BỘ GIÁO DỤC VÀ ĐÀO TẠO

TRƯỜNG ĐẠI HỌC NGOẠI NGỮ- TIN HỌC THÀNH PHỐ HỒ CHÍ MINH

KHOA CÔNG NGHỆ THÔNG TIN

ĐỒ ÁN MÔN HỌC

KHAI KHOÁNG DỮ LIỆU

ĐỀ TÀI

TÌM HIỂU CÁC PHƯƠNG PHÁP TIỀN XỬ LÝ DỮ LIỆU TRÊN CÔNG CỤ WEKA

GVHD: ThS. Nguyễn Thị Phương Trang

SVTH: Nguyễn Đức Dương 21DH110351

Đặng Võ Hoàng Văn 21DH114302

Phạm Trí Thanh 21DH114104

Tp. Hồ Chí Minh, tháng 7 năm 2024

Bảng 1. Bảng đánh giá mức độ hoàn thành

Họ và tên	**MSSV**	**Công việc được giao**	**Mức độ hoàn thành**
Nguyễn Đức Dương	21DH110351	- Tìm kiếm dataset. - Tiền xử lý dữ liệu. - Chỉnh sửa báo cáo. - Đánh giá kết quả.	100%
Đặng Võ Hoàng Văn	21DH114302	- Mô tả dữ liệu. - Trực quan dữ liệu. - Viết báo cáo.	100%
Phạm Trí Thanh	21DH114104	- Tìm kiếm dataset. - Tiền xử lý dữ liệu. - Chỉnh sửa báo cáo. - Đánh giá kết quả.	100%

MỤC LỤC

DANH MỤC HÌNH ẢNH

DANH MỤC BẢNG BIỂU

CHƯƠNG 1. GIỚI THIỆU

1.1. Giới thiệu đề tài

Trong kỷ nguyên dữ liệu lớn, việc thu thập, xử lý và phân tích dữ liệu đóng vai trò quan trọng trong việc đưa ra các quyết định chiến lược trong nhiều lĩnh vực khác nhau như kinh doanh, khoa học, và công nghệ. Tuy nhiên, dữ liệu thô thường chứa nhiều tạp chất, lỗi hoặc thiếu thông tin, làm giảm chất lượng của quá trình phân tích và dự đoán. Để khai thác dữ liệu một cách hiệu quả, bước tiền xử lý dữ liệu là vô cùng cần thiết nhằm làm sạch và chuẩn hóa dữ liệu, đảm bảo tính chính xác và nhất quán của dữ liệu đầu vào.

Đề tài "Tìm Hiểu Các Phương Pháp Tiền Xử Lý Dữ Liệu Trên Công Cụ WEKA" nhằm mục đích giới thiệu và phân tích các kỹ thuật tiền xử lý dữ liệu phổ biến được tích hợp trong WEKA, từ đó giúp người dùng hiểu rõ hơn về cách sử dụng các phương pháp này để cải thiện chất lượng dữ liệu và hiệu suất của các mô hình phân tích. Chúng em sẽ khám phá các công cụ tiền xử lý cơ bản và nâng cao như làm sạch dữ liệu, chuyển đổi thuộc tính, và lựa chọn thuộc tính. Thông qua các ví dụ minh họa và ứng dụng thực tế, đề tài hy vọng sẽ cung cấp một cái nhìn tổng quan và hướng dẫn chi tiết về cách thực hiện tiền xử lý dữ liệu hiệu quả trên WEKA.

1.2. Nội dung thực hiện

Nội dung thực hiện bao gồm các bước sau:

- Khái quát về tầm quan trọng của tiền xử lý dữ liệu
- Các phương pháp làm sạch dữ liệu (Data cleaning) như: Missing Value, Outliers, Noise, ...
- Phương pháp chuyển đổi dữ liệu (Data Transformation) bao gồm Chuẩn hóa và biến đổi dữ liệu (Normalization) và Biến đổi thuộc tính (Attribute Transformation)
- Xử lý mất cân bằng dữ liệu
- Tính độ tương quan
- Trực quan dự liệu

1.3. Giới hạn đề tài

Trong quá trình thực hiện đề tài, nhóm đã rút ra một số các giới hạn bao gồm:

- Phạm vi và phương pháp tiền xử lý dữ liệu: Giới hạn về số lượng phương pháp, chỉ một số phương pháp tiền xử lý phổ biến như làm sạch dữ liệu, chuyển đổi dữ liệu, và chọn lọc thuộc tính được phân tích chi tiết.
- Hạn chế về đánh giá và so sánh: Đề tài đánh giá hiệu quả của các phương pháp tiền xử lý chủ yếu thông qua các chỉ số cơ bản như độ chính xác và hiệu suất của mô hình.

1.4. Bố cục báo cáo

Bao gồm 4 chương và các mục nhỏ liên quan đến mỗi chương, các chương lớn bao gồm:

- Chương 1: Giới thiệu
- Chương 2: Cơ sở lý thuyết
- Chương 3: Tìm Hiểu Các Phương Pháp Tiền Xử Lý Dữ Liệu Trên Công Cụ Weka
- Chương 4: Kết quả và kết luận

CHƯƠNG 2. CƠ SỞ LÝ THUYẾT

2.1. Các khái niệm cơ bản

2.1.1 Tiền xử lý dữ liệu

Tiền xử lý dữ liệu đóng vai trò quan trọng trong việc tối ưu hóa chất lượng và hiệu suất của các quy trình phân tích và mô hình học máy. Trước hết, nó giúp cải thiện chất lượng dữ liệu bằng cách xử lý các giá trị thiếu, loại bỏ nhiễu và lỗi, từ đó làm tăng độ chính xác và đáng tin cậy của các kết quả phân tích. Điều này đặc biệt quan trọng vì dữ liệu chất lượng kém có thể dẫn đến các mô hình không chính xác, gây ra những dự đoán sai lệch. Bên cạnh đó, tiền xử lý còn tăng hiệu quả của mô hình học máy bằng cách giảm thiểu kích thước dữ liệu thông qua việc chọn lọc và rút gọn thuộc tính, tiết kiệm thời gian và tài nguyên xử lý. Chuẩn hóa và làm sạch dữ liệu cũng giúp tạo ra các tập dữ liệu nhất quán, dễ hiểu, và sẵn sàng cho các bước phân tích tiếp theo, từ đó hỗ trợ việc ra quyết định một cách hiệu quả hơn. Hơn nữa, việc tiền xử lý kỹ lưỡng còn tăng cường khả năng tổng quát hóa của mô hình, cho phép nó áp dụng trên các tập dữ liệu mới và đa dạng, đồng thời giảm nguy cơ overfitting, tức là hiện tượng mô hình khớp quá mức với dữ liệu huấn luyện.

2.1.2 WEKA

WEKA (Waikato Environment for Knowledge Analysis) là bộ phần mềm mã nguồn mở được phát triển bởi Đại học Waikato, New Zealand, nhằm cung cấp các công cụ mạnh mẽ cho khai phá dữ liệu và học máy. WEKA hỗ trợ các chức năng chính như:

- Explorer: Cho phép sử dụng các tính năng của WEKA để khai phá dữ liệu thông qua các tab Preprocess, Classify, Cluster, Associate, Select attribute, và Visualize.
- Experimenter: Cho phép tiến hành các thí nghiệm và thực hiện các bài kiểm tra thống kê giữa các mô hình học máy.
- KnowledgeFlow: Cho phép tương tác đồ họa để thiết kế các thành phần của một thí nghiệm.
- SimpleCLI: Giao diện dòng lệnh đơn giản để thực hiện các thao tác trên WEKA.

2.1.3 Làm Sạch Dữ Liệu (Data Cleaning)

Data cleaning hay còn gọi là làm sạch dữ liệu là quá trình xác định và sửa chữa các lỗi, thiếu sót, hoặc sự không nhất quán trong tập dữ liệu. Đây là một bước thiết yếu trong tiền xử

lý dữ liệu nhằm cải thiện chất lượng và độ tin cậy của dữ liệu trước khi sử dụng cho phân tích hoặc học máy. Các hoạt động chính bao gồm:

- Xử Lý Giá Trị Thiếu (Missing Values): Phát hiện và xử lý các giá trị thiếu bằng cách loại bỏ các mẫu chứa giá trị thiếu, thay thế giá trị thiếu bằng các giá trị trung bình, trung vị hoặc mode, và sử dụng các thuật toán nội suy (imputation) để dự đoán giá trị thiếu dựa trên các dữ liệu có sẵn khác.
- Loại Bỏ Dữ Liệu Ngoại Lai (Outliers): Phát hiện và loại bỏ các giá trị nằm ngoài phạm vi thông thường hoặc không tương thích với các mẫu dữ liệu khác.
- Sửa Chữa Dữ Liệu Sai Lệch (Data Errors): Phát hiện và sửa đổi các giá trị dữ liệu không chính xác, lỗi chính tả, lỗi nhập liệu hoặc dữ liệu không hợp lệ.
- Xử Lý Dữ Liệu Trùng Lặp (Duplicate Data): Phát hiện và loại bỏ các bản ghi dữ liệu bị lặp lại hoặc hợp nhất chúng thành một bản ghi duy nhất.
- Định Dạng Lại Dữ Liệu (Data Formatting): Chuẩn hóa định dạng và chuyển đổi dữ liệu để phù hợp với yêu cầu phân tích.

2.1.4 Chuyển Đổi Dữ Liệu (Data Transformation)

Chuyển đổi dữ liệu là quá trình thay đổi hoặc biến đổi dữ liệu từ một định dạng hoặc cấu trúc này sang định dạng hoặc cấu trúc khác để phù hợp với yêu cầu phân tích hoặc xử lý tiếp theo. Các hoạt động chính bao gồm:

- Chuẩn Hóa Dữ Liệu (Normalization): Điều chỉnh các giá trị dữ liệu vào một phạm vi chuẩn thường là từ 0 đến 1 hoặc -1 đến 1.
- Tiêu Chuẩn Hóa Dữ Liệu (Standardization): Điều chỉnh các giá trị dữ liệu sao cho chúng có trung bình bằng 0 và độ lệch chuẩn bằng 1.
- Chuyển Đổi Dữ Liệu Danh Mục (Categorical Data Transformation): Chuyển đổi các giá trị danh mục thành các số nguyên (Label Encoding) hoặc các cột nhị phân (One-Hot Encoding).
- Rời Rạc Hóa Dữ Liệu (Discretization): Chuyển đổi dữ liệu liên tục thành dữ liệu rời rạc bằng cách chia dữ liệu thành các khoảng hoặc lớp.
- Kết Hợp Dữ Liệu (Data Aggregation): Gộp các giá trị dữ liệu lại với nhau để tính toán các số liệu tổng hợp.

- Biến Đổi Thuộc Tính (Attribute Transformation): Tạo ra các thuộc tính mới hoặc thay đổi dạng thuộc tính hiện có để cung cấp thông tin bổ sung.
- Chuyển Đổi Thời Gian (Time-Series Transformation): Điều chỉnh và tạo các thuộc tính liên quan đến thời gian.

2.1.5 Chọn Lọc Thuộc Tính (Feature Selection)

Chọn lọc thuộc tính là quá trình xác định và lựa chọn những thuộc tính quan trọng nhất từ tập dữ liệu đầu vào nhằm tăng hiệu suất của các mô hình học máy và giảm độ phức tạp của dữ liệu. Các phương pháp chọn lọc thuộc tính bao gồm:

- Phương Pháp Dựa Trên Thống Kê (Statistical Methods): Bao gồm kiểm định Chi-Square, T-test, và hệ số tương quan.
- Phương Pháp Dựa Trên Mô Hình (Model-Based Methods): Bao gồm sử dụng cây quyết định, quan trọng của thuộc tính từ các mô hình như Random Forest, và L1 Regularization.
- Phương Pháp Dựa Trên Tìm Kiếm (Search-Based Methods): Bao gồm các phương pháp thêm tiến, lùi dần, và tiến tới lùi lại.
- Phương Pháp Dựa Trên Học Tập (Learning-Based Methods): Bao gồm sử dụng autoencoder, PCA, và hệ số từ các mô hình SVM.

2.1.6 Xử Lý Mất Cân Bằng Dữ Liệu (Handling Imbalanced Data)

Xử lý mất cân bằng dữ liệu là quá trình điều chỉnh tập dữ liệu để đảm bảo các lớp có tỷ lệ phân bố tương đồng hơn, từ đó cải thiện hiệu suất của các mô hình học máy. Các phương pháp phổ biến bao gồm:

- Undersampling**:** Giảm số lượng mẫu từ lớp chiếm đa số để cân bằng với lớp chiếm thiểu số.
- Oversampling**:** Tăng số lượng mẫu từ lớp chiếm thiểu số bằng cách sao chép hoặc tạo ra các mẫu mới.
- SMOTE (Synthetic Minority Over-sampling Technique): Tạo ra các mẫu tổng hợp mới từ lớp chiếm thiểu số bằng cách nội suy giữa các mẫu hiện có.
- Class Weight Adjustment: Điều chỉnh trọng số của các lớp trong quá trình huấn luyện để tăng trọng số cho lớp chiếm thiểu số.

2.1.7 Tính Độ Tương Quan (Correlation Analysis)

Tính độ tương quan là quá trình đo lường mức độ liên quan giữa các thuộc tính với nhau hoặc với biến mục tiêu. Các phương pháp tính độ tương quan bao gồm:

- Pearson Correlation Coefficient: Đo lường mối quan hệ tuyến tính giữa hai biến.
- Spearman Rank Correlation: Đo lường mối quan hệ không tuyến tính giữa hai biến dựa trên thứ hạng.
- Kendall Tau Correlation: Đo lường mức độ tương quan dựa trên sự sắp xếp thứ tự của các biến.

2.2. Các Công Trình Nghiên Cứu Liên Quan

- Association Rule to Increase Sales Using the Apriori Algorithm Method (2024)
- Data Pre-Processing And Its Implications In Data Mining (2024)
- Exploring Weka and Python for Educational Data Mining: Naïve Bayes vs. J48 (2024)
- Prediction of arrhythmia from MIT-BIH database using J48 and k-nearest neighbours (KNN) classifiers (2024)
- Data Mining Weka Decision Trees (2023)

CHƯƠNG 3. TÌM HIỂU CÁC PHƯƠNG PHÁP TIỀN XỬ LÝ DỮ LIỆU TRÊN CÔNG CỤ WEKA

3.1. Tổng quan phương pháp hiện thực

Đề tài tập trung vào việc nghiên cứu và áp dụng các phương pháp tiền xử lý dữ liệu sử dụng công cụ WEKA, một nền tảng phổ biến trong lĩnh vực khoa học dữ liệu và học máy. Việc tiền xử lý dữ liệu đóng vai trò quan trọng trong quá trình chuẩn bị dữ liệu trước khi áp dụng các mô hình học máy và phân tích số liệu. Các phương pháp tiền xử lý như làm sạch dữ liệu, chuẩn hóa, xử lý dữ liệu thiếu và rút gọn dữ liệu không chỉ giúp cải thiện chất lượng và độ chính xác của dữ liệu mà còn là bước quan trọng để giảm thiểu ảnh hưởng của dữ liệu nhiễu và nâng cao hiệu quả của các mô hình dự đoán.

3.2. Công cụ hiện thực

Sử dụng WEKA trong đề tài này không chỉ giúp cho việc thực hiện các phương pháp tiền xử lý dữ liệu trở nên dễ dàng mà còn cung cấp cho người dùng những công cụ mạnh mẽ để khám phá và phân tích dữ liệu một cách chính xác và hiệu quả.

3.3. Tập dữ liệu

Tập dữ liệu **heart_disease_uci.csv** là một kho thông tin phong phú và toàn diện về bệnh tim, cung cấp cái nhìn chi tiết về nhiều khía cạnh y học liên quan đến bệnh tim mạch. Đây là một loại tập dữ liệu đa biến, nghĩa là nó liên quan đến nhiều biến số toán học và thống kê khác nhau, phù hợp cho các phân tích dữ liệu số đa biến. Mỗi dòng trong tệp đại diện cho một bệnh nhân cụ thể.

Được thu thập trên nền tảng Kaggle, cung cấp các bộ dữ liệu chất lượng cao, được cộng đồng sử dụng và kiểm chứng tại địa chỉ ***heart_disease_uci.csv*** gồm 920 quan sát với 16 biến đặc trưng là:

Bảng 2. Bảng mô tả dữ liệu

STT	Thuộc tính	Kiểu dữ liệu	Ý nghĩa	Ví dụ thể hiện
1	id	Nominal	Mã định danh của bệnh nhân	1
2	age	Continuous	Tuổi của bệnh nhân	63

3	sex	Binary	Giới tính của bệnh nhân	Male, Female
4	cp	Nominal	Loại đau ngực	typical angina
5	trestbps	Continuous	Huyết áp nghỉ ngơi (mm Hg)	145
6	chol	Continuous	Nồng độ cholesterol (mg/dl)	233
7	fbs	Binary	Đường huyết khi đói > 120 mg/dl	True, False
8	restecg	Nominal	Kết quả điện tâm đồ nghỉ ngơi	lv hypertrophy
9	thalch	Continuous	Nhịp tim tối đa đạt được	150
10	exang	Binary	Đau thắt ngực do gắng sức	True, False
11	oldpeak	Continuous	Độ chênh lệch ST khi gắng sức so với khi nghỉ ngơi	2.3
12	slope	Nominal	Độ dốc của đoạn ST (upsloping, flat,downsloping)	downsloping
13	ca	Discrete	Số lượng mạch máu chính (0-3)	0, 1 , 2, 3
14	thal	Nominal	Thalassemia (normal, fixed defect, reversable defect)	fixed defect
15	dataset	Nominal	Nguồn gốc của bộ dữ liệu	Cleveland
16	num	Ordinal	Biến mục tiêu. Chẩn đoán bệnh tim (0-4, đại diện cho mức độ nghiêm trọng của bệnh)	0

- Nominal: Dữ liệu dạng danh mục, không có thứ tự. Ví dụ: id, cp, restecg, slope, thal, dataset.
- Continuous: Dữ liệu liên tục, có thể lấy mọi giá trị trong một khoảng. Ví dụ: age, trestbps, chol, thalch, oldpeak.
- Binary: Dữ liệu nhị phân, chỉ có hai giá trị. Ví dụ: sex, fbs, exang.
- Discrete: Dữ liệu rời rạc, có giá trị nguyên cụ thể. Ví dụ: ca.
- Ordinal: Dữ liệu thứ bậc, có thứ tự nhưng không có khoảng cách xác định giữa các giá trị. Ví dụ: num.

Mô tả chi tiết về cách tính hoặc đo lường các thuộc tính trong tập dữ liệu

a) id:

- Mã định danh duy nhất được gán cho mỗi bệnh nhân, có tổng số 920 bệnh nhân.

b) age (tuổi):

Selected attribute

Name: age		Type: Numeric
Missing: 0 (0%)	Distinct: 50	Unique: 3 (0%)

Statistic	Value
Minimum	28
Maximum	77
Mean	53.511
StdDev	9.425

Hình 1. Thuộc tính age

- Tuổi của bệnh nhân, độ tuổi dao động khoảng từ 28 – 77.

c) sex (giới tính):

Selected attribute

Name: sex | Type: Nominal
Missing: 0 (0%) | Distinct: 2 | Unique: 0 (0%)

No.	Label	Count	Weight
1	Male	726	726
2	Female	194	194

Hình 2. Thuộc tính sex

- Giới tính của bệnh nhân, được ghi nhận dưới dạng danh mục (Nam hoặc Nữ).

d) cp (loại đau ngực):

Selected attribute

Name: cp | Type: Nominal
Missing: 0 (0%) | Distinct: 4 | Unique: 0 (0%)

No.	Label	Count	Weight
1	typical angina	46	46
2	asymptomatic	496	496
3	non-anginal	204	204
4	atypical angina	174	174

Hình 3. Thuộc tính cp

- Loại đau ngực mà bệnh nhân trải qua, được phân loại thành 4 loại:
  - typical angina (đau thắt ngực điển hình)
  - atypical angina (đau thắt ngực không điển hình)
  - non-anginal pain (đau không phải do thắt ngực)
  - asymptomatic (không triệu chứng).

e) dataset (nguồn gốc dữ liệu):

Selected attribute

Name: dataset | Type: Nominal
Missing: 0 (0%) | Distinct: 4 | Unique: 0 (0%)

No.	Label	Count	Weight
1	Cleveland	304	304
2	Hungary	293	293
3	Switzerland	123	123
4	VA Long Beach	200	200

Hình 4. Thuộc tính dataset

- Nguồn gốc của bộ dữ liệu, chỉ ra bộ dữ liệu đến từ nguồn nào như: Cleveland, Hungary, Switzerland, hoặc VA.

f) trestbps (huyết áp lúc nghỉ ngơi):

Selected attribute

Name: trestbps | Type: Numeric
Missing: 59 (6%) | Distinct: 61 | Unique: 15 (2%)

Statistic	Value
Minimum	0
Maximum	200
Mean	132.132
StdDev	19.066

Hình 5. Thuộc tính trestbps

- Huyết áp của bệnh nhân khi nghỉ ngơi, được đo bằng mm Hg
- 59 giá trị bị thiếu (6%)
- Huyết áp tối đa là 200, giá trị trung bình là 132.132, độ lệch chuẩn là 19.066
- Huyết áp tối thiểu là 0, không hợp lý vì huyết áp không thể là 0, cần xem xét và xử lý giá trị này.

g) chol (cholesterol):

Selected attribute

Name: chol | Type: Numeric
Missing: 30 (3%) | Distinct: 217 | Unique: 67 (7%)

Statistic	Value
Minimum	0
Maximum	603
Mean	199.13
StdDev	110.781

Hình 6. Thuộc tính chol

- Nồng độ cholesterol trong máu của bệnh nhân, được đo bằng mg/dl. Dao động từ 0 đến 603mg/dl,
- 30 giá trị bị thiếu (3%)
- Giá trị lớn nhất là 603, trung bình: 199.13 mg/dl, độ lệch chuẩn: 110.78
- Giá trị tối thiểu là 0, nồng độ cholesterol không thể là 0, cần xử lý những giá trị này.

h) fbs (đường huyết lúc đói):

Selected attribute

Name: fbs | Type: Nominal
Missing: 90 (10%) | Distinct: 2 | Unique: 0 (0%)

No.	Label	Count	Weight
1	TRUE	138	138
2	FALSE	692	692

Hình 7. Thuộc tính fbs

- Đường huyết đo được sau khi bệnh nhân nhịn ăn ít nhất 8 giờ, có giá trị nhị phân: True nếu > 120 mg/dl, ngược lại là False.
- 90 giá trị bị thiếu (10%).

i) restecg (kết quả điện tâm đồ lúc nghỉ ngơi):

Selected attribute

Name: restecg		Type: Nominal
Missing: 2 (0%)	Distinct: 3	Unique: 0 (0%)

No.	Label	Count	Weight
1	lv hypertrophy	188	188
2	normal	551	551
3	st-t abnormality	179	179

Hình 8. Thuộc tính restecg

- Kết quả điện tâm đồ khi bệnh nhân ở trạng thái nghỉ ngơi, phân loại thành:
  - normal (bình thường)
  - ST-T abnormality (có bất thường sóng ST-T)
  - Lv hypertrophy (biểu hiện phì đại tâm thất trái rõ ràng hoặc có khả năng).

j) thalch (nhịp tim tối đa đạt được):

Selected attribute

Name: thalch		Type: Numeric
Missing: 55 (6%)	Distinct: 119	Unique: 20 (2%)

Statistic	Value
Minimum	60
Maximum	202
Mean	137.546
StdDev	25.926

Hình 9. Thuộc tính thalch

- Nhịp tim tối đa mà bệnh nhân đạt được trong quá trình kiểm tra gắng sức, được đo bằng nhịp mỗi phút (bpm).
- 55 giá trị bị thiếu (6%).
- Giá trị lớn nhất 202, nhỏ nhất 60, trung bình 137.546, độ lệch chuẩn 25.925.

k) exang (đau thắt ngực do gắng sức):

Selected attribute

Name: exang | Type: Nominal
Missing: 55 (6%) | Distinct: 2 | Unique: 0 (0%)

No.	Label	Count	Weight
1	FALSE	528	528
2	TRUE	337	337

Hình 10. Thuộc tính exang

- Biến nhị phân cho biết bệnh nhân có bị đau thắt ngực do gắng sức hay không, với giá trị True và False.
- 55 giá trị bị thiếu (6%).

l) oldpeak (ST chênh lệch):

Selected attribute

Name: oldpeak | Type: Numeric
Missing: 62 (7%) | Distinct: 53 | Unique: 16 (2%)

Statistic	Value
Minimum	-2.6
Maximum	6.2
Mean	0.879
StdDev	1.091

Hình 11. Thuộc tính oldpeak

- Độ chênh lệch ST (sự thay đổi về mức độ của đoạn ST trên điện tâm đồ ECG) được đo khi bệnh nhân gắng sức so với khi nghỉ ngơi, đơn vị đo là milivolt (mV).
- 62 giá trị bị thiếu (7%).
- Giá trị lơn nhất là 6.2, trung bình là 0.879, độ lệch chuẩn 1.091.
- Giá trị tối thiểu là -2.6, ở một vài trường hợp cá biệt, giá trị này có thể âm.

m) slope (độ dốc của đoạn ST):

Selected attribute

Name: slope Type: Nominal
Missing: 309 (34%) Distinct: 3 Unique: 0 (0%)

No.	Label	Count	Weight
1	downsloping	63	63
2	flat	345	345
3	upsloping	203	203

Hình 12. Thuộc tính slope

- Độ dốc của đoạn ST khi gắng sức, được phân loại thành:
  - upsloping (dốc lên)
  - flat (phẳng)
  - downsloping (dốc xuống).
- Có 309 giá trị bị thiếu (34%).

n) ca (số lượng mạch máu chính):

Selected attribute

Name: ca Type: Numeric
Missing: 611 (66%) Distinct: 4 Unique: 0 (0%)

Statistic	Value
Minimum	0
Maximum	3
Mean	0.676
StdDev	0.936

Hình 13. Thuộc tính ca

- Số lượng mạch máu chính (0-3) được nhìn thấy qua phương pháp chụp huỳnh quang.
- Số lượng giá trị bị thiếu lớn 611 giá trị (66%).
- Đây là một thuộc tính rời rạc nhưng đang bị nhầm lẫn thành thuộc tính liên tục.

o) thal (thalassemia):

Selected attribute

Name: thal Type: Nominal
Missing: 486 (53%) Distinct: 3 Unique: 0 (0%)

No.	Label	Count	Weight
1	fixed defect	46	46
2	normal	196	196
3	reversable defect	192	192

Hình 14. Thuộc tính thal

- Tình trạng thalassemia là một rối loạn di truyền ảnh hưởng đến việc sản xuất hemoglobin - protein trong hồng cầu có chức năng vận chuyển oxy của bệnh nhân, được phân loại thành:
  - normal (bình thường)
  - fixed defect (khiếm khuyết cố định)
  - reversable defect (khiếm khuyết có thể hồi phục).
- Số lượng giá trị bị thiếu lớn 486 giá trị (53%).

p) num (chẩn đoán bệnh tim):

Selected attribute

Name: num Type: Nominal
Missing: 0 (0%) Distinct: 5 Unique: 0 (0%)

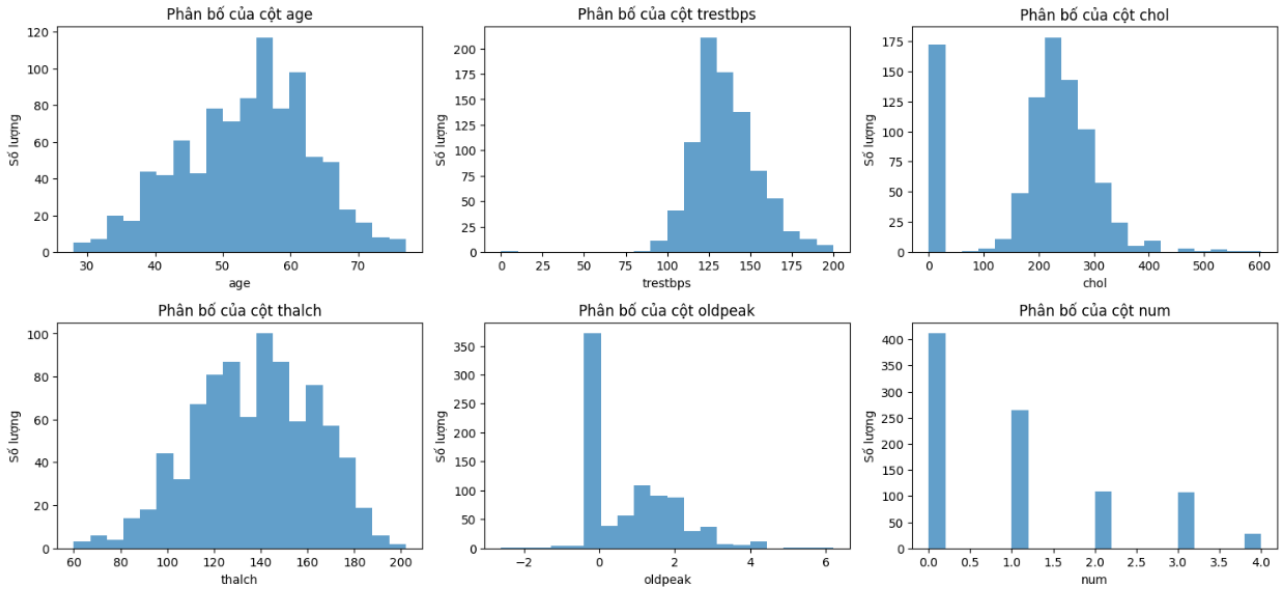
No.	Label	Count	Weight
1	0	411	411
2	1	265	265
3	2	109	109
4	3	107	107
5	4	28	28

Hình 15. Thuộc tính num

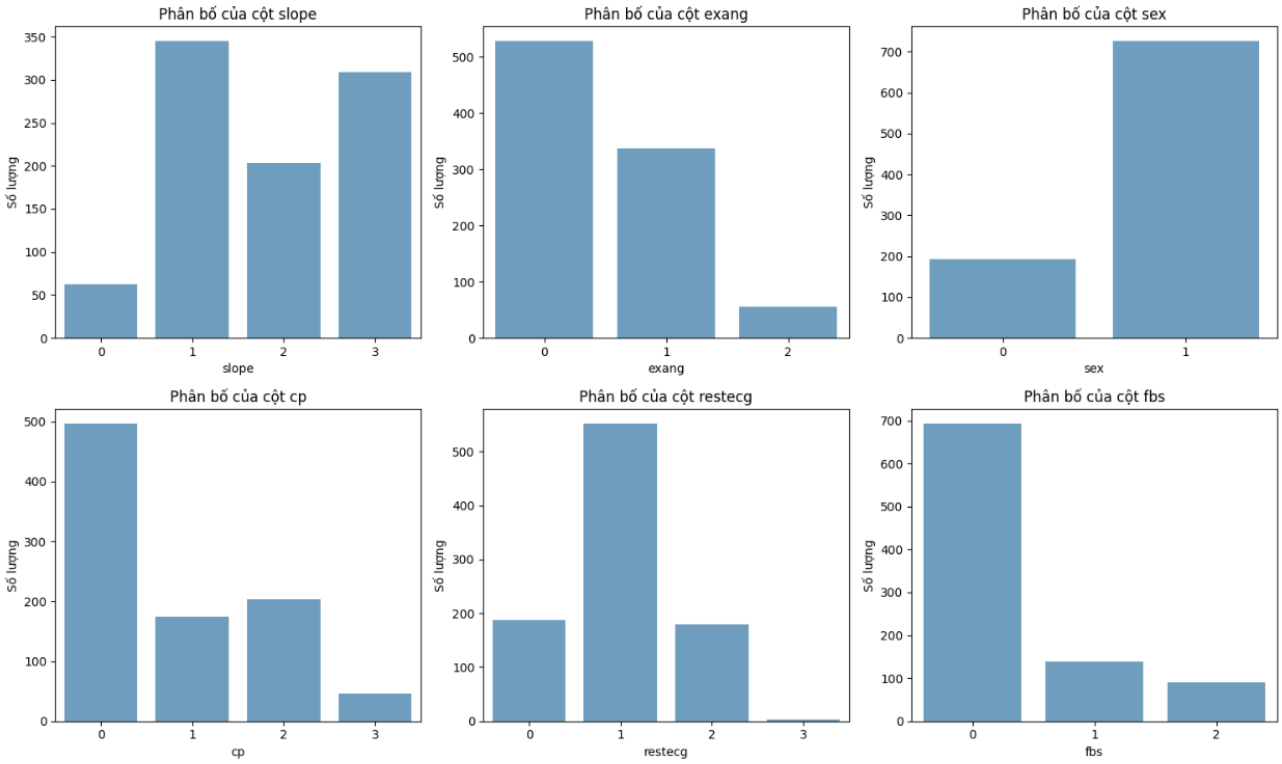
- Mức độ nghiêm trọng của bệnh tim, được đánh giá trên thang điểm từ 0 đến 4:
  - 0: không có bệnh tim

- 1 - 4: các mức độ nghiêm trọng tăng dần của bệnh.

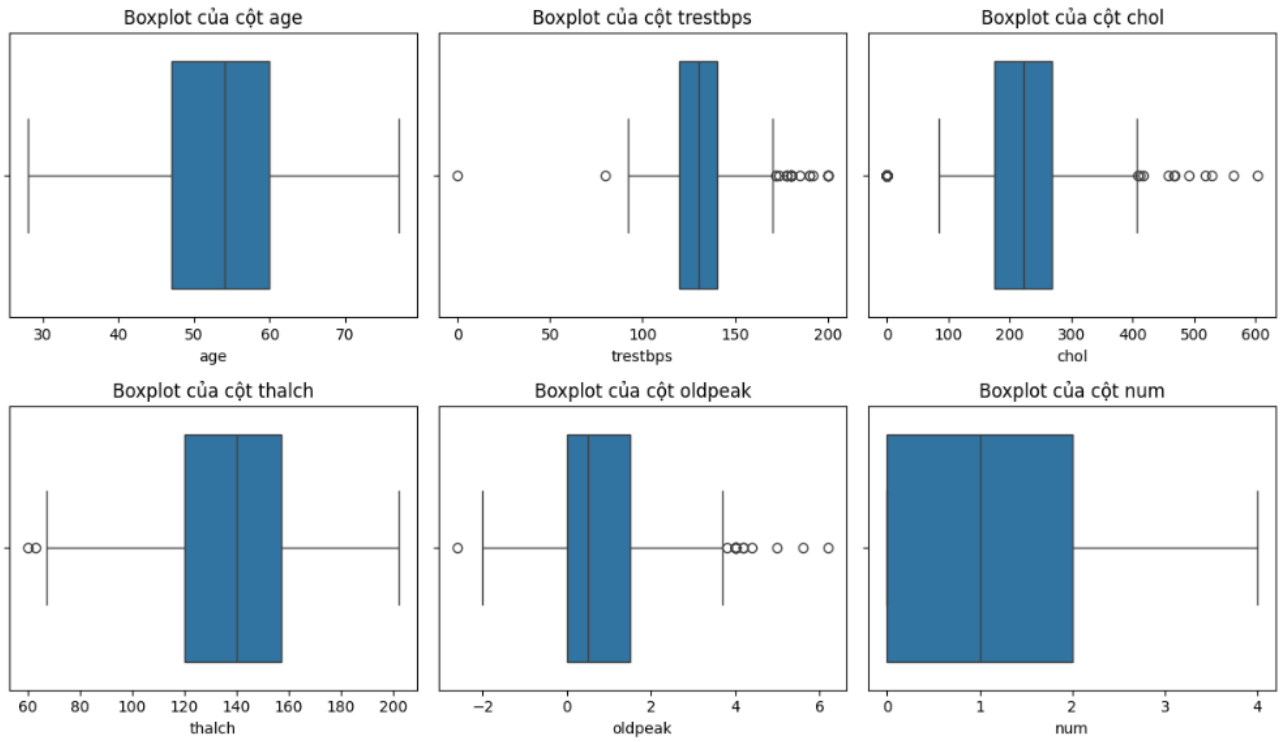
3.4. Trực quan hóa dữ liệu



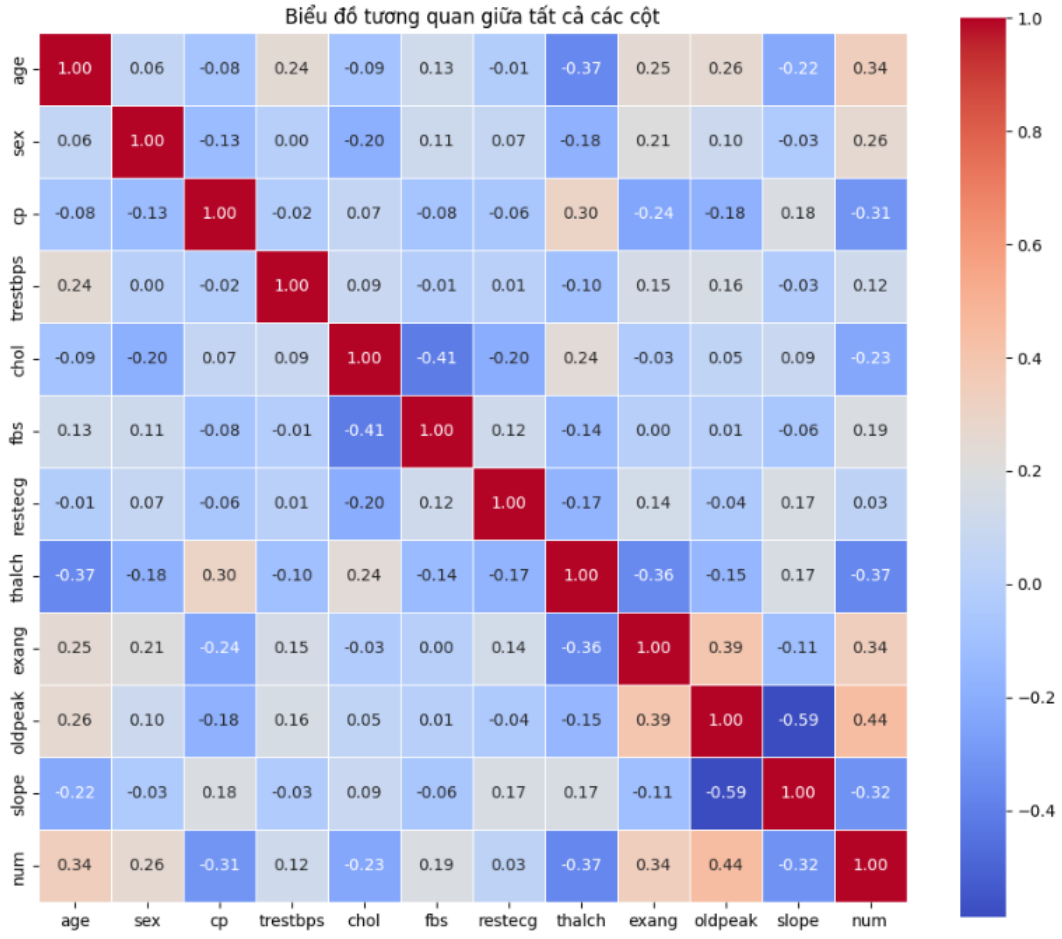
Hình 16. Phân bố giá trị các cột trong tập dữ liệu



Hình 17. Phân bố giá trị các cột trong tập dữ liệu



Hình 18. Biểu đồ boxplot



Hình 19. Biểu đồ ma trận tương quan giữa các cột

3.5. Tiền xử lý dữ liệu

3.5.1 Chọn lựa thuộc tính

Loại bỏ một số thuộc tính không cần thiết là “id” và “dataset”

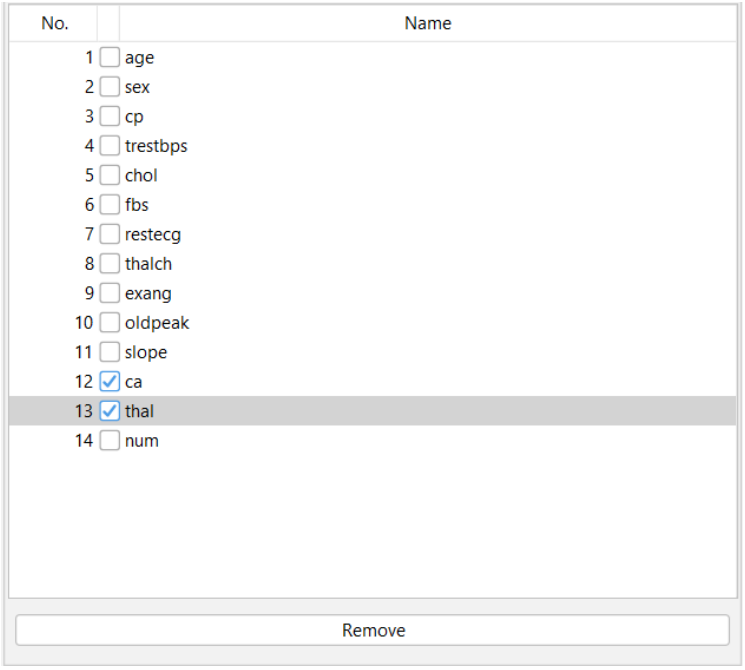
- Chọn thuộc tính cần xóa → Remove

Hình 20. Loại bỏ thuộc tính không cần thiết

3.5.2 Làm sạch dữ liệu

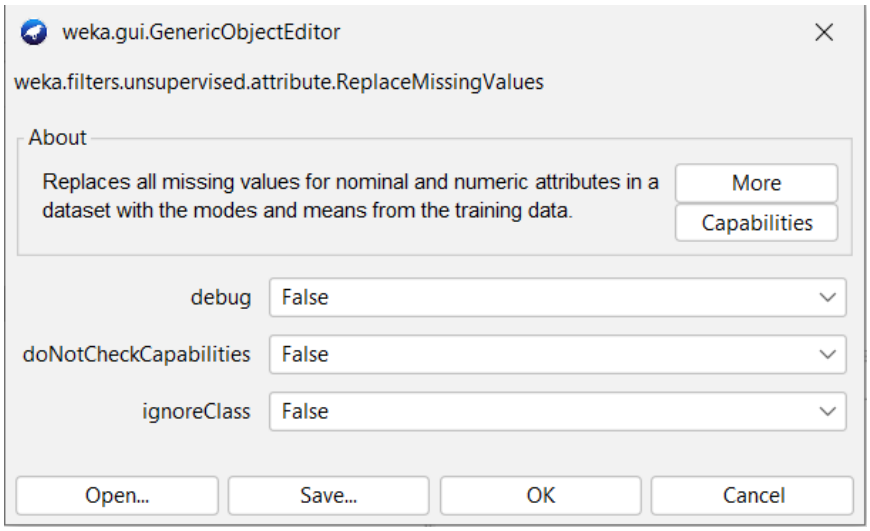
a) Xử lý dữ liệu thiếu

- Xóa 2 thuộc tính có tỉ lệ giá trị thiếu lớn hơn 50% là “ca” và “thal”:
  - Chọn thuộc tính cần xóa → Remove



Hình 21. Loại bỏ thuộc tính có nhiều giá trị thiếu

- Điền các giá trị thiếu còn lại bằng mean đối với các biến numeric và mode đối với biến nominal.
    - Sử dụng filter ReplaceMissingValue

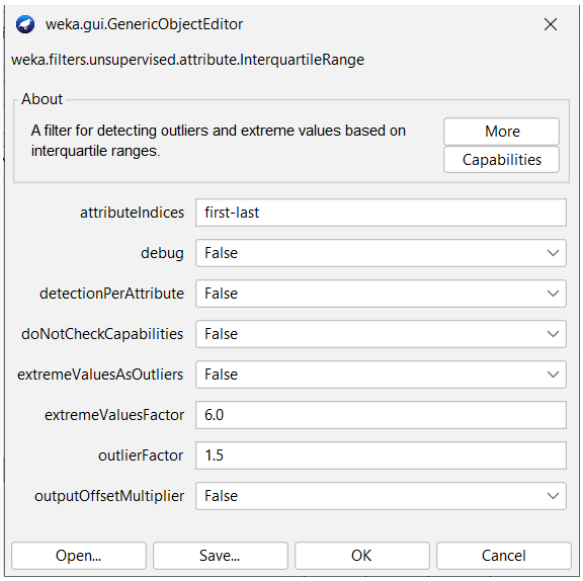


Hình 22. Điền giá trị thiếu bằng ReplaceMissingValue

b) Phát hiện giá trị ngoại lai (outlier).

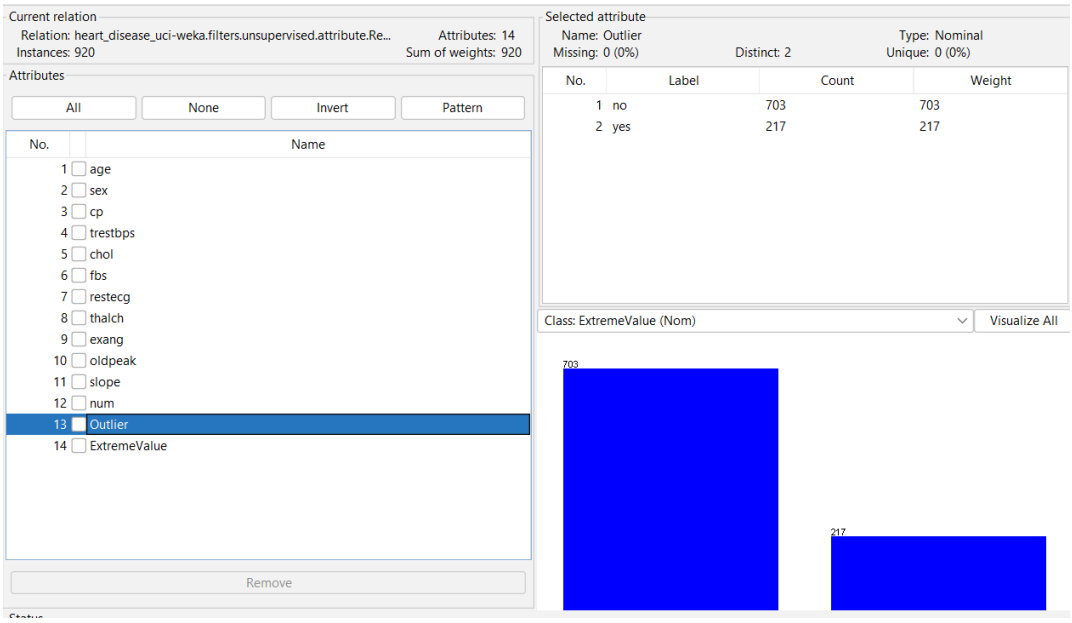
- Sử dụng filter InterquartileRange để phát hiện giá trị ngoại lai
    - **attributeIndices**: ‘first – last’: áp dụng bộ lọc cho toàn bộ thuộc tính

- **detectionPerAttribute**: Nếu True, sẽ phát hiện ngoại lai riêng cho từng thuộc tính. Nếu False, xem xét tất cả thuộc tính cùng lúc.
- **extremeValuesAsOutliers**: các giá trị cực đại là ngoại lệ.
- **outlierFactor**: để xác định hệ số cho việc phát hiện các giá trị ngoại lệ. Thông thường trong công thức chuẩn Q1−1.5×IQR thì hệ số này sẽ là 1.5



Hình 23. Sử dụng InterquartileRange để phát hiện ngoại lai

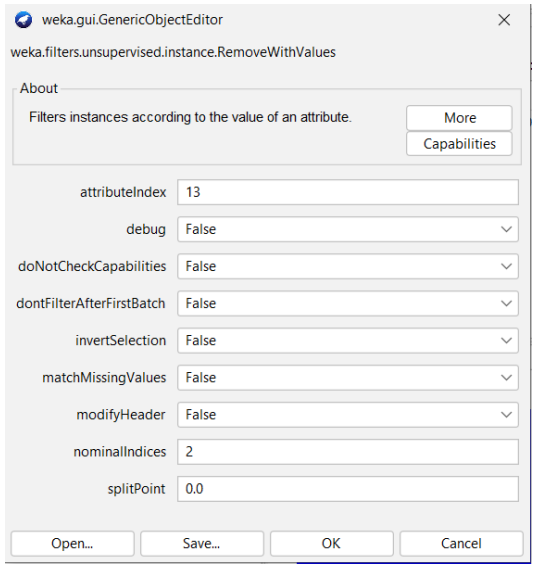
- 2 Cột mới được sinh ra là Outlier và ExtremeValue. Trong Outlier những giá trị có nhãn là yes chính là giá trị ngoại lai được tìm thấy.



Hình 24. Giá trị ngoại lai

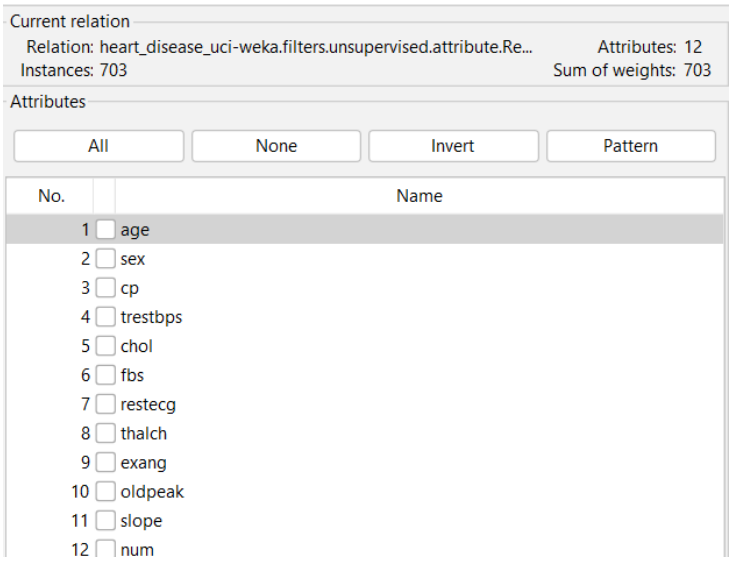
c) Loại bỏ giá trị ngoại lai.

- Sử dụng filter RemoveWithValues:
  - **attributeIndex**: giá trị cần xóa nằm ở cột 13.
  - **nominalIndices**: giá trị cần xóa có số thứ tự là 2.



Hình 25. Loại bỏ giá trị ngoại lai

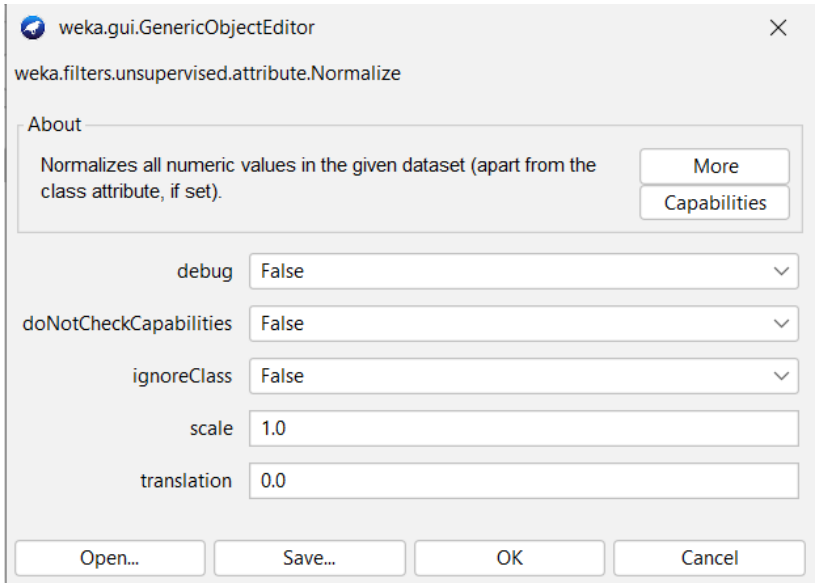
- Sau khi xử lý ngoại lai, các giá trị bằng 0 ở hai thuộc tính "trestbps" và "chol" phát hiện trước đó trong quá trình phân tích đã không còn.
- Còn lại 12 thuộc tính và 703 bản ghi sau khi làm sạch dữ liệu.



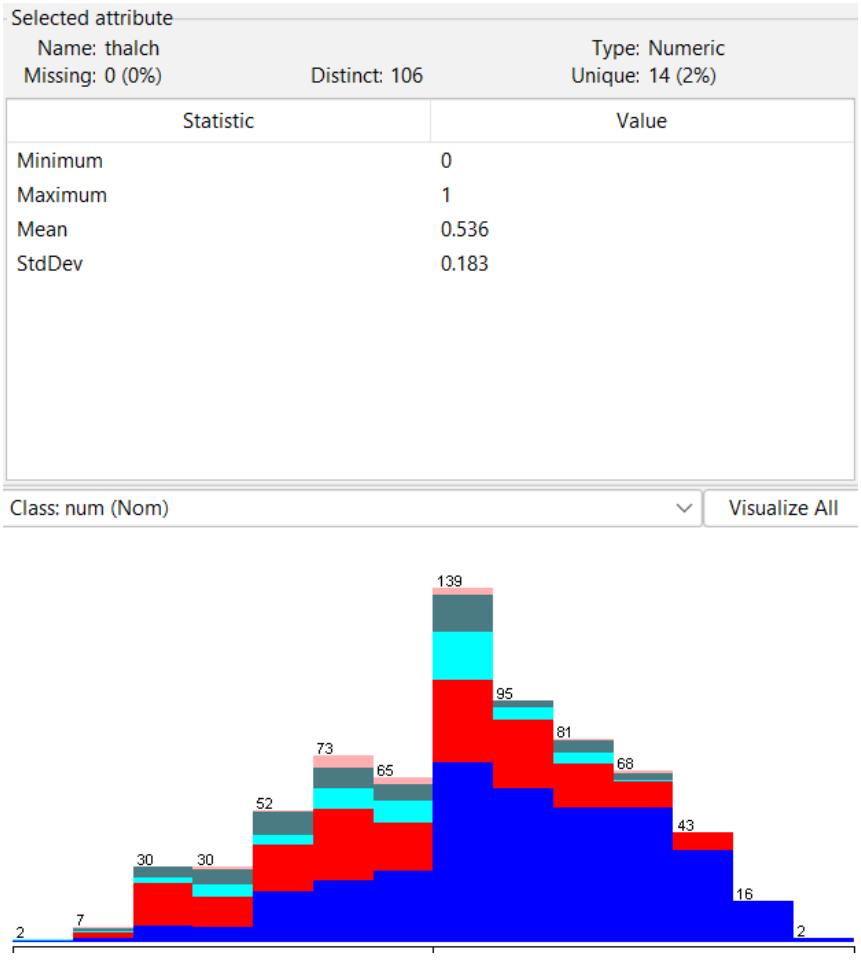
Hình 26. Dữ liệu sau khi được làm sạch

3.5.3 Chuyển đổi dữ liệu

Sử dụng filter Normalize để chuẩn hóa các dữ liệu kiểu số về khoảng [0-1].



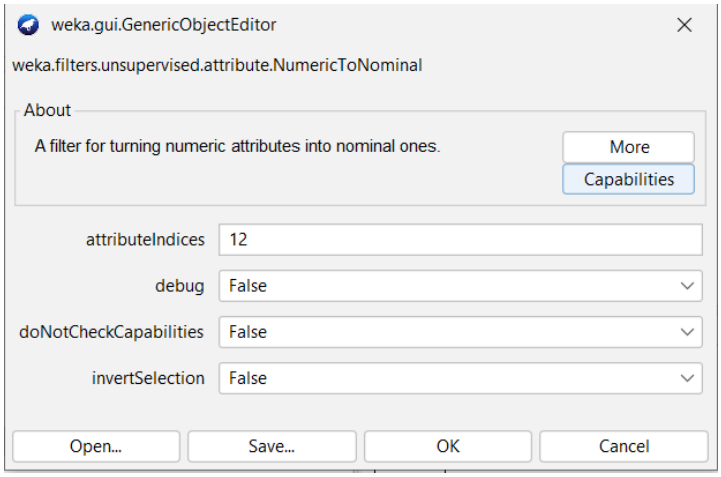
Hình 27. Chuẩn hóa dữ liệu bằng Normalize



Hình 28. Thuộc tính thalch sau khi chuẩn hóa

Chuyển đổi kiểu dữ liệu của “num” từ Numeric thành Nominal trước khi thực hiện bài toán phân lớp.

- Sử dụng filter NumericToNominal.



Hình 29. Chuyển đổi thuộc tính num bằng NumericToNominal

Selected attribute

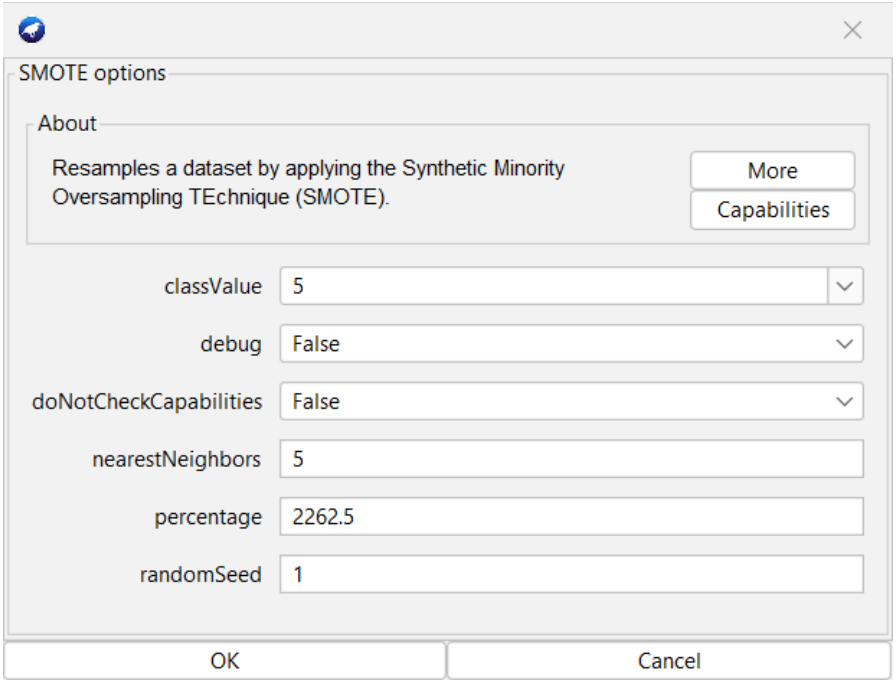
Name: num Type: Nominal
Missing: 0 (0%) Distinct: 5 Unique: 0 (0%)

No.	Label	Count	Weight
1	0	378	378
2	1	189	189
3	2	60	60
4	3	60	60
5	4	16	16

Hình 30. Thuộc tính num sau khi chuyển đổi

3.5.4 Xử lý mất cân bằng dữ liệu

Sử dụng kỹ thuật SMOTE (Synthetic Minority Over-sampling Technique) trong WEKA để tạo thêm các mẫu tổng hợp cho các lớp chiếm thiểu số, giúp cân bằng dữ liệu.



Hình 31. Sử dụng kỹ thuật SMOTE

Cấu hình bộ lọc SMOTE:

- **classValue**: Giá trị của lớp thiểu số mà bạn muốn tạo mẫu tổng hợp.
- **percentage**: Tỷ lệ phần trăm mẫu tổng hợp sẽ được tạo ra từ lớp chiếm thiểu số. Giá trị này càng cao thì số lượng mẫu tổng hợp càng lớn. Ví dụ ở lớp thứ 5 đang có 16 mẫu dữ liệu, muốn cân bằng với lớp thứ 1 thì cần (378 / 16 - 1) * 100 = 2262.5%.
- **nearestNeighbors**: Số lượng láng giềng gần nhất được sử dụng để tạo các mẫu tổng hợp. Thông thường giá trị mặc định là 5.
- **randomSeed**: Giá trị ngẫu nhiên để đảm bảo tính tái lập của kết quả.

Lặp lại quá trình này cho toàn bộ các lớp cần cân bằng.

Selected attribute

Name: num Type: Nominal
Missing: 0 (0%) Distinct: 5 Unique: 0 (0%)

No.	Label	Count	Weight
1	0	378	378
2	1	378	378
3	2	377	377
4	3	377	377
5	4	378	378

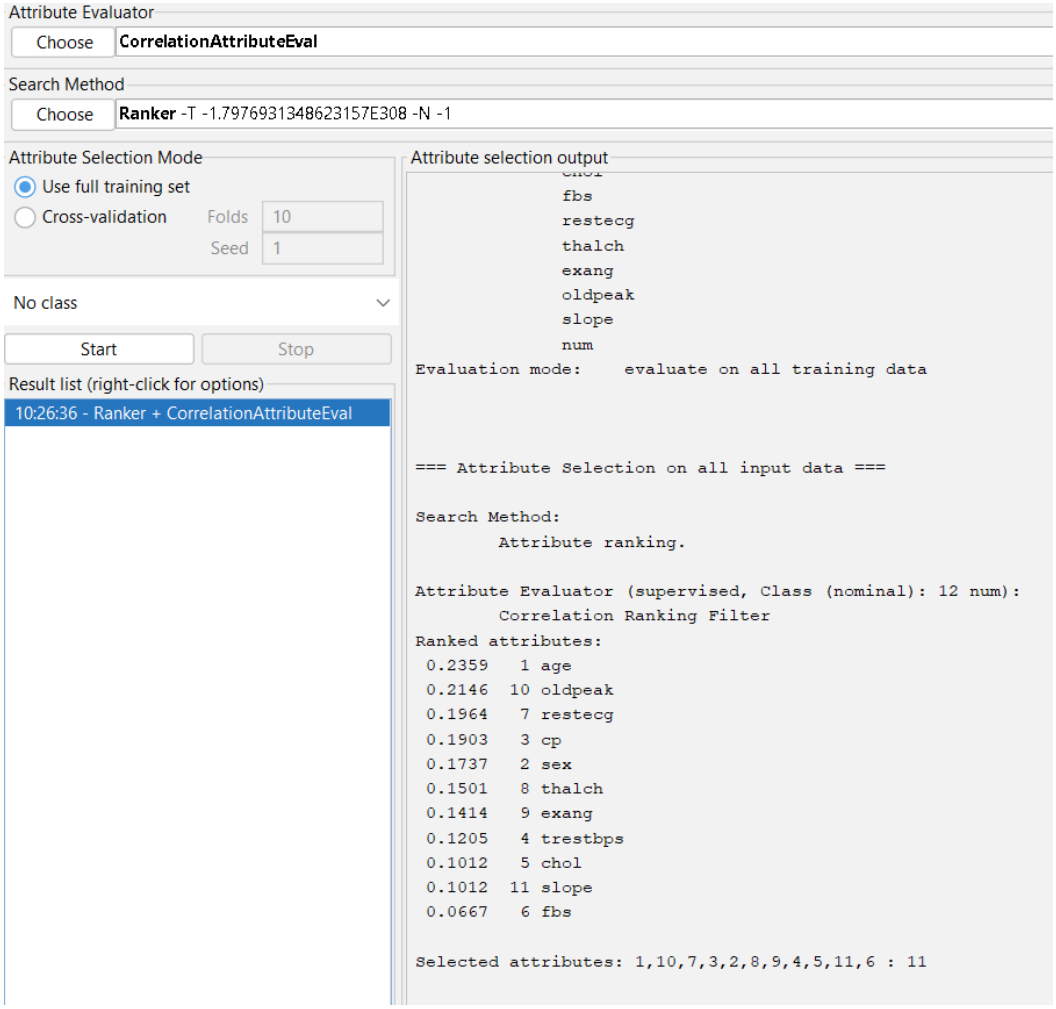
Hình 32. Các lớp của biến mục tiêu sau khi xử lý mất cân bằng

Sau quá trình cân bằng dữ liệu, tập dữ liệu tăng lên thành 1888 quan sát.

3.5.5 Tính độ tương quan

Sử Dụng Bộ Lọc CorrelationAttributeEval

Bộ lọc CorrelationAttributeEval sẽ tính toán hệ số tương quan Pearson giữa từng thuộc tính và biến mục tiêu. Hệ số này nằm trong khoảng từ -1 đến 1, với các giá trị gần 1 hoặc -1 cho thấy mối quan hệ tuyến tính mạnh mẽ giữa hai thuộc tính.



Hình 33. Sử dụng bộ lọc CorrelationAttributeEval

Nhận Xét

Kết quả tính độ tương quan cho thấy các thuộc tính sau có mối quan hệ mạnh mẽ với biến mục tiêu num (chẩn đoán bệnh tim):

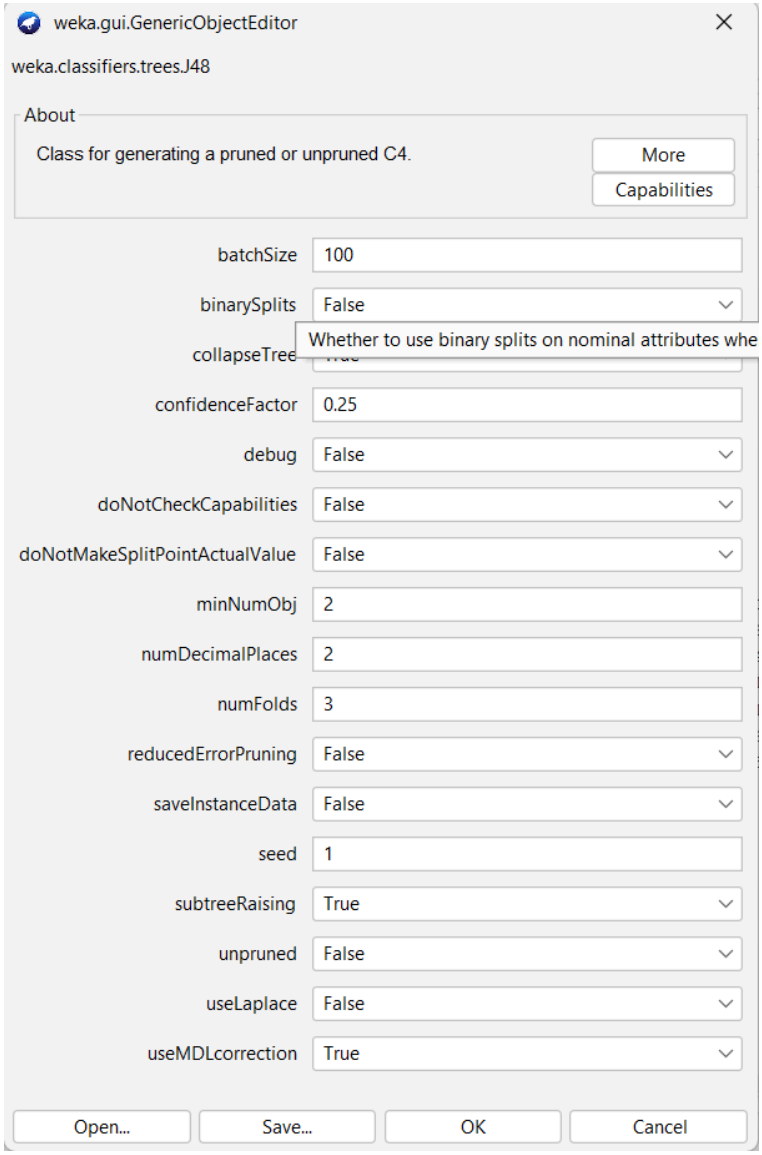
- age: 0.2359
- oldpeak: 0.2146
- restecg: 0.1964
- cp: 0.1903

- sex: 0.1737
- thalch: 0.1501

Các thuộc tính này có hệ số tương quan cao, cho thấy chúng có ảnh hưởng đáng kể đến biến mục tiêu và nên được ưu tiên trong quá trình xây dựng mô hình học máy. Các thuộc tính khác như exang, trestbps, chol, slope, và fbs có hệ số tương quan thấp hơn, nhưng vẫn có thể hữu ích trong một số bối cảnh nhất định.

3.6. Kết quả

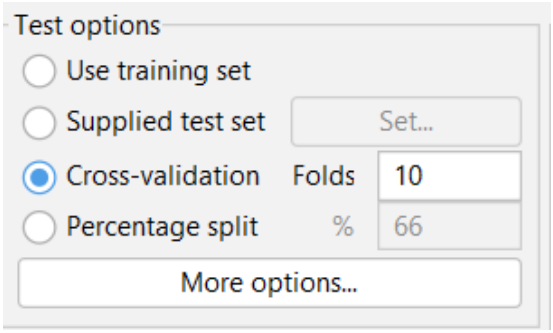
Sử dụng thuật toán phân lớp với thuật toán RandomForest trong tab Classify của Weka để so sánh kết quả trước và sau khi tiền xử lý dữ liệu.



Hình 34. Thuật toán RandomForest

Sử dụng phương pháp đánh giá là Cross-validate với k là 10.

- Chia bộ dữ liệu thành 10 phần, lặp lại việc đánh giá 10 lần:
  - Sử dụng 9 phần để huấn luyện mô hình.
  - Dùng 1 phần còn lại để kiểm tra.
  - Ghi nhận kết quả đánh giá.
- Tính trung bình kết quả từ 10 lần lặp.



Hình 35. Lựa chọn phương pháp đánh giá

```
=== Summary ===

Correctly Classified Instances         587               63.8043 %
Incorrectly Classified Instances       333               36.1957 %
Kappa statistic                          0.4561
Mean absolute error                      0.1766
Root mean squared error                  0.2979
Relative absolute error                 64.0215 %
Root relative squared error             80.2526 %
Total Number of Instances              920

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.895    0.193    0.790      0.895    0.839      0.699    0.939     0.940     0
                 0.626    0.162    0.610      0.626    0.618      0.461    0.827     0.746     1
                 0.229    0.078    0.284      0.229    0.254      0.167    0.803     0.262     2
                 0.262    0.079    0.304      0.262    0.281      0.196    0.798     0.290     3
                 0.000    0.002    0.000      0.000    0.000      -0.008   0.756     0.080     4
Weighted Avg.    0.638    0.151    0.598      0.638    0.616      0.487    0.868     0.702

=== Confusion Matrix ===

   a   b   c   d   e   <-- classified as
 368  22  10  11   0 |   a = 0
  50 166  28  19   2 |   b = 1
  24  37  25  23   0 |   c = 2
  20  38  21  28   0 |   d = 3
   4   9   4  11   0 |   e = 4
```

Hình 36. Kết quả trước khi tiền xử lý dữ liệu

```
=== Summary ===

Correctly Classified Instances        1544               81.7797 %
Incorrectly Classified Instances       344               18.2203 %
Kappa statistic                          0.7722
Mean absolute error                      0.138
Root mean squared error                  0.2416
Relative absolute error                 43.1135 %
Root relative squared error             60.3984 %
Total Number of Instances             1888

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.786    0.045    0.814      0.786    0.799      0.751    0.952     0.875     0
                 0.712    0.066    0.729      0.712    0.720      0.651    0.929     0.771     1
                 0.820    0.046    0.817      0.820    0.819      0.773    0.964     0.891     2
                 0.801    0.049    0.803      0.801    0.802      0.753    0.956     0.882     3
                 0.971    0.022    0.918      0.971    0.943      0.929    0.992     0.980     4
Weighted Avg.    0.818    0.046    0.816      0.818    0.817      0.771    0.959     0.880

=== Confusion Matrix ===

   a   b   c   d   e   <-- classified as
 297  47  13  18   3 |   a = 0
  49 269  24  32   4 |   b = 1
   6  27 309  21  14 |   c = 2
  11  24  28 302  12 |   d = 3
   2   2   4   3 367 |   e = 4
```

Hình 37. Kết quả đạt được sau khi tiền xử lý

CHƯƠNG 4. KẾT QUẢ - KẾT LUẬN

4.1. Nhận xét kết quả đề tài

Thông qua kết quả của bài toán phân lớp, có thể thấy độ chính xác đã được cải thiện đáng kể từ 60,8% lên 81,7%. Nhờ các bước tiền xử lý dữ liệu được thực hiện trên công cụ WEKA, chất lượng dữ liệu đầu vào đã được nâng cao, từ đó cải thiện hiệu suất và độ chính xác của mô hình phân lớp. Điều này minh chứng cho tầm quan trọng của việc sử dụng các phương pháp tiền xử lý dữ liệu trong học máy và khai thác dữ liệu.

4.2. Ưu - nhược điểm của đề tài

Ưu điểm:

- Giao diện thân thiện và dễ sử dụng: WEKA cung cấp giao diện người dùng thân thiện, giúp người dùng dễ dàng tiếp cận và sử dụng các công cụ tiền xử lý dữ liệu mà không cần kiến thức lập trình sâu.
- Hỗ trợ nhiều kỹ thuật tiền xử lý phổ biến: WEKA tích hợp sẵn nhiều bộ lọc và thuật toán tiền xử lý, giúp người dùng dễ dàng thực hiện các bước tiền xử lý như làm sạch, chuẩn hóa và chuyển đổi dữ liệu.
- Tích hợp sẵn nhiều bộ lọc và thuật toán không cần lập trình: Người dùng có thể dễ dàng kết hợp nhiều bước tiền xử lý thành một quy trình hoàn chỉnh mà không cần viết mã lập trình.
- Tương thích với nhiều định dạng dữ liệu phổ biến: WEKA hỗ trợ nhiều định dạng dữ liệu, giúp người dùng dễ dàng nhập và xuất dữ liệu trong quá trình phân tích.

Nhược điểm:

- Giới hạn trong việc xử lý dữ liệu lớn: WEKA có thể gặp khó khăn khi xử lý các tập dữ liệu lớn do hạn chế về hiệu suất và khả năng mở rộng.
- Ít linh hoạt hơn so với lập trình trực tiếp trong một số trường hợp đặc biệt: Đối với các tình huống yêu cầu tiền xử lý phức tạp hoặc tùy chỉnh, việc sử dụng WEKA có thể không linh hoạt bằng lập trình trực tiếp.
- Một số tùy chọn tiền xử lý nâng cao có thể không có sẵn: WEKA có thể thiếu một số tùy chọn tiền xử lý nâng cao mà người dùng có thể cần trong một số trường hợp đặc thù.

- Khó tích hợp với các công cụ bên ngoài hoặc quy trình tùy chỉnh phức tạp: Việc tích hợp WEKA với các công cụ và quy trình phân tích khác có thể gặp khó khăn do hạn chế về khả năng tùy chỉnh và mở rộng.

4.3. Hướng phát triển

Tích hợp các công cụ học máy và khai thác dữ liệu khác:

- Kết hợp WEKA với các công cụ học máy mạnh mẽ khác như Scikit-learn, TensorFlow, hoặc PyTorch để tận dụng khả năng mạnh mẽ và linh hoạt của chúng trong việc xử lý và phân tích dữ liệu.
- Phát triển các pipeline tự động hóa quá trình tiền xử lý và mô hình hóa dữ liệu giữa các công cụ khác nhau.

Nâng cao khả năng xử lý dữ liệu lớn:

- Sử dụng các phiên bản phân tán hoặc song song của WEKA để xử lý các tập dữ liệu lớn hơn.
- Kết hợp WEKA với các hệ thống xử lý dữ liệu lớn như Hadoop hoặc Spark để tăng cường khả năng mở rộng và hiệu suất xử lý.

Ứng dụng tiền xử lý dữ liệu trong các lĩnh vực khác:

- Mở rộng phạm vi nghiên cứu và ứng dụng các kỹ thuật tiền xử lý dữ liệu cho các lĩnh vực khác như tài chính, y tế, giáo dục, và nông nghiệp.
- Xây dựng các case study và minh họa cụ thể cho từng lĩnh vực để chứng minh tính hiệu quả của các phương pháp tiền xử lý dữ liệu.

Tích hợp thêm các kỹ thuật đánh giá:

- Nghiên cứu và áp dụng các kỹ thuật đánh giá mô hình tiên tiến như bootstrapping, và đánh giá dựa trên các chỉ số khác ngoài độ chính xác như AUC, F1-score.

TÀI LIỆU THAM KHẢO

[1] A. Saleem, K. H. Asif, A. Ali, S. M. Awan and M. A. Alghamdi, "Pre-processing Methods of Data Mining," 2014 IEEE/ACM 7th International Conference on Utility and Cloud Computing, London, UK, 2014, pp. 451-456, doi: 10.1109/UCC.2014.57.

[2] S. Singhal and M. Jena, "A Study on WEKA Tool for Data Preprocessing, Classification and Clustering," *International Journal of Innovative Technology and Exploring Engineering (IJITEE)*, vol. 2, no. 6, pp. 250-253, May 2013.

[3] S. Srivastava, "Weka: A Tool for Data preprocessing, Classification, Ensemble, Clustering and Association Rule Mining," *International Journal of Computer Applications*, vol. 88, no. 10, pp. 26-29, Feb. 2014, doi: 10.5120/15389-3809.