](#_Toc169055307)

[Hình 2.11 Boxplot của Serum Creatinine filter theo Elderly 14](#_Toc169055308)

[Hình 2.12 Boxplot của Age filter theo Senior Citizens 15](#_Toc169055309)

[Hình 2.13 Boxplot của Serum Creatinine filter theo Senior Citizens 15](#_Toc169055310)

[Hình 2.14 Histogram của cột Age filter theo Middle Age 16](#_Toc169055311)

[Hình 2.15 Histogram của cột Serum Creatine filter theo Middle Age 16](#_Toc169055312)

[Hình 2.16 Histogram của cột Age filter theo Elderly 17](#_Toc169055313)

[Hình 2.17 Histogram của cột Serum Creatine filter theo Elderly 17](#_Toc169055314)

[Hình 2.18 Histogram của cột Age filter theo Senior Citizens 18](#_Toc169055315)

[Hình 2.19 Histogram của cột Serum Creatine filter theo Senior Citizens 18](#_Toc169055316)

[Hình 2.20 Xếp hạng thuộc tính thứ tự 20](#_Toc169055317)

[Hình 2.21 Công thức tính ma trận tương quan 21](#_Toc169055318)

[Hình 2.22 Chuẩn hóa lại thuộc tính số (Age) 21](#_Toc169055319)

[Hình 3.1 Boxplot Age Python 29](#_Toc169055320)

[Hình 3.2 Boxplot Ejection Fraction Python 30](#_Toc169055321)

[Hình 3.3 Boxplot Platelets Python 31](#_Toc169055322)

[Hình 3.4 Q-Q Plot giữa Age và Serum Creatine Python 31](#_Toc169055323)

[Hình 3.5 Histogram của Age python 32](#_Toc169055324)

[Hình 3.6 Histogram của Serum Creatinine python 32](#_Toc169055325)

[Hình 3.7 Scatter của Serum Creatinine và Age python 33](#_Toc169055326)

[Hình 3.8 Boxplot của cột Age filter theo Middle Age python 34](#_Toc169055327)

[Hình 3.9 Boxplot của cột Serum Creatinine filter theo Middle Age python 34](#_Toc169055328)

[Hình 3.10 Histogram của cột Age filter theo Middle Age python 35](#_Toc169055329)

[Hình 3.11 Histogram của cột Serum Creatinine filter theo Middle Age python 35](#_Toc169055330)

[Hình 3.12 Boxplot của cột Age filter theo Elderly python 36](#_Toc169055331)

[Hình 3.13 Boxplot của cột Serum Creatinine filter theo Elderly python 36](#_Toc169055332)

[Hình 3.14 Histogram của cột Age filter theo Elderly python 37](#_Toc169055333)

[Hình 3.15 Histogram của cột Serum Creatinine filter theo Elderly python 37](#_Toc169055334)

[Hình 3.16 Boxplot của cột Age filter theo Senior citizens python 38](#_Toc169055335)

[Hình 3.17 Boxplot của cột Serum Creatinine filter theo Senior citizens python 38](#_Toc169055336)

[Hình 3.18 Histogram của cột Age filter theo Senior citizens python 39](#_Toc169055337)

[Hình 3.19 Histogram của cột Serum Creatinine filter theo Senior citizens python 40](#_Toc169055338)

DANH MỤC CÁC BẢNG

[Bảng 1 Field Age 5](#_Toc169055547)

[Bảng 2 Field Age Group 5](#_Toc169055548)

[Bảng 3 Field Anaemia 5](#_Toc169055549)

[Bảng 4 Field Creatinine Phosphokinase 6](#_Toc169055550)

[Bảng 5 Field Diabetes 6](#_Toc169055551)

[Bảng 6 Field Ejection Fraction 6](#_Toc169055552)

[Bảng 7 Field High Blood Pressure 6](#_Toc169055553)

[Bảng 8 Field Platelets 7](#_Toc169055554)

[Bảng 9 Field 1.1.1.1. Serum Creatinine 7](#_Toc169055555)

[Bảng 10 Field 1.1.1.1. Serum Sodium 7](#_Toc169055556)

[Bảng 11 Field Sex 7](#_Toc169055557)

[Bảng 12 Field Smoking 8](#_Toc169055558)

[Bảng 13 Field Time 8](#_Toc169055559)

[Bảng 14 Field Death 8](#_Toc169055560)

[Bảng 15 Bảng năm dòng dữ liệu dùng để đo lường 19](#_Toc169055561)

[Bảng 16 Ma trận sai phân thuộc tính nhị phân 19](#_Toc169055562)

[Bảng 17 Ma trận sai phân thuộc tính số 19](#_Toc169055563)

[Bảng 18 Thay thế khoảng giá trị 20](#_Toc169055564)

[Bảng 19 Ma trận sai phân thuộc tính thứ tự 21](#_Toc169055565)

[Bảng 20 Ma trận tương quan 22](#_Toc169055566)

[Bảng 21 Dữ liệu ban đầu của 3 thuộc tính 22](#_Toc169055567)

[Bảng 22 Chuẩn hóa thuộc tính danh nghĩa và nhị phân 23](#_Toc169055568)

[Bảng 23 Ví dụ buóc 3 phân cụm 26](#_Toc169055569)

# 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 “***Predict mortality by heart failure***”, có ý nghĩa là ***dự đoán tỷ lệ tử vong do suy tim***. Tập dataset này được em thu tập từ nguồn [Kaggle: Your Home for Data Science](https://www.kaggle.com/).

## Giới thiệu về Dataset:

Bệnh tim mạch (CVD) là nguyên nhân gây tử vong số 1 trên toàn cầu , cướp đi sinh mạng của khoảng 17,9 triệu người mỗi năm , chiếm 31% tổng số ca tử vong trên toàn thế giới .

Suy tim là một biến cố phổ biến do bệnh tim mạch gây ra và bộ dữ liệu này chứa 12 đặc điểm có thể được sử dụng để dự đoán tỷ lệ tử vong do suy tim.

Hầu hết các bệnh tim mạch có thể được ngăn ngừa bằng cách giải quyết các yếu tố nguy cơ về hành vi như sử dụng thuốc lá, chế độ ăn uống không lành mạnh và béo phì, ít hoạt động thể chất và sử dụng rượu có hại bằng các chiến lược toàn dân.

Những người mắc bệnh tim mạch hoặc có nguy cơ tim mạch cao (do có một hoặc nhiều yếu tố nguy cơ như tăng huyết áp, tiểu đường, tăng lipid máu hoặc đã mắc bệnh) cần được phát hiện và quản lý sớm trong đó mô hình học máy có thể giúp ích rất nhiều.

## Giới thiệu các thuộc tính (fields)

Tập dataset này có 299 records và 13 fields (column).

### Tên của các fields

* Age
* Anaemia
* Creatinine phosphokinase (CPK)
* Diabetes
* Ejection fraction
* Sex
* Platelets
* Serum creatinine
* Serum sodium
* Smoking
* Time
* death event

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

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Tên Field** | **Ý Nghĩa** | **Ý Nghĩa(Tiếng việt)** | **Đơn Vị** | **Phạm Vi** |
| Age | Age of the patient | Tuổi của bệnh nhân | Years | [40,…, 95] |
| Anaemia | Decrease of red blood cells or hemoglobin | Giảm hồng cầu hoặc huyết sắc tố | Boolean | 0, 1 |
| Age Group | Age distribution by group | Phân bố độ tuổi theo nhóm | Elderly/ Middle-Aged/Senior citizens | [Elderly,Middle-Aged,Senior citizens] |
| High blood pressure | If a patient has hypertension | Nếu bệnh nhân bị tăng huyết áp | Boolean | 0, 1 |
| Creatinine phosphokinase (CPK) | Level of the CPK enzyme in the blood | Nồng độ enzym CPK trong máu | mcg/L | [23,…, 7861] |
| Diabetes | If the patient has diabetes | Nếu bệnh nhân mắc bệnh tiểu đường | Boolean | 0, 1 |
| Ejection fraction | Percentage of blood leaving | Tỷ lệ máu đi | % | [14,…, 80] |
| Sex | Woman or man | Nữ hay nam | Boolean | 0, 1 |
| Platelets | Platelets in the blood | Tiểu cầu trong máu | kilotiểu cầu/mL | [25.01,…, 850.00] |
| Serum creatinine | Level of creatinine in the blood | Nồng độ creatinine trong máu | mg/dL | [0.50,…, 9.40] |
| Serum sodium | Level of sodium in the blood | Nồng độ natri trong máu | mEq/L | [114,…, 148] |
| Smoking | If the patient smokes | Nếu bệnh nhân hút thuốc | Boolean | 0, 1 |
| Time | Follow-up period | Thời gian theo dõi | Days | [4,…,285] |
| Death event | If the patient died during the follow-up period | Nếu bệnh nhân tử vong trong thời gian theo dõi | Boolean | 0, 1 |

### Số giá trị null của các fields

|  |  |
| --- | --- |
| Thuộc tính (field) | Số giá trị null |
| Age | 0 |
| Age Group | 0 |
| Anaemia | 0 |
| Creatinine Phosphokinase (CPK) | 0 |
| Diabetes | 0 |
| High Blood Pressure | Boolean |
| Ejection Fraction | 0 |
| Sex | 0 |
| Platelets | 0 |
| Serum Sreatinine | 0 |
| Serum Sodium | 0 |
| Smoking | 0 |
| Time | 0 |
| Death Event | 0 |

### Số giá trị unique của các fields

|  |  |
| --- | --- |
| Thuộc tính (field) | Số giá trị unique |
| Age | 0 |
| Age Group | 0 |
| Anaemia | 0 |
| Creatinine Phosphokinase (CPK) | 0 |
| Diabetes | 0 |
| High Blood Pressure | Boolean |
| Ejection Fraction | 0 |
| Sex | 0 |
| Platelets | 0 |
| Serum Sreatinine | 0 |
| Serum Sodium | 0 |
| Smoking | 0 |
| Time | 0 |
| Death Event | 0 |

### Đặc điểm của Boolean :

* Sex : Giới tính của bệnh nhân Nam = 1, Nữ =0
* Diabetes : 0 = Không, 1 = Có
* Anaemia : 0 = Không, 1 = Có
* High\_blood\_pressure : 0 = Không, 1 = Có
* Smoking : 0 = Không, 1 = Có
* DEATH\_EVENT : 0 = Không, 1 = Có

### Kiểu dữ liệu của các fields

|  |  |
| --- | --- |
| Thuộc tính (field) | Kiểu Dữ Liệu |
| Age | Int |
| Age Group | String |
| Anaemia | Boolean |
| Creatinine Phosphokinase (CPK) | Int |
| Diabetes | Boolean |
| High Blood Pressure | Boolean |
| Ejection Fraction | Int |
| Sex | Boolean |
| Platelets | Int |
| Serum Sreatinine | Float |
| Serum Sodium | Int |
| Smoking | Boolean |
| Time | Int |
| Death Event | Int |

#### Field Age

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **AGE** | | | | | | | |
| Mean | Midrange | Mode | Five-number summary | | | | |
| 60.83 | 67.5 | 60 | Min | Q1 | Median | Q3 | Max |
|  |  | Xuất hiện 33 lần | 40 | 51 | 60 | 70 | 95 |

Bảng 1 Field Age

#### Field Age Group

|  |  |  |  |
| --- | --- | --- | --- |
| **REGION** | | | |
| Giá trị | Số lượng | Chiếm (%) | Mode |
| Middle-aged | 162 | 54 | ✓ |
| Elderly | 119 | 40 |  |
| Senior Citizens | 18 | 6 |  |
| Tổng | 1338 | 100 |  |

Bảng 2 Field Age Group

#### Field Anaemia

|  |  |  |  |
| --- | --- | --- | --- |
| **ANAMIA** | | | |
| Giá trị | Số lượng | Chiếm (%) | Mode |
| 0 | 170 | 56.86 | ✓ |
| 1 | 129 | 43.14 |  |
| Tổng | 299 | 100 |  |

Bảng 3 Field Anaemia

#### Field creatinine\_phosphokinase(CPK )

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **CREATININE\_PHOSPHOKINASE** | | | | | | | |
| Mean | Midrange | Mode | Five-number summary | | | | |
| 581.83 | 3942 | 582 | Min | Q1 | Median | Q3 | Max |
|  |  | Xuất hiện 47 lần | 23 | 116.5 | 250 | 582 | 7861 |

Bảng 4 Field Creatinine Phosphokinase

#### Field Diabetes

|  |  |  |  |
| --- | --- | --- | --- |
| **DIABETES** | | | |
| Giá trị | Số lượng | Chiếm (%) | Mode |
| 0 | 174 | 58.19 | ✓ |
| 1 | 125 | 41.81 |  |
| Tổng | 299 | 100 |  |

Bảng 5 Field Diabetes

#### Field Ejection Fraction

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **EJECTION\_FRACTION** | | | | | | | |
| Mean | Midrange | Mode | Five-number summary | | | | |
| 38.08 | 47 | 35 | Min | Q1 | Median | Q3 | Max |
|  |  | Xuất hiện 49 lần | 14 | 30 | 38 | 45 | 80 |

Bảng 6 Field Ejection Fraction

#### Field High Blood Pressure

|  |  |  |  |
| --- | --- | --- | --- |
| **HIGH BLOOD PRESSURE** | | | |
| Giá trị | Số lượng | Chiếm (%) | Mode |
| 0 | 194 | 64.88 | ✓ |
| 1 | 105 | 35.12 |  |
| Tổng | 299 | 100 |  |

Bảng 7 Field High Blood Pressure

#### Field Platelets

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **PLATELETS** | | | | | | | |
| Mean | Midrange | Mode | Five-number summary | | | | |
| 263358 | 437550 | 263358.03 | Min | Q1 | Median | Q3 | Max |
|  |  | Xuất hiện 25 lần | 25100 | 212500 | 262000 | 303500 | 850000 |

Bảng 8 Field Platelets

#### Field Serum Creatinine

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **SERUM CREATININE** | | | | | | | |
| Mean | Midrange | Mode | Five-number summary | | | | |
| 1.39 | 4.95 | 1 | Min | Q1 | Median | Q3 | Max |
|  |  | Xuất hiện 50 lần | 0.5 | 0.9 | 1.1 | 1.4 | 9.4 |

Bảng 9 Field 1.1.1.1. Serum Creatinine

#### Field Serum Sodium

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **SERUM SODIUM** | | | | | | | |
| Mean | Midrange | Mode | Five-number summary | | | | |
| 136.62 | 130.5 | 136 | Min | Q1 | Median | Q3 | Max |
|  |  | Xuất hiện 40 lần | 113 | 134 | 137 | 140 | 148 |

Bảng 10 Field 1.1.1.1. Serum Sodium

#### Field Sex

|  |  |  |  |
| --- | --- | --- | --- |
| **SEX** | | | |
| Giá trị | Số lượng | Chiếm (%) | Mode |
| 0 | 105 | 35.12 |  |
| 1 | 194 | 64.88 | ✓ |
| Tổng | 299 | 100 |  |

Bảng 11 Field Sex

#### Field Smoking

|  |  |  |  |
| --- | --- | --- | --- |
| **SMOKING** | | | |
| Giá trị | Số lượng | Chiếm (%) | Mode |
| 0 | 203 | 67.89 | ✓ |
| 1 | 96 | 32.1 |  |
| Tổng | 299 | 100 |  |

Bảng 12 Field Smoking

#### Field Time

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **TIME** | | | | | | | |
| Mean | Midrange | Mode | Five-number summary | | | | |
| 130.26 | 144.5 | 187 | Min | Q1 | Median | Q3 | Max |
|  |  | Xuất hiện 7 lần | 4 | 73 | 115 | 203 | 285 |

Bảng 13 Field Time

#### Field Death

|  |  |  |  |
| --- | --- | --- | --- |
| **DEATH** | | | |
| Giá trị | Số lượng | Chiếm (%) | Mode |
| 0 | 203 | 67.89 | ✓ |
| 1 | 96 | 32.1 |  |
| Tổng | 299 | 100 |  |

Bảng 14 Field Death

# PHÂN TÍCH – THỐNG KÊ TRÊN CƠ SỞ DỮ LIỆU ĐÃ CHỌN

## Tìm hiểu dữ liệu

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

3 thuộc tính em chọn để vẽ các đồ thị là: Age, Ejection Fraction và Platelets

##### Boxplot dựa trên five-number summary

###### Age

A screenshot of a graph

Description automatically generated

Hình 2.1 Boxplot Age

###### Ejection fraction

A screenshot of a graph

Description automatically generated

Hình 2.2 Boxplot Ejection Fraction

###### PLATELETS

A screenshot of a computer

Description automatically generated

Hình 2.3 Boxplot Platelets

#### 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: Serum Creatinine và Age.

Mức độ serum creatinine trong máu là một chỉ số quan trọng để đánh giá chức năng thận. Nồng độ cao của serum creatinine có thể chỉ ra rằng thận không hoạt động hiệu quả và có thể là dấu hiệu của bệnh thận hoặc suy thận.

Age (Tuổi tác): Tuổi tác có thể ảnh hưởng đáng kể đến chức năng thận. Khi con người già đi, chức năng thận thường giảm dần. Điều này có nghĩa là ở người lớn tuổi, mức độ serum creatinine trong máu có thể tăng lên một cách tự nhiên, ngay cả khi không có bệnh lý cụ thể về thận.

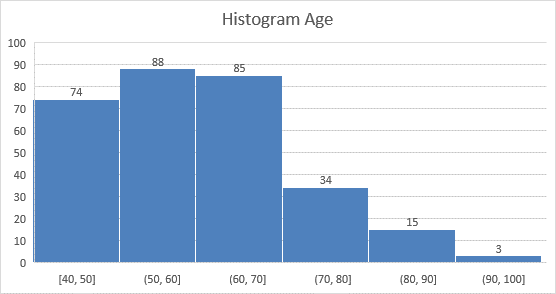
Serum Creatinine và Age là hai chỉ số quan trọng có mối quan hệ mật thiết trong lĩnh vực y tế. Việc theo dõi và đánh giá cả hai chỉ số này giúp các bác sĩ hiểu rõ hơn về tình trạng sức khỏe của bệnh nhân và đưa ra các quyết định điều trị phù hợp.

Hình 2.4 Q-Q Plot giữa Age và Serum Creatine

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

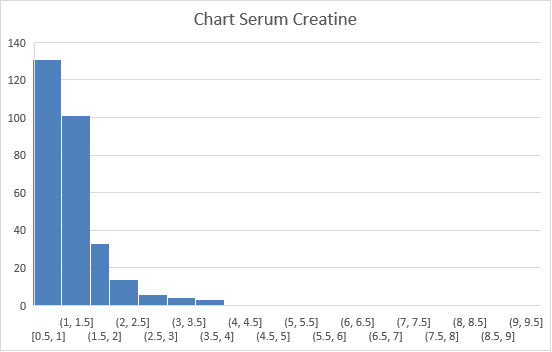
Ở đây 2 thuộc tính liên quan về ý nghĩa em chọn tiếp tục là: Serum Creatinine và Age .

###### Histogram của Age



Hình 2.5 Histogram của Age

###### Histogram của Serum Creatine



Hình 2.6 Histogram của Serum Creatinine

#### 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 vẫn là: Serum Creatinine và Age. Serum Creatinine và Age là hai chỉ số quan trọng có mối quan hệ mật thiết trong lĩnh vực y tế. Việc theo dõi và đánh giá cả hai chỉ số này giúp các bác sĩ hiểu rõ hơn về tình trạng sức khỏe của bệnh nhân và đưa ra các quyết định điều trị phù hợp.

Hình 2.7 Scatter của Serum Creatinine và Age

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

Ở đây em chọn thuộc tính age\_group gồm 3 giá trị là: Middle-aged, Elderly, Senior citizens.

#### Boxplot dựa trên five-number summary

Filter theo field age\_group gồm 3 giá trị là: Middle-aged, Elderly , Senior citizens

##### Sau khi đã filter theo Middle-aged :

Boxplot của field Age sau khi đã filter theo Middle-aged:

A screenshot of a graph

Description automatically generated

Hình 2.8 Boxplot của Age filter theo Middle-aged

Boxplot của field Serum Creatinine sau khi đã filter theo Middle-aged:

A screenshot of a graph

Description automatically generated

Hình 2.9 Boxplot của Serum Creatine filter theo Middle-aged

##### Sau khi đã filter theo Elderly:

Boxplot của field Age sau khi đã filter theo Elderly:

A screenshot of a computer

Description automatically generated

Hình 2.10 Boxplot của Age filter theo Elderly

Boxplot của field Serum Creatinine sau khi đã filter theo Elderly:

A screenshot of a graph

Description automatically generated

Hình 2.11 Boxplot của Serum Creatinine filter theo Elderly

##### Sau khi đã filter theo Senior Citizens:

Boxplot của field Age sau khi đã filter theo Senior Citizens:

A graph with a blue bar

Description automatically generated

Hình 2.12 Boxplot của Age filter theo Senior Citizens

Boxplot của field Serum Creatinine sau khi đã filter theo Senior Citizens:

A screenshot of a graph

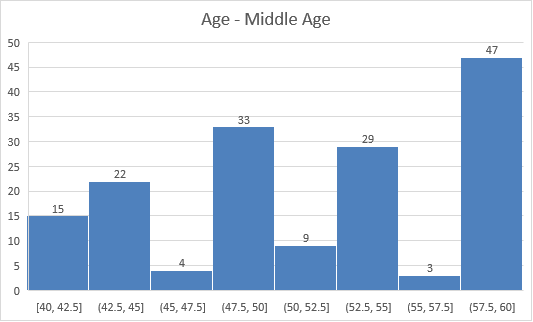
Description automatically generated

Hình 2.13 Boxplot của Serum Creatinine filter theo Senior Citizens

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

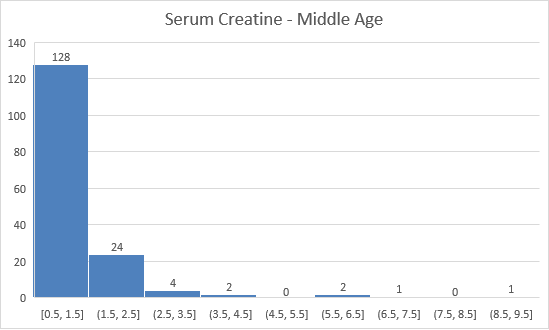
##### Filter theo Middle Age

Histogram của cột Age filter theo Middle Age



Hình 2.14 Histogram của cột Age filter theo Middle Age

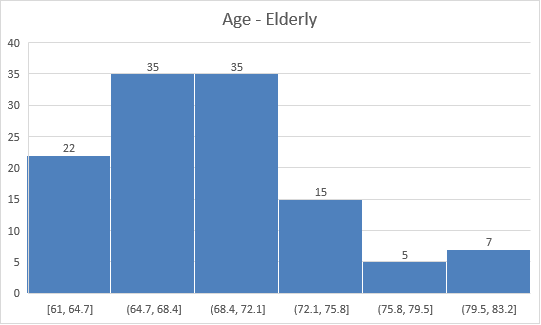
Histogram của cột Serum Creatine filter theo Middle Age



Hình 2.15 Histogram của cột Serum Creatine filter theo Middle Age

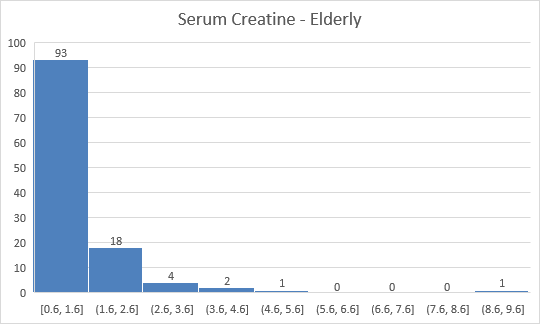
##### Filter theo Elderly

Histogram của cột Age filter theo Elderly



Hình 2.16 Histogram của cột Age filter theo Elderly

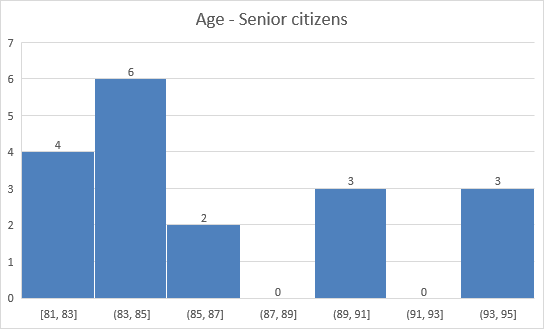
Histogram của cột Serum Creatine filter theo Elderly



Hình 2.17 Histogram của cột Serum Creatine filter theo Elderly

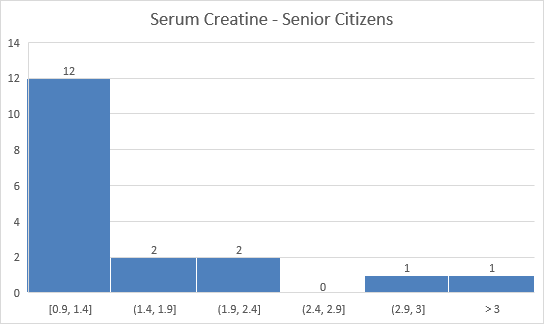
##### Filter theo Senior Citizens

Histogram của cột Age filter theo Senior Citizens



Hình 2.18 Histogram của cột Age filter theo Senior Citizens

Histogram của cột Serum Creatine filter theo Senior Citizens



Hình 2.19 Histogram của cột Serum Creatine filter theo Senior Citizens

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

* Vì không có thuộc tính dạng danh nghĩa cho nên em xin dùng 3 thuộc tính để vẽ ma trận tương quan và tính độ đo Cosin em chọn là: Age (dạng số ), Anaemia (dạng nhị phân) và Age Group (dạng thứ tự) .
* Chọn tối thiểu 4 dòng của dữ liệu đã lọc:

|  |  |  |  |
| --- | --- | --- | --- |
| Thứ tự | Age | Anaemia | Age Group |
| A | 55 | 0 | Middle-aged |
| B | 65 | 0 | Elderly |
| C | 50 | 1 | Middle-aged |
| D | 65 | 1 | Elderly |
| E | 90 | 1 | Senior citizens |

Bảng 15 Bảng năm dòng dữ liệu dùng để đo lường

#### Ma trận tương quan :

Trước tiên, ta cần tính ma trận sai phân của từng thuộc tính:

##### Ma trận sai phân thuộc tính dạng nhị phân (Anaemia)

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

Bảng 16 Ma trận sai phân thuộc tính nhị phân

##### Ma trận sai phân thuộc tính dạng số (age)

Sử dụng khoảng cách Manhattan:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **A** | **B** | **C** | **D** | **E** |
| **A** | 0 |  |  |  |  |
| **B** | 10 | 0 |  |  |  |
| **C** | 5 | 15 | 0 |  |  |
| **D** | 10 | 0 | 15 | 0 |  |
| **E** | 35 | 25 | 40 | 25 | 0 |

Bảng 17 Ma trận sai phân thuộc tính số

##### Ma trận sai phân thuộc tính dạng thứ tự (Age Group)

A screenshot of a computer

Description automatically generated

Hình 2.20 Xếp hạng thuộc tính thứ tự

* Giá trị Middle Age (Rank =1) = 0
* Giá trị Elderly (Rank =2) = 0,5
* Giá trị Senior Citizens (Rank =3) = 1

**Suy ra chia được 3 khoảng giá trị:**

0 0,5 1

Rank =1 Rank =2 Rank =3

**Thay thế giá trị:**

|  |  |
| --- | --- |
| ID | Age Group |
| A | 0 |
| B | 0,5 |
| C | 0 |
| D | 0.5 |
| E | 1 |

Bảng 18 Thay thế khoảng giá trị

* Ma trận sai phân (dùng khoảng cách Manhattan):

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **A** | **B** | **C** | **D** | **E** |
| **A** | 0 |  |  |  |  |
| **B** | 0,5 | 0 |  |  |  |
| **C** | 0 | 0,5 | 0 |  |  |
| **D** | 0,5 | 0 | 0,5 | 0 |  |
| **E** | 1 | 0,5 | 1 | 0,5 | 0 |

Bảng 19 Ma trận sai phân thuộc tính thứ tự

##### Ma trận tương quan (hỗn hợp)

Dựa vào các ma trận sai phân của 4 thuộc tính trên, ta có thể tính được ma trận hỗn hợp dựa vào công thức:

A mathematical equation with numbers and symbols

Description automatically generated

Hình 2.21 Công thức tính ma trận tương quan

A white sheet with black text and numbers

Description automatically generated

Hình 2.22 Chuẩn hóa lại thuộc tính số (Age)

Dựa vào công thức trên, ta suy ra được ma trận hỗn hợp với các giá trị của ma trận như sau:

* d(B,A) = = 0.25
* d(C,A) = = 0.375
* d(D,A) = = 0.583
* d(E,A) = = 0.958
* d(C,B) = = 0.625
* d(D,B) = = 0.333
* d(E,B) = = 0.708
* d(D,C) = = 0.291
* d(E,C) = = 0.666
* d(E,D) = = 0.375

Suy ra, ta có ma trận tương quan như sau:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **A** | **B** | **C** | **D** | **E** |
| **A** | 0 |  |  |  |  |
| **B** | 0.25 | 0 |  |  |  |
| **C** | 0.375 | 0.625 | 0 |  |  |
| **D** | 0.583 | 0.333 | 0.291 | 0 |  |
| **E** | 0.958 | 0.708 | 0.666 | 0.375 | 0 |

Bảng 20 Ma trận tương quan

#### Độ đo Cosin

Ta cần chuẩn hóa các thuộc tính như danh nghĩa và nhị phần về dạng số, và chuyển các thuộc tính thành các vector:

Đây là dữ liệu ban đầu của 3 thuộc tính:

|  |  |  |
| --- | --- | --- |
| Age | Anaemia | Age Group |
| 55 | 0 | Middle-aged |
| 65 | 0 | Elderly |
| 50 | 1 | Middle-aged |
| 65 | 1 | Elderly |
| 90 | 1 | Senior citizens |

Bảng 21 Dữ liệu ban đầu của 3 thuộc tính

##### Chuẩn hóa thuộc tính danh nghĩa và nhị phân về dạng số :

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Age | Anaemia | Middle Age | Elderly | Senior Citizens |
| 55 | 0 | 1 | 0 | 0 |
| 65 | 0 | 0 | 1 | 0 |
| 50 | 1 | 1 | 0 | 0 |
| 65 | 1 | 0 | 1 | 0 |
| 90 | 1 | 0 | 0 | 1 |

Bảng 22 Chuẩn hóa thuộc tính danh nghĩa và nhị phân

##### Các vecto thu được

* Vecto A = [55,0,1,0,0]
* Vecto B = [65,0,0,1,0]
* Vecto C = [50,1,1,0,0]
* Vecto D = [65,1,0,1,0]
* Vecto E = [90,1,0,0,1]

##### Thực hiện tính độ tương quan giữa các vecto:

**Áp dụng công thức** : sin(a,b) =

sin(A,B) = = 0,99971

sin(A,C) = = 0,9998

sin(A,D) = = 0,9996

sin(A,E) = = 0,9997

sin(B,C) = = 0,9994

sin(B,D) = = 0,999648

sin(B,E) = = 0.99975

sin(C,D) = = 0.99967

sin(C,E) = = 0.9997

sin(D,E) = = 0.99981

###### Kết luận

* Cặp D,E có độ tương đồng cao nhất với giá trị độ tương tự cosin =0.99981
* Cặp A,D có độ tương đồng thấp nhất với giá trị độ tương tự cosin = 0,9996

## Thực hiện khai thác dữ liệu

### Sử dụng các phương pháp khai phá dữ liệu đã biết để khai thác dữ liệu đã chọn trong phần 1 (tập phổ biến, phân lớp, phân cụm) với yêu cầu thực hiện tối thiểu 2 phương pháp bất kỳ do SV tự chọn (ví dụ sử dụng Apriori và FP-growth)

#### Thực hiện phân loại bằng Naive Bayes

Để thực hiện phân loại bằng Naïve Bayes , Em chọn ra 4 cột có liên quan nhất với cột Dead Event (Target) là : Age , Anaemia , Serum Creatinine , High blood pressure. Trong đó , có Age và Serum Creatinine ở dạng số , em sẽ chuyển đổi các giá trị này về các khoảng để có thể dễ dàng thực hiện phân loại , phân cụm ở các phần dưới , cụ thể các giá trị thành các khoảng như sau :

Thuộc tính Age :

|  |  |
| --- | --- |
| **Giá Trị** | **Giá trị thay thế** |
| 40 - 60 | Middle-aged |
| 60 - 80 | Elderly |
| 80 - 100 | Senior Citizens |

Thuộc tính Serum Creatinine :

|  |  |
| --- | --- |
| **Giá trị** | **Giá trị thay thế** |
| < 1.0 | Low |
| 1.0 – 1.5 | Normal |
| 1.5 – 2.0 | High |
| > 2.0 | Very High |

##### Với giá trị Input không trùng với bất kì dòng nào trong dataset là :

X = ( Age = Senior citizens , Anaemia = 0 , serum\_creatinine : Normal , High Blood\_Pressure = 1 )

##### Xác suất của thuộc tính Dead Event như sau :

P (Yes| Dead Event ) = = 0.321

P(No | Dead Event) = = 0.679

##### Thực hiện tính các giá trị xác suất có điều kiện :

**Age**

(Age = Senior Citizens có 18 dòng )

P(Age = Senior Citizens | Dead Event = Yes ) = = 0.1354

P(Age = Senior Citizens | Dead Event = No ) = = 0.0246

**Anaemia**

(Anaemia = 0 có 170 dòng )

P(Anaemia = 0 | Dead Event = Yes ) = = 0.255

P(Anaemia = 0 | Dead Event = No ) = = 0.591

**Serum Creatinine**

( Serum Creatinine : Normal có 146 dòng )

P(Serum Creatinine = Normal | Dead Event = Yes ) = = 0.4375

P(Serum Creatinine = Normal | Dead Event = No ) = = 0.1523

**High Blood Pressure**

( High Blood Pressure: 1 có 105 dòng )

P(High Blood Pressure= 1 | Dead Event = Yes ) = = 0.4

P(High Blood Pressure= 1 | Dead Event = No ) = = 0.325

P(Dead Event = Yes | X ) = P (Yes| Dead Event ) \* P(Age = Senior Citizens | Dead Event = Yes ) \* P(Anaemia = 0 | Dead Event = Yes ) \* P(Serum Creatinine = Normal | Dead Event = Yes ) \* P(High Blood Pressure= 1 | Dead Event = Yes )

= 0.321 \* 0.1354 \* 0.255 \* 0.4375 \* 0.4

= 0.0019

P(Dead Event = No | X ) = P (No | Dead Event ) \* P(Age = Senior Citizens | Dead Event = No) \* P(Anaemia = 0 | Dead Event = No) \* P(Serum Creatinine = Normal | Dead Event = No) \* P(High Blood Pressure= 1 | Dead Event = No)

= 0.679 \* 0.0246 \* 0.591 \* 0.1523 \* 0.325

= 0.0004886

P(Dead Event = Yes | X ) > P(Dead Event = No | X ) => Dead Event = Yes

#### Thực hiện phân cụm với K-Means

Để thực hiện phân cụm bằng K-means thủ công ở Excel, em thực hiện các bước như sau:

***Bước 1:*** Đầu tiên em sẽ chọn giá trị **k** để có thể thực hiện được việc chọn số centroid của dataset. Và giá trị **k** em chọn là bằng 3 vì vậy em sẽ chọn ngẫu nhiên 3 dòng centroid từ dataset ra làm centroid.

***Bước 2:*** Sau khi đã lấy ra được 3 centroid, em thực hiện tính khoảng cách của từng dòng (từng datapoint) trong dữ liệu đến từng centroid **bằng khoảng cách Manhattan.**

***Bước 3:*** Tiếp đến, sau khi đã tính khoảng cách của từng dòng đến 3 centroid, em thực hiện việc phân cụm ra bằng cách xét giá trị nào nhỏ nhất sẽ nằm ở cụm đã được phân. Ví dụ, em có bảng gồm các cột và 2 dòng với các giá trị như sau:

**Cụm**

**Dòng**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **1** | **2** | **3** | **Thuộc** |
| Giá trị datapoint 1 | 3 | 5 | 7 | Cụm 1 (vì 3 là giá trị nhỏ nhất và nằm ở cụm 1) |
| Giá trị datapoint 2 | 4 | 8 | 1 | Cụm 3 (vì 1 là giá trị nhỏ nhất và nằm ở cụm 3) |

Bảng 23 Ví dụ buóc 3 phân cụm

***Bước 4:*** Sau khi đã phân cụm cho các datapoint ở lần đầu tiên, em thực hiện tính lại 3 centroid bằng cách dùng hàm **AVERAGE()** trong excel để tính lại 3 centroid cho dataset, các lần phân cụm tiếp theo cũng cần tính lại centroid sau khi đã phân cụm trước đó.

***Bước 5:*** Sau khi đã thực hiện phân cụm cho các datapoint, em thực hiện việc so sánh cụm đã phân so với lần trước đó (bỏ qua lần đầu tiên). Nếu cụm của các datapoint được phân ở lần mới nhất trùng với cụm của các datapoint ở lần trước đó thì em sẽ dừng việc phân cụm. Còn nếu các cụm vẫn còn khác nhau thì em vẫn tiếp tục phân cụm cho những lần tiếp theo và sau đó so sánh các cụm.

Ở dataset của em, em dừng phân cụm ở lần thứ 15 vì chỉ còn 1 dòng khác cụm so với lần thứ 14.

**Để xem cụ thể phần phân cụm trong excel, em xin mời thầy click vào link sau đây:**

### Thực hiện đánh giá các mẫu thu được bằng các phương pháp đã biết bằng cách chọn 2 trong số các phương pháp đánh giá để đánh giá kết quả của việc thực hiện ở phần 2.2.1

#### Đánh giá phương pháp phân loại :

Trong quá trình đánh giá phương pháp phân loại, em đã thực hiện các bước sau:

1. **Lựa chọn các đặc trưng liên quan**:
   * Em đã chọn ra 4 đặc trưng có liên quan nhất với Dead Event (mục tiêu), bao gồm:
     + Age Group (Nhóm tuổi)
     + Anaemia (Thiếu máu)
     + Serum Creatinine (Nồng độ creatinine trong huyết thanh)
     + High Blood Pressure (Huyết áp cao)
2. **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)
     + 20% dữ liệu còn lại được sử dụng làm tập kiểm tra (test set)
3. **Tính toán xác suất có điều kiện P(X|Dead Event)**:
   * Sử dụng tập huấn luyện (train set), em đã tính toán xác suất có điều kiện P(X|Dead Event) cho từng đặc trưng đã chọn.
4. **Dự đoán trên tập kiểm tra**:
   * Dựa vào các xác suất có điều kiện đã tính toán từ tập huấn luyện (train set), em thực hiện dự đoán Dead Event trên tập kiểm tra (test set).
5. **So sánh với tập dữ liệu ban đầu**:
   * Sau khi dự đoán Dead Event cho các dòng trong tập kiểm tra, em so sánh các dự đoán này với giá trị thực tế của Dead Event trong tập dữ liệu ban đầu. Từ đó, em xác định các giá trị:
     + True Positives (TP)
     + False Negatives (FN)
     + False Positives (FP)
     + True Negatives (TN)
6. **Vẽ Confusion Matrix**:
   * Sau khi có được các giá trị True Positives (TP), False Negatives (FN), False Positives (FP), và True Negatives (TN), em đã vẽ được Confusion Matrix như sau :

A table with numbers and letters

Description automatically generated

# PHÂN TÍCH – THỐNG KÊ BẰNG PYTHON

## Chọn 3 thuộc tính để vẽ đồ thị trên python :

### Vẽ đồ thị Boxplot

3 thuộc tính em chọn để vẽ các đồ thị là: Age, Ejection Fraction và Platelets .

#### Boxplot của Age :

A graph with a line and a rectangle

Description automatically generated

Hình 3.1 Boxplot Age Python

#### Boxplot của Ejection Fraction :

A graph with a line and a line

Description automatically generated with medium confidence

Hình 3.2 Boxplot Ejection Fraction Python

#### Boxplot của Platelets :

A graph with numbers and lines

Description automatically generated

Hình 3.3 Boxplot Platelets Python

### Quantile–Quantile Plot 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 là: Serum Creatinine và Age.

Serum Creatinine và Age là hai chỉ số quan trọng có mối quan hệ mật thiết trong lĩnh vực y tế. Việc theo dõi và đánh giá cả hai chỉ số này giúp các bác sĩ hiểu rõ hơn về tình trạng sức khỏe của bệnh nhân và đưa ra các quyết định điều trị phù hợp.

A graph with a line

Description automatically generated

Hình 3.4 Q-Q Plot giữa Age và Serum Creatine Python

### 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: Serum Creatinine và Age .

#### Histogram của Age :

A blue graph with black text

Description automatically generated

Hình 3.5 Histogram của Age python

#### Histogram của Serum Creatine :

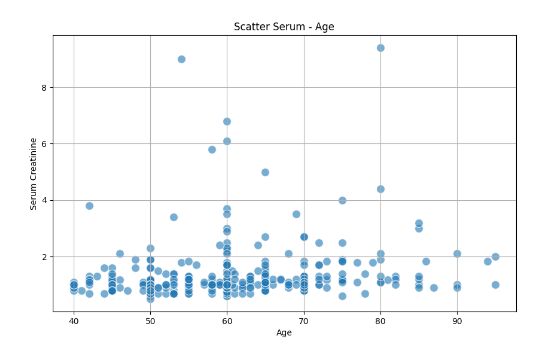
A graph of serum creatine

Description automatically generated

Hình 3.6 Histogram của Serum Creatinine python

#### 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 vẫn là: Serum Creatinine và Age. Serum Creatinine và Age là hai chỉ số quan trọng có mối quan hệ mật thiết trong lĩnh vực y tế. Việc theo dõi và đánh giá cả hai chỉ số này giúp các bác sĩ hiểu rõ hơn về tình trạng sức khỏe của bệnh nhân và đưa ra các quyết định điều trị phù hợp.



Hình 3.7 Scatter của Serum Creatinine và Age python

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

Ở đây em chọn thuộc tính age\_group gồm 3 giá trị là: Middle-aged, Elderly, Senior citizens.

Filter theo field age\_group gồm 3 giá trị là: Middle-aged, Elderly , Senior citizens

#### Filter theo Middle-aged :

##### Boxplot Age :

A screenshot of a computer

Description automatically generated

Hình 3.8 Boxplot của cột Age filter theo Middle Age python

##### Boxplot Serum Creatinine :

A screenshot of a video game

Description automatically generated

Hình 3.9 Boxplot của cột Serum Creatinine filter theo Middle Age python

##### Histogram Age :

A graph of a graph

Description automatically generated with medium confidence

Hình 3.10 Histogram của cột Age filter theo Middle Age python

##### Histogram Serum Creatinine :

A graph with a line going up

Description automatically generated

Hình 3.11 Histogram của cột Serum Creatinine filter theo Middle Age python

#### Filter theo Elderly :

##### Boxplot Age :

A screenshot of a computer

Description automatically generated

Hình 3.12 Boxplot của cột Age filter theo Elderly python

##### Boxplot Serum Creatinine :

A screenshot of a video game

Description automatically generated

Hình 3.13 Boxplot của cột Serum Creatinine filter theo Elderly python

##### Histogram Age :

A graph on a screen

Description automatically generated

Hình 3.14 Histogram của cột Age filter theo Elderly python

##### Histogram Serum Creatinine :

A graph of a person with a line

Description automatically generated

Hình 3.15 Histogram của cột Serum Creatinine filter theo Elderly python

#### Filter theo Senior citizens:

##### Boxplot Age :

A screen shot of a computer

Description automatically generated

Hình 3.16 Boxplot của cột Age filter theo Senior citizens python

##### Boxplot Serum Creatinine :

A screenshot of a computer

Description automatically generated

Hình 3.17 Boxplot của cột Serum Creatinine filter theo Senior citizens python

##### Histogram Age :

A graph on a screen

Description automatically generated

Hình 3.18 Histogram của cột Age filter theo Senior citizens python

##### Histogram Serum Creatinine

A graph on a screen

Description automatically generated

Hình 3.19 Histogram của cột Serum Creatinine filter theo Senior citizens python

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

* Vì không có thuộc tính dạng danh nghĩa cho nên em xin dùng 3 thuộc tính để vẽ ma trận tương quan và tính độ đo Cosin em chọn là: Age (dạng số ), Anaemia (dạng nhị phân) và Age Group (dạng thứ tự) .

#### Ma trận tương quan :

A red and blue squares with numbers

Description automatically generated

#### Độ đo Cosin :

A diagram of numbers and a grid

Description automatically generated with medium confidence

## Thực hiện khai thác dữ liệu :

### Thực hiện phân loại bằng Naive Bayes

Để thực hiện phân loại bằng Naïve Bayes , Em chọn ra 4 cột có liên quan nhất với cột Dead Event (Target) là : Age , Anaemia , Serum Creatinine , High blood pressure. Trong đó , có Age và Serum Creatinine ở dạng số , em sẽ chuyển đổi các giá trị này về các khoảng để có thể dễ dàng thực hiện phân loại , phân cụm ở các phần dưới , cụ thể các giá trị thành các khoảng như sau :

Thuộc tính Age :

|  |  |
| --- | --- |
| **Giá Trị** | **Giá trị thay thế** |
| 40 - 60 | Middle-aged |
| 60 - 80 | Elderly |
| 80 - 100 | Senior Citizens |

Thuộc tính Serum Creatinine :

|  |  |
| --- | --- |
| **Giá trị** | **Giá trị thay thế** |
| < 1.0 | Low |
| 1.0 – 1.5 | Normal |
| 1.5 – 2.0 | High |
| > 2.0 | Very High |

Dự đoán với dữ liệu : X = ( Age = Senior citizens , Anaemia = 0 , serum\_creatinine : Normal , High Blood\_Pressure = 1 )

Kết quả khi chạy chương trình , dự đoán Dead Event của dòng trên : 0 . Trùng với phương pháp làm bằng tay mà em đã làm ở trên .

### Thực hiện phân cụm với K-Means :

A screen shot of a computer

Description automatically generated

## Đánh giá phương pháp phân loại :

A diagram of a test

Description automatically generated

## So sánh kết quả thực hiện của việc sử dụng công cụ với việc thực hiện thủ công :

 **Thời gian và công sức:** Sử dụng công cụ thường giảm thiểu thời gian và công sức so với thực hiện thủ công. Em có thể sử dụng các thư viện có sẵn , điều đó giúp giảm bớt thời gian và công sức cần thiết so với việc thực hiện từng bước một bằng tay.

 **Độ chính xác:** Công cụ thường cung cấp kết quả chính xác và nhất quán hơn so với việc thực hiện thủ công . Tuy nhiên nếu ta viết sai hàm hay sai 1 lỗi gì đó nhỏ thì có thể kéo theo cả chương trình lỗi hoặc không chính xác theo .

 **Tính tái sử dụng và mở rộng:** Công cụ thường có thể tái sử dụng và mở rộng dễ dàng cho các tác vụ tương tự, trong khi thực hiện thủ công thì cần phải thực hiện lại từ đầu mỗi khi có yêu cầu mới hoặc dữ liệu mới.

 **Khả năng hiểu biết và sáng tạo:** Tuy thực hiện bằng công cụ có những đặc điểm nổi bật như trên tuy nhiên thực hiện thủ công có thể giúp người thực hiện hiểu rõ hơn về quá trình và dữ liệu, đồng thời khuyến khích sự sáng tạo và phát triển kỹ năng. Tuy nhiên, công cụ cung cấp tính đồng nhất và có thể dễ dàng được sử dụng mà không cần kiến thức chuyên sâu về lĩnh vực cụ thể.

PHỤ LỤC

* File Excel Phân Cụm :



* File Excel Đánh Giá Phân Loại :



TÀI LIỆU THAM KHẢO

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

[2]. <https://machinelearningcoban.com/>

[3]. Các video hướng dẫn làm thủ công trên youtube

[4].<https://www.kaggle.com/datasets/andrewmvd/heart-failure-clinical-data/data>