

**ĐẠI HỌC KHOA HỌC TỰ NHIÊN – ĐHQG HCM**  
**KHOA CÔNG NGHỆ THÔNG TIN**



**REPORT**

**LAB01: DATA PREPROCESSING AND DATA EXPLORATION**  
**MÔN: KHAI THÁC DỮ LIỆU VÀ ỨNG DỤNG**

**SINH VIÊN THỰC HIỆN:**

Nguyễn Thế Hiển - 20120078

Nguyễn Thị Cẩm Lai - 20120128

# MỤC LỤC

I.	Thông tin chung .....	3
1.	Thông tin thành viên .....	3
2.	Phân công công việc .....	3
3.	Đánh giá mức độ hoàn thành .....	4
II.	Cài đặt WEKA .....	4
1.	Yêu cầu 1 .....	4
2.	Yêu cầu 2 .....	5
III.	Làm quen với WEKA .....	8
1.	Khám phá tập dữ liệu Breast Cancer.....	8
2.	Khám phá tập dữ liệu Weather.....	13
3.	Khám phá tập dữ liệu Credit in Germany .....	18
IV.	Tiền xử lý dữ liệu trong Python .....	30
0.	Đặc tả chung .....	30
1.	Extract columns with missing values.....	31
2.	Count the number of lines with missing data.....	31
4.	Deleting rows containing more than a particular number of missing values (Example: delete rows with the number of missing values is more than 50% of the number of attributes).....	32
5.	Deleting columns containing more than a particular number of missing values (Example: delete columns with the number of missing values is more than 50% of the number of samples) .....	33
6.	Delete duplicate samples.....	33
7.	Normalize a numeric attribute using min-max and Z-score methods .....	33
8.	Performing addition, subtraction, multiplication, and division between two numerical attributes.....	34
V.	Tài liệu tham khảo .....	35

## I. Thông tin chung

### 1. Thông tin thành viên

STT	Họ tên	MSSV	Email
1	Nguyễn Thế Hiển	20120078	20120078@student.hcmus.edu.vn
2	Nguyễn Thị Cẩm Lai	20120128	20120128@student.hcmus.edu.vn

### 2. Phân công công việc

Mục	Nhiệm vụ	Người thực hiện
3.1 Install WEKA	Requirement 1	Thế Hiển Cẩm Lai
	Requirement 2	Thế Hiển
3.2 Getting Acquainted With WEKA	3.2.1 Exploring Breast Cancer data set	Cẩm Lai
	3.2.2 Exploring Weather data set	Cẩm Lai
	3.2.3 Exploring Credit in Germany data set	Thế Hiển
3.3 Preprocessing Data in Python	1. Extract columns with missing values	Thế Hiển
	2. Count the number of lines with missing data.	Thế Hiển
	3. Fill in the missing value using mean, median (for numeric properties) and mode (for the categorical attribute).	Thế Hiển
	4. Deleting rows containing more than a particular number of missing values	Thế Hiển
	5. Deleting columns containing more than a particular number of missing values	Cẩm Lai
	6. Delete duplicate samples.	Cẩm Lai

	7. Normalize a numeric attribute using min-max and Z-score methods.	Cầm Lai
	8. Performing addition, subtraction, multiplication, and division between two numerical attributes.	Cầm Lai
Trình bày báo cáo		Thế Hiển Cầm Lai

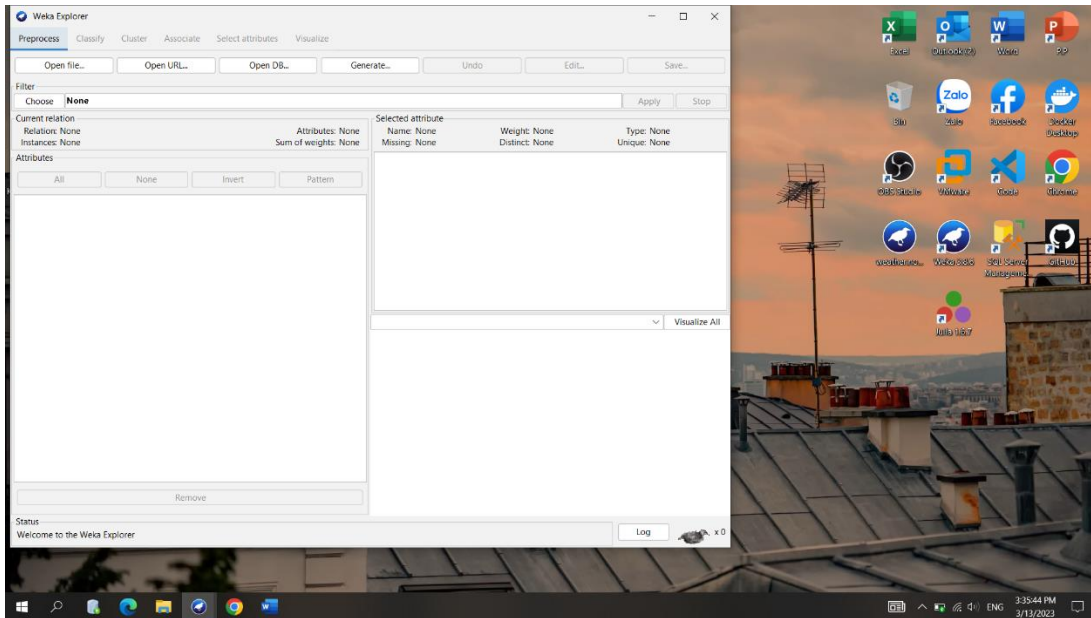
### 3. Đánh giá mức độ hoàn thành

Mục	Nội dung	Mức độ hoàn thành
3.1 Install WEKA	Requirement 1	100%
	Requirement 2	100%
3.2 Getting Acquainted With WEKA	3.2.1 Exploring Breast Cancer data set	100%
	3.2.2 Exploring Weather data set	100%
	3.2.3 Exploring Credit in Germany data set	100%
3.3 Preprocessing Data in Python	Set 8 data preprocessing functions	100%
	Test with house-prices.csv data set	100%
Tổng		100%

## II. Cài đặt WEKA

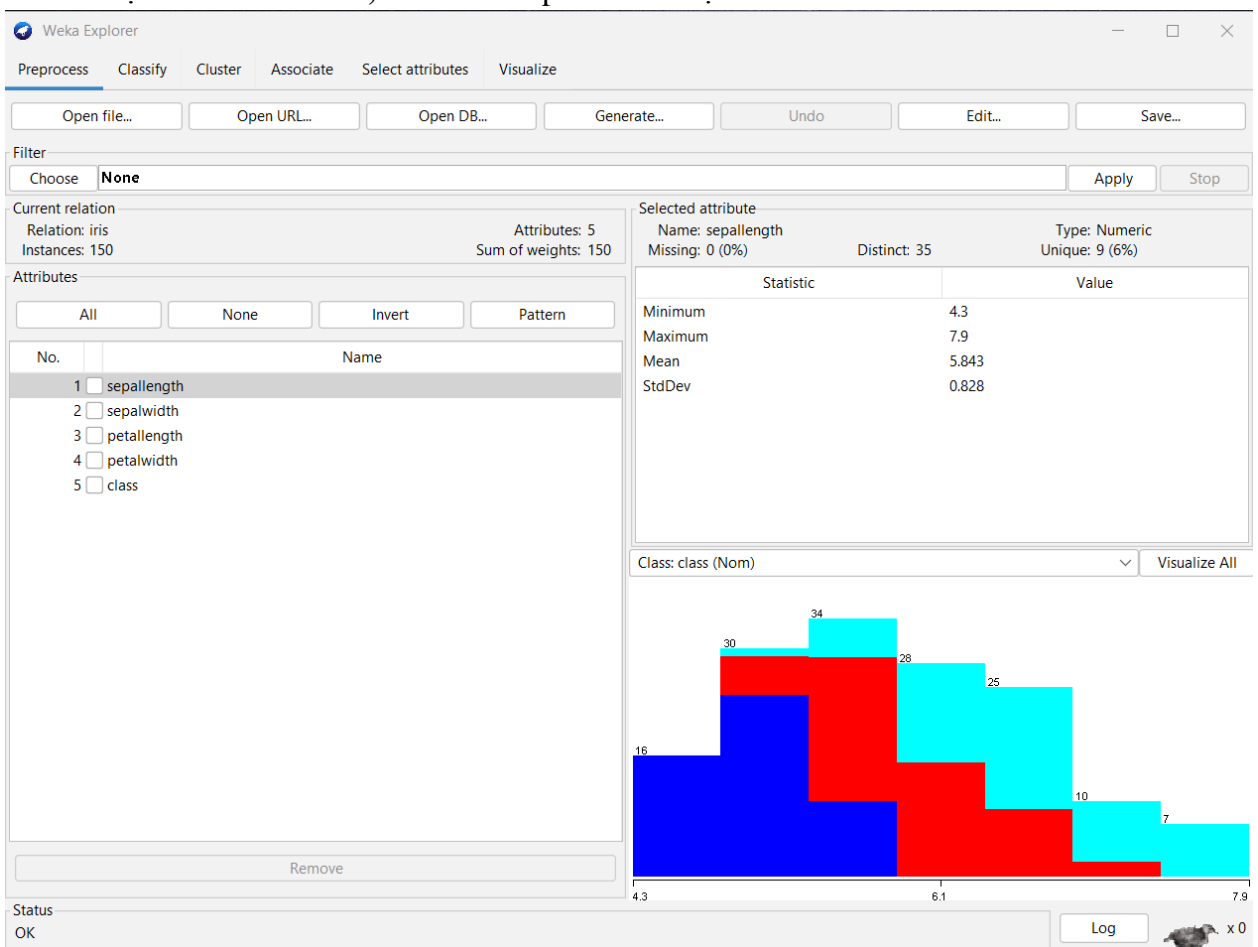
### 1. Yêu cầu 1

Màn hình Explorer sau khi cài đặt Weka thành công:



## 2. Yêu cầu 2

Nhóm chọn mở file **iris.arff**, màn hình explore hiển thị:



a) Giải thích ý nghĩa các mục trong thẻ Preprocess

- Mục **Current Relation**: cho biết các thông tin chung về tập dữ liệu hiện tại như: tên tập dữ liệu, số mẫu, số thuộc tính.
  - **Relation** (tên thể hiện): iris
  - **Instances** (số mẫu): 150
  - **Attributes** (số thuộc tính của bảng): 5
  - **Sum of weights** (tổng trọng số của bảng): 150

Current relation	
Relation: iris	Attributes: 5
Instances: 150	Sum of weights: 150

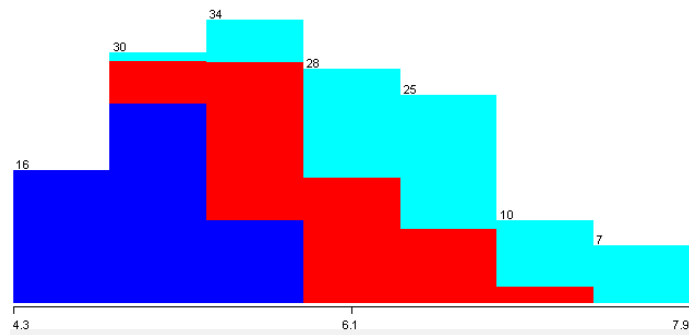
- Mục **Attributes**: hiển thị các thuộc tính có trong tập dữ liệu. Và khi chọn một thuộc tính bất kỳ trong danh sách thuộc tính này thì thông tin chi tiết về từng thuộc tính sẽ được thể hiện trong mục **Selected Attribute**.

Attributes	
<div>AllNoneInvertPattern</div>	
No.	Name
1 <input checked="" type="checkbox"/>	sepalength
2 <input type="checkbox"/>	sepalwidth
3 <input type="checkbox"/>	petallength
4 <input type="checkbox"/>	petalwidth
5 <input type="checkbox"/>	class

- Mục **Selected Attribute**: khi một thuộc tính bất kỳ được chọn tại mục Attributes thì thông tin chi tiết về thuộc tính đó sẽ được hiển thị tại đây, cụ thể bao gồm:
  - Name (tên thuộc tính)
  - Type (loại thuộc tính)
  - Missing (số lượng các giá trị bị thiếu)
  - Distinct (số lượng các giá trị khác nhau)
  - Unique (số lượng và tỷ lệ phần trăm các giá trị của thuộc tính này mà khác với bất kỳ giá trị của bất kỳ thuộc tính nào trong tập dữ liệu)

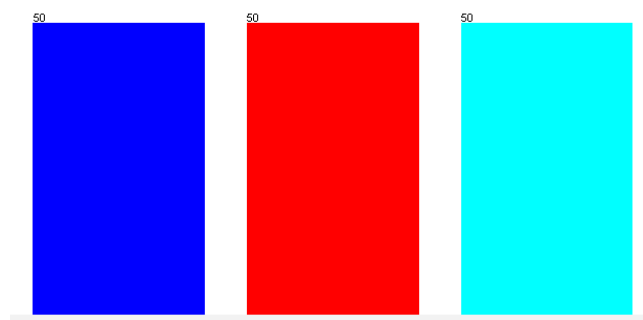
Đối với dữ liệu dạng số (numeric): cho biết thông tin chung liên quan đến thống kê như: trung bình cộng, giá trị min, giá trị max. Và một biểu đồ trực quan hóa cho thuộc tính đó.

Selected attribute		
Name: sepallength		Type: Numeric
Missing: 0 (0%)	Distinct: 35	Unique: 9 (6%)
	Statistic	Value
	Minimum	4.3
	Maximum	7.9
	Mean	5.843
	StdDev	0.828
Class: class (Nom) <span>Visualize All</span>		



Đối với dữ liệu dạng định danh (nominal): cung cấp danh sách các định danh và số lượng mỗi định danh có trong thuộc tính. Và một biểu đồ trực quan hóa cho thuộc tính đó.

Selected attribute			
Name: class		Type: Nominal	
Missing: 0 (0%)		Distinct: 3	
		Unique: 0 (0%)	
No.	Label	Count	Weight
1	Iris-setosa	50	50
2	Iris-versicolor	50	50
3	Iris-virginica	50	50
Class: class (Nom) <span>Visualize All</span>			



### b) Giải thích các tag còn lại trong WEKA Explorer

- **Preprocess Tag:** khi mới mở trình Explorer thì tag Preprocess là tag mặc định sẽ được bật trước. Dùng để thực hiện các bước tiền xử lý dữ liệu phục vụ cho các mục đích khác.
- **Classify Tag:** cung cấp một số thuật toán máy học phục vụ cho việc phân loại dữ liệu, chẳng hạn như: Linear Regression, Logistic Regression, Decision Trees, RandomTree, RandomForest, NaiveBayes, ...
- **Cluster Tag:** cung cấp một số thuật toán phân cụm, chẳng hạn như: SimpleKMeans, FilteredClusterer, ...
- **Associate Tag:** được sử dụng để khám phá các luật kết hợp từ dữ liệu, chứa các thuật toán như FPGrowth, ...
- **Select Attributes Tag:** hỗ trợ việc làm nổi bật các thuộc tính liên quan của dữ liệu, dựa trên một số thuật toán như ClassifierSubsetEval, ...
- **Visualize Tag:** hỗ trợ việc trực quan hóa dữ liệu đã xử lý để phân tích.

## III. Làm quen với WEKA

### 1. Khám phá tập dữ liệu Breast Cancer

#### a) How many instances does this data set have?

- Tập dữ liệu có 286 mẫu

Filter	
Choose	None
Current relation	
Relation: breast-cancer	Attributes: 10
Instances: 286	Sum of weights: 286

#### b) How many attributes does this data set have?

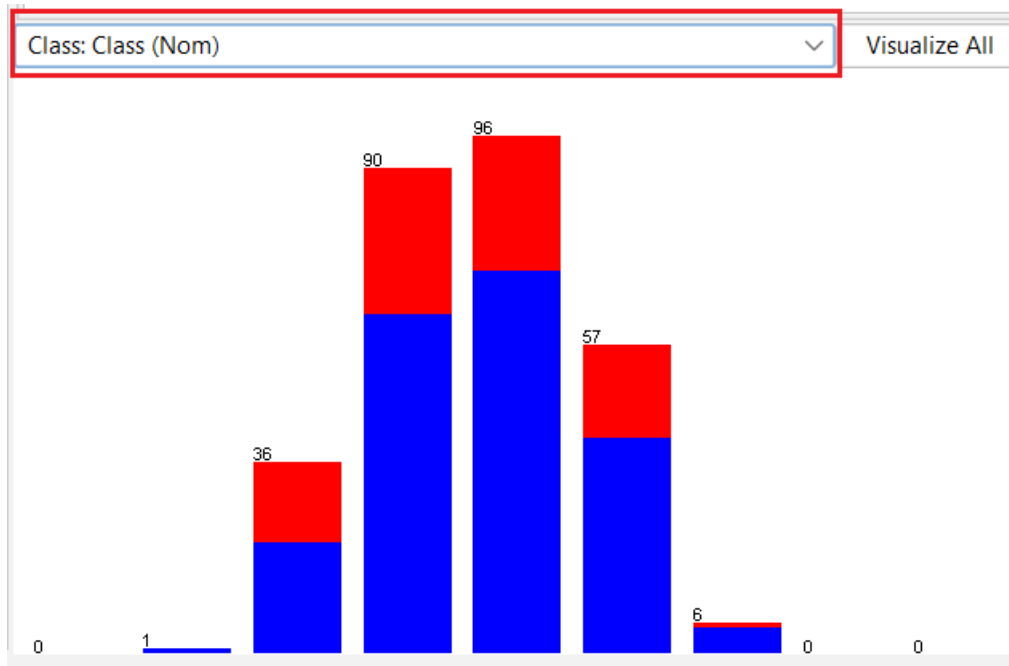
- Tập dữ liệu có 10 thuộc tính.

Filter	
Choose	None
Current relation	
Relation: breast-cancer	Attributes: 10
Instances: 286	Sum of weights: 286

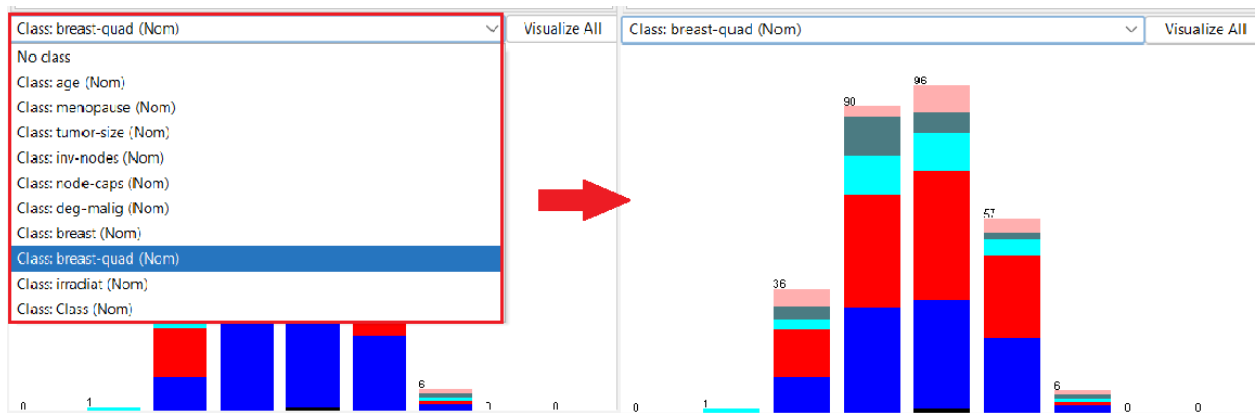
#### c) Which attribute is used for the label? Can it be changed? How?

- Thuộc tính mặc định làm nhãn (class) của tập dữ liệu này là "Class".





- Ta có thể thay đổi thuộc tính làm nhãn (class), bằng cách chọn trong tag class như hình dưới:



#### d) What is the meaning of each attribute?

STT	Thuộc tính	Ý nghĩa thuộc tính
1	age	Tuổi của bệnh nhân tại thời điểm chẩn đoán
2	menopause	Bệnh nhân đang trong giai đoạn tiền hoặc hậu mãn kinh tại thời điểm chẩn đoán
3	tumor-size	Đường kính lớn nhất (tính bằng mm) của khối u bị cắt bỏ
4	inv-nodes	Số lượng (phạm vi 0 - 39) của các hạch bạch huyết nách có chứa ung thư vú di căn có thể nhìn thấy khi kiểm tra mô học
5	node-caps	Khối u có di căn đến hạch bạch huyết khác hay không?
6	deg-malig	Cấp độ mô học (phạm vi 1-3) của khối u. Các khối u độ 1 chủ yếu bao gồm các tế bào, trong khi tân sinh, vẫn giữ được nhiều đặc điểm thông thường của chúng. Các khối u độ 3 chủ yếu bao gồm các tế bào rất bất thường

7	breast	Bệnh nhân bị ung thư ở vú bên trái hay bên phải?
8	breast-quad	Góc phần tư của vú (vú có thể được chia thành bốn góc phần tư, sử dụng núm vú làm điểm trung tâm)
9	irradiat	Bệnh nhân có xạ trị hay không? Xạ trị là một phương pháp điều trị sử dụng tia X năng lượng cao để phá hủy các tế bào ung thư.
10	class	Phân loại bệnh nhân có bị tái phát bệnh sau điều trị hay không?

**e) Let's investigate the missing value status in each attribute and describe in general ways to solve the problem of missing values.**

- Gồm có hai thuộc tính có chứa giá trị thiếu: node-caps và breast-quad
  - Node-caps: có 8 giá trị thiếu (chiếm ~ 3%)
  - Breast-quad: có 1 giá trị thiếu (chiếm ~ 0%)

Selected attribute			
Name: node-caps		Distinct: 2	Type: Nominal
Missing: 8 (3%)			Unique: 0 (0%)
No.	Label	Count	Weight
Selected attribute			
Name: breast-quad		Distinct: 5	Type: Nominal
Missing: 1 (0%)			Unique: 0 (0%)
No.	Label	Count	Weight

- Có 2 cách chính để giải quyết các giá trị thiếu:
  - Loại bỏ giá trị thiếu:
    - Xóa dòng có chứa giá trị thiếu
    - Xóa toàn bộ cột có chứa giá trị thiếu

Đây là một trong những kỹ thuật đơn giản và nhanh chóng mà người ta có thể sử dụng để xử lý các giá trị thiếu. Tuy nhiên phương pháp này không được khuyến khích vì nó sẽ làm giảm số lượng mẫu/thuộc tính, mất đi các mẫu/thuộc tính quan trọng,...

- Áp đặt giá trị cho giá trị thiếu: tùy vào đặc tính phân bố các giá trị của thuộc tính, có nhiều phương pháp khác nhau để thay thế các giá trị còn thiếu:
  - Thay thế bằng một giá trị tùy ý
  - Thay thế bằng giá trị trung bình
  - Thay thế bằng giá trị mode
  - Thay thế bằng trung vị
  - Thay thế bằng giá trị trước đó
  - Thay thế bằng giá trị sau đó
  - Nội suy

**f) Let's propose solutions to the problem of missing values in the specific attribute.**

- Thuộc tính "node-cap":

- Nhận xét: “node-cap” là thuộc tính phân loại (yes/no), giá trị “no” nhiều gấp 4 lần giá trị “yes”. Do đó ta nên điền giá trị mode là “no” (giá trị xuất hiện với tần số nhiều nhất) cho các giá trị thiếu.

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

- Thực hiện trên Weka:  
Ở tính năng Filter ta tiến hành: Choose -> filters -> unsupervised -> attribute -> ReplaceMissingValues -> Apply

Filter

Choose **ReplaceMissingValues**

Apply Stop

Current relation

Relation: breast-cancer-weka.filters.unsupervised.attribute.Repla...  
Instances: 286

Attributes: 10  
Sum of weights: 286

Attributes

All None Invert Pattern

No.	Name
1	<input type="checkbox"/> age
2	<input type="checkbox"/> menopause
3	<input type="checkbox"/> tumor-size
4	<input type="checkbox"/> inv-nodes
5	<input checked="" type="checkbox"/> node-caps
6	<input type="checkbox"/> deg-malig
7	<input type="checkbox"/> breast

Selected attribute

Name: node-caps

Missing: 0 (0%)

Distinct: 2

Type: Nominal

Unique: 0 (0%)

No.	Label	Count	Weight
1	yes	56	56
2	no	230	230

- Kết quả:

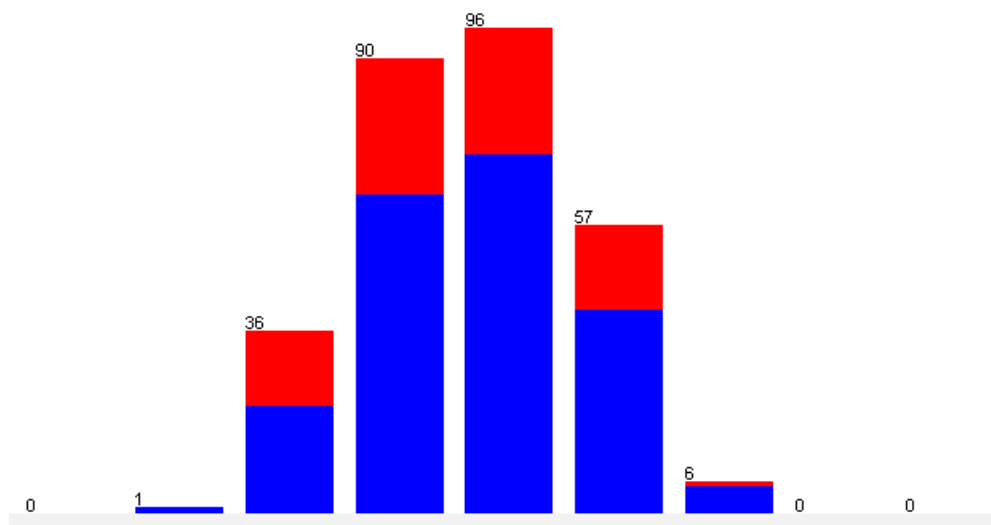
Selected attribute			
Name: node-caps		Type: Nominal	
Missing: 0 (0%)		Distinct: 2	
		Unique: 0 (0%)	
No.	Label	Count	Weight
1	yes	56	56
2	no	230	230

- Thuộc tính “breast-quad”: cách xử lý tương tự như thuộc tính “node-cap”, “breast-quad” là thuộc tính định danh, ta nên điền giá trị mode là “left\_low” cho các giá trị thiếu. Ta có kết quả sau:

Selected attribute			
Name: breast-quad Missing: 0 (0%)		Distinct: 5	Type: Nominal Unique: 0 (0%)
No.	Label	Count	Weight
1	left_up	97	97
2	left_low	111	111
3	right_up	33	33
4	right_low	24	24
5	central	21	21

**g) Let's explain the meaning of the chart in the WEKA Explorer. Setting the title for it and describing its legend**

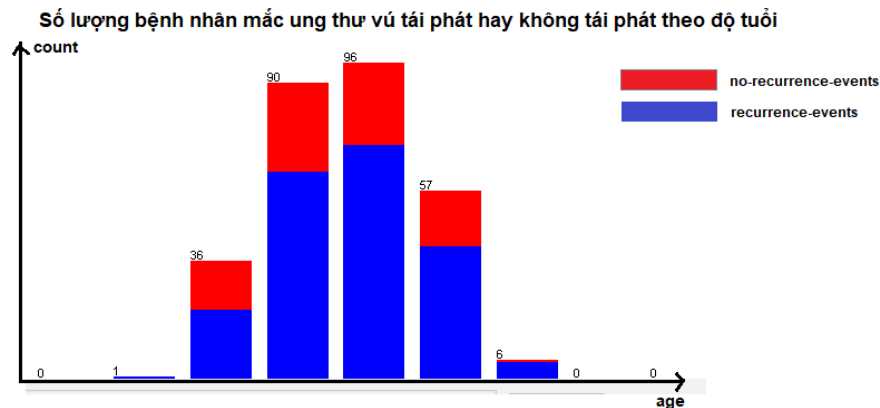
- Ý nghĩa biểu đồ trên màn hình WEKA Explorer: đây là biểu đồ thể hiện thuộc tính ta đang chọn, được tô màu theo thuộc tính chọn để phân loại là lớp (class) và chỉ có thuộc tính định danh mới có thể tô màu. Tô màu theo tỷ lệ/số lượng mà giá trị định danh đó chiếm trong thuộc tính.
  - Ví dụ: với thuộc tính chọn để xét là “age”, thuộc tính phân lớp là “class”. Ta có biểu đồ ở màn hình WEKA Explorer:



- Giải thích ý nghĩa:
 

Đồ thị biểu diễn số bệnh nhân bị mắc ung thư vú theo độ tuổi với phân loại là đã tái phát hay chưa tái phát khối u. Màu xanh dương trên biểu đồ đại diện cho số lượng các bệnh nhân không bị tái phát khối u (recurrence-events), ngược lại màu đỏ là các bệnh nhân bị tái phát (no-recurrence-events).
- Cài đặt tiêu đề và chú giải

- Tiêu đề (title): “Số lượng bệnh nhân mắc ung thư vú tái phát hay không tái phát theo độ tuổi”.
- Chú giải (legend):
  - + Màu xanh: không tái phát (no-recurrence-events).
  - + Màu đỏ: tái phát (recurrence-events).
- Hình vẽ minh họa:



## 2. Khám phá tập dữ liệu Weather

a) How many attributes does this data set have? How many samples? Which attributes have data type categorical? Which attributes have a data type that is numerical? Which attribute is used for the label?

- Tập dữ liệu này có 5 thuộc tính, 14 mẫu dữ liệu.

Filter
 

Choose
 None

Current relation
 

Relation: weather
 Instances: 14

Attributes: 5
 Sum of weights: 14

- Thuộc tính có kiểu dữ liệu phân loại: outlook, windy, play.

Viewer					
Relation: weather					
No.	1: outlook Nominal	2: temperature Numeric	3: humidity Numeric	4: windy Nominal	5: <b>play</b> Nominal
1	sunny	85.0	85.0	FALSE	no
2	sunny	80.0	90.0	TRUE	no
3	overcast	83.0	86.0	FALSE	yes
4	rainy	70.0	96.0	FALSE	yes
5	rainy	68.0	80.0	FALSE	yes
6	rainy	65.0	70.0	TRUE	no
7	overcast	64.0	65.0	TRUE	yes
8	sunny	72.0	95.0	FALSE	no
9	sunny	69.0	70.0	FALSE	yes
10	rainy	75.0	80.0	FALSE	yes
11	sunny	75.0	70.0	TRUE	yes
12	overcast	72.0	90.0	TRUE	yes
13	overcast	81.0	75.0	FALSE	yes
14	rainy	71.0	91.0	TRUE	no

- Thuộc tính có kiểu dữ liệu là số: temperature, humidity.

Viewer					
Relation: weather					
No.	1: outlook Nominal	2: temperature Numeric	3: humidity Numeric	4: windy Nominal	5: <b>play</b> Nominal
1	sunny	85.0	85.0	FALSE	no
2	sunny	80.0	90.0	TRUE	no
3	overcast	83.0	86.0	FALSE	yes
4	rainy	70.0	96.0	FALSE	yes
5	rainy	68.0	80.0	FALSE	yes
6	rainy	65.0	70.0	TRUE	no
7	overcast	64.0	65.0	TRUE	yes
8	sunny	72.0	95.0	FALSE	no
9	sunny	69.0	70.0	FALSE	yes
10	rainy	75.0	80.0	FALSE	yes
11	sunny	75.0	70.0	TRUE	yes
12	overcast	72.0	90.0	TRUE	yes
13	overcast	81.0	75.0	FALSE	yes
14	rainy	71.0	91.0	TRUE	no

- Thuộc tính được chọn làm nhãn (label/class) của tập dữ liệu này là “play” (có 2 giá trị yes/no, liệu có tổ chức chơi trong điều kiện thời tiết đó hay không).

Class: play (Nom)

Visualize All

b) Let's list five-number summary of two attributes temperature and humidity.  
Does WEKA provide these values?

- Định nghĩa: file-number summary là một tập thống kê mô tả thông tin cho một tập dữ liệu. Nó bao gồm các giá trị thống kê sau: minimum, maximum, median, Q1, Q3.
- file-number summary của thuộc tính **temperature** và **humidity**:

Five-number summary	Temperature	Humidity
Minimum	64	65
Lower quartile	69.25	71.25
Median	72	82.5
Upper quartile	78.75	90
Maximum	85	96

- Trong Weka đối với thuộc tính có kiểu dữ liệu dạng số (numeric), Weka liệt kê sẵn 2 giá trị trong file-number summary là giá trị lớn nhất (minimum), giá trị nhỏ nhất (maximum). Ngoài ra còn có giá trị trung bình (mean) và độ lệch chuẩn (StdDev).

○ Thuộc tính nhiệt độ “temperature”:

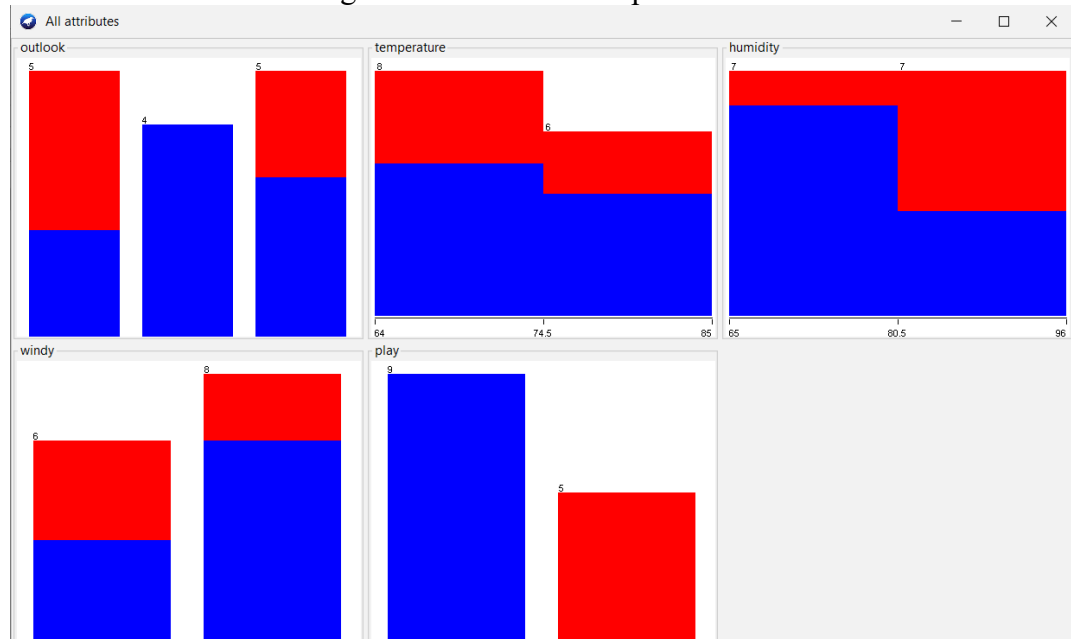
Selected attribute		
Name: temperature	Type: Numeric	
Missing: 0 (0%)	Distinct: 12	Unique: 10 (71%)
Statistic	Value	
Minimum	64	
Maximum	85	
Mean	73.571	
StdDev	6.572	

○ Thuộc tính độ ẩm “humidity”:

Selected attribute		
Name: humidity	Type: Numeric	
Missing: 0 (0%)	Distinct: 10	Unique: 7 (50%)
Statistic	Value	
Minimum	65	
Maximum	96	
Mean	81.643	
StdDev	10.285	

c) Let's explain the meaning of all charts in the WEKA Explorer. Setting the title for it and describing its legend.

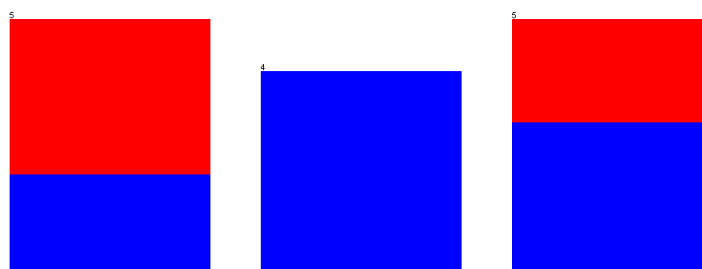
- Tất cả các biểu đồ có trong màn hình WEKA Explorer:



- Giải thích ý nghĩa:

- Gồm có 5 biểu đồ tương ứng với 5 thuộc tính của tập dữ liệu: outlook, temperature, humidity, windy, play. Các biểu đồ mô tả số lượng các giá trị trong thuộc tính đang xét theo thuộc tính phân lớp.
- Màu sắc mô tả cho tỷ lệ số lượng mà giá trị thuộc tính đó được phân lớp theo thuộc tính “play” (thuộc tính được chọn làm lớp). Màu xanh ứng với “play” có giá trị “yes”, màu đỏ ứng với “play” có giá trị “no”.
- Ví dụ: với thuộc tính “outlook” ta có biểu đồ sau:

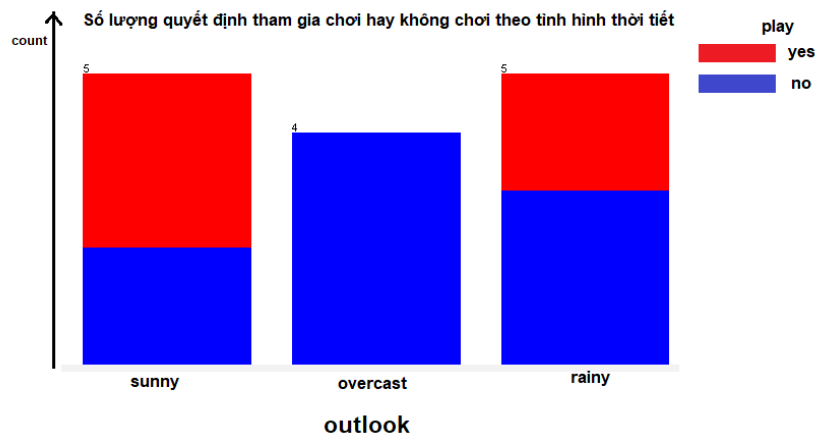
Selected attribute				
Name: outlook				
Missing: 0 (0%)				
Distinct: 3				
Type: Nominal				
Unique: 0 (0%)				
No.	Label	Count	Weight	
1	sunny	5	5	
2	overcast	4	4	
3	rainy	5	5	



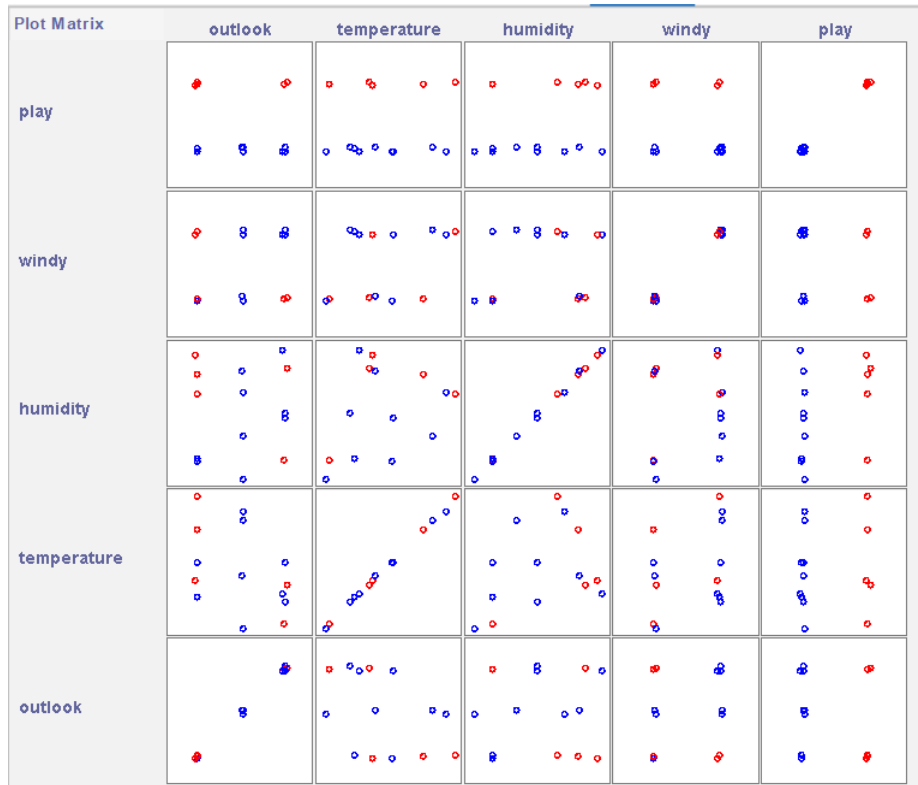


- + Thuộc tính “outlook” là thuộc tính định danh có 3 giá trị phân biệt: sunny, overcast, rainy.
- + Mỗi cột của đồ thị từ trái qua phải lần lượt đại diện cho các giá trị: sunny, overcast, rainy.
- + Độ cao các cột thể hiện cho số lượng giá trị đó có trong thuộc tính như: cột thứ nhất thể hiện có 5 giá trị “sunny”, cột thứ 2 thể hiện có 4 giá trị “overcast”, cột thứ 3 thể hiện có 5 giá trị “rainy”.
- + Màu sắc sẽ phân lớp cho giá trị. Ví dụ với cột “sunny”, màu xanh thể hiện có 2 giá trị được phân lớp theo “play” là “yes”, màu đỏ thể hiện có 3 giá trị có “play” là “no”.

- Cài đặt tiêu đề và chú giải: thực hiện với thuộc tính “outlook”.
  - Tiêu đề (title): “Số lượng quyết định tham gia chơi hay không chơi theo tình hình thời tiết”.
  - Chú giải (legend):
    - + Màu xanh: có tham gia chơi (yes).
    - + Màu đỏ: không tham gia chơi (no).
  - Hình vẽ minh họa:



- d) Let's move to the Visualize tag. What's the name of this chart? Do you think there are any pairs of different attributes that have correlated?
- Biểu đồ ở thẻ Visualize:



- Tên của biểu đồ: scatter-plot matrix (ma trận biểu đồ phân tán).
- Nhận xét sự tương quan giữa các thuộc tính: quan sát từ biểu đồ trên, dường như không có thuộc tính nào có sự tương quan với nhau. Vì biểu đồ phân tán của mỗi cặp thuộc tính khác nhau không có dạng hướng lên (positive correlation) hay hướng xuống (negative correlation).

### 3. Khám phá tập dữ liệu Credit in Germany

a) What is the content of the comments section in credit-g.arff (when opened with any text editor) about? How many samples does the data set have? How many attributes? Describe any five attributes (must have both discrete and continuous attributes).

- Hình ảnh mô tả bộ dữ liệu mở bằng Notepad:

```

credit-g - Notepad
File Edit Format View Help
% Description of the German credit dataset.
%
% 1. Title: German Credit data
%
% 2. Source Information
%
% Professor Dr. Hans Hofmann
% Institut f"ur Statistik und "Okonometrie
% Universit"at Hamburg
% FB Wirtschaftswissenschaften
% Von-Melle-Park 5
% 2000 Hamburg 13
%
% 3. Number of Instances: 1000
%
% Two datasets are provided. the original dataset, in the form provided
% by Prof. Hofmann, contains categorical/symbolic attributes and
% is in the file "german.data".
%
% For algorithms that need numerical attributes, Strathclyde University
% produced the file "german.data-numeric". This file has been edited
% and several indicator variables added to make it suitable for
% algorithms which cannot cope with categorical variables. Several
% attributes that are ordered categorical (such as attribute 17) have
% been coded as integer. This was the form used by StatLog.
%
%
% 6. Number of Attributes german: 20 (7 numerical, 13 categorical)
%    Number of Attributes german.numeric: 24 (24 numerical)
%
%
% 7. Attribute description for german
%
% Attribute 1: (qualitative)
%              Status of existing checking account
%              A11 :      ... <  0 DM
%              A12 : 0 <= ... < 200 DM
%              A13 :      ... >= 200 DM /
%                   salary assignments for at least 1 year
%              A14 : no checking account
%
% Attribute 2: (numerical)

```

- Nội dung của ghi chú là một bản mô tả tóm tắt về tập dữ liệu, bao gồm các thông tin như: (1) tiêu đề, (2) nguồn thông tin, (3) số lượng mẫu, (6) số lượng các thuộc tính, (7) mô tả các thuộc tính. Một số thông tin cơ bản của tập dữ liệu:
  - Có hai bộ dữ liệu được cung cấp, bản gốc do giáo sư Hofmann cung cấp, chứa các thuộc tính phân loại/biểu tượng nằm trong tệp german.data.
  - Đối với thuật toán cần thuộc tính số Đại học Strathclyde tạo ra tệp german.data-numeric. Tệp này đã được chỉnh sửa và thêm một số biến chỉ báo để phù hợp với các thuật toán không thể sử dụng các biến phân loại. Một số thuộc tính được sắp xếp theo thứ tự phân loại (chẳng hạn như thuộc tính 17) đã được mã hóa thành số nguyên. Đây là hình thức được StatLog sử dụng.
- Tập dữ liệu có 1000 mẫu và 21 thuộc tính.

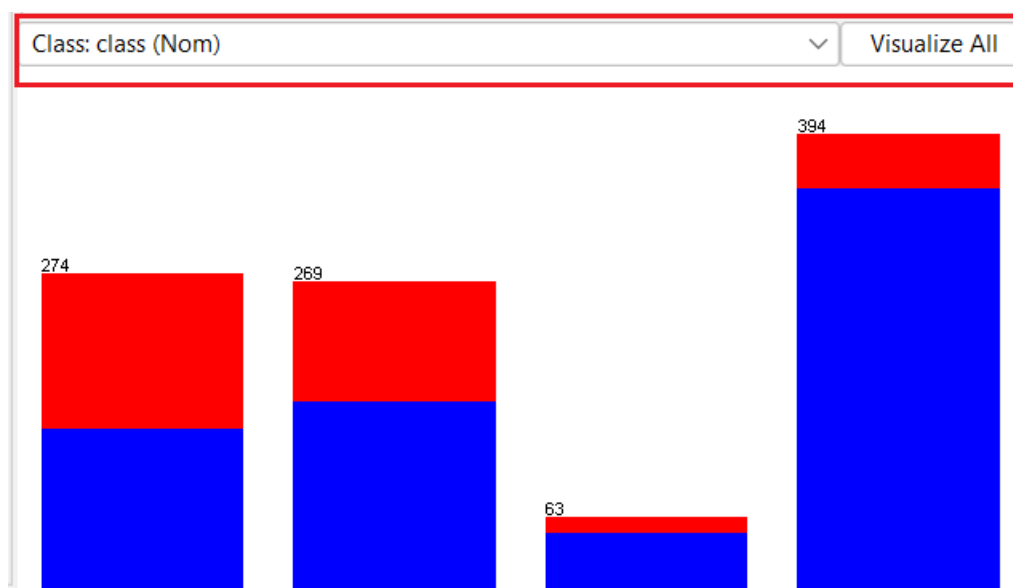
Current relation	
Relation: german_credit	Attributes: 21
Instances: 1000	Sum of weights: 1000

- Mô tả 5 thuộc tính bất kì:

STT	Tên thuộc tính	Loại thuộc tính	Kiểu dữ liệu	Ý nghĩa
1	age	Liên tục	numeric	Thông tin tuổi của khách hàng
2	job	Rời rạc	nominal	Công việc của chủ tài khoản
3	purpose	Rời rạc	nominal	Mục đích của việc vay tín dụng
4	credit_amount	Liên tục	numeric	Số dư trong thẻ tín dụng
5	persional_status	Rời rạc	nominal	Giới tính và trạng thái hiện tại của một người

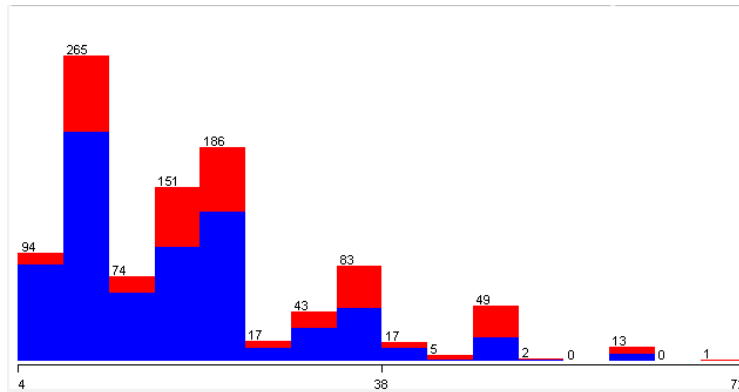
**b) Which attribute is used for the label?**

Tên thuộc tính lớp: class (bao gồm 2 giá trị là good và bad). Cân bằng lệch về phía good.

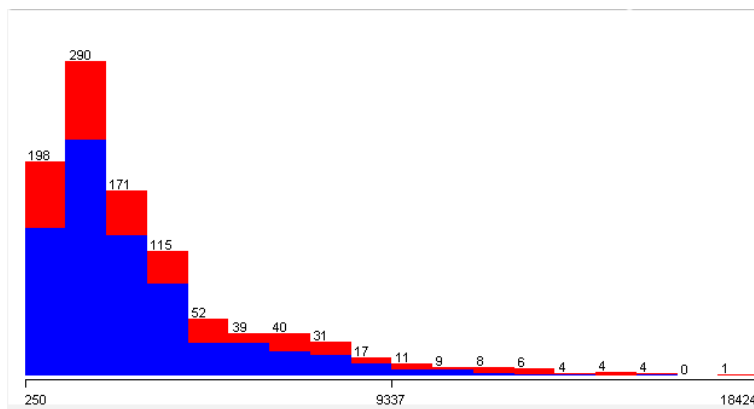


**c) Let's describe the distribution of continuous attributes? (Left skewed or right skewed?).**

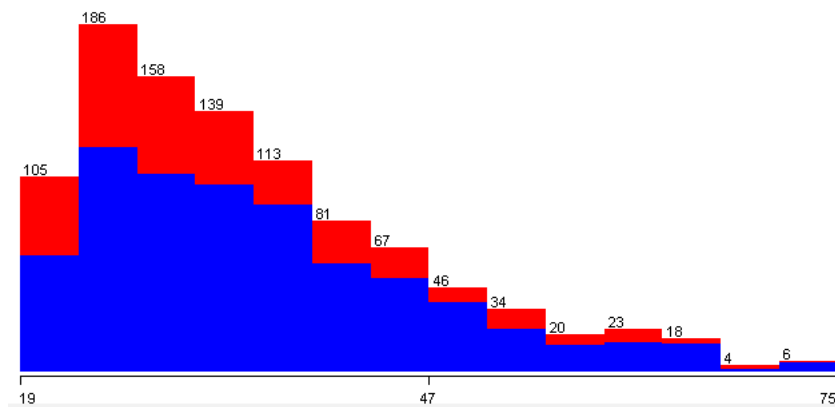
- Có 3 thuộc tính chứa kiểu dữ liệu liên tục: duration, credit\_amount, age
  - Thuộc tính “duration”: dữ liệu phân bố lệch trái



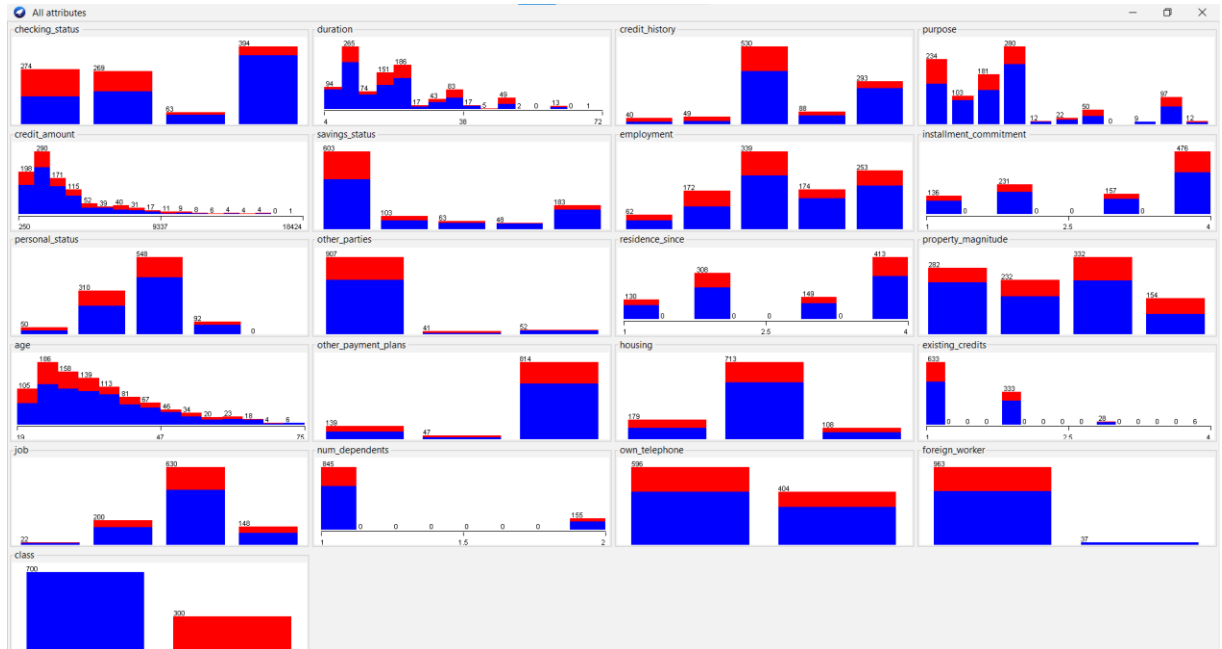
- Thuộc tính “credit\_amount”: dữ liệu phân bố lệch trái



- Thuộc tính “age”: dữ liệu phân bố lệch trái



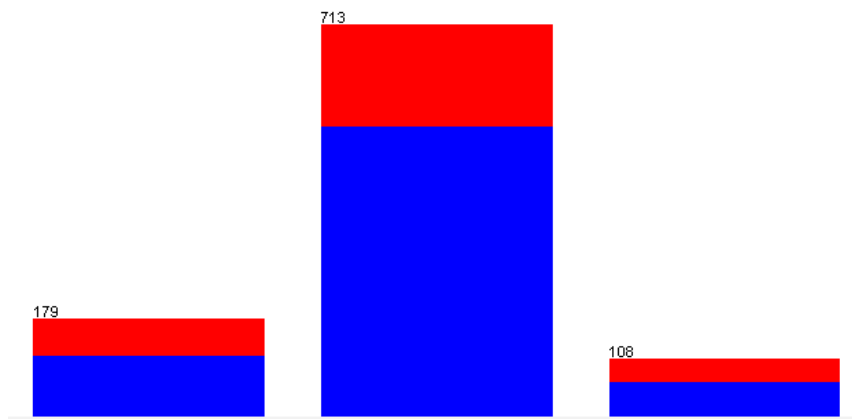
- d) Let's explain the meaning of all charts in the WEKA Explorer. Setting the title for it and describing its legend.
- Tất cả các biểu đồ có trong màn hình WEKA Explorer:



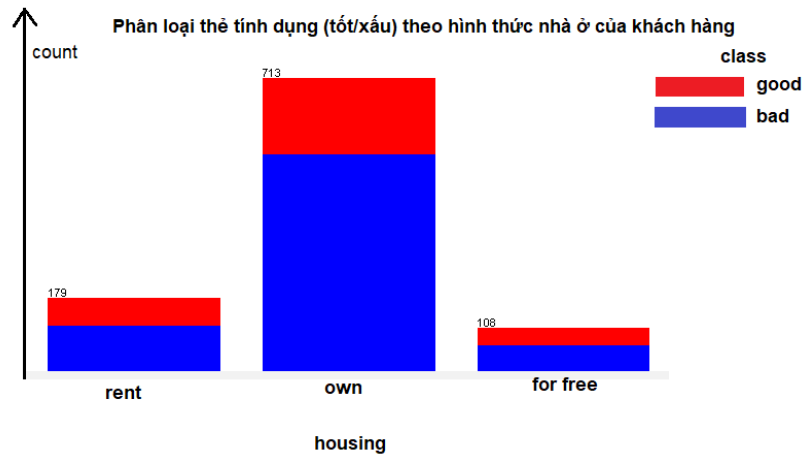
- Giải thích ý nghĩa:
  - Gồm có 21 biểu đồ tương ứng với 21 thuộc tính của tập dữ liệu. Các biểu đồ mô tả số lượng các giá trị trong thuộc tính đang xét theo thuộc tính phân lớp là 'class'.
  - Màu sắc mô tả cho tỷ lệ số lượng mà giá trị thuộc tính đó được phân lớp theo thuộc tính "class" (thuộc tính được chọn làm lớp). Màu xanh ứng với "class" có giá trị "good", màu đỏ ứng với "class" có giá trị "bad".
  - Ví dụ:  
Với thuộc tính "housing" ta có biểu đồ sau:

Selected attribute			
Name: housing		Type: Nominal	
Missing: 0 (0%)		Distinct: 3	
		Unique: 0 (0%)	
No.	Label	Count	Weight
1	rent	179	179
2	own	713	713
3	for free	108	108

Class: class (Nom) Visualize All

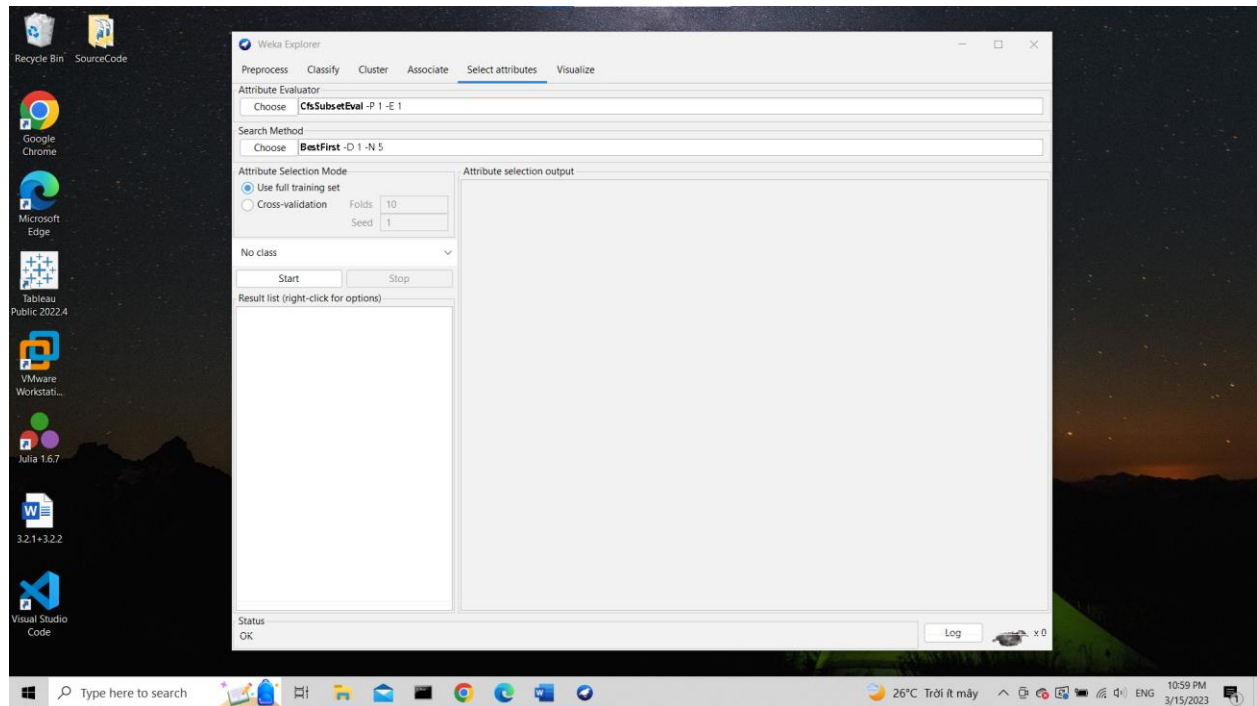


- + Thuộc tính “housing” là thuộc tính định danh có 3 giá trị phân biệt: rent, own, for free.
- + Mỗi cột của đồ thị từ trái qua phải lần lượt đại diện cho các giá trị: rent, own, for free.
- + Độ cao các cột thể hiện cho số lượng giá trị đó có trong thuộc tính như: cột thứ nhất thể hiện có 179 giá trị “rent”, cột thứ 2 thể hiện có 713 giá trị “own”, cột thứ 3 thể hiện có 108 giá trị “for rent”.
- + Màu sắc sẽ phân lớp cho giá trị. Ví dụ với cột “rent”, màu xanh thể hiện giá trị được phân lớp theo “class” là “good”, màu đỏ thể hiện giá trị có “class” là “bad”.
- Cài đặt tiêu đề và chú giải: thực hiện với thuộc tính “housing”
  - Tiêu đề (title): “Phân loại thể tính dụng (tốt/xấu) theo hình thức nhà ở của khách hàng”.
  - Chú giải (legend):
    - + Màu xanh: tín dụng tốt (good).
    - + Màu đỏ: tín dụng xấu (bad).
  - Hình vẽ minh họa:



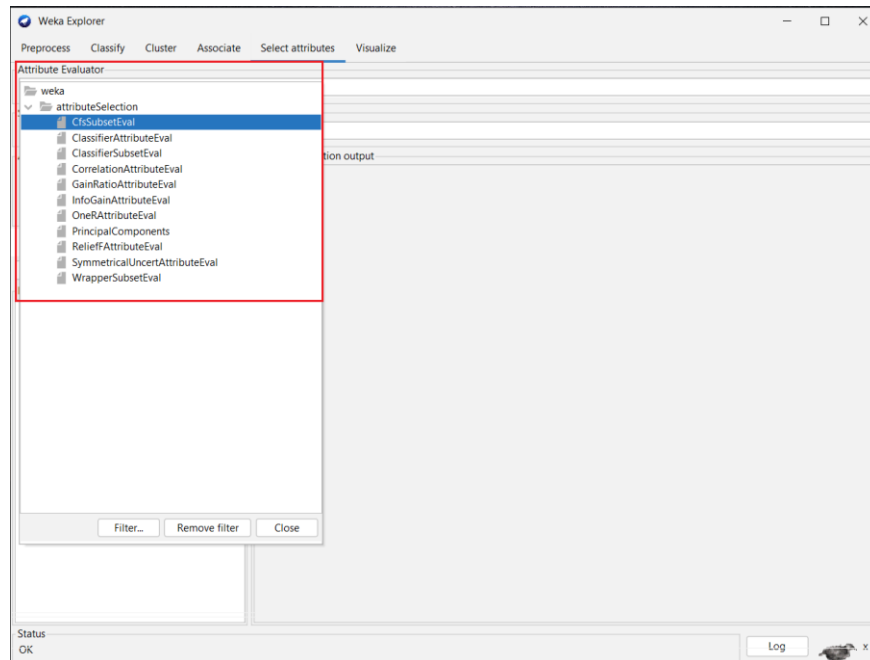
e) Let's move to the Select attributes tag. Describe all of the options for attribute selection

- Hình ảnh ở Select attributes tag:



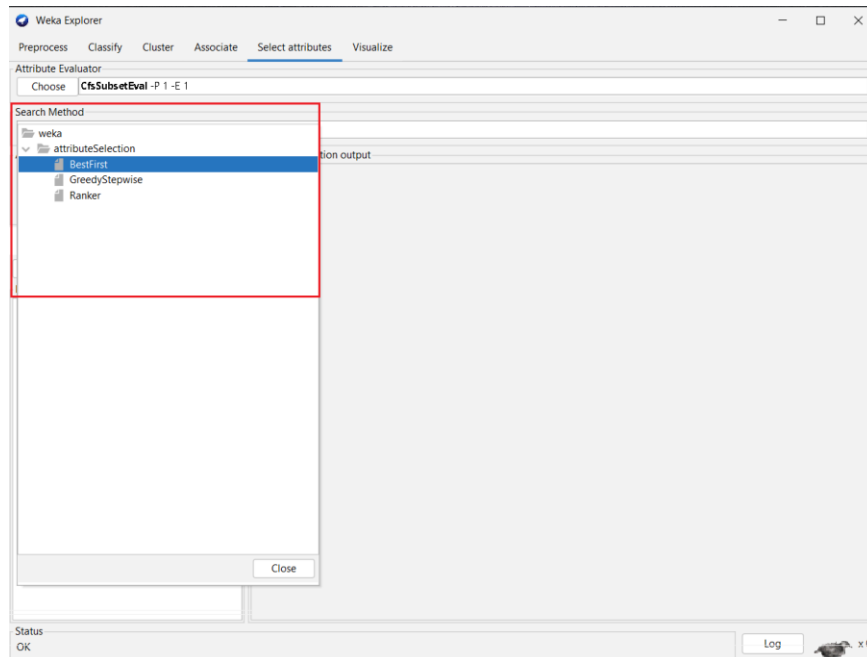
- Bộ đánh giá thuộc tính (Attribute Evaluator): để đánh giá tập các thuộc tính của tập dữ liệu. WEKA cung cấp 11 phương pháp đánh giá thuộc tính.





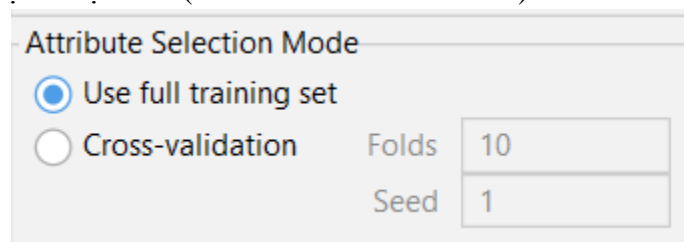
Tên bộ đánh giá	Mô tả
CfsSubsetEval	Đánh giá tập thuộc tính bằng cách xem xét khả năng dự đoán của từng thuộc tính riêng lẻ và mức độ dư thừa giữa chúng
ClassifierSubsetEval	Đánh giá tập thuộc tính con trong tập huấn luyện hoặc tập kiểm tra riêng biệt
ClassifierAttributeEval	Đánh giá thuộc tính bằng cách sử dụng bộ phân lớp do người dùng chọn
CorrelationAttributeEval	Đánh giá một thuộc tính dựa trên sự tương quan với lớp
GainRatioAttributeEval	Đánh giá một thuộc tính dựa trên tỷ lệ gia tăng
InfoGainAttributeEval	Đánh giá một thuộc tính dựa trên thông tin thu được
OneRAttributeEval	Đánh giá một thuộc tính bằng cách sử dụng bộ phân loại OneR
PrincipalComponents	Thực hiện phân tích thành phần chính và chuyển đổi dữ liệu
ReliefFAttributeEval	Đánh giá thuộc tính dựa trên các thể hiện
SymmetricalUncertAttributeEval	Đánh giá một thuộc tính dựa trên bất đối xứng
WrapperSubsetEval	Đánh giá tập thuộc tính dựa trên một bộ phân loại cùng với xác nhận chéo

- Phương pháp tìm kiếm (Search Method): để xác định phương pháp tìm kiếm được thực hiện. WEKA cung cấp 3 phương thức tìm kiếm, gồm: BestFirst, GreedyStepwise, Ranker.

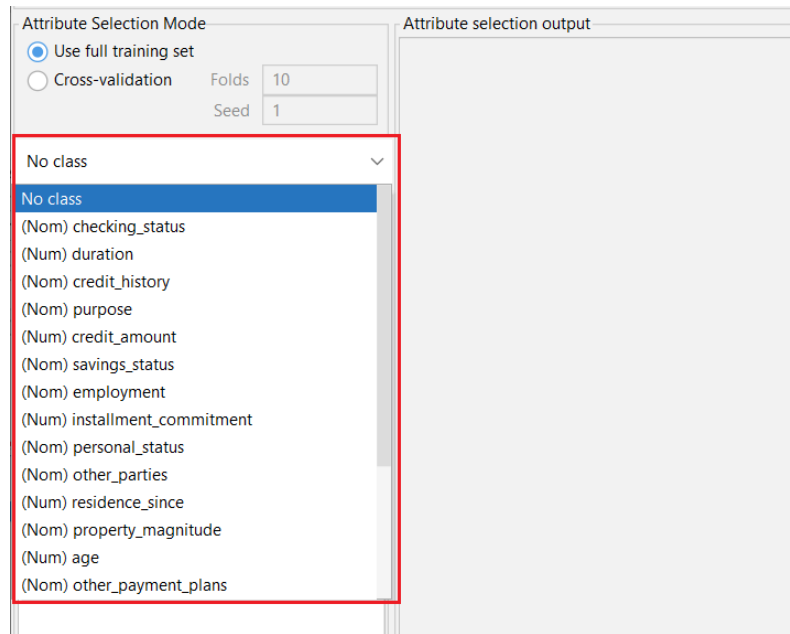


Phương pháp	Mô tả
BestFirst	Thực hiện leo đồi tham lam (greedy hill climbing) với quay lui (backtracking). Nó có thể tìm kiếm tiến (forward) từ một tập attribute rỗng, lui (backward) từ tập chứa toàn bộ attribute hoặc có thể bắt đầu từ một trạng thái cụ thể nào đó và tìm kiếm theo 2 hướng.
GreedyStepwise	Tìm kiếm tham lam trong không gian các tập attribute. Nó cũng có thể tìm kiếm tới và lui. Tuy nhiên, nó không sử dụng quay lui mà dừng lại ngay khi thêm hoặc xóa đi thuộc tính tốt nhất còn lại mà làm giảm số liệu đánh giá.
Ranker	Phương pháp này không chỉ xếp hạng các thuộc tính (attributes) mà còn thực hiện chọn các thuộc tính bằng cách loại bỏ những thuộc tính xếp hạng thấp.

- Chế độ lựa chọn thuộc tính (Attribute Selection Mode):

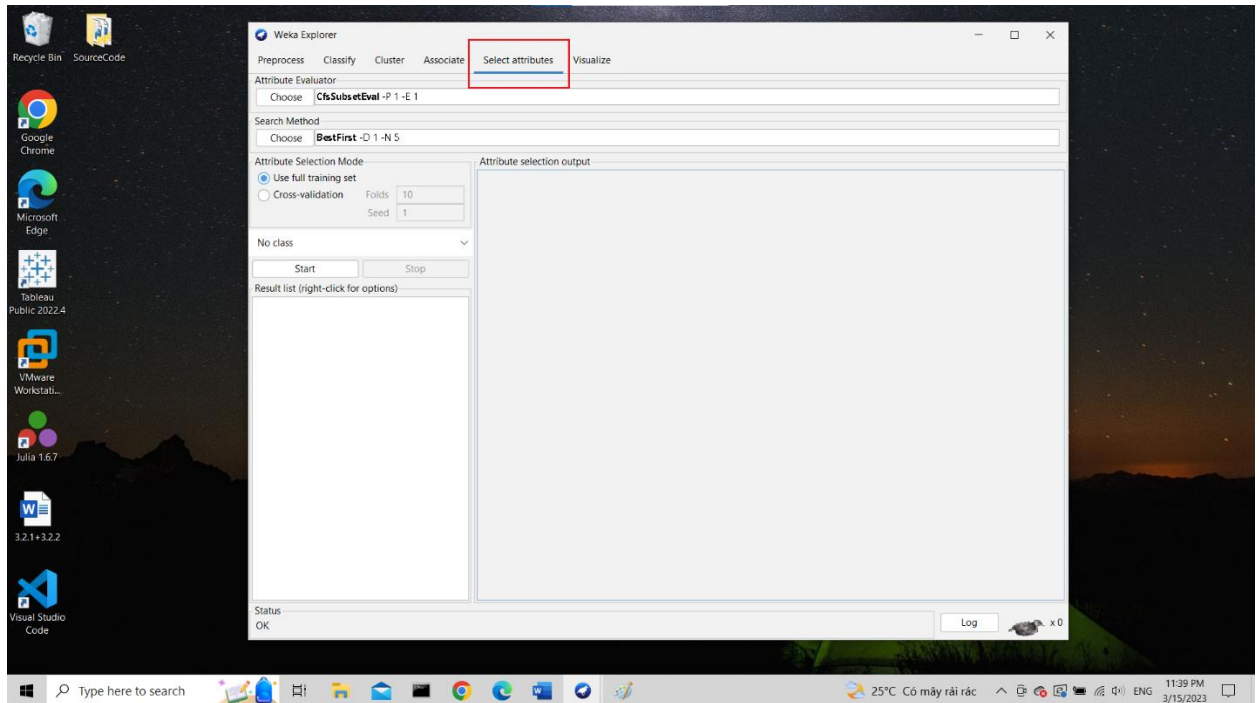


- Lựa chọn thuộc tính dự đoán/phân loại:

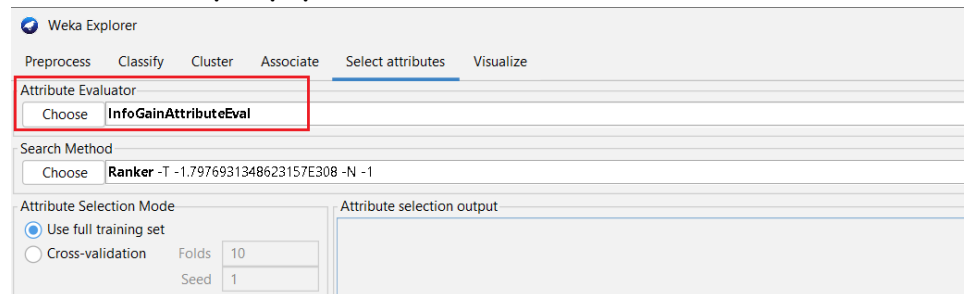


f) Which options should be used to select the 5 attributes with the highest correlation?(Step-by-step description, with step-by-step photos and final results).

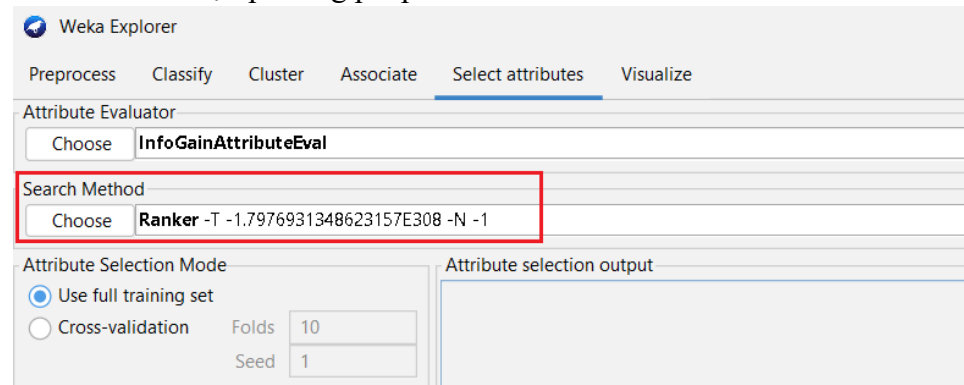
- Cần sử dụng bộ lọc (Attribute Evaluator) là **InfoGainAttributeEval** kết hợp với Search Method là **Ranker** để chọn ra 5 thuộc tính có tương quan cao nhất với thuộc tính lớp vì bộ lọc **InfoGainAttributeEval** đánh giá giá trị của một thuộc tính bằng cách đo lường thu được thông tin liên quan đến lớp kết hợp với **Ranker** sẽ xếp hạng độ từ cao đến thấp độ tương quan của các thuộc tính với thuộc tính lớp.
- Các bước tiến hành:
  - Bước 1: Mở tag **Select attributes**



○ Bước 2: Chọn bộ lọc **InfoGainAttributeEval** ở **Attribute Evaluator**



○ Bước 3: Chọn phương pháp **Ranker** ở **Search Method**



○ Bước 4: Chọn thuộc tính làm lớp là **“class”**

Attribute Selection Mode

☒ Use full training set

☐ Cross-validation

Folds: 10

Seed: 1

(Nom) class

(Nom) employment

(Num) installment\_commitment

(Nom) personal\_status

(Nom) other\_parties

(Num) residence\_since

(Nom) property\_magnitude

(Num) age

(Nom) other\_payment\_plans

(Nom) housing

(Num) existing\_credits

(Nom) job

(Num) num\_dependents

(Nom) own\_telephone

(Nom) foreign\_worker

(Nom) class

Attribute selection output

- Bước 5: nhấn nút **Start** để xem kết quả

Attribute Selection Mode

☒ Use full training set

☐ Cross-validation

Folds: 10

Seed: 1

(Nom) class

Start

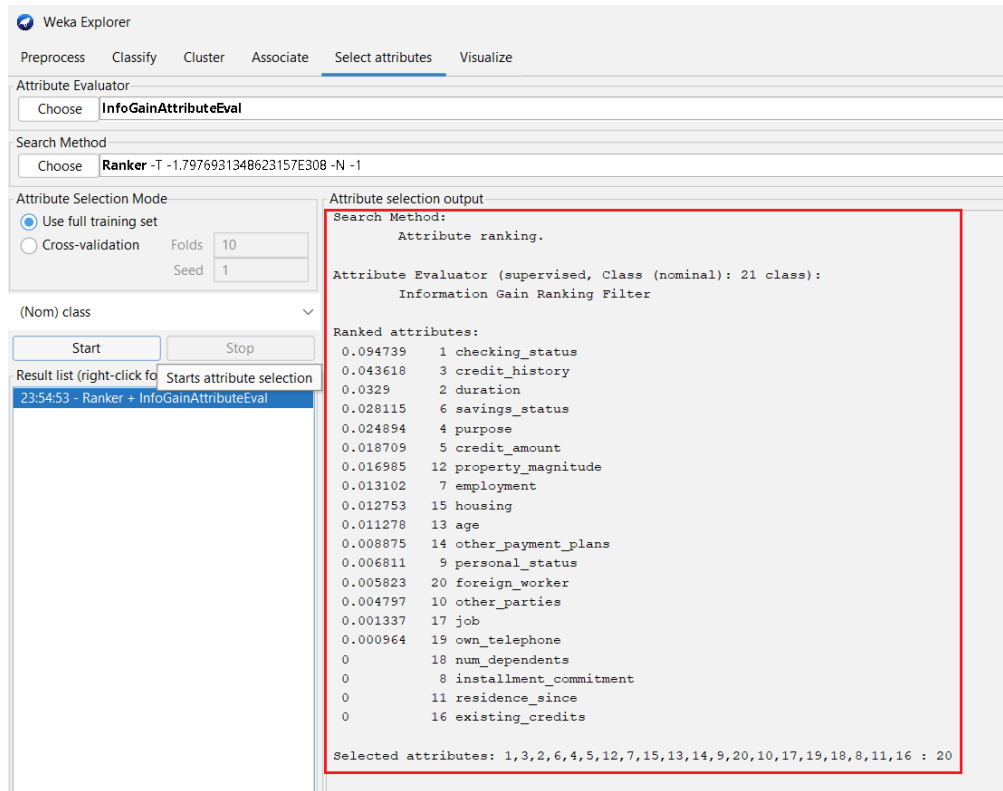
Stop

Result list (right-click for options)

23:54:53 - Ranker

Starts attribute selection

- Bước 7: xem kết quả ở khung **Attribute selection output**



- Như vậy bằng phương pháp trên, ta có 5 thuộc tính có mối tương quan cao nhất đối với thuộc tính lớp “class” theo thứ tự là: checking\_status, credit\_history, duration, savings\_status, purpose.

```
Ranked attributes:
0.094739    1 checking_status
0.043618    3 credit_history
0.0329      2 duration
0.028115    6 savings_status
0.024894    4 purpose
```

## IV. Tiền xử lý dữ liệu trong Python

### 0. Đặc tả chung

- Chương trình hoạt động theo cơ chế console và các yêu cầu người dùng được đặc tả thông qua tham số dòng lệnh.
- Một số quy định chung về tham số dòng lệnh của cả chương trình:
  - Tham số thứ nhất là tên file thực thi, mặc định là main.py
  - Tham số thứ hai là tên file dữ liệu cần xử lý
  - Tham số thứ ba là tên chức năng tiền xử lý, bao gồm các chức năng:
    - + ListMissingValue
    - + CountMissingRow
    - + ImputeMissingValue

- + DeleteMissingColumn
- + DeleteMissingRow
- + DeleteDuplicateInstance
- + StandardizeData
- + CalculateAttribute
- Đối với các chức năng có xuất file, thì tham số cuối cùng là tên file ở định dạng .csv

## 1. Extract columns with missing values

### • Cú pháp

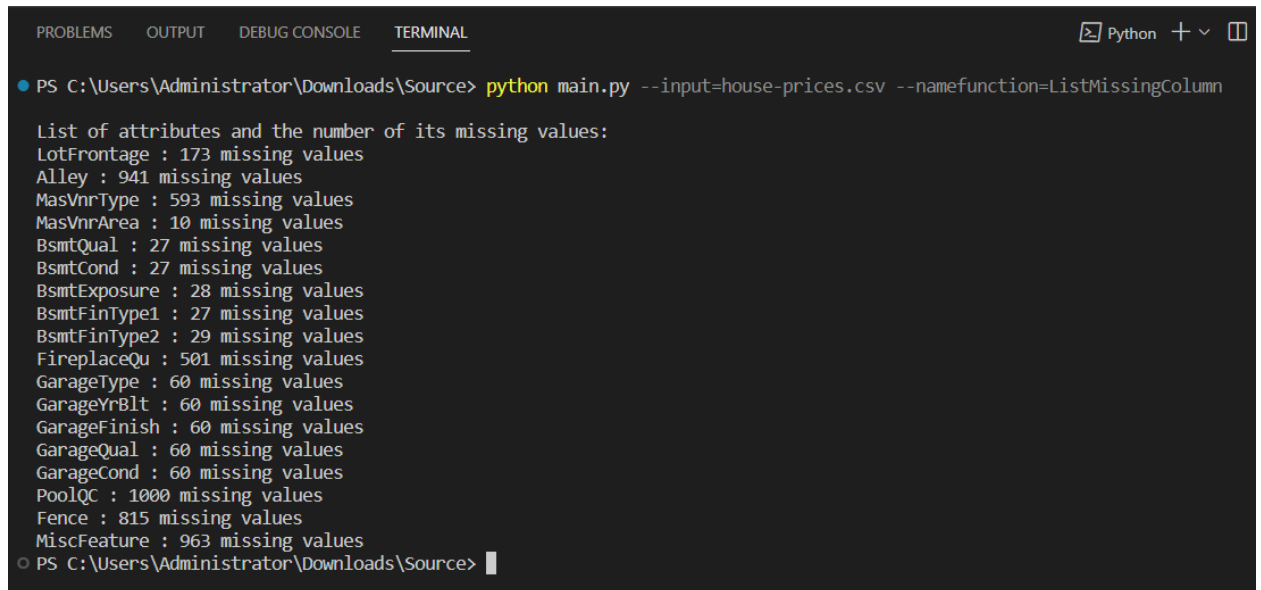
Argument syntax:

```
python main.py --input=input.csv --namefunction=A
```

Example:

```
python main.py --input=house-prices.csv --namefunction=ListMissingColumn
```

### • Kết quả



```

PROBLEMS  OUTPUT  DEBUG CONSOLE  TERMINAL
Python + v □

PS C:\Users\Administrator\Downloads\Source> python main.py --input=house-prices.csv --namefunction=ListMissingColumn

List of attributes and the number of its missing values:
LotFrontage : 173 missing values
Alley : 941 missing values
MasVnrType : 593 missing values
MasVnrArea : 10 missing values
BsmtQual : 27 missing values
BsmtCond : 27 missing values
BsmtExposure : 28 missing values
BsmtFinType1 : 27 missing values
BsmtFinType2 : 29 missing values
FireplaceQu : 501 missing values
GarageType : 60 missing values
GarageYrBlt : 60 missing values
GarageFinish : 60 missing values
GarageQual : 60 missing values
GarageCond : 60 missing values
PoolQC : 1000 missing values
Fence : 815 missing values
MiscFeature : 963 missing values
PS C:\Users\Administrator\Downloads\Source>

```

## 2. Count the number of lines with missing data

### • Cú pháp

Argument syntax:

```
python main.py --input=input.csv --namefunction=A
```

Example:

```
python main.py --input=house-prices.csv --namefunction=CountMissingRow
```

### • Kết quả

```
PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL Python + v
PS C:\Users\Administrator\Downloads\Source> python main.py --input=house-prices.csv --namefunction=CountMissingRow
Number of rows with missing data: 1000
PS C:\Users\Administrator\Downloads\Source> |
```

### 3. Fill in the missing value using mean, median (for numeric properties) and mode (for the categorical attribute)

Áp dụng được cho điền giá trị thiếu toàn bộ các thuộc tính (kể cả dạng numeric và categorical).

- **Cú pháp**

```
Argument syntax:
python main.py --input=input.csv --namefunction=A --method=B --output=result.csv
Example:
python main.py --input=house-prices.csv --namefunction=ImputeMissingValue --method=MEAN --output=result.csv
Note: tên method phải viết hoa hoặc viết thường toàn bộ
```

- **Kết quả**

- Để kiểm nghiệm kết quả, ta sẽ liệt kê các thuộc tính có chứa giá trị thiếu của file “result.csv” có được sau khi thực hiện chức năng “ImputeMissingValue”.

```
PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL Python + v
PS C:\Users\Administrator\Downloads\Source> python main.py --input=house-prices.csv --namefunction=ImputeMissingValue --method=MEAN --output=result.csv
PS C:\Users\Administrator\Downloads\Source> python main.py --input=result.csv --namefunction=ListMissingColumn
List of attributes and the number of its missing values:
PoolQC : 1000 missing values
PS C:\Users\Administrator\Downloads\Source> |
```

- Nhận xét: file “result.csv” chỉ còn mỗi thuộc tính “PoolQC” là có 1000 giá trị thiếu. Nguyên nhân là do ở dữ liệu gốc “house-prices.csv” thuộc tính này có tỉ lệ missing 100%, do đó không có cơ sở nào để điền giá trị thiếu cho thuộc tính này.

### 4. Deleting rows containing more than a particular number of missing values (Example: delete rows with the number of missing values is more than 50% of the number of attributes)

- **Cú pháp**

```
Argument syntax:
python main.py --input=input.csv --namefunction=A --rate=B --output=output.csv
Example:
python main.py --input=house-prices.csv --namefunction=DeleteMissingRow --rate=10 output=result.csv
```

- **Kết quả**

- Để kiểm nghiệm kết quả, ta sẽ đếm số dòng có chứa giá trị thiếu của file “result.csv” có được sau khi thực hiện chức năng “DeleteMissingRow”.

```
PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL Python + v
PS C:\Users\Administrator\Downloads\Source> python main.py --input=house-prices.csv --namefunction=DeleteMissingRow --rate=10 output=result.csv
PS C:\Users\Administrator\Downloads\Source> python main.py --input=result.csv --namefunction=CountMissingRow
Number of rows with missing data: 920
PS C:\Users\Administrator\Downloads\Source> |
```



- Nhận xét: số dòng chứa giá trị thiếu của file “result.csv” là 920, đã giảm so với file dữ liệu gốc “house-prices.csv” (chứa 1000 dòng có dữ liệu thiếu).

## 5. Deleting columns containing more than a particular number of missing values (Example: delete columns with the number of missing values is more than 50% of the number of samples)

### • Cú pháp

```
Argument syntax:
python main.py --input=input.csv --namefunction=A --rate=B --output=output.csv
Example:
python main.py --input=house-prices.csv --namefunction=DeleteMissingColumn --rate=10 --output=result.csv
```

### • Kết quả

- Để kiểm nghiệm kết quả, ta sẽ liệt kê các thuộc tính có chứa giá trị thiếu của file “result.csv” có được sau khi thực hiện chức năng “DeleteMissingColumn”.

```
PROBLEMS  OUTPUT  DEBUG CONSOLE  TERMINAL
PS C:\Users\Administrator\Downloads\Source> python main.py --input=house-prices.csv --namefunction=DeleteMissingColumn --rate=10 --output=result.csv
PS C:\Users\Administrator\Downloads\Source> python main.py --input=result.csv --namefunction=ListMissingColumn

List of attributes and the number of its missing values:
MasVnrArea : 10 missing values
BsmtQual : 27 missing values
BsmtCond : 27 missing values
BsmtExposure : 28 missing values
BsmtFinType1 : 27 missing values
BsmtFinType2 : 29 missing values
GarageType : 60 missing values
GarageYrBlt : 60 missing values
GarageFinish : 60 missing values
GarageQual : 60 missing values
GarageCond : 60 missing values
PS C:\Users\Administrator\Downloads\Source> █
```

- Nhận xét: các thuộc tính và số lượng giá trị thiếu của nó ở file “result.csv” đã giảm rất nhiều so với file dữ liệu gốc “house-prices.csv” (xem hình ảnh ở chức năng 1).

## 6. Delete duplicate samples

### • Cú pháp

```
Argument syntax:
python main.py --input=input.csv --namefunction=A --output=output.csv
Example:
python main.py --input=house-prices.csv --namefunction=DeleteDuplicateInstance --output=result.csv
```

### • Kết quả

- Để kiểm nghiệm kết quả, ta sẽ đếm số dòng có chứa giá trị thiếu của file “result.csv” có được sau khi thực hiện chức năng “DeleteDuplicateInstance”.

```
PROBLEMS  OUTPUT  DEBUG CONSOLE  TERMINAL
PS C:\Users\Administrator\Downloads\Source> python main.py --input=house-prices.csv --namefunction=DeleteDuplicateInstance --output=result.csv
PS C:\Users\Administrator\Downloads\Source> python main.py --input=result.csv --namefunction=CountMissingRow

Number of rows with missing data: 716
PS C:\Users\Administrator\Downloads\Source> █
```

- Nhận xét: số dòng chứa giá trị thiếu của file “result.csv” là 716, đã giảm so với file dữ liệu gốc “house-prices.csv” (chứa 1000 dòng có dữ liệu thiếu).

## 7. Normalize a numeric attribute using min-max and Z-score methods

- **Cú pháp**

```
Argument syntax:
python main.py --input=input.csv --namefunction=A --method=B --columns: C D F --output=result.csv
Example:
python main.py --input=house-prices.csv --namefunction=StandardizedData --method=MINMAX --columns: ID alley --output=result.csv
Note: + nếu cột yêu cầu chuẩn hóa không phải dạng số (numeric) sẽ có thông báo ko chuẩn hóa được cho thuộc tính đó
      + tên các cột yêu cầu chuẩn hóa phải cách nhau 1 dấu cách
      + tên method phải viết hoa hoặc viết thường toàn bộ
```

- **Kết quả (chuẩn hóa MINMAX)**

```
PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL
PS C:\Users\Administrator\Downloads\Source> python main.py --input=house-prices.csv --namefunction=StandardizedData --method=MINMAX --columns: ID alley --output=result.csv
Alley isn't numeric, can't standardize!
PS C:\Users\Administrator\Downloads\Source>
```

	Id	MSSubClass	MSZoning
0	0.85	20	RL
1	0.844	90	RL
2	0.96	50	RM
3	0.943	30	RL
4	0.141	20	RL
5	0.953	90	RL
6	0.671	20	RL
7	0.331	120	RM
8	0.267	60	RL
9	0.499	30	RM

- Giá trị cột “Id” đã được chuẩn hóa min-max trong phạm vi [0:1].
- Cột “Alley” có kiểu dữ liệu nominal do đó xuất hiện thông báo không thể chuẩn hóa.

- **Kết quả (chuẩn hóa Zscore)**

```
PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL
PS C:\Users\Administrator\Downloads\Source> python main.py --input=house-prices.csv --namefunction=StandardizedData --method=ZSCORE --columns: ID alley --output=result.csv
Alley isn't numeric, can't standardize!
PS C:\Users\Administrator\Downloads\Source>
```

	Id	MSSubClass	MSZoning	LotFrontage
0	1.196	20	RL	83
1	1.174	90	RL	70
2	1.583	50	RM	50
3	1.525	30	RL	52
4	-1.321	20	RL	
5	1.562	90	RL	65
6	0.558	20	RL	80
7	-0.649	120	RM	32
8	-0.873	60	RL	71


- Giá trị cột “Id” đã được chuẩn hóa z-score.
- Cột “Alley” có kiểu dữ liệu nominal do đó xuất hiện thông báo không thể chuẩn hóa.

## 8. Performing addition, subtraction, multiplication, and division between two numerical attributes

- **Cú pháp**

```
Argument syntax:
python main.py --input=house-prices.csv --namefunction=A --expression=B --output=result.csv
Example:
python main.py --input=house-prices.csv --namefunction=CalculateAttribute --expression=Id+LotFrontage --output=result.csv
Note: thuộc tính chứa kết quả phép tính sẽ có tên giống expression
```

- **Kết quả**



The screenshot shows a terminal window with the command: `python main.py --input=house-prices.csv --namefunction=CalculateAttribute --expression=Id+LotFrontage --output=result.csv`. Below the terminal, a table represents the output file 'result.csv'.

	A	D	CD	CE
1	Id	LotFrontage	Id+LotFrontage	
2	1242	83	1325	
3	1233	70	1303	
4	1401	50	1451	
5	1377	52	1429	
6	208			
7	1392	65	1457	
8	980	80	1060	
9	484	32	516	
10	392	71	463	
11	730	52	782	
12	255	70	325	
13	1094	71	1165	

- Sau khi thực hiện phép tính “Id+LotFrontage”, mở file “result.csv” sẽ thấy có thêm cột “Id+LotFrontage” chứa kết quả phép tính.

## V. Tài liệu tham khảo

<https://thinkingneuron.com/german-credit-risk-classification-case-study-in-python/>

<http://bis.net.vn/forums/p/505/942.aspx>