BÁO CÁO ĐỒ ÁN 1

PREPROCESSING

Môn: Khai thác dữ liệu và ứng dụng
Lớp: 18_21
Giáo viên: Lê Hoài Bắc
Dương Nguyễn Thái Bảo

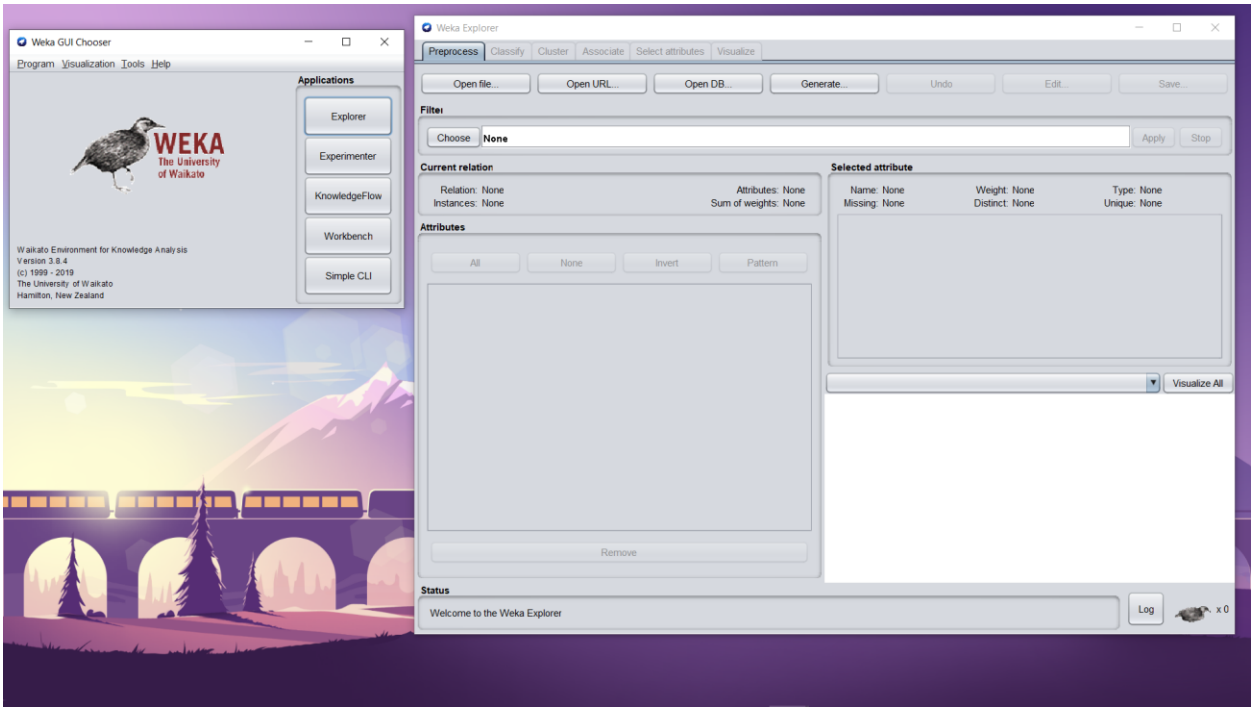
SINH VIÊN

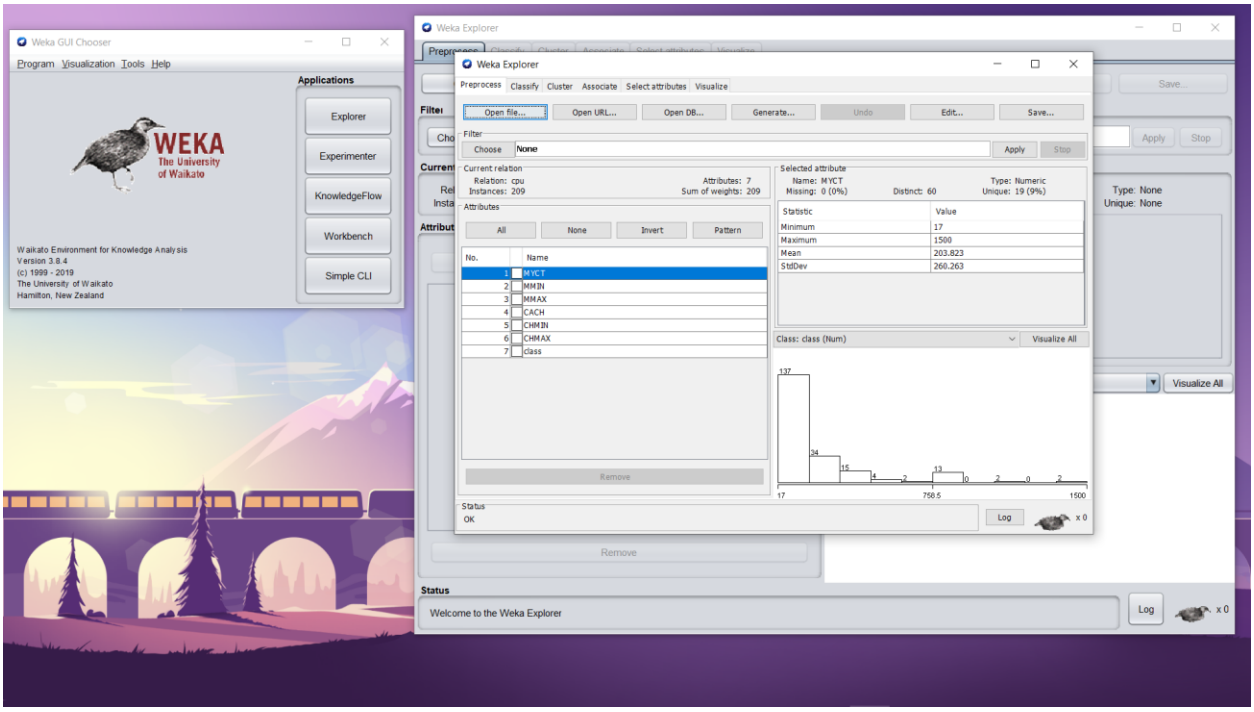
18120555 – Đặng Hữu Thắng

Mục lục

Các yêu cầu	Mức độ hoàn thành	Ghi chú
Cài đặt Weka	100%	
Làm quen với Weka	100%	
Cài đặt tiền xử lí dữ liệu	87.5%	Không hoàn thành chức năng 8

1. Cài đặt Weka





- Ý nghĩa các nhóm điều khiển:
    - **Current relation**: thông tin về tập dữ liệu: tên tập dữ liệu, số lượng mẫu (records), số thuộc tính (features)
    - **Attributes**: danh sách các thuộc tính của tập dữ liệu

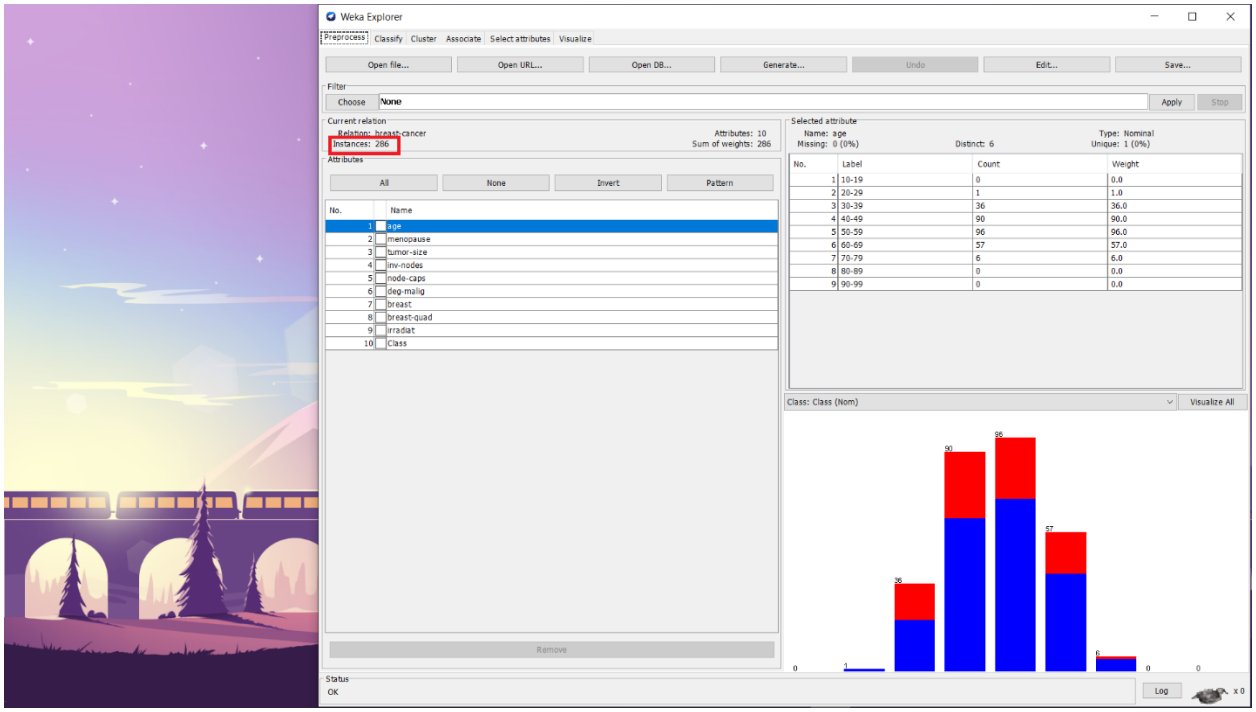
- **Selected attribute**: liệt kê các thông tin chi tiết thuộc tính được chọn trong phần Attributes

- Giải thích 5 tab trong giao diện Explorer của Weka:
  - **Preprocess:** tiền xử lí dữ liệu (cho phép mở, điều chỉnh, lưu 1 tập tin dữ liệu, thẻ này chứa các thuật toán áp dụng trong tiền xử lí dữ liệu)
  - **Classify:** phân lớp dữ liệu (cung cấp các mô hình phân loại dữ liệu hoặc hồi quy)
  - **Cluster:** phân cụm dữ liệu (cũng cấp các mô hình gom cụm)
  - **Associate:** khai thác tập phổ biến và luật kết hợp
  - **Select attributes:** lựa chọn các thuộc tính thích hợp nhất trong tập dữ liệu
  - **Visualize:** trực quan hóa dữ liệu (thể hiện dữ liệu dưới dạng biểu đồ)

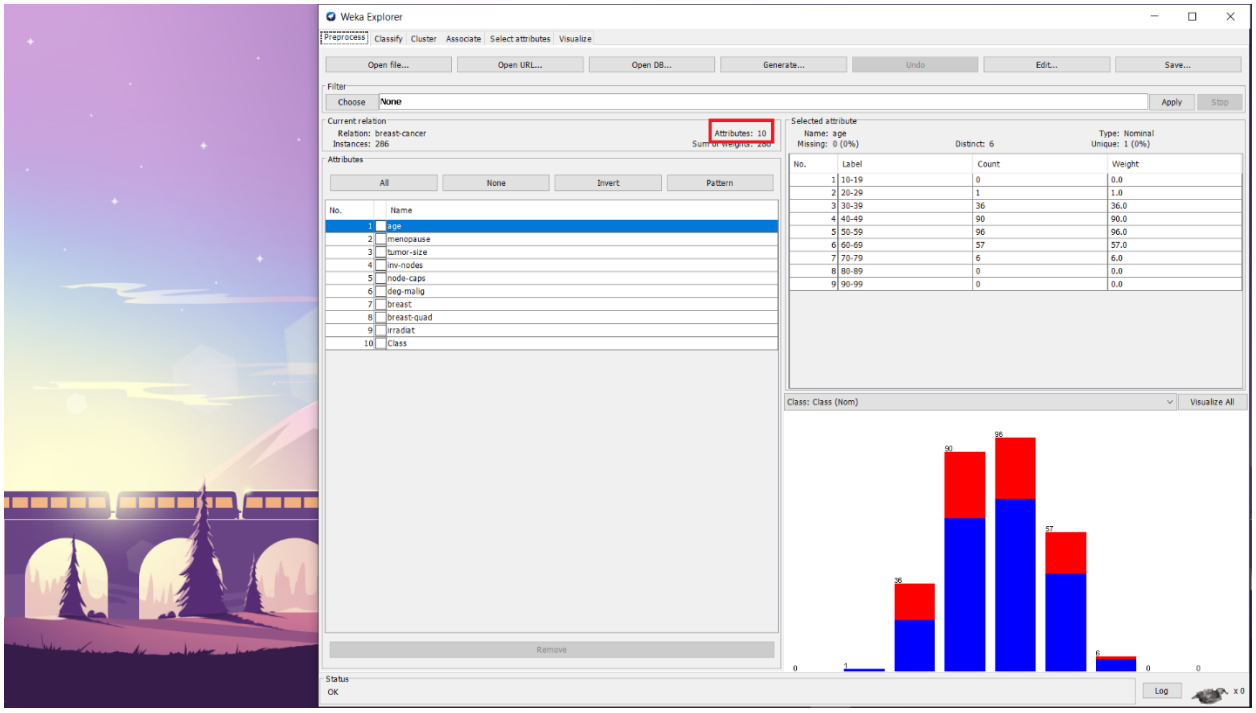
2. Làm quen với weka

2.1 Đọc dữ liệu vào weka

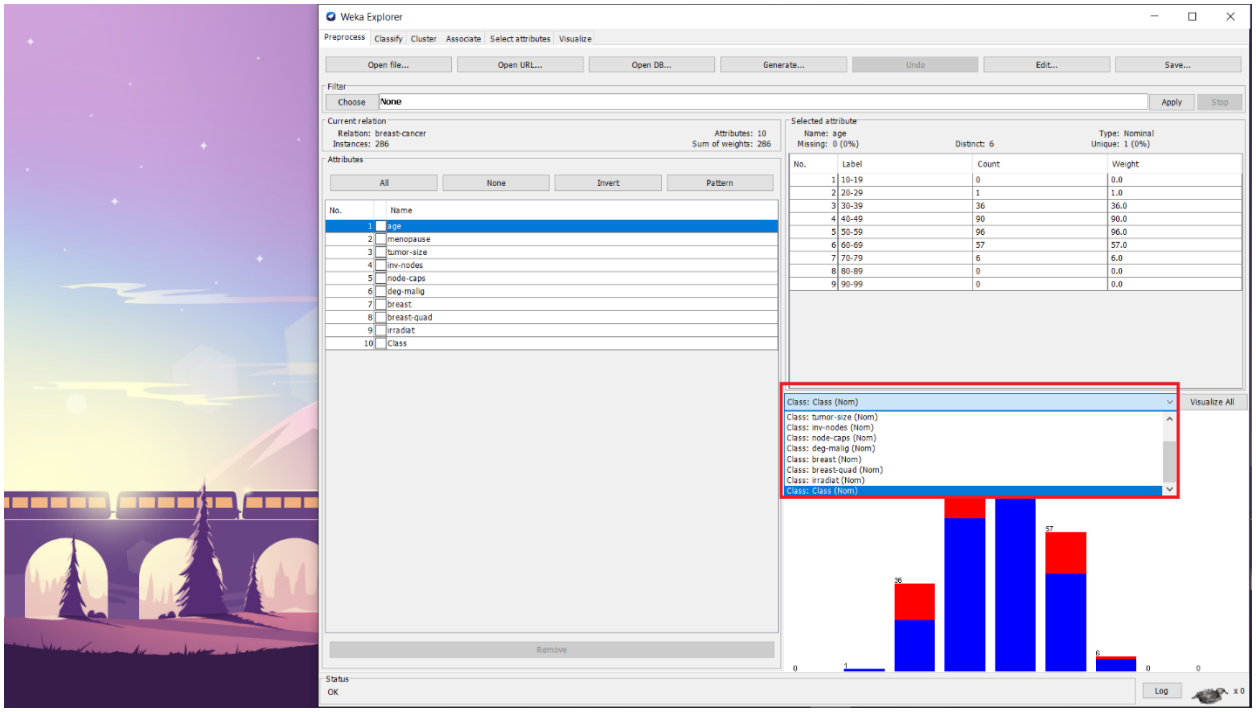
- Tập dữ liệu có bao nhiêu *mẫu* (instances)?
  ➔ Tập dữ liệu có 286 mẫu



- Tập dữ liệu có bao nhiêu *thuộc tính* (attributes)?
  ➔ Tập dữ liệu có 10 thuộc tính

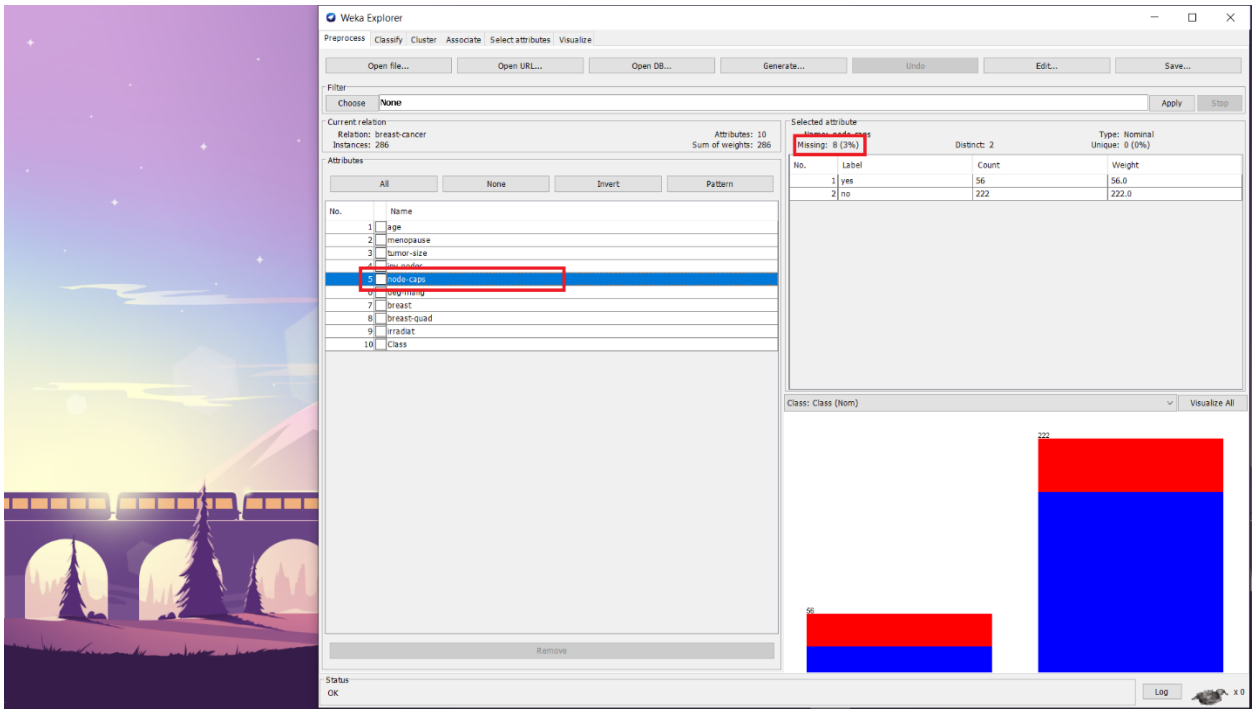


- Thuộc tính nào được dùng làm *lớp* (class)? Có thể thay đổi thuộc tính làm lớp không? Nếu có thì bằng cách nào?
  ➔ Thuộc tính Class được dùng làm lớp. Các thuộc tính nominal đều có thể được dùng làm class. Có thể thay đổi thuộc tính làm class. Thay đổi bằng cách click vào box ở ngay phía trên histogram

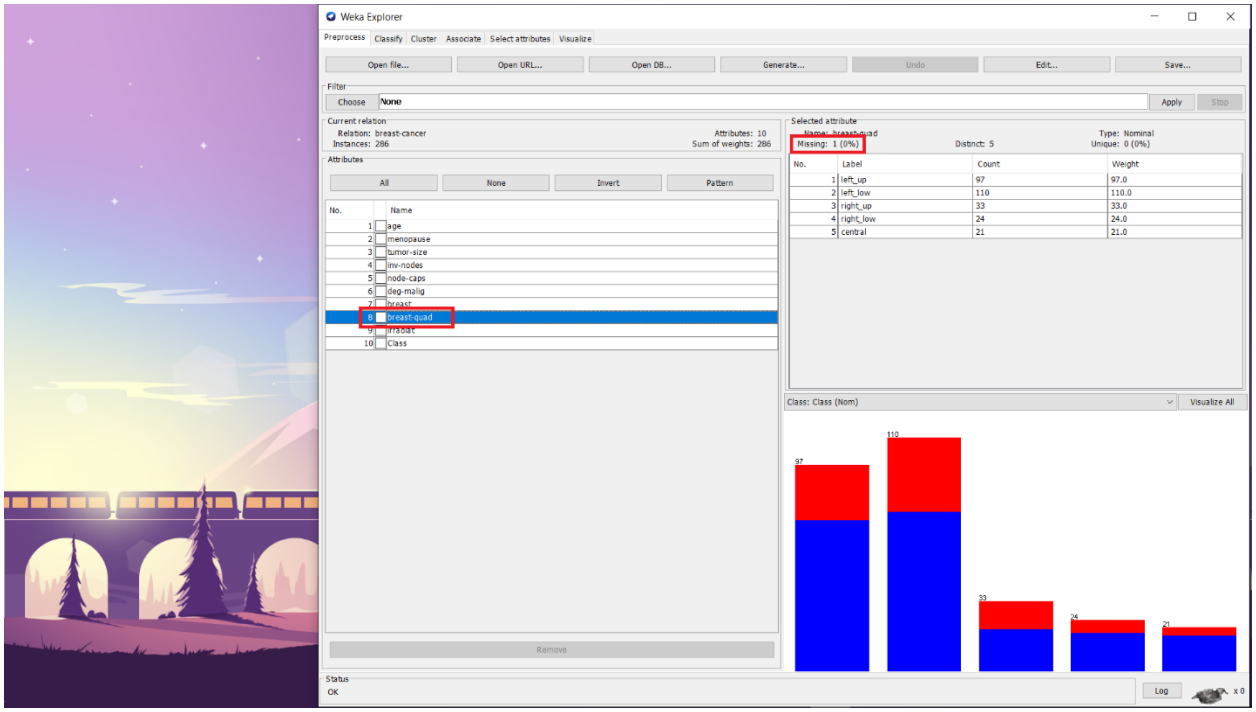


- Tìm hiểu chi tiết từng thuộc tính trong khung Attributes:

- Có 2 thuộc tính bị mất dữ liệu: *node-caps* và *breast-quad*
- Thuộc tính thiếu dữ liệu nhiều nhất: *node-caps*



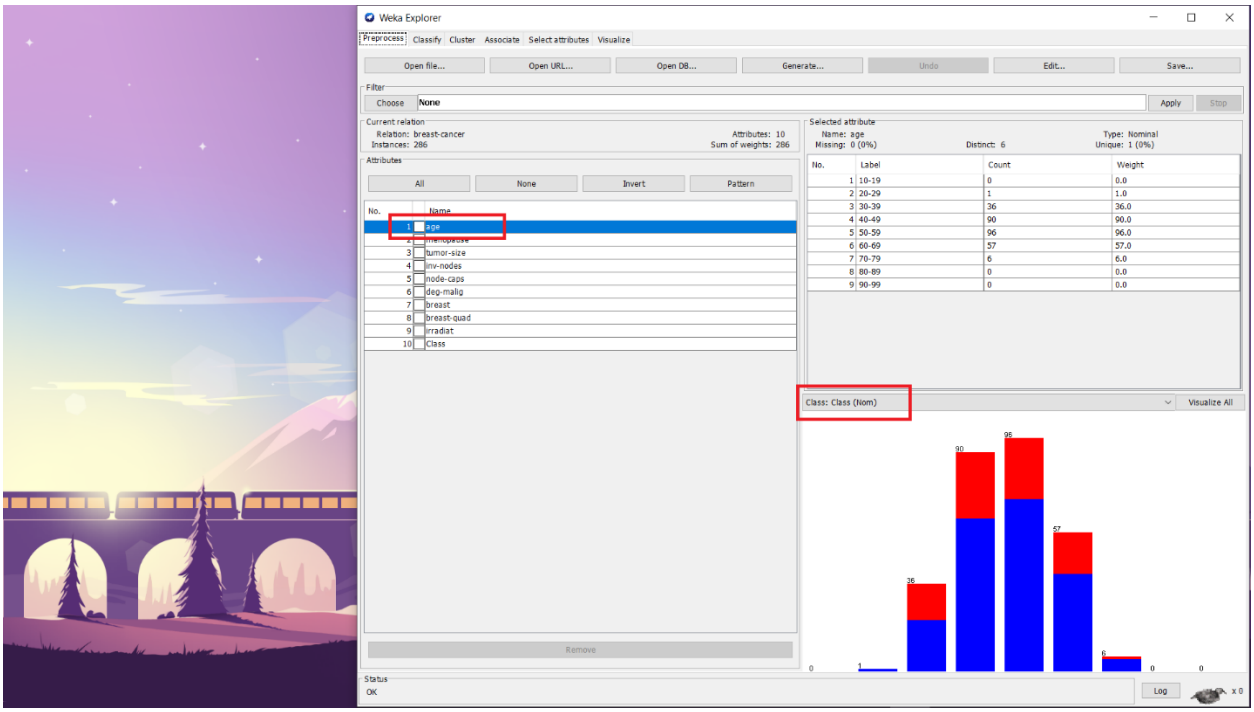
- Thuộc tính thiếu dữ liệu ít nhất: *breast-quad*



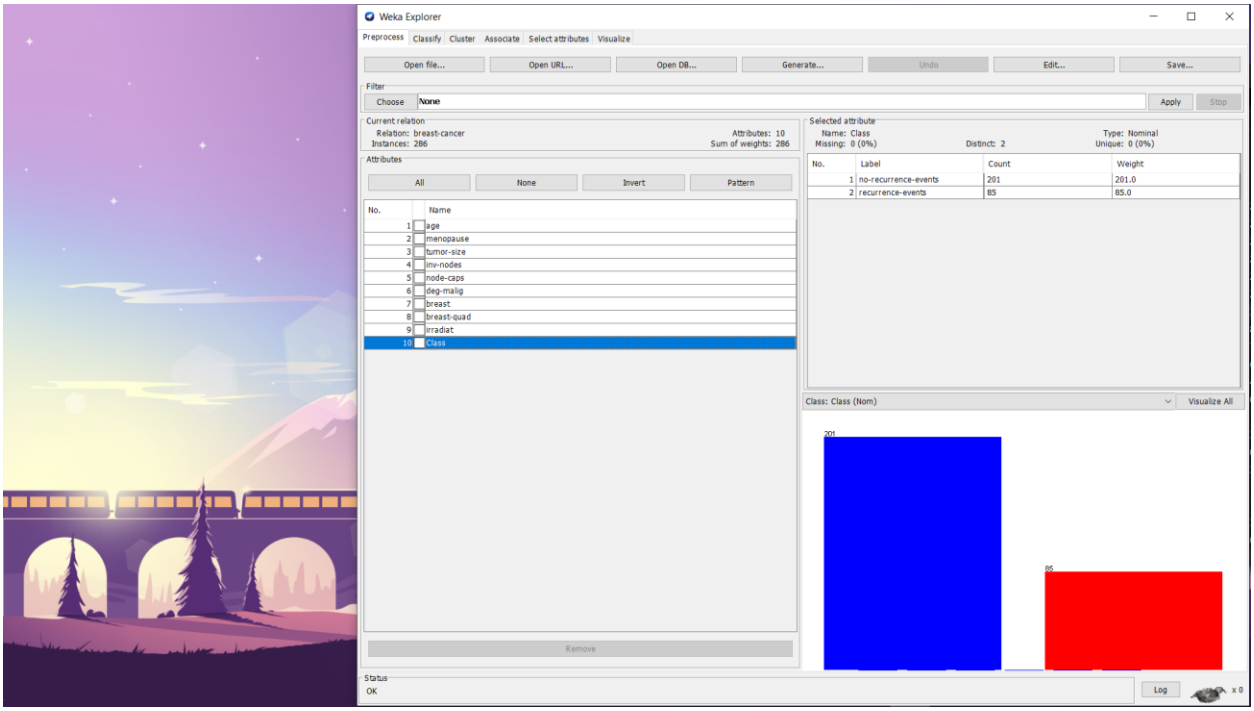
- Các cách giải quyết vấn đề missing value:
    - Bỏ qua dữ liệu bị thiếu (ignore the tuple)

- Điền các dữ liệu thiếu bằng phương pháp thủ công
- Dán nhãn một giá trị khác cho dữ liệu bị thiếu (ví dụ như “Unknown”)
- Sử dụng giá trị trung tâm như meadian, mean

- Đồ thị này thể hiện sự phân bố giá trị của thuộc tính đang chọn ở khung Attributes được phân lớp theo thuộc tính được chọn ở box Class
- Tên đồ thị: Đồ thị phân bố giá trị của thuộc tính theo lớp (class)



Đồ thị 1



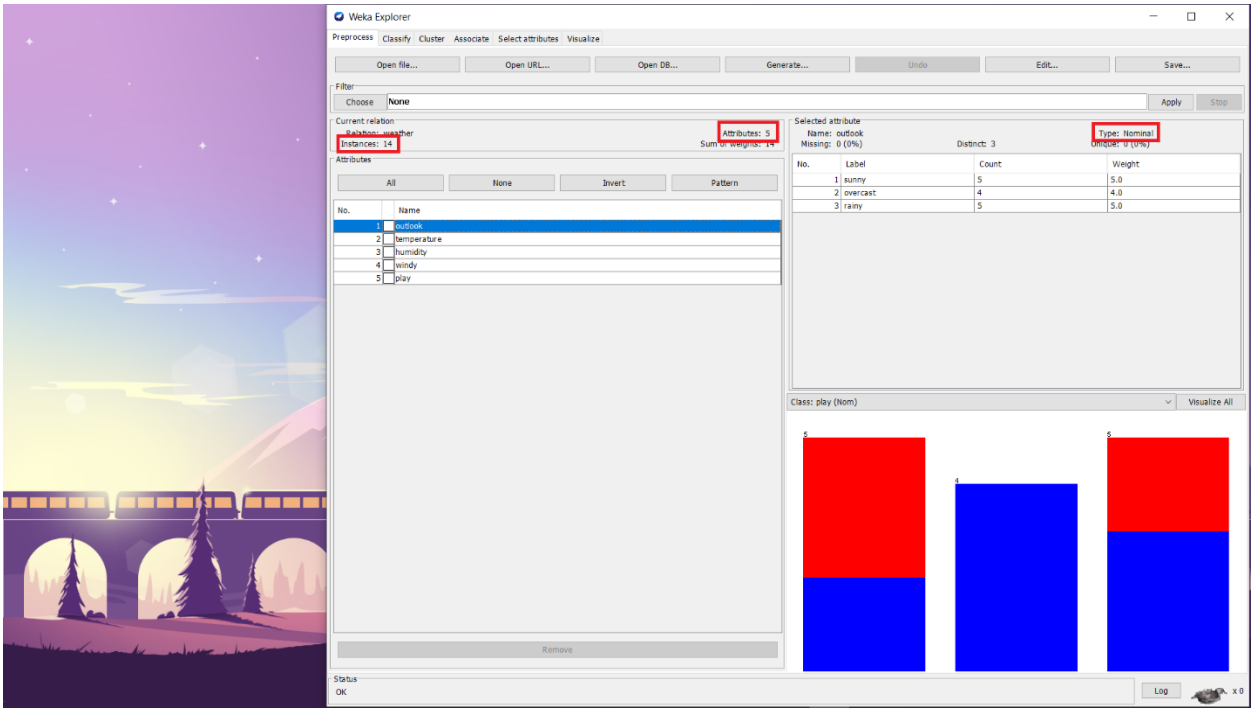
Đồ thị 2

- Như trong hình trên thì màu xanh thể hiện cho label "no-recurrence-events", màu đỏ thể hiện cho label "recurrence-events"

- Đồ thị 1 biểu diễn phân bố tuổi và tỉ lệ giữa người tái phát bệnh (recurrence-events) và không tái phát bệnh (no-recurrence-events) trên mỗi khoảng tuổi được chia

2.2 Khám phá tập dữ liệu Weather

- Tập dữ liệu có 5 *thuộc tính*, 14 *mẫu*. Thuộc tính "play" là *lớp* (class). Phân loại:
  - **Categorical:** outlook, windy, play
  - **Numeric:** temperature, humidity



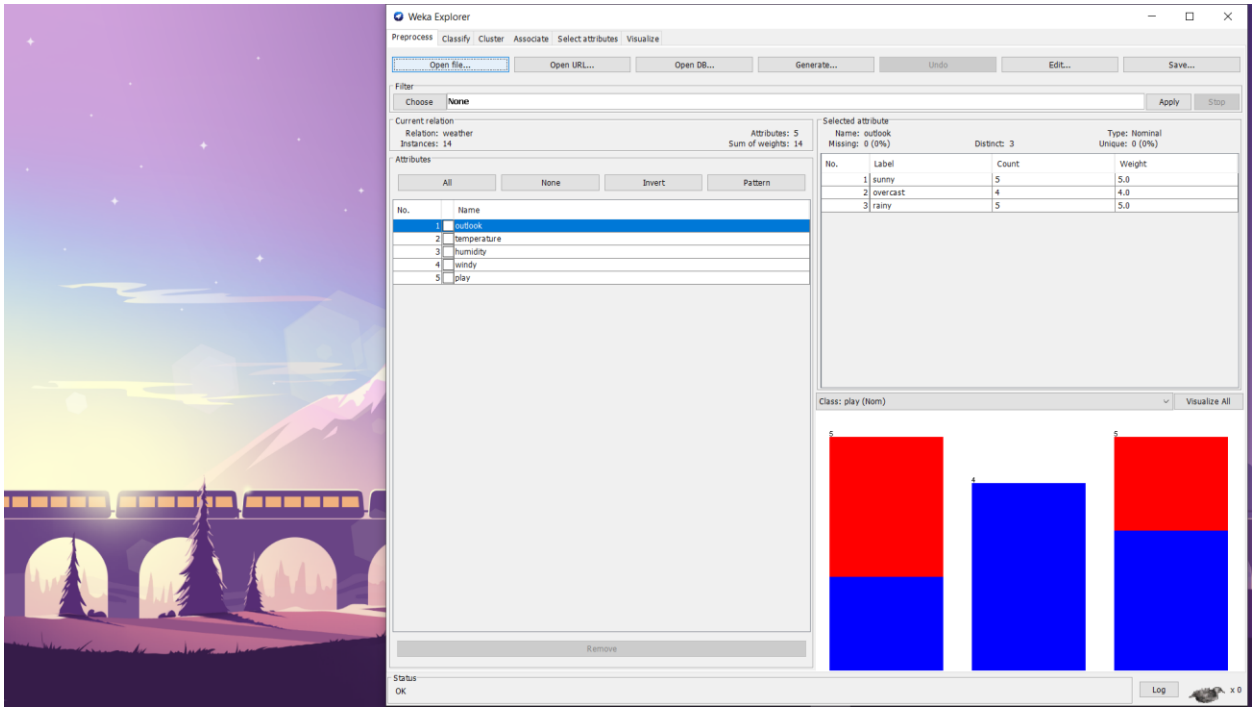
- Five-number summary:

	Temperature	**Humidity**
Minimum	64	65
1'st quartile	69.5	72.5
Median	72	82.5
3'rd quartile	77.5	90
Maximum	85	96

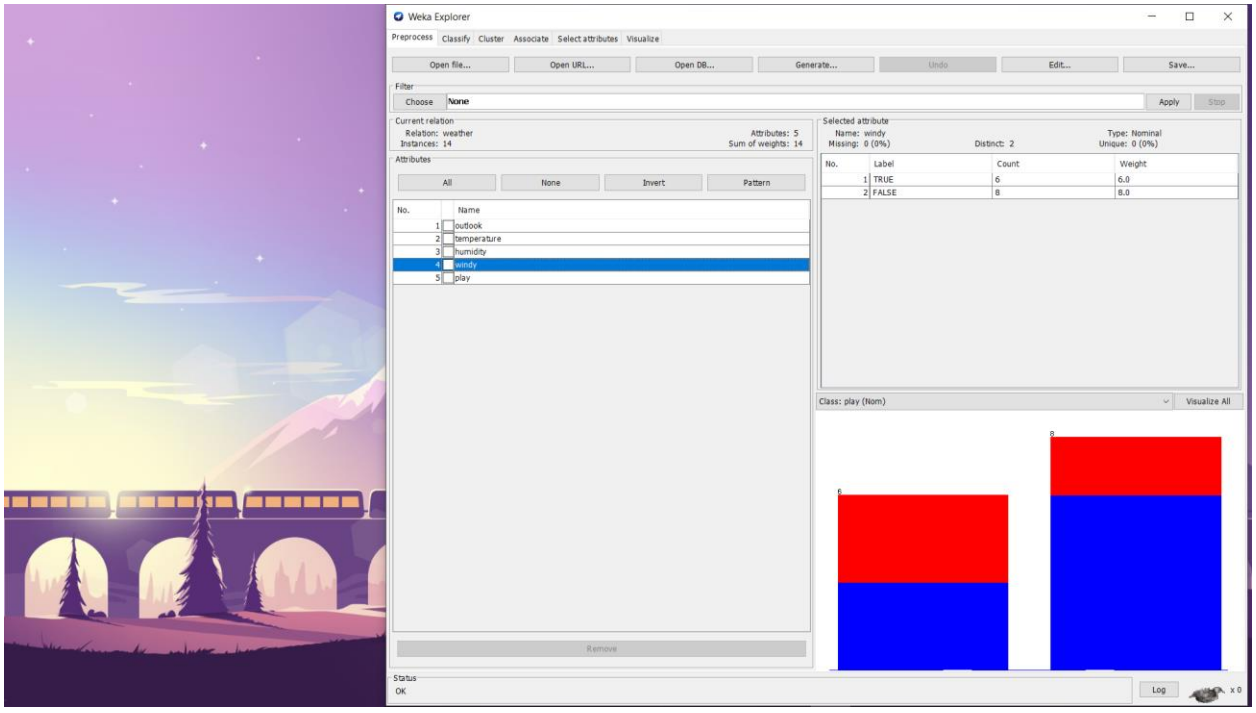
  - Weka chỉ cung cấp 2 giá trị Min và Max và không cung cấp 3 giá trị còn lại

- Đồ thị của các thuộc tính khác:

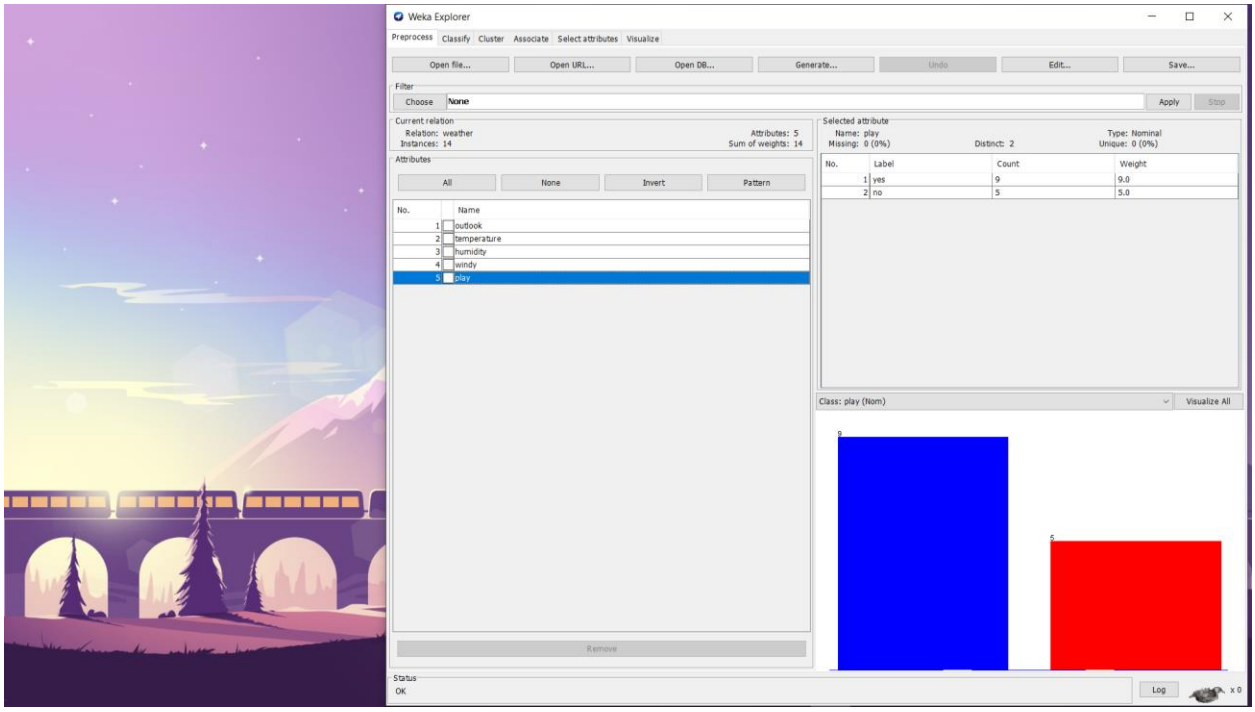
- **Outlook:**



- **Windy:**



- **Play:**



- Thuật ngữ : "Scatter plot" hay cụ thể là "A matrix of two-dimensional scatter plots of every pair of attributes". Không có cặp thuộc tính nào có vẻ tương quan với nhau.

2.3 Khám phá tập dữ liệu tín dụng Đức

- Nội dung phần ghi chú nói về thông tin sơ lược của dataset: title, nguồn, số lượng mẫu và thuộc tính và mô tả sơ lược về từng thuộc tính.
  - Tập dữ liệu có 1000 *mẫu*, 20 *thuộc tính*
  - Mô tả 5 *thuộc tính*:
    - **Checking_status:**

```
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
```

    - **Duration:**

```
Attribute 2:  (numerical)
              Duration in month
```

    - **Credit history:**

```
Attribute 3:  (qualitative)
              Credit history
              A30 : no credits taken/
                    all credits paid back duly
                A31 : all credits at this bank paid back duly
              A32 : existing credits paid back duly till now
                A33 : delay in paying off in the past
              A34 : critical account/
                    other credits existing (not at this bank)
```

- **Purpose:**

```
Attribute 4:  (qualitative)
              Purpose
              A40 : car (new)
              A41 : car (used)
              A42 : furniture/equipment
              A43 : radio/television
              A44 : domestic appliances
              A45 : repairs
              A46 : education
              A47 : (vacation - does not exist?)
              A48 : retraining
              A49 : business
              A410 : others
```

- **Credit_amount:**

```
Attribute 5:  (numerical)
              Credit amount
```

- Tên của thuộc tính lớp: Class
- Các lựa chọn của Weka để chọn lọc thuộc tính là:
  - ❑ *Các phương pháp đánh giá:*
    - **CfsSubsetEval:** Đánh giá giá trị của một tập hợp con các thuộc tính bằng cách xem xét khả năng dự đoán riêng của từng tính năng cùng với mức độ dư thừa giữa chúng.
    - **ClassifierSubsetEval:** Đánh giá các tập hợp thuộc tính trên dữ liệu huấn luyện hoặc một bộ kiểm tra riêng biệt.
    - **CorrelationAttributeEval:** Đánh giá giá trị của một thuộc tính bằng cách đo lường mối tương quan giữa nó và lớp. Các thuộc tính danh nghĩa được xem xét trên một giá trị theo cơ sở giá trị bằng cách coi mỗi giá trị là một chỉ báo.
    - **GainRatioAttributeEval:** Đánh giá theo độ đo tỉ lệ đạt được với các lớp. với công thức tính gain là GainR(Class, Attribute) = (H(Class) - H(Class | Attribute)) / H(Attribute).
    - **OneRAttributeEval:** Đánh giá giá trị của một thuộc tính bằng cách sử dụng trình phân loại OneR.
    - **PrincipalComponents:** Thực hiện phân tích thành phần chính (chọn ra các thành phần chính) nhằm biến đổi dữ liệu (transformation data).

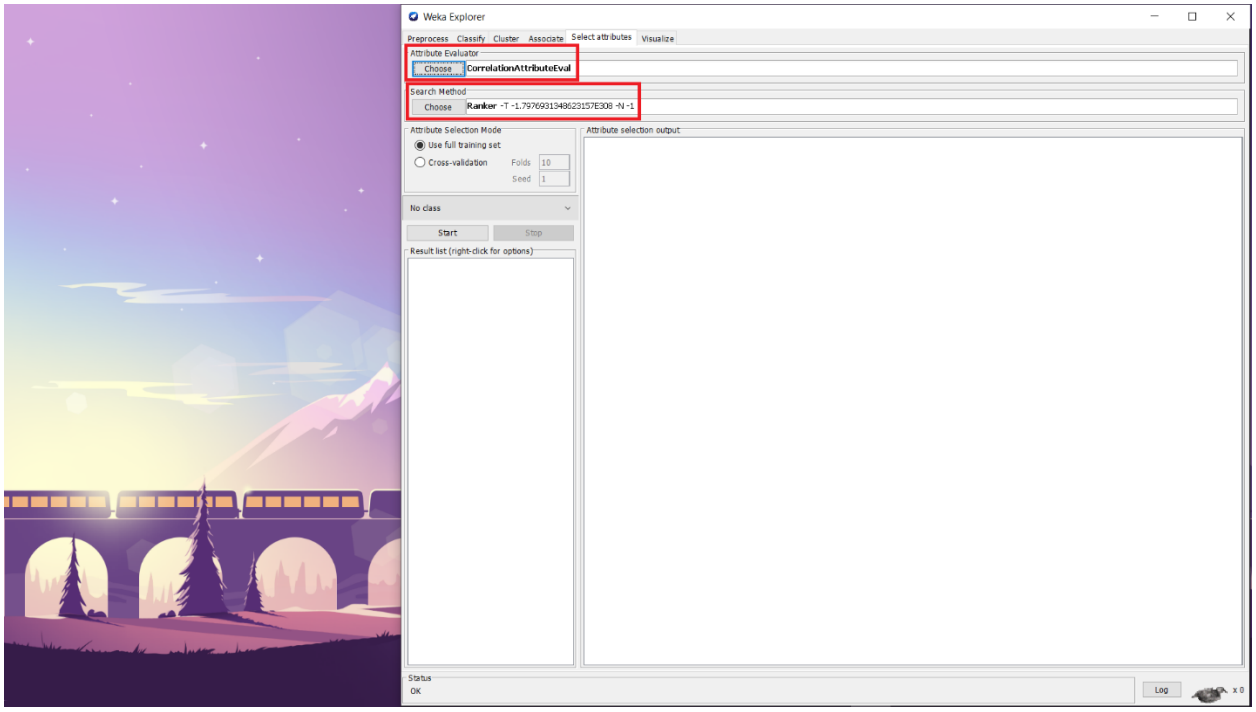
- **ReliefFAttributeEval:** Đánh giá giá trị của một thuộc tính bằng cách lặp lại việc lấy mẫu (sampling) của một instance và xem xét giá trị của thuộc tính đã cho với instance gần nhất của cùng một lớp và khác lớp.
- **SymmetricalUncertAttributeEval:** Đánh giá giá trị của một thuộc tính bằng cách đo độ không đảm bảo đối xứng đối với lớp.
- **WrapperSubsetEval:** đánh giá bằng tập bao các phân loại ("wrapper" method wraps a classifier in cross-validation loop.

❑ *Các phương pháp tìm kiếm (chọn thuộc tính theo nhu cầu):*

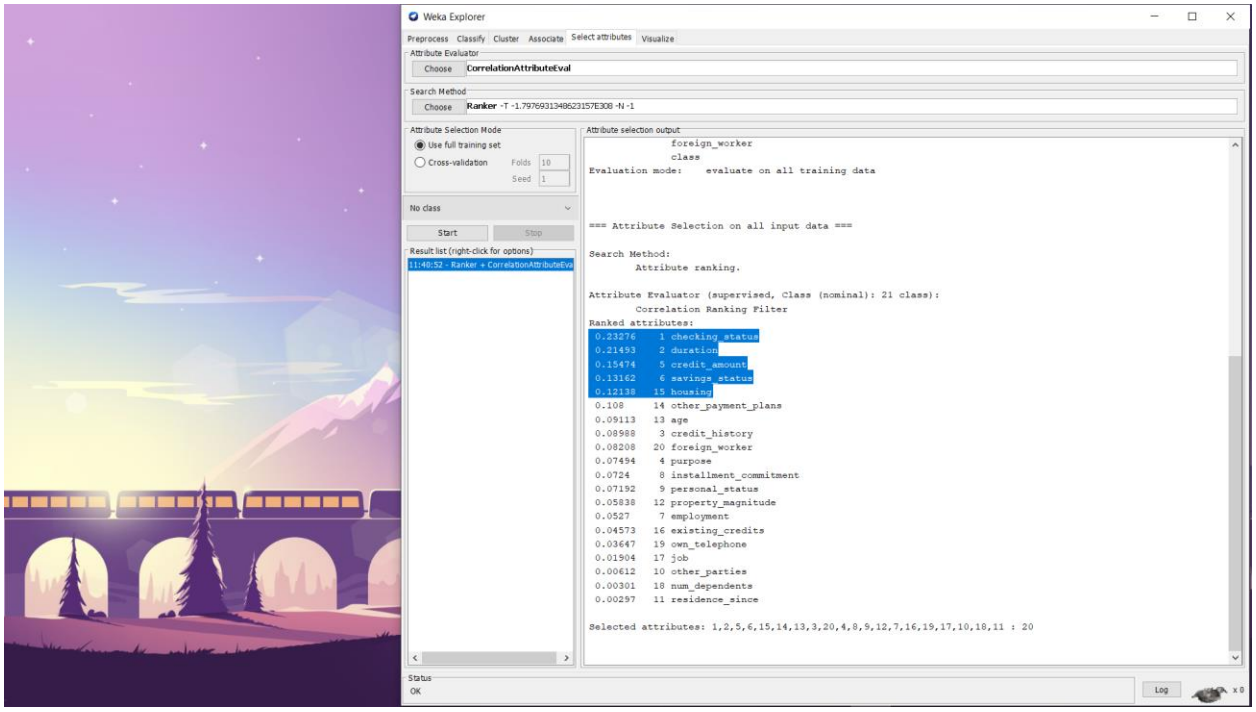
- **BestFirst:** Tìm kiếm tập con không gian các thuộc tính bằng tham lam tăng cường với cơ sở quay lui.
- **GreedyStepwise:** Phương pháp tìm kiếm tham lam tiến hay lùi thông qua tập con không gian các thuộc tính.
- **Ranker:** xếp hạng các thuộc tính theo giá trị nó được đánh giá

- Cần sử dụng bộ lọc **CorrelationAttributeEval**

- **B1:** lựa chọn *Search method* là Ranker và *Attribute Evaluator* là CorrelationAttributeEval, sau đo click **Start**

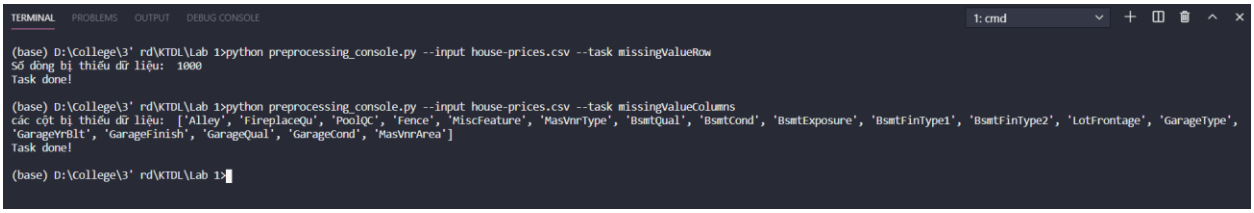


- **B2:** Kết quả cuối cùng

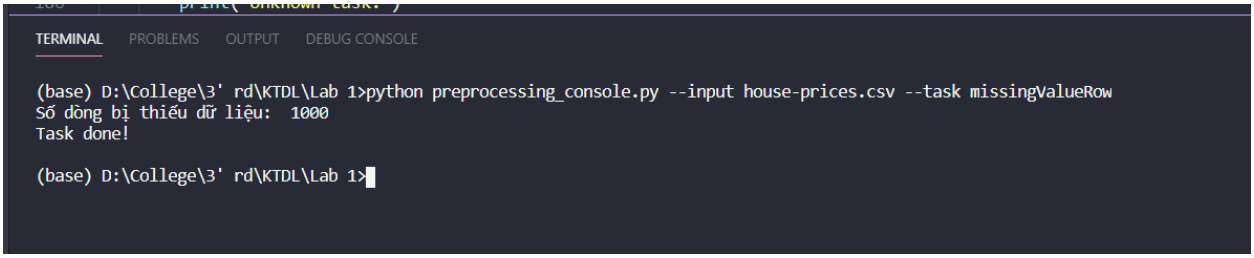


3. Cài đặt tiền xử lí dữ liệu

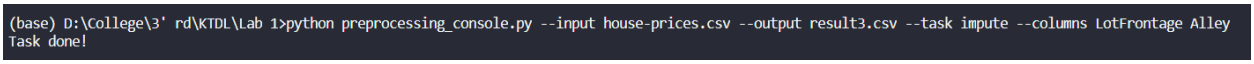
- Chức năng 1:



- Chức năng 2:



- Chức năng 3:



- Chức năng 4:

```
(base) D:\College\3' rd\KTDL\Lab 1>python preprocessing_console.py --input house-prices.csv --output result4.csv --task delete_row --ratio 50
Task done!
```

- Chức năng 5:

```
(base) D:\College\3' rd\KTDL\Lab 1>python preprocessing_console.py --input house-prices.csv --output result5.csv --task delete_columns --ratio 50
Task done!
```

- Chức năng 6:

```
(base) D:\College\3' rd\KTDL\Lab 1>python preprocessing_console.py --input house-prices.csv --output result6.csv --task duplicateRow
Task done!
```

- Chức năng 7:

```
(base) D:\College\3' rd\KTDL\Lab 1>python preprocessing_console.py --input house-prices.csv --output result7.csv --task minMaxNorm --newMinMax 0 1 --columns MSSubClass
Task done!
```