

Trường Đại học Khoa học Tự nhiên
Đại học Quốc gia TP.HCM

TIỀN XỬ LÝ DỮ LIỆU

Khai thác dữ liệu & Ứng dụng

Nguyễn Bảo Long - MSSV: 18120201

Huỳnh Long Nam - MSSV: 18120212

TP Hồ Chí Minh, ngày 20/10/2020

Contents

1	Thông tin chung	2
2	Yêu cầu 1: Cài đặt Weka	3
2.1	Giao diện chức năng Explorer sau khi cài đặt Weka	3
2.2	Giải thích ý nghĩa các nhóm điều khiển và các tabs trên giao diện	3
3	Yêu cầu 2: Làm quen với Weka	4
3.1	Đọc dữ liệu vào Weka	4
3.2	Khám phá tập dữ liệu Weather	6
3.3	Khám phá tập dữ liệu Tín dụng Đức	7
4	Yêu cầu 3: Cài đặt tiền xử lý dữ liệu	11

1 Thông tin chung

1. Link GitHub: <https://github.com/baolongnguyenmac/DataMining-Lab1>

2. Thông tin thành viên nhóm

STT	Họ tên	MSSV	Email
1	Nguyễn Bảo Long	18120201	18120201@student.hcmus.edu.vn
2	Huỳnh Nam Long	18120212	18120212@student.hcmus.edu.vn

Table 1: Bảng thông tin thành viên nhóm

3. Tỷ lệ tham gia công việc

STT	MSSV	Công việc		Tỷ lệ hoàn thành
1	18120201 18120212	Yêu cầu 1	Cài đặt Weka	100%
2	18120212	Yêu cầu 2	Đọc dữ liệu vào Weka	100%
3	18120212		Khám phá tập dữ liệu Weather	100%
3	18120201		Khám phá tập dữ liệu tín dụng Đức	100%
4	18120201	Yêu cầu 3	Liệt kê các cột bị thiếu dữ liệu	100%
5	18120201		Đếm số dòng bị thiếu dữ liệu	100%
6	18120201		Điền giá trị bị thiếu	100%
7	18120201		Xoá các dòng bị thiếu với ngưỡng cho trước	100%
8	18120201		Xoá các cột bị thiếu với ngưỡng cho trước	100%
9	18120212		Xoá các mẫu bị trùng lặp	100%
10	18120212		Chuẩn hoá một thuộc tính	100%
11	18120212		Tính giá trị biểu thức	100%
12	18120201	Viết		100%
13	18120201	Trình bày báo cáo		100%

Table 2: Bảng phân chia công việc và mức độ hoàn thành

2 Yêu cầu 1: Cài đặt Weka

2.1 Giao diện chức năng Explorer sau khi cài đặt Weka

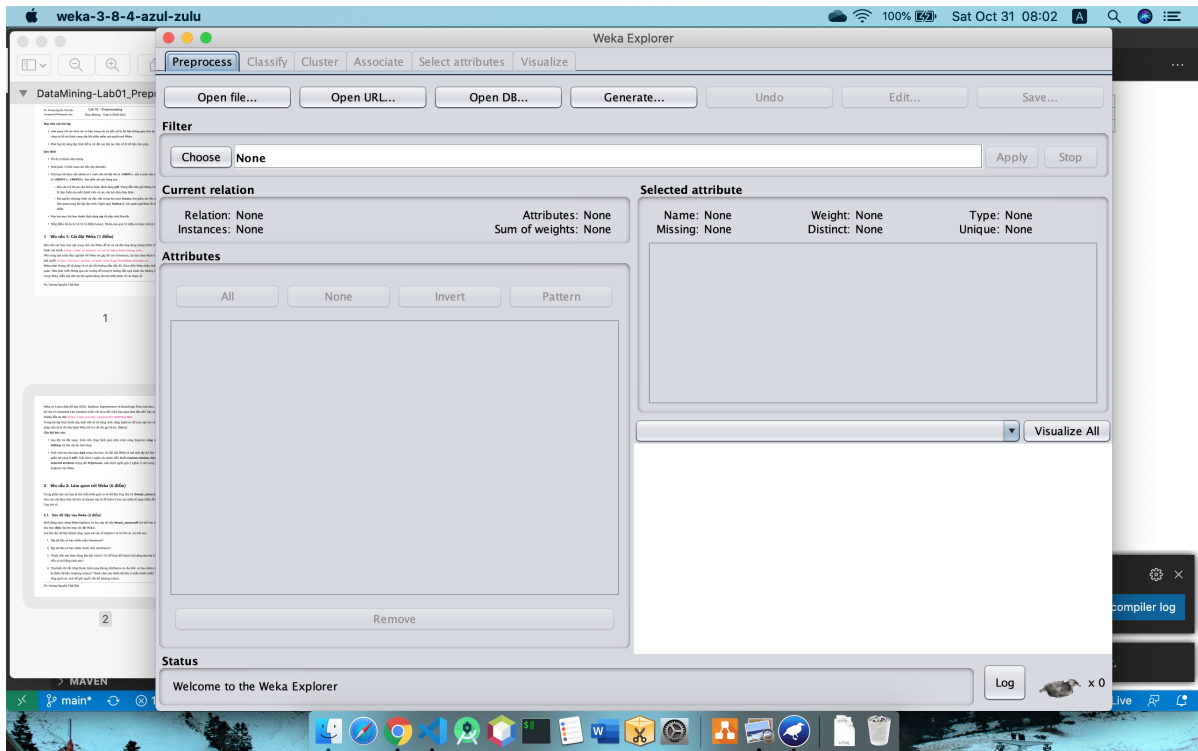


Figure 1: Giao diện chức năng Explorer của Weka

2.2 Giải thích ý nghĩa các nhóm điều khiển và các tabs trên giao diện

1. Ý nghĩa nhóm lệnh điều khiển trong tab PreProcess

- Current relation (tạm dịch ‘Quan hệ hiện tại’): Cho biết các thông tin chung về tập dữ liệu hiện tại như tên tập dữ liệu, số mẫu, số thuộc tính
- Attributes (tạm dịch ‘Thuộc tính’): Cho biết danh sách các thuộc tính hiện tại trong tập dữ liệu
- Selected attribute (tạm dịch ‘Thuộc tính được chọn’): Cho biết thông tin chung liên quan đến thống kê (trong TH dữ liệu là dạng số) như trung bình cộng, giá trị min, giá trị max của thuộc tính đã được chọn trước trong phần **Attributes**. Đối với dữ liệu dạng định danh, Weka sẽ cung cấp danh sách các định danh và số lượng mỗi định danh

2. Ý nghĩa của các tabs

- PreProcess: Tiền xử lý dữ liệu (tab mặc định)
- Classify: Phân lớp dữ liệu
- Cluster: Gom cụm dữ liệu
- Associate: Khai phá các luật kết hợp

- Select Attributes: Lựa chọn thuộc tính. Sử dụng khi cần xem xét mối tương quan giữa các thuộc tính
- Visualize: Trực quan hoá dữ liệu

3 Yêu cầu 2: Làm quen với Weka

3.1 Đọc dữ liệu vào Weka

- Mở tập dữ liệu **breast_cancer.arff** trong giao diện Weka Explorer
- Explore dữ liệu
 - Tập dữ liệu có 286 mẫu
 - Tập dữ liệu có 10 thuộc tính

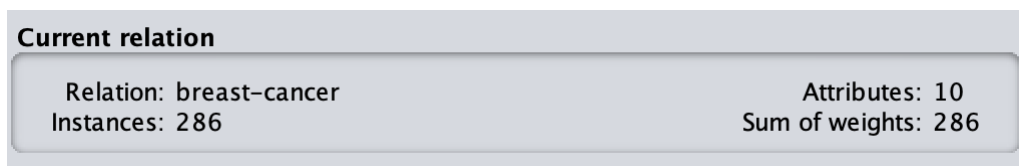


Figure 2: Số lượng mẫu và thuộc tính của tập dữ liệu **breast_cancer.arff**

- Thuộc tính "Class" là thuộc tính class của tập dữ liệu với 2 giá trị recurrence-events và no-recurrence-events được dùng làm lớp. Có thể thay đổi thuộc tính dùng làm lớp bằng cách chọn nút **Edit**, nhấp phải chuột tại thuộc tính mới và chọn **Attribute as class**

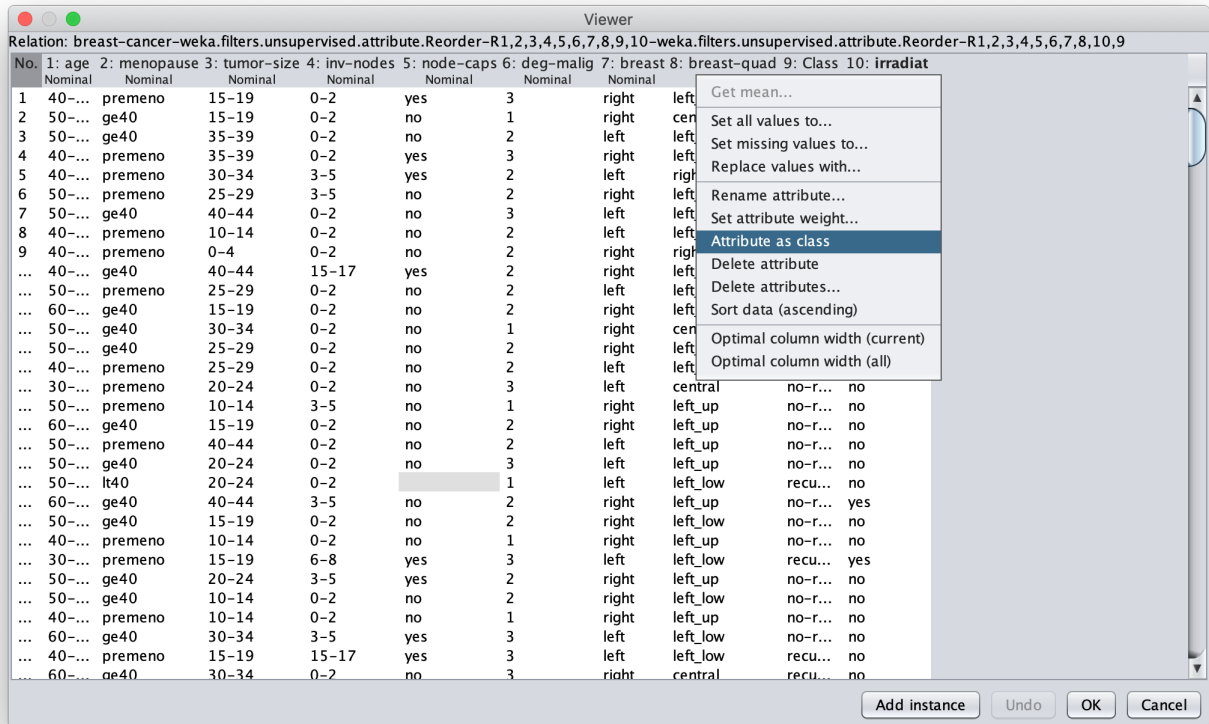


Figure 3: Thay đổi thuộc tính được dùng làm lớp

- Có 2 thuộc tính bị thiếu dữ liệu là **node_caps** và **breast_quad**. Trong đó, **node_caps** thiếu nhiều nhất (thiếu 8 mẫu, tương đương 3%) và **breast_quad** thiếu ít nhất (thiếu 1 mẫu, tương đương gần 0%)

Selected attribute			
Name: breast-quad		Type: Nominal	
Missing: 1 (0%)		Unique: 0 (0%)	
Distinct: 5			
No.	Label	Count	Weight
1	left_up	97	97.0
2	left_low	110	110.0
3	right_up	33	33.0
4	right_low	24	24.0
5	central	21	21.0

Selected attribute			
Name: node-caps		Type: Nominal	
Missing: 8 (3%)		Unique: 0 (0%)	
Distinct: 2			
No.	Label	Count	Weight
1	yes	56	56.0
2	no	222	222.0

Figure 4: Thuộc tính **node_caps** và **breast_quad** bị thiếu dữ liệu

- Khi gặp tình trạng thiếu dữ liệu, ta có thể loại bỏ dữ liệu thiếu đó hoặc bổ sung dữ liệu thiếu (bằng phương pháp lấy trung bình hoặc kNN)
- Có thể đặt tên cho đồ thị này là đồ thị phân bố lớp. Màu xanh biểu thị tại mỗi khoảng

dữ liệu của attribute được chọn, có bao nhiêu mẫu cho kết quả no-recurrence-events.
Tương tự với màu đỏ, cho kết quả recurrence-events

3.2 Khám phá tập dữ liệu Weather

- Tập dữ liệu có 5 thuộc tính và 14 mẫu
 - Thuộc tính định danh: outlook, windy, play
 - Thuộc tính số: temperature, humidity
- Five-number summary của thuộc tính temperature và humidity

	Min	Q1	Mean	Q3	Max
Temperature	64	69	73.571	80	85
Humidity	65	70	81.643	90	96

Table 3: Bảng Five-number summary của thuộc tính temperature và humidity

- Đồ thị biểu diễn các thuộc tính của tập dữ liệu

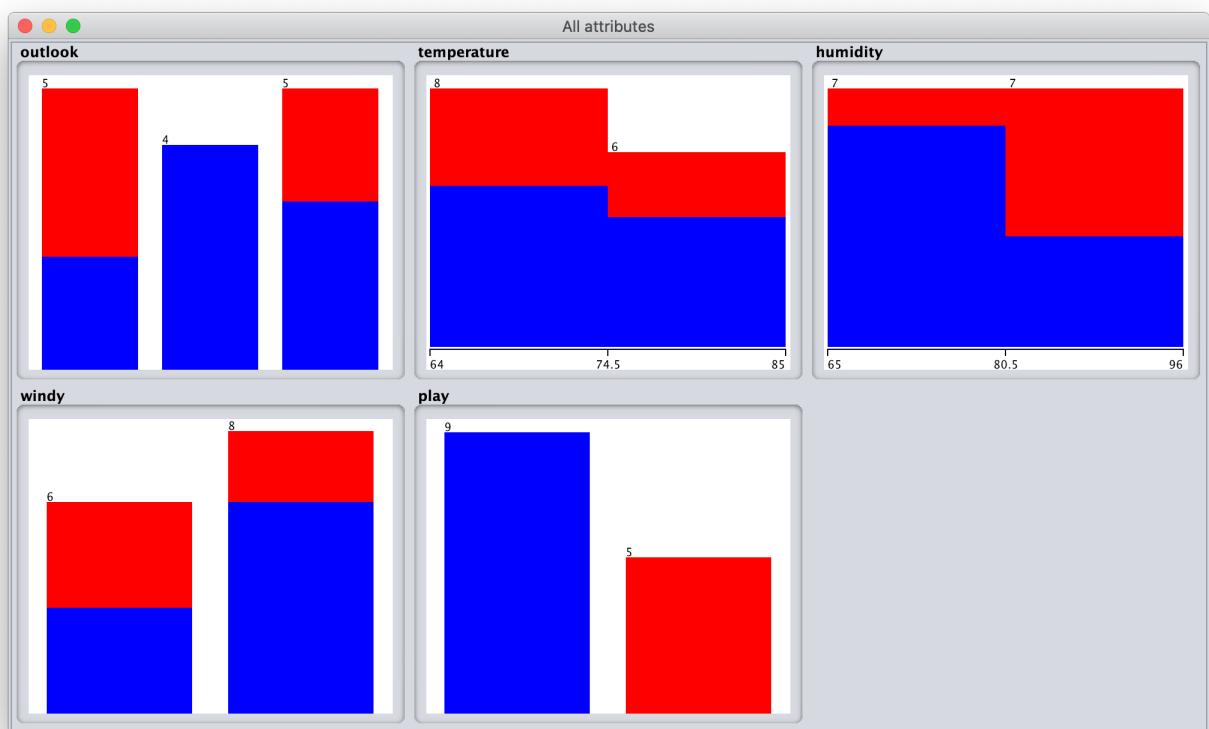


Figure 5: Đồ thị biểu diễn các thuộc tính của tập dữ liệu của tập weather

- Thuật ngữ sử dụng cho các đồ thị ở tab Visualize là đồ thị phân tán (scatter plot)

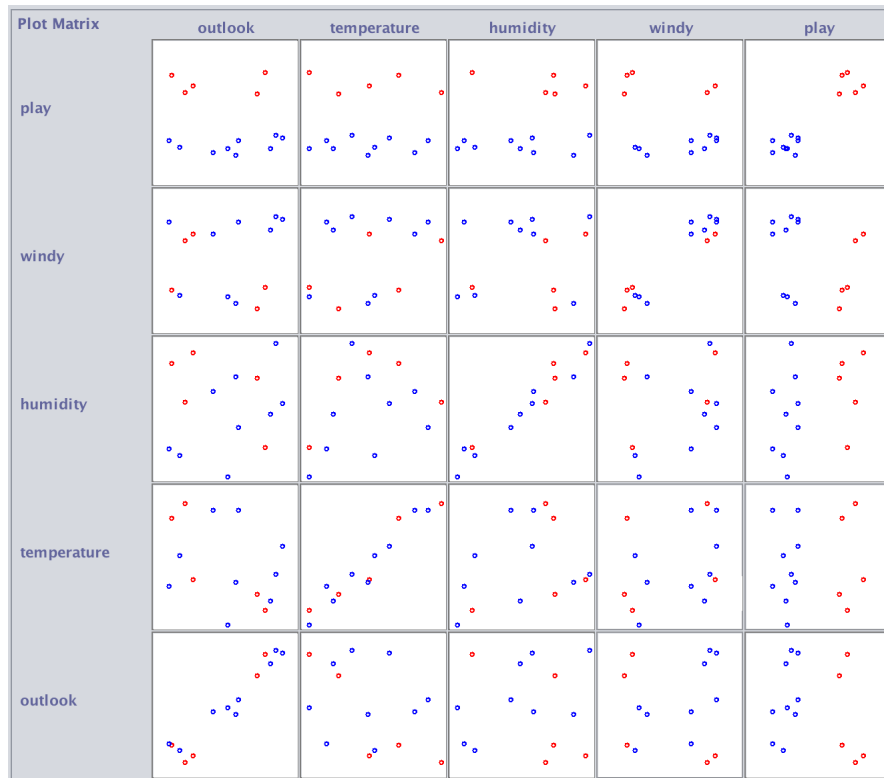


Figure 6: Đồ thị phân tán của các thuộc tính trong tập dữ liệu Weather

- Dựa vào các đồ thị quan sát được, theo phán đoán của nhóm, cặp dữ liệu *outlook* - *play* có vẻ như tương quan với nhau

3.3 Khám phá tập dữ liệu Tín dụng Đức

- Nội dung chú thích ở đầu tập tin mô tả về tập dữ liệu tín dụng Đức. Phần mô tả bao gồm tiêu đề, thông tin về nguồn gốc của tập tin, số lượng mẫu, số lượng thuộc tính và loại dữ liệu của các thuộc tính và mô tả chi tiết về thuộc tính. Ngoài ra, phần chú thích còn cho biết thêm về ma trận chi phí và ánh xạ ý nghĩa của giá trị của thuộc tính với ký hiệu hiển thị trên giao diện. Tập thuộc tính có 1000 mẫu với 21 thuộc tính. Dưới đây là thông tin về 5/21 thuộc tính có trong tập dữ liệu

- duration (thuộc tính rời rạc): thời hạn vay tín dụng (tính theo tháng)
- credit_history (thuộc tính rời rạc): lịch sử tín dụng, bao gồm 5 trạng thái
 - * Không có tín dụng nào được thực hiện hoặc các khoản tín dụng được trả một cách hợp lệ
 - * Tất cả các tín dụng ở ngân hàng này đều được trả một cách hợp lệ
 - * Các khoản tín dụng đã được hoàn trả hợp lệ cho đến nay
 - * Trong quá khứ đã từng hoàn trả tín dụng muộn
 - * Tài khoản tín dụng quan trọng hoặc đã tồn tại tài khoản tín dụng ở ngân hàng khác
- purpose (thuộc tính rời rạc): mục đích của việc vay tín dụng
 - * Mua xe mới
 - * Mua xe đã qua sử dụng

- * Mua đồ nội thất
- * Mua radio/TV
- * Mua thiết bị gia dụng
- * Sửa chữa
- * Chi trả cho giáo dục
- * Đi nghỉ dưỡng
- * Chi trả chi phí đào tạo lại
- * Đầu tư kinh doanh
- * Khác
- saving_status (thuộc tính liên tục): tài khoản tiết kiệm, được chia ra các mức
 - * Nhỏ hơn 100 Mark Đức¹
 - * Nằm trong khoảng 100 và 500 Mark Đức
 - * Nằm trong khoảng 500 và 1000 Mark Đức
 - * Lớn hơn 1000 Mark Đức
- personal_status (thuộc tính rời rạc): giới tính và trạng thái hiện tại của một người
 - * Nam, ly thân
 - * Nữ, ly thân hoặc đã có gia đình
 - * Nam, độc thân
 - * Nữ độc thân
 - * Nam đã có gia đình hoặc góa vợ
- Tên thuộc tính lớp: class (bao gồm 2 giá trị là good và bad). Cân bằng lệch về phía good
- Phương pháp chọn lọc thuộc tính của Weka trong tab Select attributes
 - **GainRatioAttributeEval**: Độ đo này được sử dụng trong thuật toán C4.5 do Quinlan đưa ra năm 1993. Ý tưởng của thuật toán là xét tất cả các phép thử có thể phân chia tập dữ liệu đã cho và chọn ra một phép thử cho GainRatio tốt nhất. GainRatio cũng là một độ đo sự hiệu quả của một thuộc tính trong thuật toán triển khai cây quyết định

$$GainRatio(S, A) = \frac{Gain(S, A)}{SplitInformation(S, A)}$$

$$SplitInformation(S, A) = - \sum_{i=1}^c \left(\frac{|S_i|}{|S|} \log_2 \frac{|S_i|}{|S|} \right)$$

* S_i là tập con của S với A có giá trị v_i

- **InfoGainAttributeEval**: Đo mức hiệu quả của một thuộc tính trong bài toán phân lớp dữ liệu

$$Gain(S, A) = Entropy(S) - \sum_{v \in Value(A)} \left[\frac{S_v}{S} Entropy(S_v) \right]$$

$$Entropy(S) = \sum_{i=1}^c (-p_i \log_2 p_i)$$

¹Đơn vị tiền tệ của Đức

- * p_i là tỉ lệ của các mẫu thuộc lớp i trong tập S
- * $Value(A)$ là tập tất cả các giá trị có thể có đối với thuộc tính A
- * S_v là tập con của S mà A có giá trị là v
- Ngoài ra còn rất nhiều options để lọc thuộc tính tùy theo ý đồ của người sử dụng.
- Theo như mô tả trên thì việc chọn **GainRatioAttributeEval** hay **InfoGainAttributeEval** đều cho biết về các thuộc tính có tương quan cao nhất đối với thuộc tính lớp. Trong báo cáo, người viết chọn **InfoGainAttributeEval** để chọn các thuộc tính này.
 - Bước 1: Chọn vào tab Select attributes. Tại mục Attribute Evaluator, chọn **InfoGainAttributeEval**. Tại mục Search Method, chọn **Ranker**

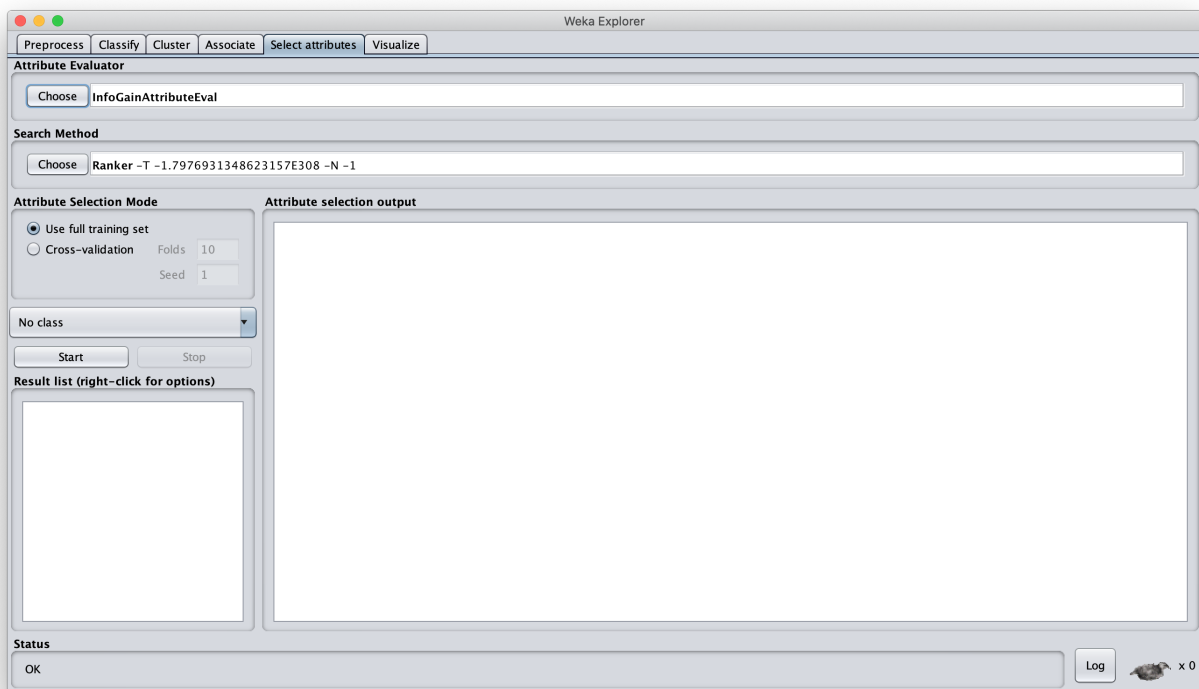


Figure 7: Chọn Attribute Evaluator và Search Method

- Bước 2: Nhấn nút Start
- Bước 3: Đọc kết quả

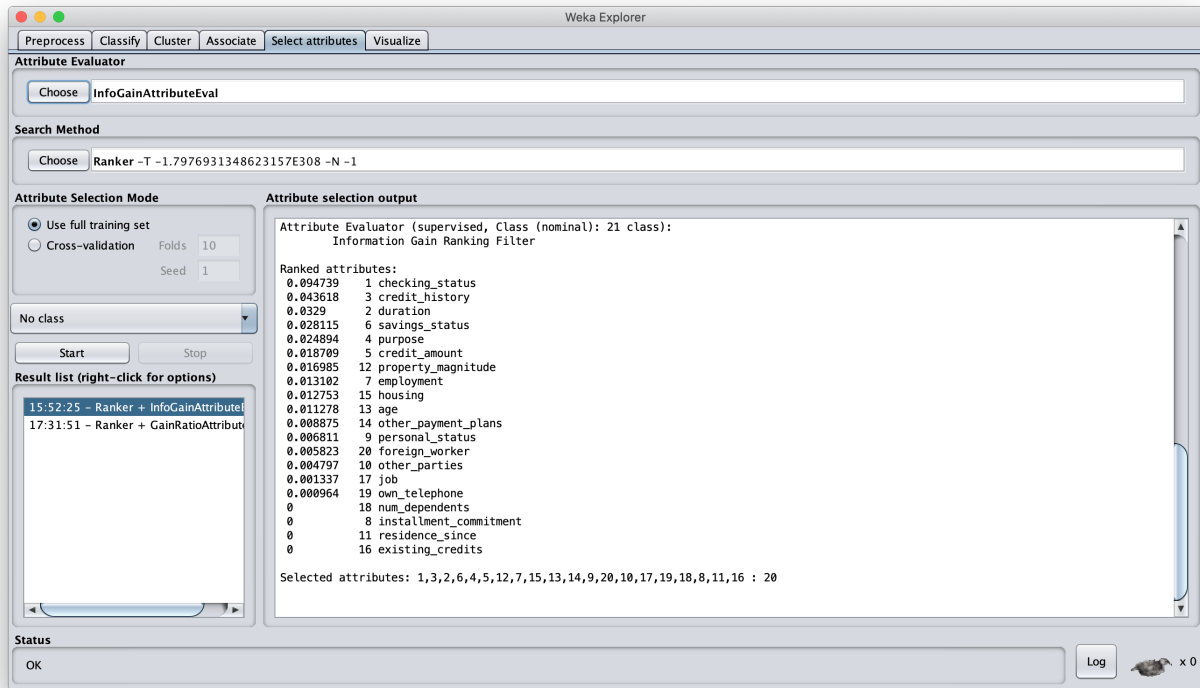


Figure 8: Danh sách các thuộc tính có tương quan cao nhất đối với thuộc tính lớp

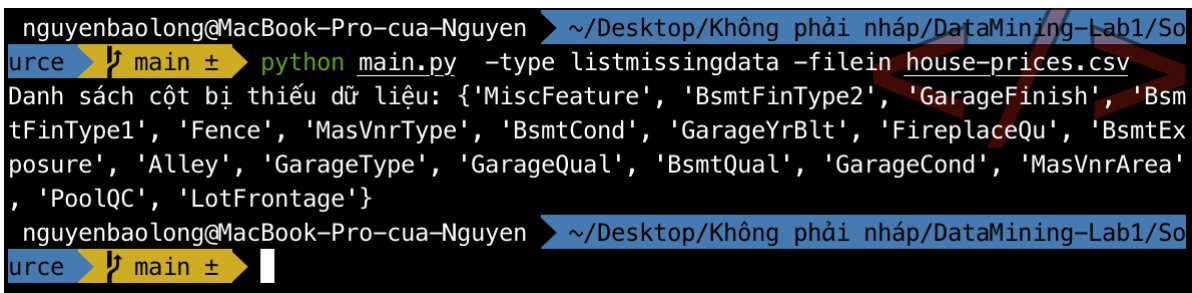
- Từ hình trên, ta có thể thấy 5 thuộc tính có tương quan cao nhất với thuộc tính lớp là `checking_status`, `credit_history`, `duration`, `saving_status`, `purpose`

4 Yêu cầu 3: Cài đặt tiền xử lý dữ liệu

- Trong thư mục **Result**, nhóm đề án đã có đính kèm tất cả các dữ liệu đầu ra của chương trình kèm theo mô tả trong file **README.md**
- Đường dẫn: <https://github.com/baolongnguyenmac/DataMining-Lab1/tree/main/Result>

1. Liệt kê các cột bị thiếu dữ liệu

- Cú pháp: `python main.py -type listmissingdata -filein <fileName>`
- Kết quả: Chương trình in ra danh sách các cột bị thiếu dữ liệu

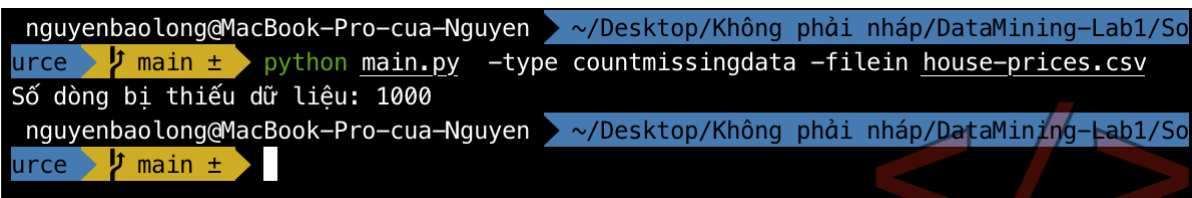


```
nguyenbaolong@MacBook-Pro-cua-Nguyen > ~/Desktop/Không phải nháp/DataMining-Lab1/Source
urce > main ± python main.py -type listmissingdata -filein house-prices.csv
Danh sách cột bị thiếu dữ liệu: {'MiscFeature', 'BsmtFinType2', 'GarageFinish', 'BsmtFinType1', 'Fence', 'MasVnrType', 'BsmtCond', 'GarageYrBlt', 'FireplaceQu', 'BsmtExposure', 'Alley', 'GarageType', 'GarageQual', 'BsmtQual', 'GarageCond', 'MasVnrArea', 'PoolQC', 'LotFrontage'}
nguyenbaolong@MacBook-Pro-cua-Nguyen > ~/Desktop/Không phải nháp/DataMining-Lab1/Source
urce > main ±
```

Figure 9: Kết quả sau khi liệt kê các thuộc tính bị mất dữ liệu

2. Đếm số dòng bị thiếu dữ liệu

- Cú pháp: `python main.py -type countmissingdata -filein <fileName>`
- Kết quả: Chương trình in ra số dòng bị thiếu dữ liệu



```
nguyenbaolong@MacBook-Pro-cua-Nguyen > ~/Desktop/Không phải nháp/DataMining-Lab1/Source
urce > main ± python main.py -type countmissingdata -filein house-prices.csv
Số dòng bị thiếu dữ liệu: 1000
nguyenbaolong@MacBook-Pro-cua-Nguyen > ~/Desktop/Không phải nháp/DataMining-Lab1/Source
urce > main ±
```

Figure 10: Kết quả sau khi đếm số dòng bị thiếu dữ liệu

3. Điền giá trị thiếu

- Cú pháp
 - `python main.py -type fillmissingdata -method mean -filein <fileName> -fileout <fileName>`
 - `python main.py -type fillmissingdata -method median -filein <fileName> -fileout <fileName>`
- Kết quả: Dữ liệu sau khi xử lý được lưu trong file kết quả

4. Xoá các dòng bị thiếu dữ liệu

- Cú pháp: `python main.py -type eraserow -rate <rate> -filein <fileName> -fileout <fileName>`

- Kết quả: Dữ liệu sau khi xử lý được lưu trong file kết quả

5. Xóa các cột bị thiếu dữ liệu

- Cú pháp: `python main.py -type erasecolumn -rate <rate> -filein <fileName> -fileout <fileName>`
- Kết quả: Dữ liệu sau khi xử lý được lưu trong file kết quả

6. Xóa các mẫu bị trùng lặp

- Cú pháp: `python main.py -type eraseduplicaterow -filein <fileName> -fileout <fileName>`
- Kết quả: Dữ liệu sau khi xử lý được lưu trong file kết quả

7. Chuẩn hoá một thuộc tính

- Cú pháp
 - `python main.py -type standardize -method minmax -attribute <attribute> -filein <fileName> -fileout <fileName>`
 - `python main.py -type standardize -method zscore -attribute <attribute> -filein <fileName> -fileout <fileName>`
- Kết quả: Dữ liệu sau khi chuẩn hoá được lưu trong file kết quả

8. Tính giá trị biểu thức

- Cú pháp: `python main.py -type expression`
- Kết quả: Chương trình cho người dùng nhập biểu thức trên màn hình Console. Kết quả của biểu thức được lưu thành 1 thuộc tính mới (có tên trùng với biểu thức nhập vào) trong file kết quả