**TRƯỜNG ĐẠI HỌC THỦY LỢI**

**KHOA CÔNG NGHỆ THÔNG TIN**

-----🙡🕮🙣-----

**A blue and white logo

Description automatically generated**

**ĐỀ TÀI:**

**KHAI PHÁ DỮ LIỆU BỆNH ĐỘT QUỴ BẰNG PHƯƠNG PHÁP HỒI QUY VÀ PHÂN LỚP**

**Nhóm sinh viên thực hiện: Nhóm 15**

                          1. Nguyễn Văn Xuân - 2151060305 - 63 CNTT2

                                    2. Trương Quốc Nam - 2151060299 - 63 CNTT2

**Giảng viên hướng dẫn:** TS. Trần Mạnh Tuấn

**Hà Nội, 10/2024**

MỞ ĐẦU

MỞ ĐẦU

# CHƯƠNG 1: TỔNG QUAN VỀ KHAI PHÁ DỮ LIỆU

* 1. Phát triển tri thức và khai phá dữ liệu.
  2. Quy trình khám phá tri thức trong CSDL.
  3. Mô tả bài toán.
     1. Tồng quan bài toán.
     2. Phân tích dữ liệu thô.

# CHƯƠNG 2: TIỀN XỬ LÝ DỮ LIỆU

2.1 Làm sạch dữ liệu.

2.2 Tích hợp dữ liệu.

2.3 Biến đổi dữ liệu.

# CHƯƠNG 3: KHAI PHÁ DỮ LIỆU BẰNG MÔ HÌNH HỒI QUY

3.1 Giới thiệu về Mô hình hồi quy.

3.2 Thuật toán hồi quy logistic nhị phân

# CHƯƠNG 4: KHAI PHÁ DỮ LIỆU BẰNG THUẬT TOÁN PHÂN LỚP

# CHƯƠNG 5: KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

# TÀI LIỆU THAM KHẢO

Mở Đầu

Báo cáo gồm 5 chương:

Chương 1: Tổng quan về khai phá dữ liệu.

Chương 2: Tiền xử lý dữ liệu.

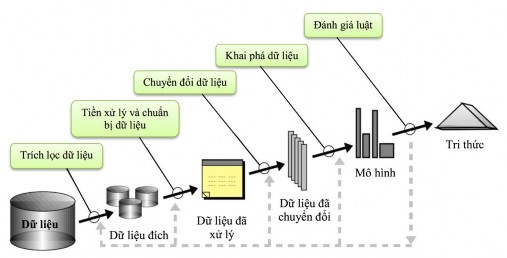
Chương 3: Khai phá dữ liệu bằng mô hình hồi quy.

Chương 4: Khai phá dữ liệu bằng thuật toán phân lớp

Chương 5: Kết luận và hướng phát triển

**CHƯƠNG 1: TỔNG QUAN VỀ KHAI PHÁ DỮ LIỆU**

* 1. **Phát triển tri thức và khai phá dữ liệu.**
* *Phát hiện tri thức (Knowledge Discovery):* là quá trình tìm kiếm và trích xuất thông tin hữu ích từ 1 lượng lớn dữ liệu. Quá trình này bao gồm các bước như: Chọc lọc dữ liệu, Tiền xử lý dữ liệu, Chuyển đổi dữ liệu, Khai phá dữ liệu, Đánh giá kết quả mẫu. Trở thành một phần quan trọng trong các lĩnh vực như kinh doanh, y tế và khoa học.
* *Khai phá dữ liệu (Data mining):* Là quá trình phân tích và khám phá các mẫu, mối quan hệ, và thông tin hữu ích từ một lượng lớn dữ liệu.
* Khai phá dữ liệu có thể được sử dụng cho các lĩnh vực y tế, phân tích thị trường, xây dựng ... có thể được xem như là kết quả của sự tiến triển tự nhiên của công nghệ thông tin.
  1. **Quy trình khám phá tri thức trong CSDL.**



*Hình 1.1: Quá trình khai phá dữ liệu từ cơ sở dữ liệu*

***Các giai đoạn khai phá tri thức***

* Trích chọn dữ liệu: Trích chọn những tập dữ liệu cần khai phá từ các tập dữ liệu khác nhau theo một tiêu chí nhất định
* Tiền xử lý:
* Làm sạch dữ liệu.
* Rút gọn dữ liệu.
* Rời rạc hóa dữ liệu.

Sau các bước này dữ liệu sẽ được nhất quán và đồng nhất

* Biến đổi dữ liệu: Là bước chuẩn hóa và làm mịn dữ liệu để đưa dữ liệu về dạng thuận lợi phục vụ cho các kỹ thuật khai phá ở bước sau.
* Khai phá dữ liệu: Áp dụng các ký thuật phân tích, nhằm:
* Khai thác dữ liệu
* Trích chọn mẫu thông tin
* Xây dựng tri thức
* Đánh giá
* Những mẫu thông tin và mã liên hệ trong bảng dữ liệu đã được khám phá ở bước trên được chuyển về biểu diễn ở một dạng gần với thế giới thực của người sử dụng như: Đồ thị, cây, bảng biểu, luật, …
* Đánh giá những tri thức khám phá được theo những tiêu chí nhất định
  1. **Mô tả bài toán.**
     1. **Lý do chọn đề tài.**
* Đột quỵ là một trong những nguyên nhân gây tử vong hàng đầu trên thế giới. Việc phân tích dữ liệu để phát hiện các yếu tố rủi ro và dự đoán khả năng mắc bệnh có thể giúp tăng cường phòng ngừa và điều trị sớm, giảm thiểu tổn thất về sức khỏe và tài chính. Kết quả của nghiên cứu có thể giúp xây dựng các hệ thống hỗ trợ ra quyết định y khoa, giúp bác sĩ và bệnh nhân nhận biết sớm nguy cơ đột quỵ, từ đó tăng hiệu quả điều trị và phòng ngừa. Vì vậy nhóm chúng em chọn đề tài “**Khai phá dữ liệu bệnh đột quỵ bằng phương pháp hồi quy và phân lớp**” để thực hiện dự đoán về bệnh đột quỵ.
  + 1. **Tổng quan bài toán.**
* Dataset gồm các mô tả về các thuộc tính tương ứng với chuẩn đoán nguy cơ bệnh đột quỵ. Áp dụng các thuật toán để xác định xem đối tượng có nguy cơ mắc bệnh đột quỵ hay không: 1(có), 0(không).

**1.3.3 Phân tích dữ liệu thô.**

\* Nguồn dữ liệu thô: [stoke prediction dataset (kaggle.com)](https://www.kaggle.com/datasets/kukuroo3/stoke-prediction-dataset)

+ *Hiểu dữ liệu:* Dữ liệu xét nghiệm chuẩn đoán bệnh huyết áp.

+ *Dữ liệu gồm:* Dữ liệu bao gồm 1137 bản ghi cùng 12 thuộc tính chuẩn đoán bệnh huyết áp.

A screenshot of a computer

Description automatically generated

*Hình 1.2: Dữ liệu ban đầu.*

***Hiểu các thuộc tính:***

|  |  |  |
| --- | --- | --- |
| **STT** | **Thuộc Tính** | **Ý nghĩa thuộc tính** |
| 1 | Id | Mã |
| 2 | Gender | Giới Tính |
| 3 | Age | Tuổi |
| 4 | Hypertension | Tình trạng huyết áp cao  (1 = có; 0 = không) |
| 5 | Heart\_disease | Bệnh tim  (1 = có; 0 = không) |
| 6 | Ever\_married | Tình trạng hôn nhân  (Yes/No) |
| 7 | Work\_type | Loại công việc |
| 8 | Residence\_type | Loại nơi cư trú (Thành thị hoặc Nông thôn) |
| 9 | Avg\_Glucose\_level | Mức đường huyết trung bình   * Dưới 7,8 mmol/l (140,4 mg/dl) là chỉ số đường huyết trung bình và an toàn * Từ 7,9 – 11,1 mmol/l (142,2 mg/dl – 199,8 mg/dl) thì nguy cơ cao bạn đã mắc bệnh tiểu đường * Nếu >11,1 mmol/l (199,8 mg/dl) thì nguy cơ cao bạn đã mắc bệnh tiểu đường |
| 10 | BMI (Body Mass Index) | Chỉ số khối cơ thể - Đánh giá tình trạng cơ thể dựa trên cân nặng và chiều cao   * BMI < 18,5: Gầy * 18,5 ≤ BMI < 24.9: Bình thường * 25 ≤ BMI < 29.9: Thừa cân * 30 ≤ BMI < 34.9: Béo phì độ 1 * 35 ≤ BMI < 39,9: Béo phì độ 2 * BMI ≥ 40: Béo phì độ 3 |
| 11 | Smoking\_status | Tình trạng hút thuốc   * Never smoked: chưa từng hút thuốc. * Unknown: không rõ. * Formerly smoked: trước đây đã hút thuốc. * Smokes: hút thuốc. |
| 12 | Stroke | Đột quỵ  (1: Có, 0: Không) |

**CHƯƠNG 2: TIỀN XỬ LÝ DỮ LIỆU**

* Là quá trình xử lý dữ liệu thô/gốc nhằm cải thiện chất lượng dữ liệu và chất lượng của kết quả KPDL
  1. **Làm sạch dữ liệu.**

**-** Là quá trình nhận dạng dữ liệu đã có để tiến hành xử lý các dữ liệu bị thiếu (missing data) xử lý dữ liệu bị nhiễu (noisy data) và không nhất quán.

* + 1. **Loại bỏ thuộc tính dư thừa**

- Đọc dữ liệu vào Weka:

A screenshot of a computer

Description automatically generated

*Hình 2.1: Chọn file dữ liệu vào Weka*

* Loại bỏ thuộc tính dư thừa, không cần thiết: **ID(Mã).**
* Tích chọn thuộc tính **ID** và nhấn chọn **Remove** để loại bỏ.

A screenshot of a computer

Description automatically generated

*Hình 2.2: Xóa bỏ thuộc tính ID*

* + 1. **Loại bỏ nhiễu**
* Xử lý dữ liệu nhiễu, không hợp lý có thể ảnh hưởng đến kết quả phân tích:
* Thuộc tính **Gender** có 1 giá trị **Other** xuất hiện duy nhất 1 lần là giá trị ngoại lai, không rõ ràng:
* Xử lý bằng cách thay thế nó bằng giá trị phổ biến trong thuộc tính **Gender**
  + + **Edit -> Replace values with**

A screenshot of a computer

Description automatically generated

*Hình 2.3: Thay thế giá trị phổ biến trong thuộc tính Gender*

* Tại thuộc tính age loại bỏ các bản ghi với độ tuổi nhỏ hơn 18 vì đột quỵ hiếm gặp ở trẻ em và trẻ sơ sinh:

**Filter -> Unsupervised -> Instance -> RemoveWithValues**

**A screenshot of a computer

Description automatically generated**

Hình*2.4: Loại bỏ bản ghi có tuổi nhỏ hơn 18*

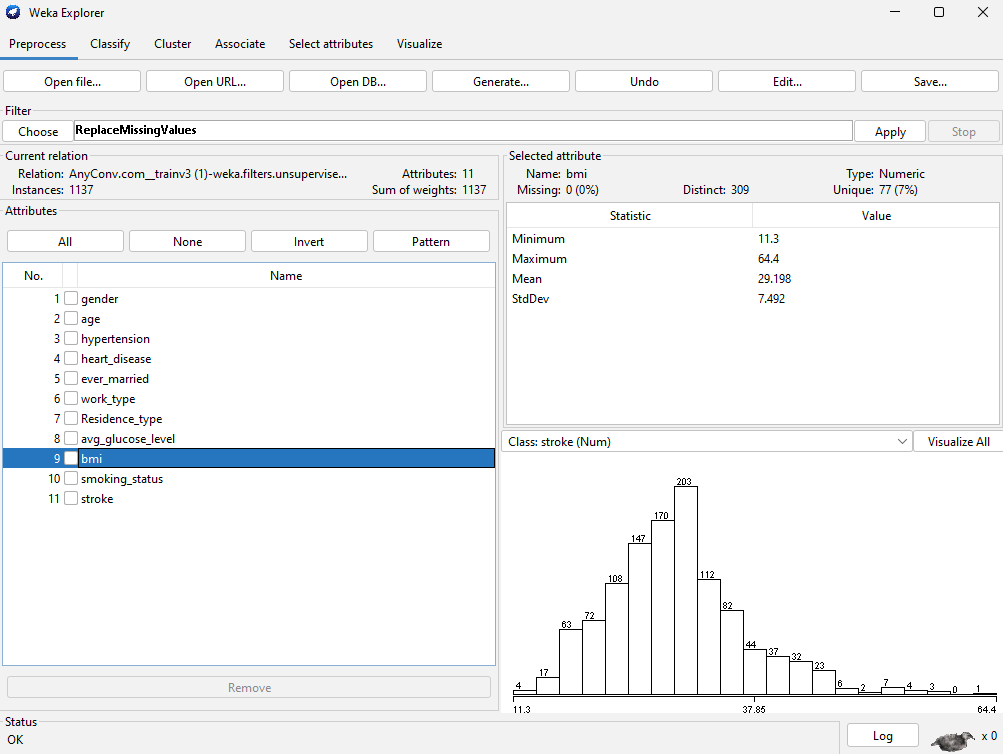
* + 1. **Xử lý dữ liệu bị thiếu**
* Trường BMI thiếu 52(5%) bản ghi:

*A screenshot of a computer

Description automatically generated*

*Hình 2.5: Trường bmi đang bị thiếu bản ghi*

* *Thay* thế các dữ liệu bị thiếu tại các thuộc tính: **Filter -> unsupervised -> attribute -> ReplaceMissingValue.**
* Với các dữ liệu bị thiếu dạng numeric sẽ được thay thế bằng giá trị trung bình tại thuộc tính đó.



*Hình 2.6: Trường bmi sau khi xử lý Missing value*

* 1. **Tịch hợp dữ liệu.**
* Là quá trình trộn dữ liệu từ các nguồn khác nhau vào một kho dữ liệu sẵn sàng cho quá trình khai phá dữ liệu.
* Liên quan đến cấu trúc và tính không thuần nhất (heterogeneity) về ngữ nghĩa (semantics) của dữ liệu
* Hỗ trợ việc giảm và tránh dư thừa và không nhất quán về dữ liệu > cải thiện tính chính xác và tốc độ quá trình khai phá dữ liệu

(1) Tích hợp lược đồ và so trùng đối tượng.

(2) Vấn đề dư thừa.

(3) Phát hiện và xử lý mẫu thuẫn giá trị dữ liệu.

=> **Dữ liệu lấy từ một nguồn nên không cần thực hiện quá trình này.**

* 1. **Biến đổi dữ liệu**
* Là quá trình biến đổi hay kết hợp dữ liệu vào những dạng thích hợp cho quá trình khai phá dữ liệu.

Các chiến lược thu giảm:

+ Làm trơn dữ liệu.

+ Kết hợp dữ liệu.

+ Tổng quát hóa dữ liệu.

+ Chuẩn hóa dữ liệu.

+ Xây dựng thuộc tính đặc tính.

* Chuẩn hóa với mô hình hồi quy logistic.

+ Nhìn vào tập dữ liệu ta thấy có 5 thuộc tính rời rạc (nominal):

Gender (Female: 0, Male: 1), ever\_married (no: 0, yes: 1), word\_type (govt\_job: 0, children: 1, private: 2, self\_employee: 3, never worked: 4), Residence\_type(Rural: 0, Urbal: 1), smoking\_status(never smoked: 0, unknown: 1, formerly smoker: 2, smoker: 3).

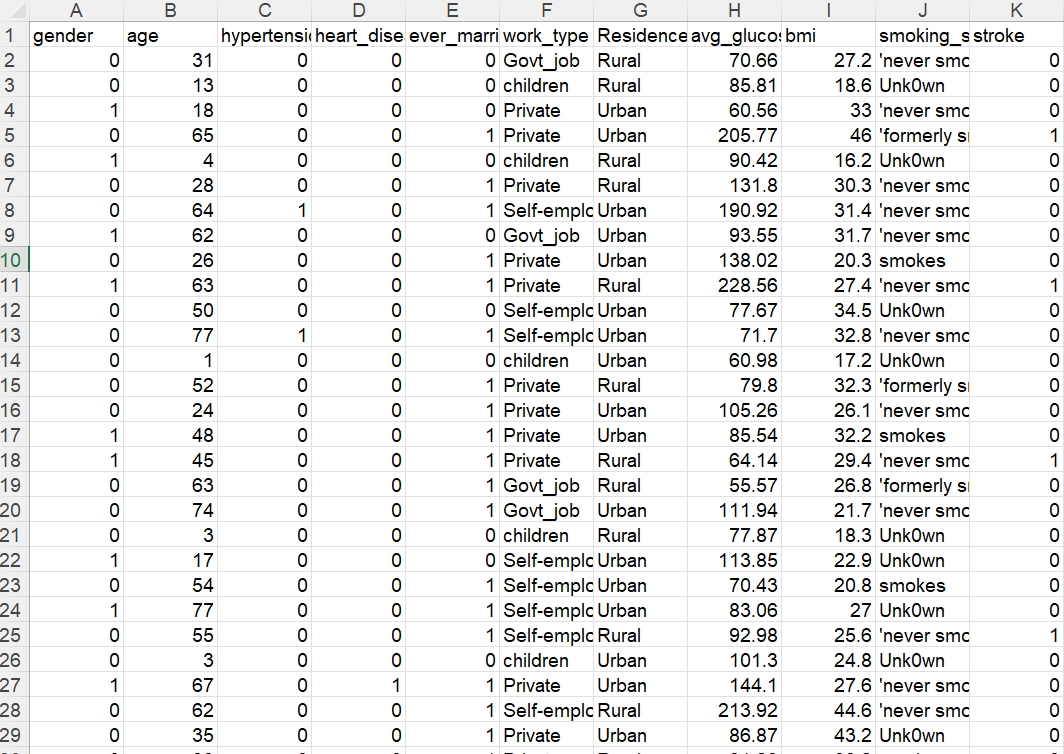
|  |  |  |
| --- | --- | --- |
| **STT** | **Thuộc Tính** | **Miền giá trị** |
| 1 | Gender | {0,1} |
| 2 | Age | [18:82] |
| 3 | Hypertension | {0,1} |
| 4 | Heart\_disease | {0,1} |
| 5 | Ever\_married | {0,1} |
| 6 | Work\_type | {0,2,3} |
| 7 | Residence\_type | {0,1} |
| 8 | Avg\_Glucose\_level | [55.27:266.69] |
| 9 | BMI (Body Mass Index) | [11.3:64.4] |
| 10 | Smoking\_status | {0,1,2,3} |
| 11 | Stroke | {0,1} |

* Xử lý trên Excel:

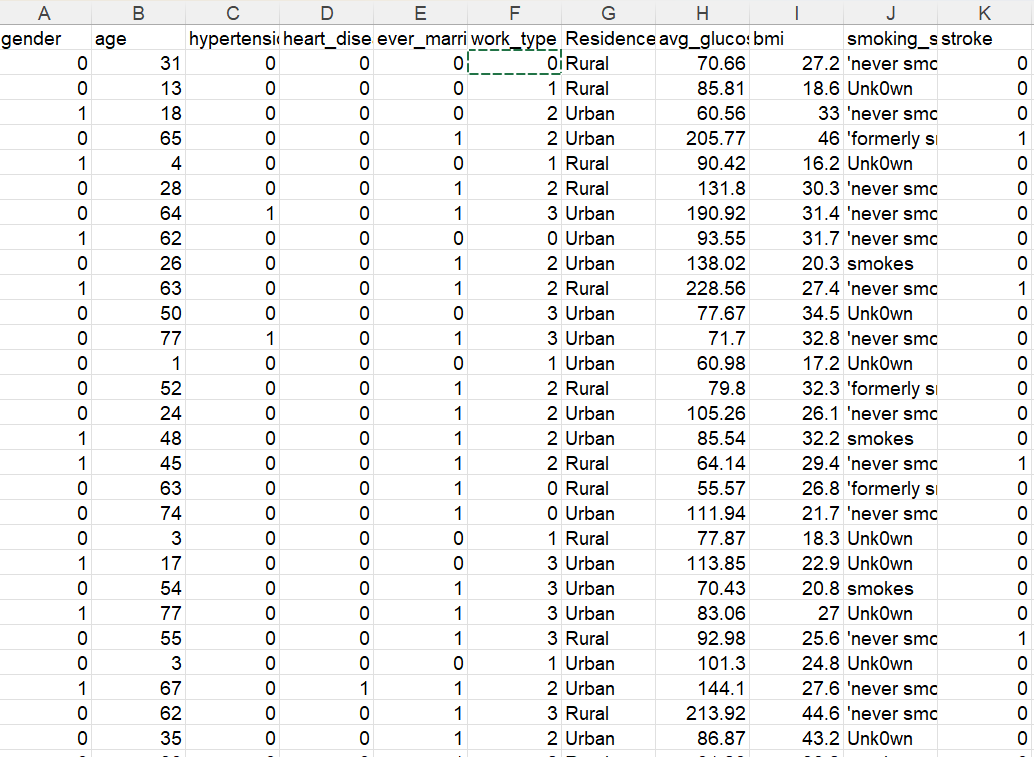
A screenshot of a computer

Description automatically generated

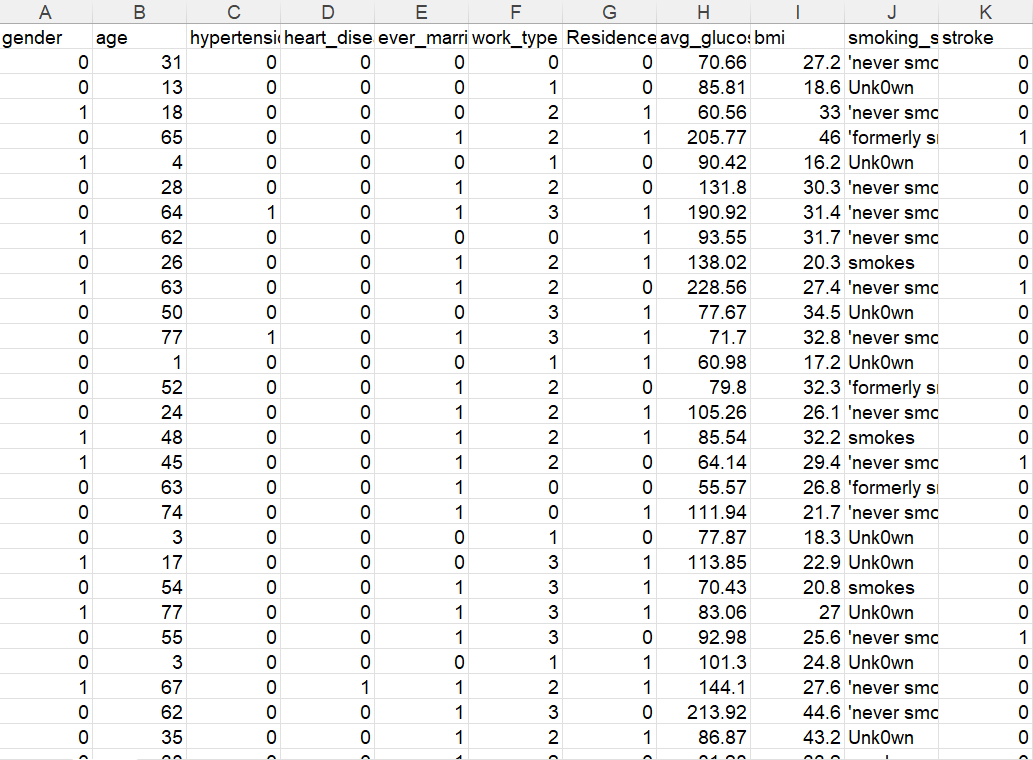
*Hình 2.7: Trường gender sau khi chuẩn hóa*

**

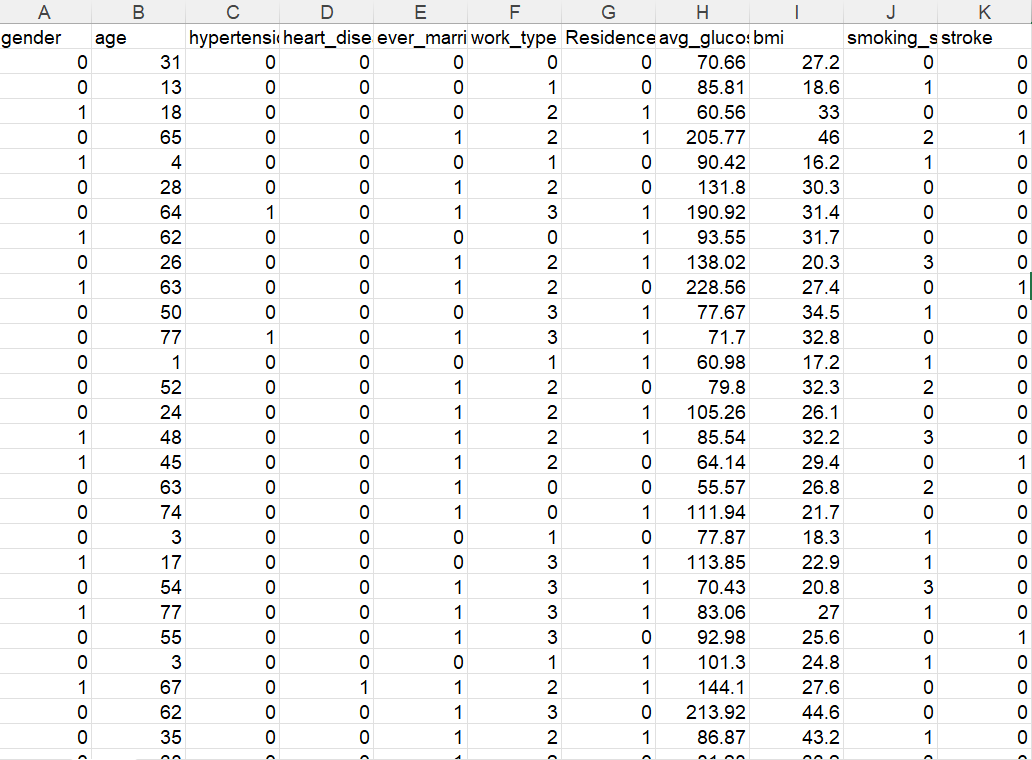
*Hình 2.8: trường ever\_married sau khi chuẩn hóa*

**

*Hình 2.9: trường work\_type sau khi chuẩn hóa*

**

*Hình 2.10: trường* Residence\_type sau khi chuẩn hóa

**

*Hình 2.11: trường smoking\_status sau khi chuẩn hóa*

**CHƯƠNG 3: KHAI PHÁ DỮ LIỆU BẰNG MÔ HÌNH HỒI QUY**

* 1. **Giới thiệu về kỹ thuật hồi quy.** [1]
* ***Khái niệm:***
* J. Han et al (2001, 2006): Hồi quy là kỹ thuật thống kê cho phép dự đoán các trị (số) liên tục.
* Wiki (2009): Là kỹ thuật thống kê cho phép ước lượng các mối liên kết giữa các biến.
* Theo R.D. snee (1977): Hồi quy là kỹ thuật thống kê trong lĩnh vực phân tích dữ liệu và xây dựng các mô hình từ thực nghiệm.
  1. **Thuật toán hồi quy logistic nhị phân.** [2,3]
* **Khái niệm:**
* Hồi quy logistic nhị phân được sử dụng để dự đoán mô hình mà biến phụ thuộc dạng nhị phân 0 hoặc 1.
* Hồi quy logistic hoạt động dựa trên hàm Sigmoid, được biểu diễn như sau:

+ s(z): Đầu ra trong khoảng từ 0 đến 1{giá trị xác suất ước lượng}

+ z: Đầu vào của hàm (giá trị dự đoán của thuật toán)

+ e: Hàm số Euler

* Ranh giới quyết định (Decision boundary)
* Hàm dự đoán lúc này sẽ trả về giá trị trong khoảng 0 đến 1. Ta cần chọn 1 giá trị ngưỡng để chia các giá trị thành 2 danh mục

(VD: p )

* Các bước thực hiện

**Bước 1: Xác định thuộc tính và biến mục tiêu**

* Chọn các thuộc tính (features) quan trọng từ tập dữ liệu sẽ sử dụng để phân loại đối tượng.
* Xác định biến mục tiêu (target variable) cần dự đoán (ví dụ: lớp 0 hoặc 1).

**Bước 2: Chuẩn bị dữ liệu**

* Chia tập dữ liệu thành hai phần: tập huấn luyện (training set) và tập kiểm tra (testing set) để đánh giá hiệu suất mô hình.

**Bước 3: Khởi tạo trọng số**

* Khởi tạo trọng số (weights) ban đầu bằng 0 hoặc các giá trị nhỏ ngẫu nhiên.

**Bước 4: Định nghĩa hàm sigmoid**

* Hàm sigmoid chuyển đổi tổng trọng số của các đặc trưng thành giá trị xác suất trong khoảng từ 0 đến 1: \[ \sigma(z) = \frac{1}{1 + e^{-z}} \]

**Bước 5: Định nghĩa hàm mất mát (Loss Function)**

* Sử dụng hàm binary cross-entropy để đánh giá chất lượng dự đoán của mô hình: \[ L(y, p) = - \left[ y \log(p) + (1 - y) \log(1 - p) \right] \]

**Bước 6: Thực hiện Gradient Descent**

* Cập nhật các trọng số để giảm thiểu hàm mất mát:
  + Tính gradient của hàm mất mát đối với các trọng số.
  + Cập nhật trọng số dựa trên gradient và tốc độ học (learning rate): \[ w = w - \alpha \nabla L \]

**Bước 7: Huấn luyện mô hình**

* Lặp lại việc cập nhật trọng số cho đến khi hàm mất mát hội tụ hoặc không thay đổi nhiều qua các vòng lặp.

**Bước 8: Dự đoán**

* Áp dụng hàm sigmoid lên tổng trọng số của các đặc trưng để tính xác suất.
* Chuyển đổi xác suất thành dự đoán nhị phân (0 hoặc 1) dựa trên ngưỡng (thường là 0.5).
* Kết quả thực hiện
* Test options với Use training set:

A screenshot of a computer

Description automatically generated

*Hình 3.1: Khi sử dụng Use training set*

* Test options với Cross-validation:

A screenshot of a computer

Description automatically generated

*Hình 3.2: Khi sử dụng Cross-validation*

* Test options với Percentage split(70/30):

A screenshot of a computer

Description automatically generated

*Hình 3.3: Khi sử dụng Percentage split*

* Biểu đồ biểu diễn sự khác biệt giữa các xác suất dự đoán của lớp đúng và các lớp sai

A screen shot of a graph

Description automatically generated

Hình 3.3: Sự khác biệt giữa các xác suất dự đoán của lớp đúng và các lớp sai

* Triển khai code thuật toán

A screenshot of a computer

Description automatically generated

* Kết quả thu được

A screen shot of a computer screen

Description automatically generated

* Tri thức thu được:
* **Độ chính xác**: Mô hình đạt khoảng **85%** (trên cả training set và cross-validation), cho thấy khả năng phân loại tương đối tốt.
* **Kappa statistic**: Thấp (khoảng **0.13**), cho thấy mô hình có sự phân biệt giữa các lớp nhưng chưa mạnh.
* **Lỗi tuyệt đối trung bình (MAE)**: ~**0.20**, cho thấy sai số dự đoán trung bình tương đối nhỏ.
* **Hiệu suất theo lớp**:
  + Với lớp "không đột quỵ": Recall cao (**0.97**), precision tốt (**0.87**).
  + Với lớp "bị đột quỵ": Recall rất thấp (**0.12**), mô hình khó phát hiện chính xác các trường hợp đột quỵ.

Mô hình chủ yếu phân loại tốt lớp "không đột quỵ", nhưng gặp khó khăn với lớp "đột quỵ".

**CHƯƠNG 4: KHAI PHÁ DỮ LIỆU BẰNG THUẬT TOÁN PHÂN LỚP**

# Giới thiệu về bài toán phân lớp.

# Khái niệm: Là để xây dựng 1 mô hình mà có thể dự đoán được trên lớp của những phần tử vào những đặc điểm của nó.

# Kỹ thuật phân lớp:

# B1: Xây dựng mô hình:

# + Mô tả lớp được định nghĩa trước đó

# + Tập train/test: Các mẫu dành cho xây dựng mô hình

# + Tìm luật phân lớp, cây quyết định mô tả lớp

# B2: Sử dụng mô hình:

# + Đánh giá tính chính xác của mô hình (Nếu độ chính xác chấp nhận -> áp dụng mô hình)

# + Phân lớp DL

# Thuật toán phân lớp J48.

# Khái niệm: Là để xây dựng 1 mô hình mà có thể dự đoán được trên lớp của những phần tử vào những đặc điểm của nó.

# Kỹ thuật phân lớp:

# B1: Xây dựng mô hình:

# + Mô tả lớp được định nghĩa trước đó

# + Tập train/test: Các mẫu dành cho xây dựng mô hình

# + Tìm luật phân lớp, cây quyết định mô tả lớp

# B2: Sử dụng mô hình:

# + Đánh giá tính chính xác của mô hình (Nếu độ chính xác chấp nhận -> áp dụng mô hình)

# + Phân lớp DL

# Cây quyết định (Decision Tree): Dùng để phân lớp các đối tượng dựa vào dãy các luật.

# Các thuộc tính của đối tượng có thể thuộc các kiểu dữ liệu khác nhau:

# Nhị phân (Binary)

# Định danh (Nominal)

# Thứ tự (Ordinal)

# Số lượng (Quantiative)

# (Trong khi đó thuộc tính phân lớp phải có kiểu dữ liệu là Binary hoặc Ordinal)

# Cây quyết định sẽ sinh ra luật để dự đoán lớp của các dữ liệu chưa biết

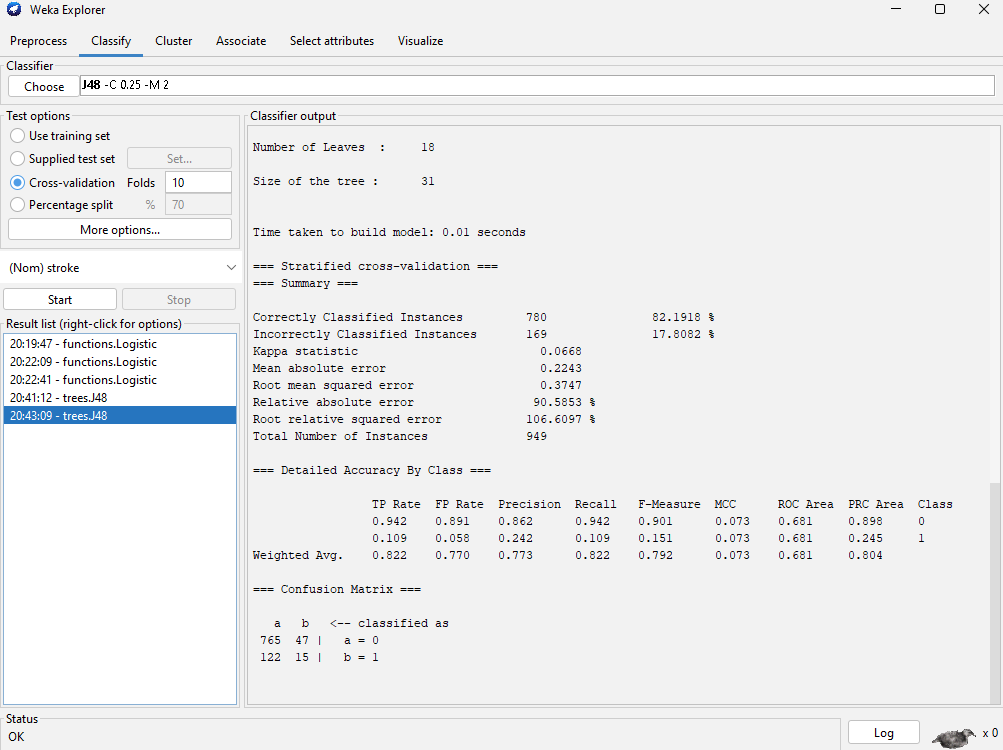
* Kết quả thực hiện
* Test options với Use training set:

A screenshot of a computer

Description automatically generated

*Hình 3.1: Khi sử dụng Use training set*

* Test options với Cross-validation:



*Hình 3.2: Khi sử dụng Cross-validation*

* Test options với Percentage split(70/30):

A screenshot of a computer

Description automatically generated

*Hình 3.3: Khi sử dụng Percentage split*

* Cây quyết định

A screenshot of a computer

Description automatically generated

*Hình 3.3: Cây quyết định*

* Triển khai code thuật toán

A screen shot of a computer code

Description automatically generated

* Kết quả thu được

A screenshot of a computer

Description automatically generated

**Sử dụng tập huấn luyện (Use training set):**

* **Hiệu suất:** Với tập huấn luyện, mô hình J48 thường cho tỷ lệ chính xác cao hơn do dữ liệu đã được thấy trong quá trình huấn luyện.
* **Tỷ lệ đúng:** Cao, thường trên 80% đến 90%.
* **Lưu ý:** Kết quả này có thể bị thiên lệch vì mô hình đã học từ chính tập dữ liệu này, dẫn đến hiện tượng quá khớp (overfitting).

**Kiểm tra chéo (Cross-validation):**

* **Hiệu suất:** Chia dữ liệu thành k phần, mỗi phần lần lượt được sử dụng làm tập kiểm tra trong khi các phần còn lại được sử dụng để huấn luyện.
* **Tỷ lệ đúng:** Thường dao động từ 80% đến 85%.
* **Lưu ý:** Đây là phương pháp đánh giá đáng tin cậy hơn vì nó kiểm tra mô hình trên nhiều tập con khác nhau của dữ liệu, giảm thiểu hiện tượng quá khớp và cho kết quả tổng quát hơn.

**Chia tỷ lệ (Percentage split 70/30):**

* **Hiệu suất:** Chia 70% dữ liệu làm tập huấn luyện và 30% làm tập kiểm tra.
* **Tỷ lệ đúng:** Thường dao động từ 82% đến 85%.
* **Lưu ý:** Phương pháp này cân bằng giữa việc kiểm tra mô hình trên dữ liệu mới và sử dụng dữ liệu để huấn luyện, mang lại kết quả tương đối đáng tin cậy.

**Kết luận tổng quát:**

* **Use training set:** Hiệu suất cao nhưng dễ bị quá khớp, không phản ánh khả năng tổng quát của mô hình.
* **Cross-validation:** Phương pháp đánh giá tốt nhất, giúp giảm thiểu quá khớp và cho kết quả tổng quát chính xác hơn.
* **Percentage split (70/30):** Kết quả đáng tin cậy nhưng không đa dạng bằng cross-validation.

**Đánh giá chi tiết:**

* **Sử dụng tập huấn luyện:** Kết quả có thể bị thiên lệch, dẫn đến hiện tượng quá khớp.
* **Kiểm tra chéo:** Kết quả tổng quát hơn, giảm thiểu hiện tượng quá khớp và phản ánh đúng khả năng của mô hình trên dữ liệu chưa thấy.
* **Chia tỷ lệ:** Mang lại kết quả đáng tin cậy, nhưng không phản ánh đa dạng bằng cross-validation.

**Kết luận:**

* Mô hình J48 hoạt động tốt với các phương pháp kiểm thử khác nhau. Sử dụng cross-validation là cách tốt nhất để đánh giá khả năng tổng quát của mô hình, giảm thiểu hiện tượng quá khớp và mang lại kết quả đáng tin cậy.

**Kết luận tổng quát:**

* **J48**: Mô hình J48 phù hợp với dữ liệu có mối quan hệ logic rõ ràng giữa các đặc trưng và biến mục tiêu. Tuy nhiên, nó có thể bị quá khớp với dữ liệu phức tạp và mất cân bằng. Sử dụng cross-validation là phương pháp đánh giá tốt nhất để giảm thiểu hiện tượng quá khớp và cho kết quả đáng tin cậy.
* **Hồi quy logistic**: Phù hợp với dữ liệu nhị phân, hiệu suất tốt khi các đặc trưng có mối quan hệ tuyến tính với biến mục tiêu. Sử dụng cross-validation giúp giảm thiểu quá khớp và đánh giá đúng khả năng tổng quát của mô hình. Mô hình hồi quy logistic có nhược điểm khi làm việc với dữ liệu phức tạp và đa lớp.

Nhìn chung, cả hai mô hình đều hoạt động tốt với các phương pháp kiểm thử khác nhau. Sử dụng crossvalidation là phương pháp đánh giá tốt nhất cho cả hai mô hình, giúp giảm thiểu hiện tượng quá khớp và mang lại kết quả tổng quát đáng tin cậy.

# Giải thích các thống kê:

# - Correctly Classified Instances: Số lượng các mẫu trong tập dữ liệu được phân loại chính

# xác bởi mô hình.

# - Incorrectly Classified Instances: Số lượng các mẫu trong tập dữ liệu bị phân loại sai bởi

# mô hình .

# - Kappa statistic (Thống kê kappa): Đây là một chỉ số đo lường sự đồng ý giữa nhãn lớp

# được dự đoán và thực tế, được hiệu chỉnh để tính đến sự trùng hợp ngẫu nhiên.

# - Mean absolute error (Sai số trung bình tuyệt đối): Sai số trung bình giữa xác suất dự

# đoán của mô hình và xác suất thực tế tương ứng với mỗi mẫu.

# - Root mean squared error (Căn bậc hai của sai số trung bình bình phương): là căn bậc

# hai của sai số trung bình bình phương giữa xác suất dự đoán của mô hình và xác suất thực

# tế tương ứng với mỗi mẫu.

# - Relative absolute error (Sai số tuyệt đối tương đối): Đây là sai số trung bình giữa xác

# suất dự đoán của mô hình và xác suất thực tế, được chuẩn hóa bởi xác suất thực tế trung

# bình.

# - Root relative squared error (Căn bậc hai của sai số bình phương tương đối): Đây là căn

# bậc hai của sai số trung bình bình phương giữa xác suất dự đoán của mô hình và xác suất

# thực tế, được chuẩn hóa bởi xác suất thực tế trung bình.

# - Total Number of Instances: Tổng số lượng mẫu trong tập dữ liệu

# Giải thích các độ đo:

# - TP Rate: Tỷ lệ số lượng các mẫu được phân loại chính xác vào nhãn Positive (dương tính) so với tổng số mẫu Positive trong tập dữ liệu

# - FP Rate: Tỷ lệ số lượng các mẫu bị phân loại sai vào nhãn Positive so với tổng số mẫu

# Negative trong tập dữ liệu.

# - Precision: Tỷ lệ số lượng các mẫu được phân loại chính xác vào nhãn Positive so với

# tổng số các mẫu được phân loại vào nhãn Positive.

# - Recall: Tỷ lệ số lượng các mẫu được phân loại chính xác vào nhãn Positive so với tổng số

# mẫu Positive trong tập dữ liệu.

# - F-Measure: Kết hợp giữa Precision và Recall để đánh giá hiệu quả phân loại. F-Measure

# càng lớn thì phân loại càng chính xác.

# - MCC: Độ đo tính tương đồng của hai chuỗi số. MCC = 1 tương đương với việc phân loại

# hoàn hảo và MCC = -1 tương đương với việc phân loại hoàn toàn ngược lại.

# - ROC Area: Đường cong ROC được sử dụng để đánh giá hiệu quả của thuật toán phân

# loại trong bài toán dự đoán nhị phân. ROC Area là diện tích dưới đường cong ROC.

# - PRC Area: Đường cong Precision-Recall được sử dụng để đánh giá hiệu quả của thuật

# toán phân loại trong bài toán dự đoán nhị phân. PRC Area là diện tích dưới đường cong

# Precision-Recall

# Tài liệu tham khảo.

# Tài liệu tiếng việt

* [1]: TS.Vo\_Thi\_Ngoc\_Trau\_Hoi\_Quy.pdf
* [2]: [Hồi quy tuyến tình logistic](https://ml-glossary-vn.readthedocs.io/vi/latest/logistic_regression_vn.html#hoi-quy-logistic-nhi-phan).
* [3]: [Hồi quy Logistic (Logistic Regression) — Tài liệu ML Glossary (ml-glossary-vn.readthedocs.io)](https://ml-glossary-vn.readthedocs.io/vi/latest/logistic_regression_vn.html#hoi-quy-logistic-nhi-phan).
* [4]: [Phân lớp dữ liệu - Description - Phân lớp dữ liệu I. Khái niệm Phân lớp dữ liệu là 1 kỹ thuật trong - Studocu](https://www.studocu.com/vn/document/dai-hoc-kinh-te-quoc-dan/principles-of-data-mining/phan-lop-du-lieu-description/49319171).
* Tài liệu tiếng anh

# []: [Building a Machine Learning Model Using J48 Classifier - GeeksforGeeks](https://www.geeksforgeeks.org/building-a-machine-learning-model-using-j48-classifier/)