ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH

ĐẠI HỌC KHOA HỌC TỰ NHIÊN

KHOA CÔNG NGHỆ THÔNG TIN



DATA MINING

LAB 1: PREPROCESSING

*Lớp*: KTDL & UD 18\_21

*Nhóm thực hiện*:

18120078 – Ngô Phù Hữu Đại Sơn

18120253 – Mai Ngọc Tú

MỤC LỤC

[A. Thông tin khái quát 2](#_Toc54517123)

[I. Thông tin nhóm 2](#_Toc54517124)

[II. Bảng phân công công việc 2](#_Toc54517125)

[B. Nội dung 3](#_Toc54517126)

[I. Mục tiêu của đồ án: 3](#_Toc54517127)

[II. Yêu cầu: 3](#_Toc54517128)

[III. Triển khai 4](#_Toc54517129)

[1. Cài đặt Weka: 4](#_Toc54517130)

[2. Làm quen với Weka: 4](#_Toc54517131)

[3. Cài đặt tiền xử lý dữ liệu: 4](#_Toc54517132)

[IV. Đánh giá đồ án 4](#_Toc54517133)

[1. Mức độ hoàn thành của các thành viên 4](#_Toc54517134)

[2. Mức độ hoàn thành đồ án: 4](#_Toc54517135)

[VII. Nguồn tham khảo 4](#_Toc54517136)

# Thông tin khái quát

## I. Thông tin nhóm

|  |  |  |
| --- | --- | --- |
| **MSSV** | **Họ tên** | **Vai trò** |
| 18120078 | Ngô Phù Hữu Đại Sơn | Nhóm trưởng |
| 18120253 | Mai Ngọc Tú | Thành viên |

## II. Bảng phân công công việc

|  |  |  |
| --- | --- | --- |
| **MSSV** | **Công việc phụ trách** | **Thời gian thực hiện** |
|  |  |  |
|  |  |  |

# B. Nội dung

## I. Mục tiêu của đồ án:

* Làm quen với các thao tác cơ bản trong tác vụ tiền xử lý dữ liệu thông qua việc áp dụng các công cụ hỗ trợ được cung cấp bởi phần mềm mã nguồn mở Weka.
* Phát huy kỹ năng lập trình để tự cài đặt các thủ tục tiền xử lý dữ liệu đơn giản.

## II. Yêu cầu:

1. **Cài đặt Weka.**

* Chụp hình giao diện chức năng Exploere cùng màn hình desktop.
* Tìm thư mục data trong thư mục cài đặt của Weka và mở một tập dữ liệu bất kì (có phần mở rộng là arff). Giải thích ý nghĩa các nhóm điều khiển Current relation, Attributes và Selected attribute trong tab Preprocess. Giải thích ngắn gọn ý nghĩa 5 tab trong giao diện Explorer của Weka.

1. **Làm quen với Weka.**

* Khám phá bộ dữ liệu Breast Cancer
* Khám phá bộ dữ liệu Weather
* Khám phá bộ dữ liệu Tín dụng Đức

1. **Cài đặt tiền xử lý dữ liệu**

* Cài đặt chương trình khám phá bộ dữ liệu house\_prices gồm có các chức năng:

+ Liệt kê các cột bị thiếu dữ liệu.

+ Đếm số dòng bị thiếu dữ liệu.

+ Điền giá trị bị thiếu.

+ Xóa các dòng bị thiếu dữ liệu với ngưỡng tỉ lệ thiếu cho trước.

+ Xóa các cột bị thiếu dữ liệu với ngưỡng tỉ lệ thiếu cho trước.

+ Xóa các mẫu bị trung lặp.

+ Chuẩn hóa một thuộc tính numeric bằng phương pháp min-mã và Z-score.

+ Tính giá trị biểu thức thuộc tính.

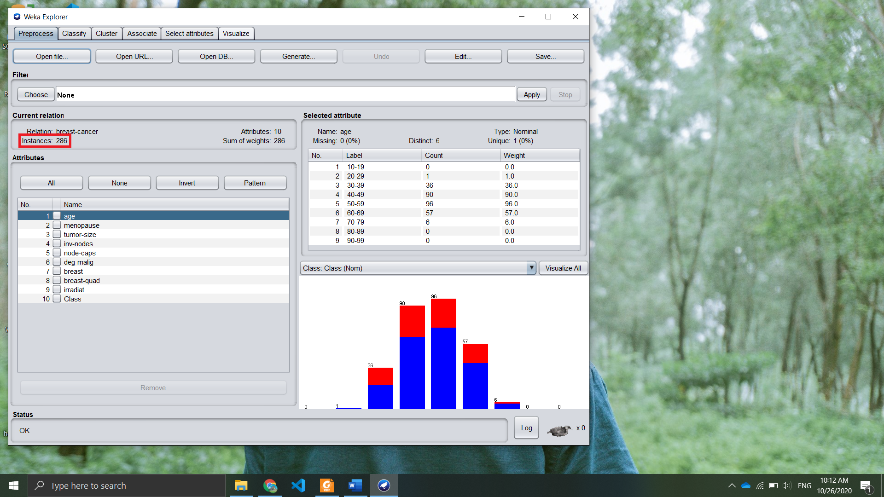
## III. Triển khai

### Cài đặt Weka:

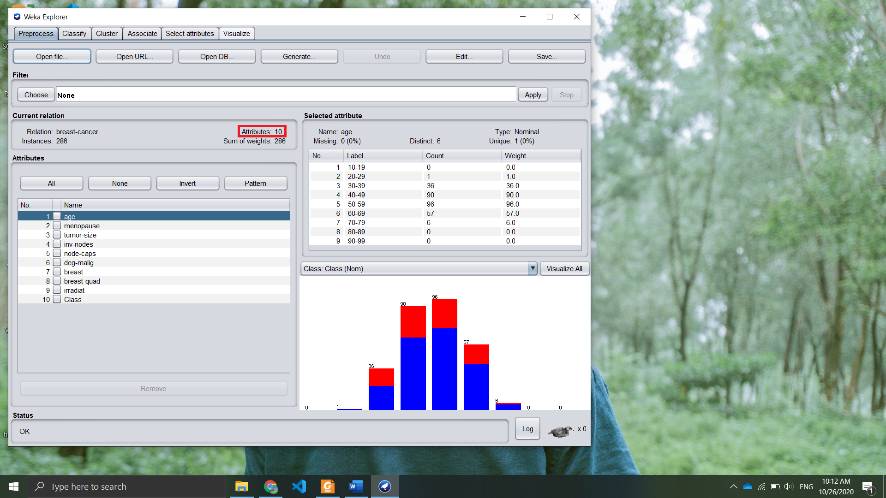
### Làm quen với Weka:

* 1. **Đọc dữ liệu vào Weka:**

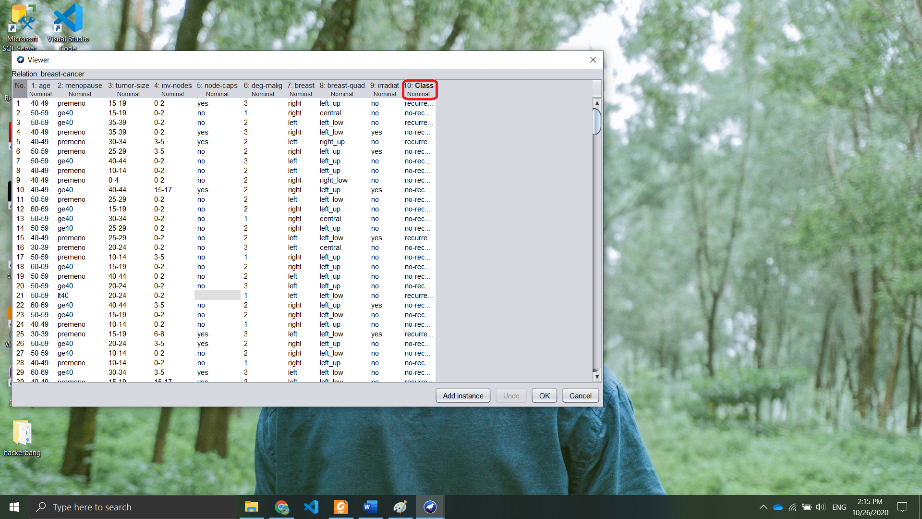
1/ - Tập dữ liệu có 286 mẫu (Instances)



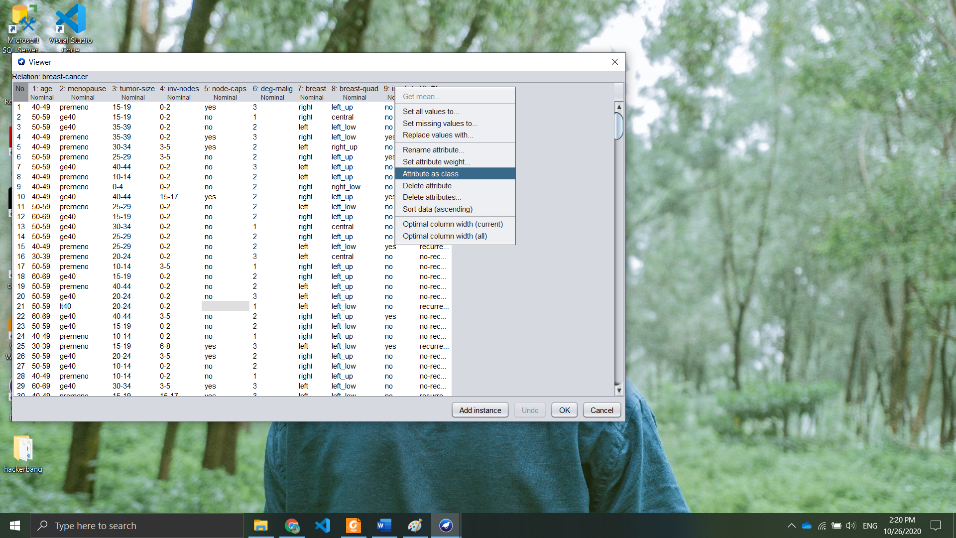
2/ - Tập dữ liệu có 10 thuộc tính (attributes)



3/ - Thuộc tính Class dùng làm lớp (class)



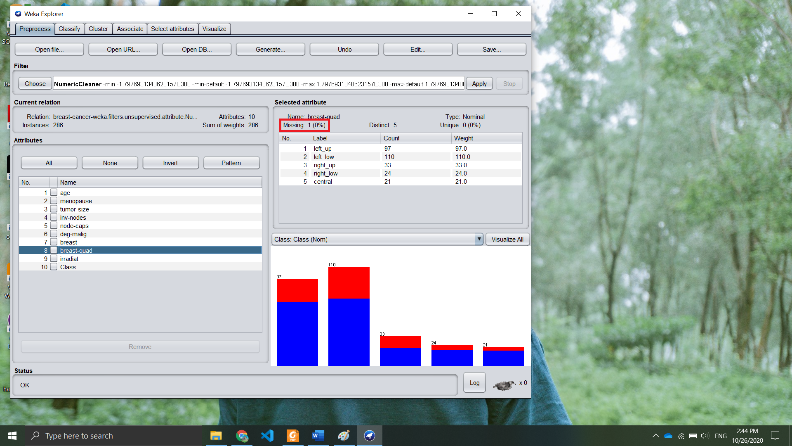
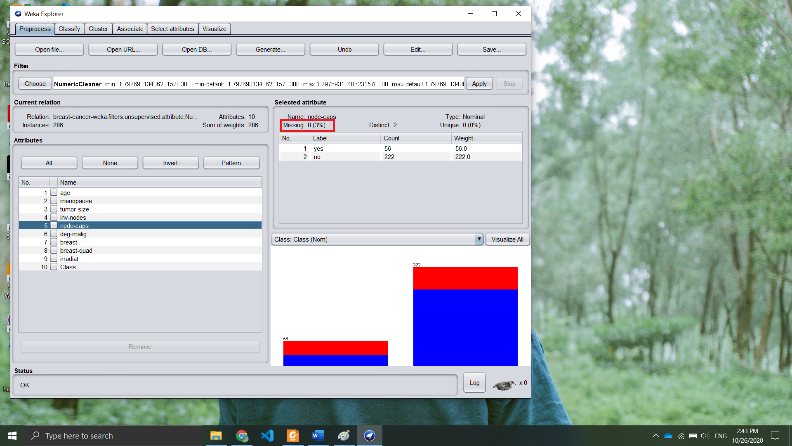
- Có thể thay đổi thuộc tính dùng làm lớp. Ta có thể thay đổi bằng cách chọn một thuộc tính khác, sau đó nhấn chuột phải và chọn **Attribute as class**



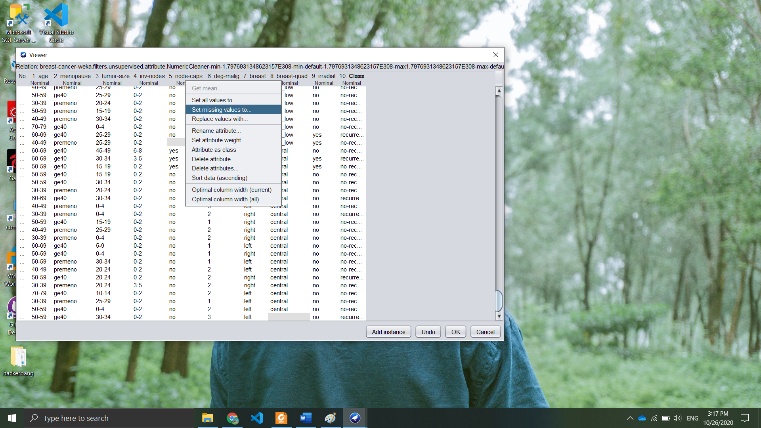
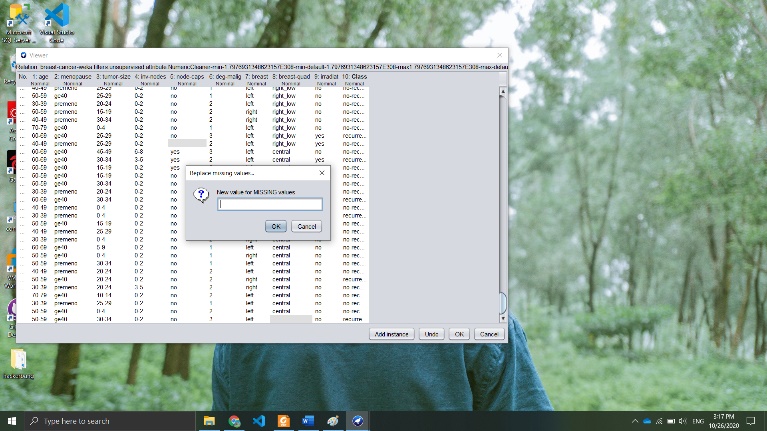
4/ - Có 2 thuộc tính bị thiếu dữ liệu: node-caps, breast-quad

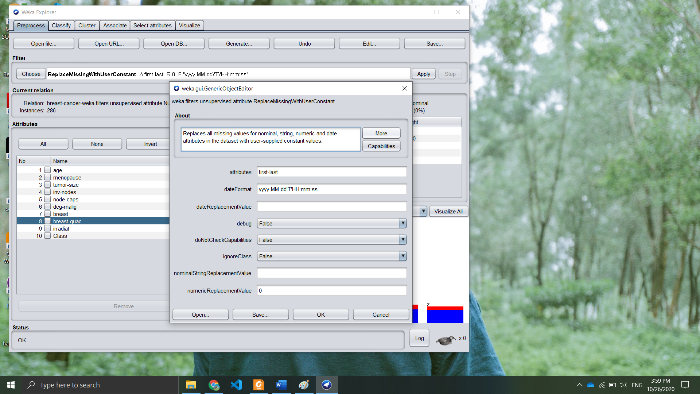
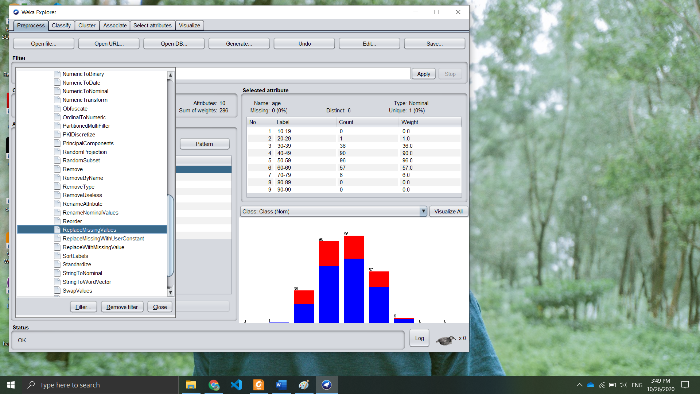
+ Thuộc tính thiếu dữ liệu ít nhất: breast-quad (thiếu 1)

+ Thuộc tính thiếu dữ liệu nhiều nhất: node-caps (thiếu 8)

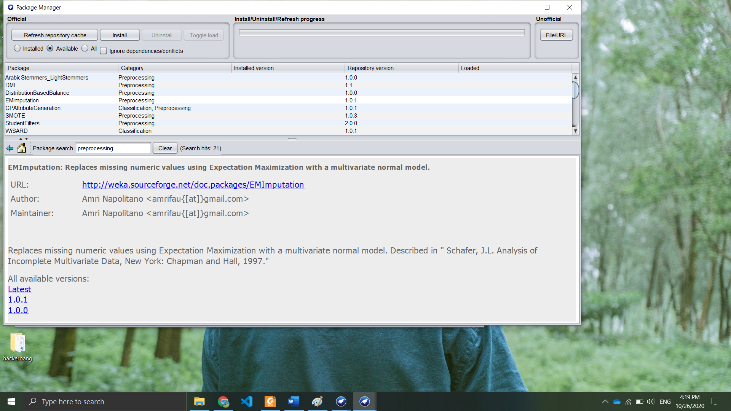


- Các cách giải quyết vấn đề missing values:

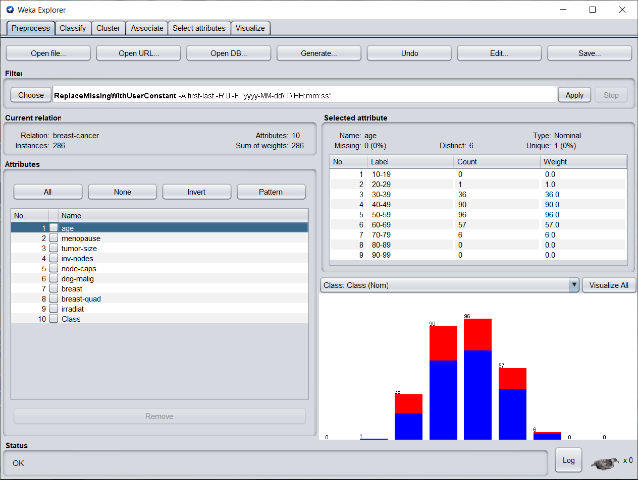
+ Cách 1: Chọn Edit, chọn các thuộc tính thiếu dữ liệu, sau đó nhấn chuột phải và chọn **Set missing value to…** Gán các ô bị thiếu dữ liệu bằng giá trị nhập trong **New value for MISSING values**. Lưu ý giá trị nhập vào trong **New value for MISSING values** là giá trị được cho phép của thuộc tính đó thì mới gán thành công. Ví dụ: thuộc tính node-caps chỉ được gán [yes,no] cho các ô bị thiếu dữ liệu.

 + Cách 2: dùng filter **ReplaceMissingValues** hoặc **ReplaceMissingWithUserConstant** trong Weka. Filter **ReplaceMissingValues** giúp thêm tự động các ô bị thiếu dữ liệu trong các thuộc tính, dữ liệu được thêm vào là giá trị được cho phép của thuộc tính đó. Còn filter **ReplaceMissingWithUserConstant** sẽ gán cho những ô thiếu dữ liệu của các thuộc tính một giá trị mặc định mà người dùng cho. Ta có thể gán giá trị mặc định dành cho thuộc tính dạng numeric, nomial, date…

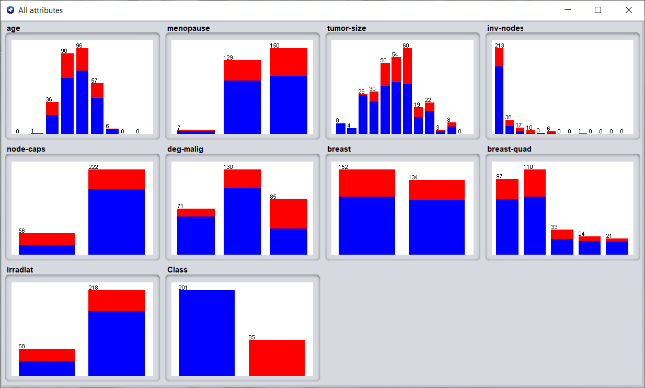
+ Cách 3: Vào Weka GUI chọn **Tools 🡪 Package manager**. Trong ô **Package search** nhập “preprocessing” và kết quả là các filter dành cho quá trình tiền xử lý dữ liệu, trong đó có các filter dùng để giải quyết vấn đề missing values. Chọn filter phù hợp, nhấn Install và sử dụng để giải quyết misssing values.



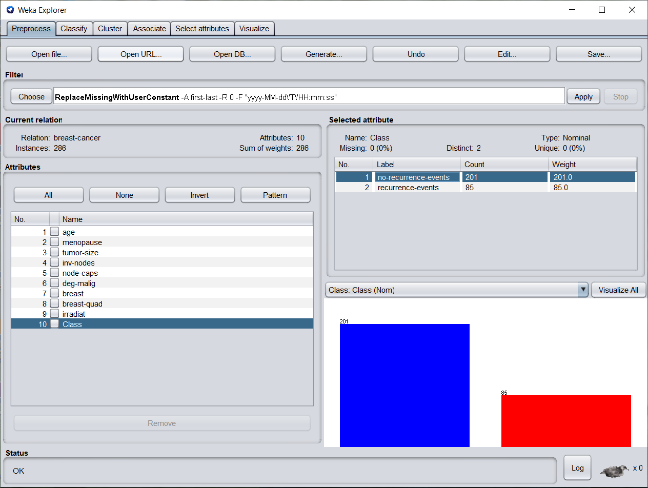
5/ - Ý nghĩa của đồ thị trong của sổ Explorer: đồ thị thể hiện các giá trị Count của từng Label trong cùng một thuộc tính. Các cột sắp xếp theo thứ tự của các Label và chiều cao của cột là giá trị Count tương ứng.



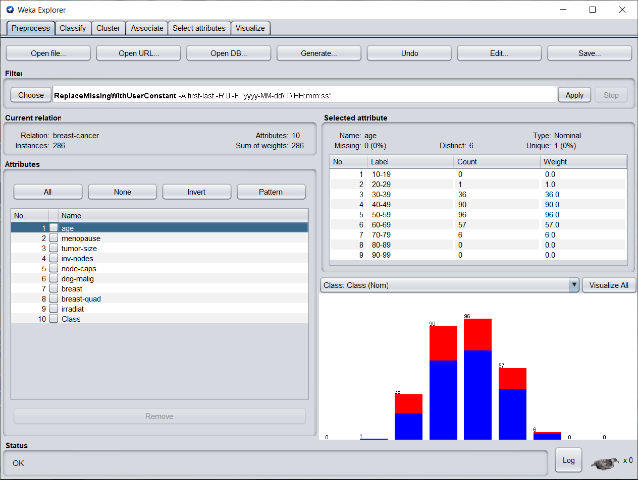
- Đặt tên cho đồ thị là tên của thuộc tính tương ứng với đồ thị đó.



- Màu xanh của đồ thị tương ứng với Label no-recurrence-events của attribute Class (attribute dùng làm lớp). Màu đỏ của đồ thị tương ứng với Label recurrence-events của attribute Class.



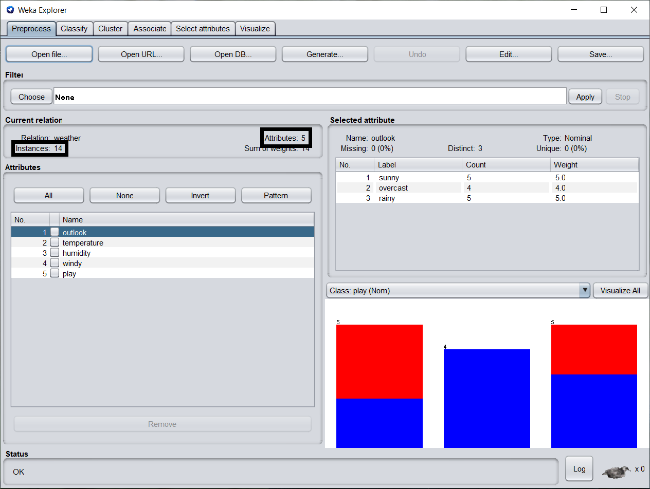
- Đồ thị biểu diễn từng giá trị Count của mỗi Label trong từng thuộc tính theo từng cột đồ thị và chia từng cột theo từng màu khác nhau ứng với các giá trị trong cùng một Label thỏa mãn các giá trị khác nhau của attribute đóng vai trò là lớp (attribute Class). Ví dụ, trong attribute age, các cột ứng với các Label, màu xanh ứng với thỏa mãn no-recurrence-events của attribute Class, màu đỏ ứng với thỏa mãn recurrence-events của attribute Class.



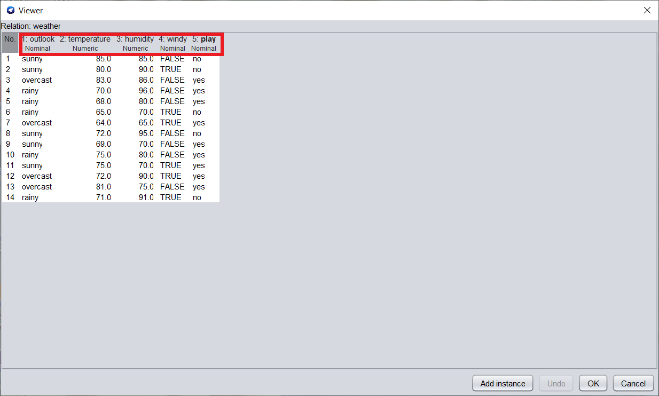
* 1. **Khám phá tập dữ liệu Weather:**

**1/**

* Tập dữ liệu có 5 thuộc tính, 14 mẫu.



* Thuộc tính kiểu dữ liệu nominal: outlook, windy, play
* Thuộc tính kiểu dữ liệu numeric: temperature, humidity
* Thuộc tính lớp: play



**2/**

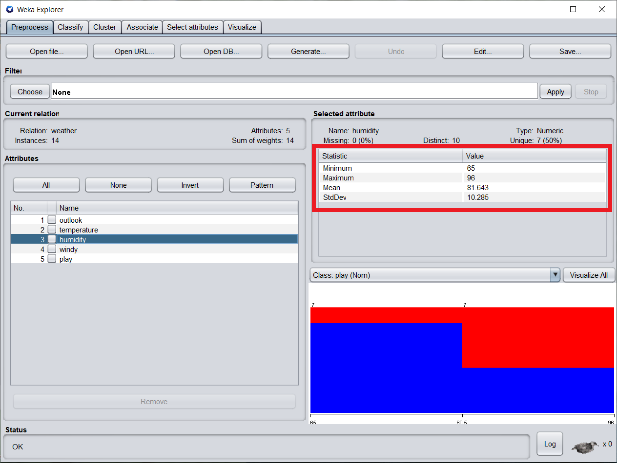
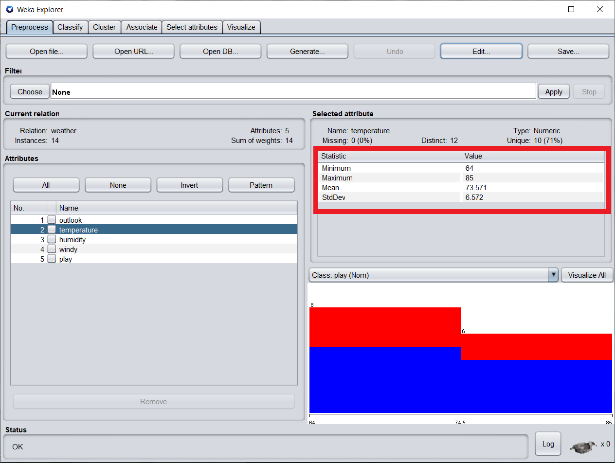
* Five-number summary của thuộc tính temperature:

+ Minimum: 64

+ Maximum: 85

+ Mean: 73.571

+ StdDev: 6.572



* Five-number summary của thuộc tính :

+ Minimum: 65

+ Maximum: 96

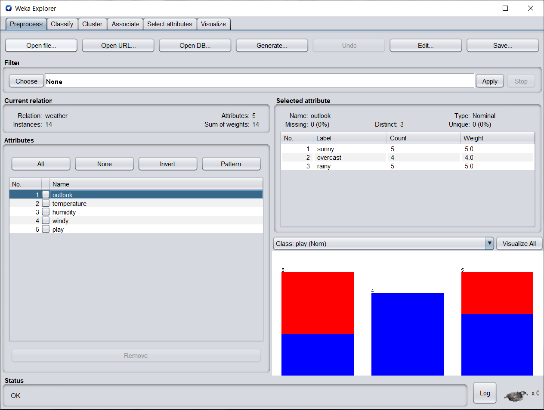
+ Mean: 81.6434

+ StdDev: 10.285

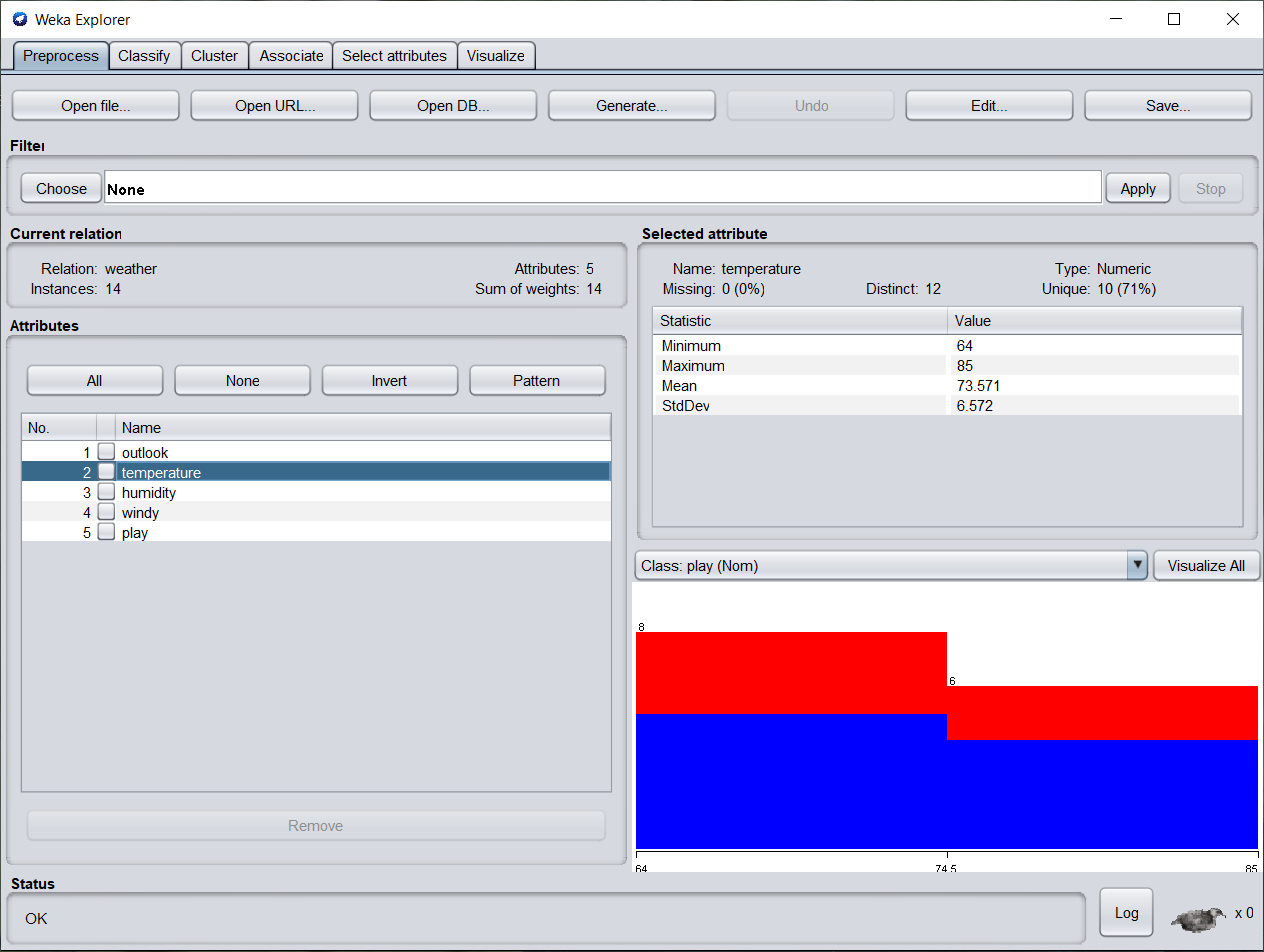
* Weka cung cấp những giá trị này có sẵn trong các file arff.

**3/**

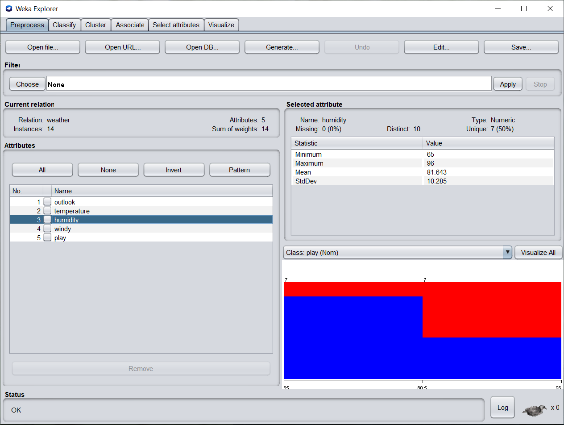
* Thuộc tính outlook:



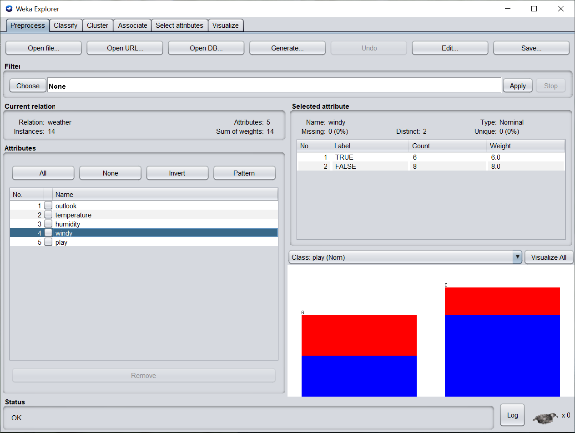
* Thuộc tính temperature:



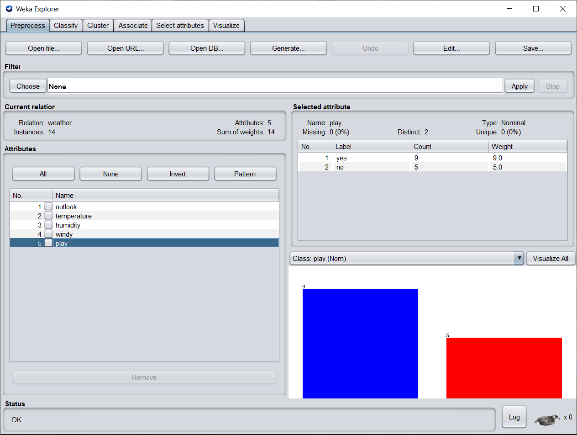
* Thuộc tính humidity:



* Thuộc tính windy:



* Thuộc tính play:



* 1. **Khám phá tập dữ liệu Tín dụng Đức:**

### Cài đặt tiền xử lý dữ liệu:

* 1. **Liệt kê các cột bị thiếu dữ liệu**
  2. **Đếm số dòng bị thiếu dữ liệu**
  3. **Điền giá trị bị thiếu**
  4. **Xóa các dòng bị thiếu dữ liệu với ngưỡng tỉ lệ thiếu cho trước**
  5. **Xóa các cột bị thiếu dữ liệu với ngưỡng tỉ lệ thiếu cho trước**
  6. **Xóa các mẫu bị trùng lặp**
  7. **Chuẩn hóa một thuộc tính numeric bằng phương pháp min-mã và Z-score**
  8. **Tính giá trị biểu thức thuộc tính:**

## IV. Đánh giá đồ án

### 1. Mức độ hoàn thành của các thành viên

|  |  |  |
| --- | --- | --- |
| **MSSV** | **Mức độ hoàn thành công việc** | **Đóng góp** |
| 18120078 |  |  |
| 18120253 |  |  |

### 2. Mức độ hoàn thành đồ án:

## VII. Nguồn tham khảo