

ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH  
TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN

----------



BÁO CÁO LAB 1  
PREPROCESSING

**Bộ môn:** Khai Thác Dữ Liệu Và Ứng Dụng  
**Giảng viên hướng dẫn:** Dương Nguyễn Thái Bảo

2020 - 2021

**MỤC LỤC**

I. Thông tin nhóm và phân công công việc .....	2
II. Các câu hỏi tự luận.....	2
1. Yêu cầu 1: Cài đặt Weka.....	2
2. Yêu cầu 2: Làm quen với Weka.....	4
2.1. Đọc dữ liệu vào Weka .....	4
2.2. Khám phá tập dữ liệu Weather.....	13
2.3. Khám phá tập dữ liệu Tín dụng Đức.....	19
III. Cài đặt tiền xử lý dữ liệu.....	25
IV. Tài liệu tham khảo.....	27

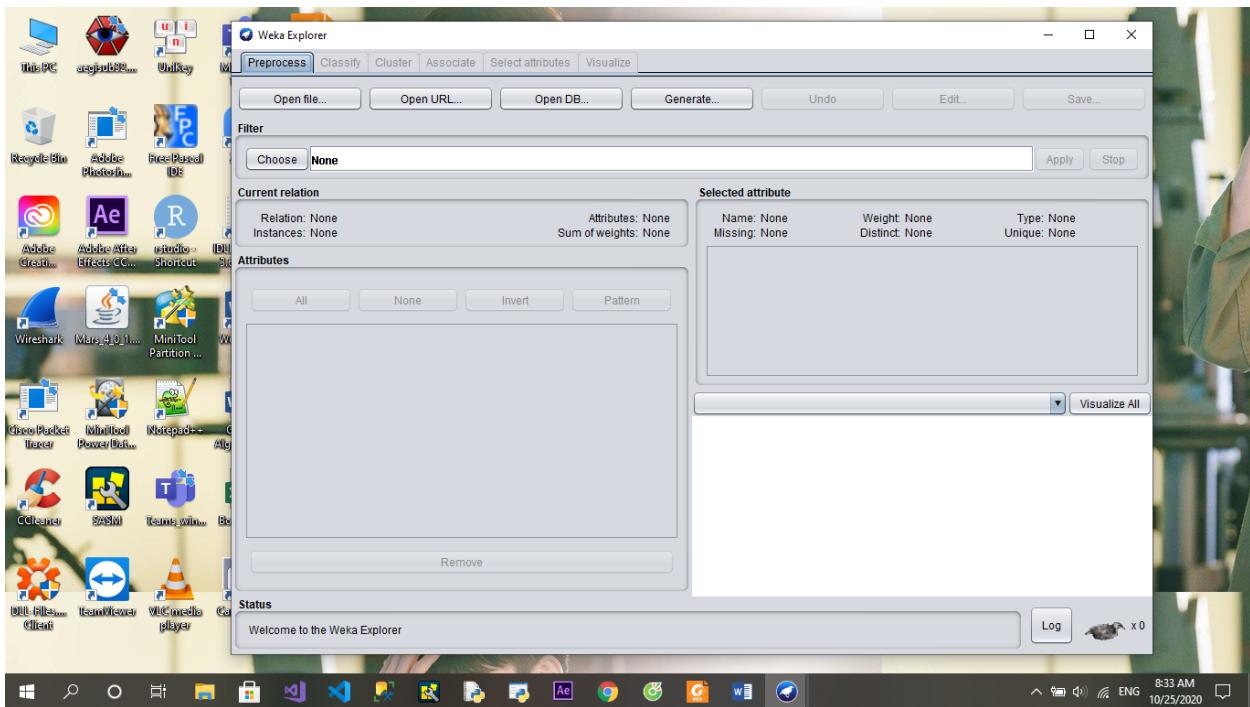
**I. THÔNG TIN NHÓM VÀ PHÂN CÔNG CÔNG VIỆC**

Họ và tên	MSSV	Công việc
Lê Hoàng Phương Nhi	18120496	1. Cài đặt Weka (giải thích ngắn gọn ý nghĩa 5 tab trong giao diện Explorer của Weka) 2.2. Khám phá tập dữ liệu Weather 2.3. Khám phá tập dữ liệu Tín dụng Đức (mục 1, 2) 3. Cài đặt tiền xử lý dữ liệu (Các chức năng: 1, 3, 5, 7)
Lê Thị Như Quỳnh	18120530	1. Cài đặt Weka (Giải thích ý nghĩa các nhóm điều khiển Current relation, Attributes và Selected attribute trong tab Preprocess) 2.1. Đọc dữ liệu vào Weka 2.3. Khám phá tập dữ liệu Tín dụng Đức (mục 3, 4) 3. Cài đặt tiền xử lý dữ liệu (Các chức năng: 2, 4, 6, 8)

**II. CÁC CÂU HỎI TỰ LUẬN****1. Yêu cầu 1: Cài đặt Weka**

*Câu hỏi báo cáo:*

\* *Giao diện chức năng Explorer:*



\* **Ý nghĩa các nhóm điều khiển Current relation, Attributes và Selected attribute trong tab Preprocess:**

- Current relation: cung cấp thông tin về tập dữ liệu
- + Relation: hiển thị tên của mối quan hệ
- + Instances: số lượng của các mẫu bản ghi trong dữ liệu
- + Attributes: số lượng của các thuộc tính trong dữ liệu
- Attributes: chứa các thuộc tính của quan hệ được thể hiện theo danh sách gồm 3 cột
- + No: số thứ tự của thuộc tính được sắp xếp trong tệp dữ liệu
- + Ô tích chọn thuộc tính: ô này cho phép sử dụng được tùy chọn quyết định thuộc tính có được xuất hiện trong mối quan hệ hay không.
- + Name: hiển thị tên thuộc tính theo danh sách như ở trong tệp dữ liệu
- Selected Attribute: hiển thị các đặc tính, thông tin riêng của từng thuộc tính attributes
- + Name: tên của thuộc tính, giống như tên trong danh sách thuộc tính attributes
- + Type: loại thuộc tính, thường hiển thị dưới hai dạng là Nominal hoặc Numeric
- + Missing: số lượng và tỷ lệ phần trăm của các cá thể trong dữ liệu mà thuộc tính này bị thiếu
- + Distinct: số lượng giá trị khác nhau mà dữ liệu chứa cho thuộc tính

+ Unique: số lượng và tỷ lệ phần trăm của các cá thể trong dữ liệu có giá trị cho thuộc tính này mà có trường hợp nào khác có.

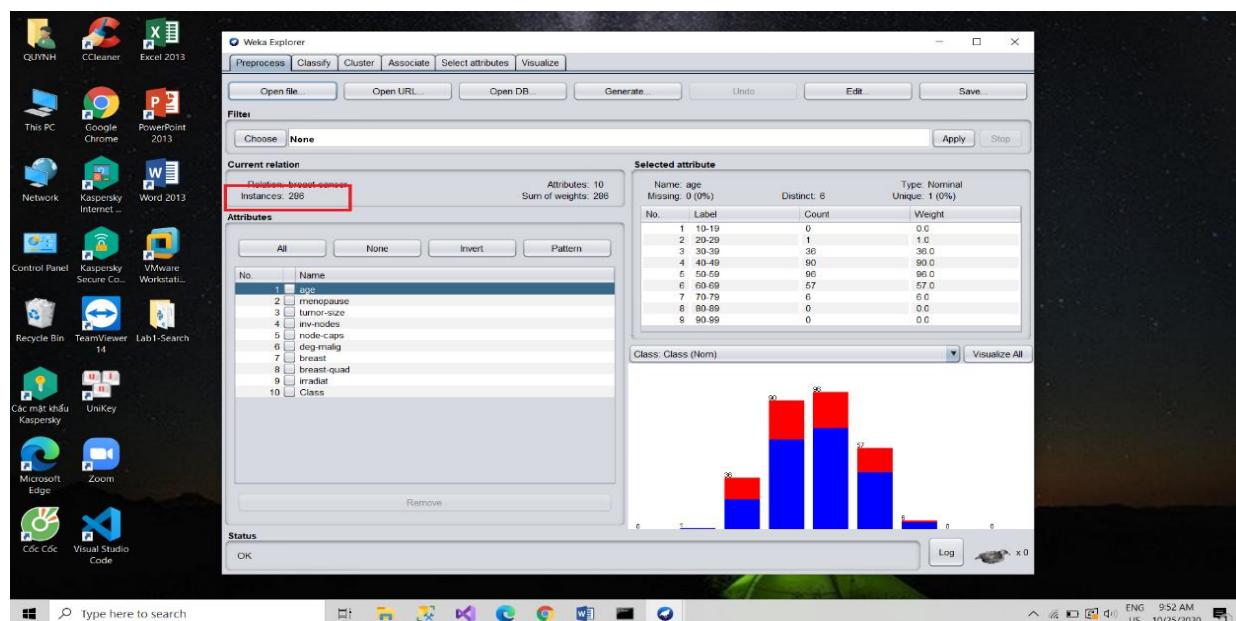
**\* Ý nghĩa 5 tab trong giao diện Explorer:**

- Preprocess (tiền xử lý dữ liệu): Để chọn và thay đổi (xử lý) dữ liệu làm việc
- Classify: Để huấn luyện và kiểm tra các mô hình học máy (phân loại, hoặc hồi quy/dự đoán)
- Cluster: Để học các nhóm từ dữ liệu (phân cụm)
- Associate: Để khám phá các luật kết hợp từ dữ liệu
- Select attributes: Để xác định và lựa chọn những thuộc tính liên quan nhất trong tập dữ liệu
- Visualize: Để xem (hiển thị) biểu đồ tương tác 2 chiều đối với dữ liệu

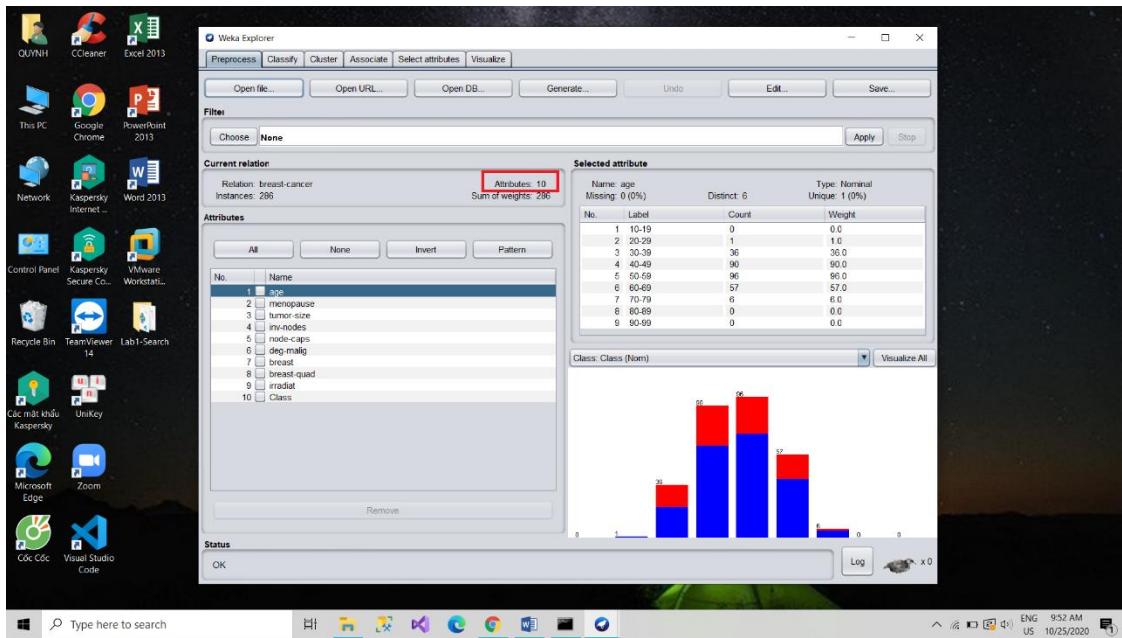
## **2. Yêu cầu 2: Làm quen với Weka**

### **2.1. Đọc dữ liệu vào Weka**

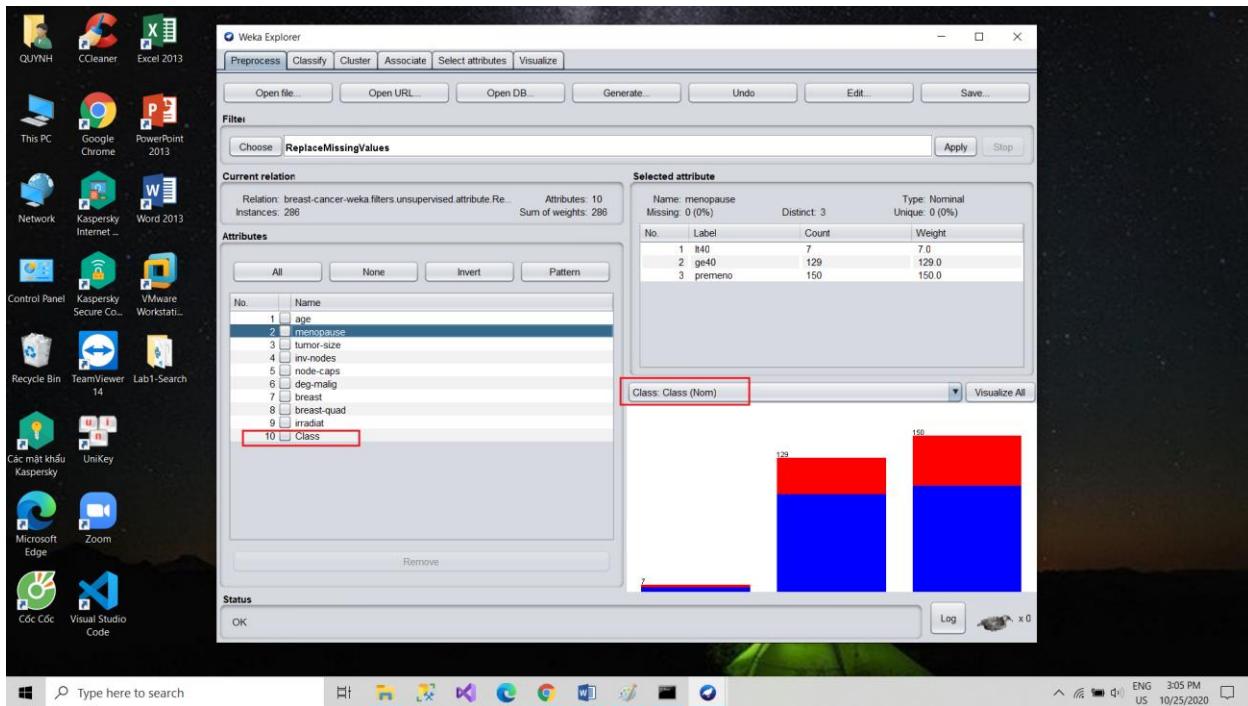
**2.1.1.** Tập dữ liệu có 286 mẫu:

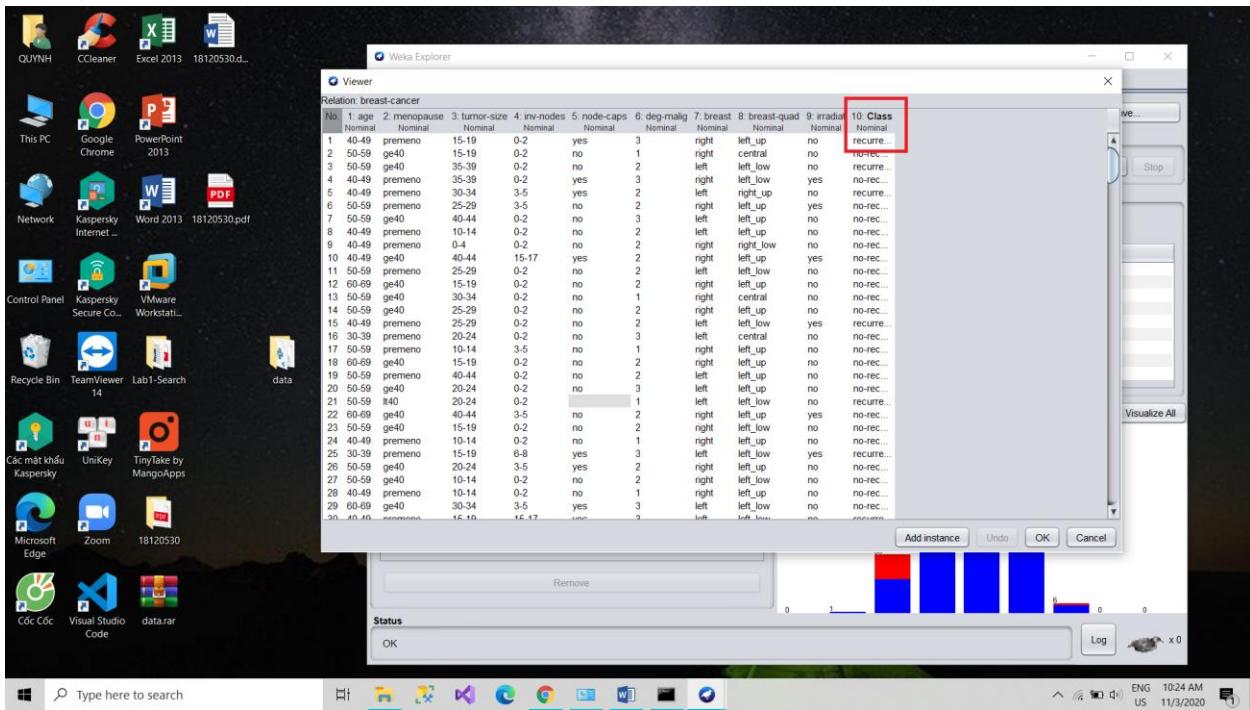


**2.1.2.** Tập dữ liệu có 10 thuộc tính:

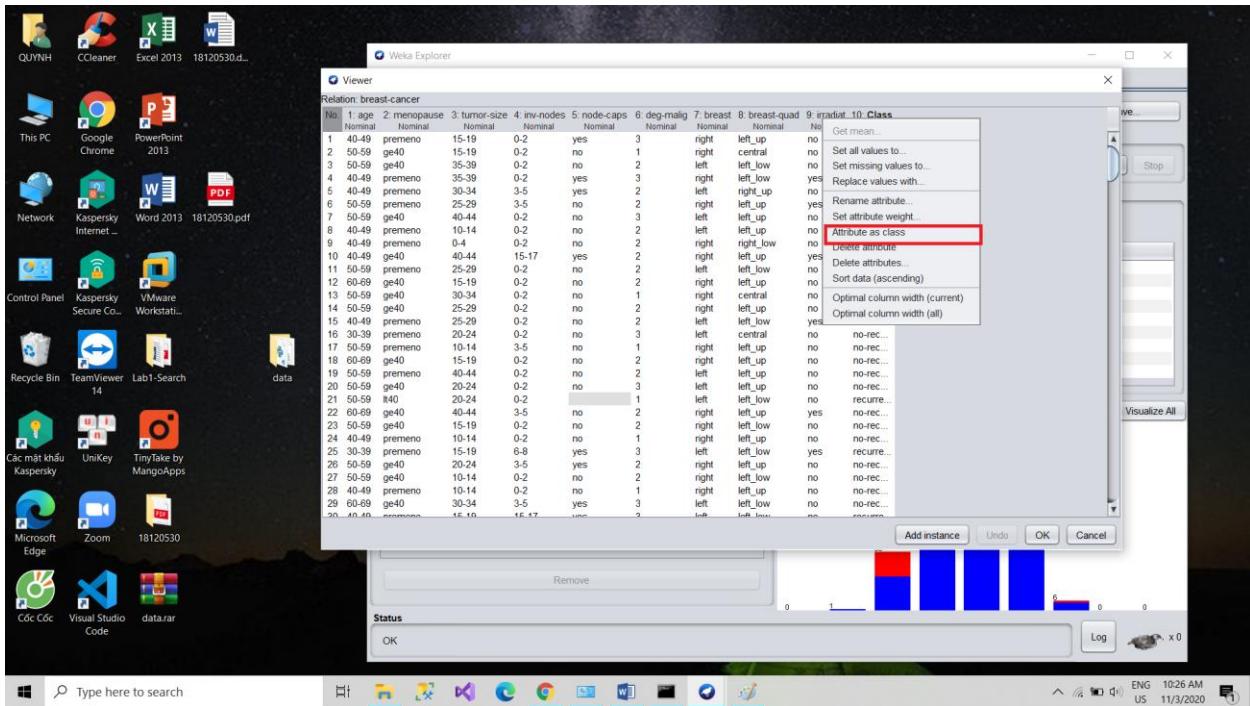


### 2.1.3. Thuộc tính làm lớp là class (Mặc định là thuộc tính cuối cùng)





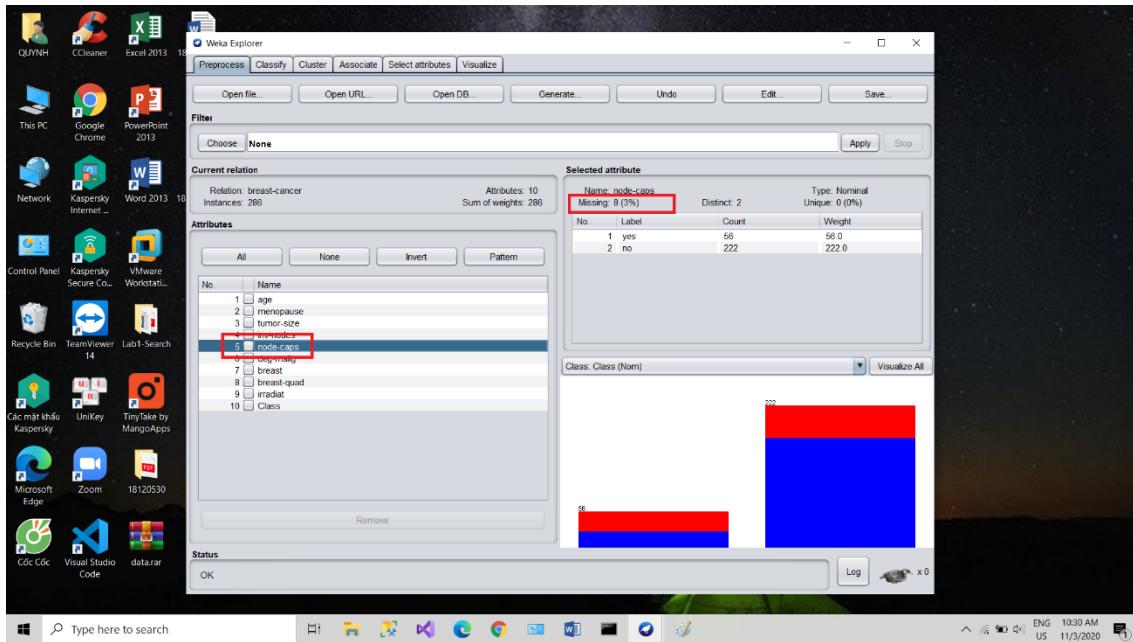
- Có thể thay đổi thuộc tính dùng làm lớp. Bằng cách kích chuột phải vào thuộc tính muốn chọn làm class, chọn attribute as class:



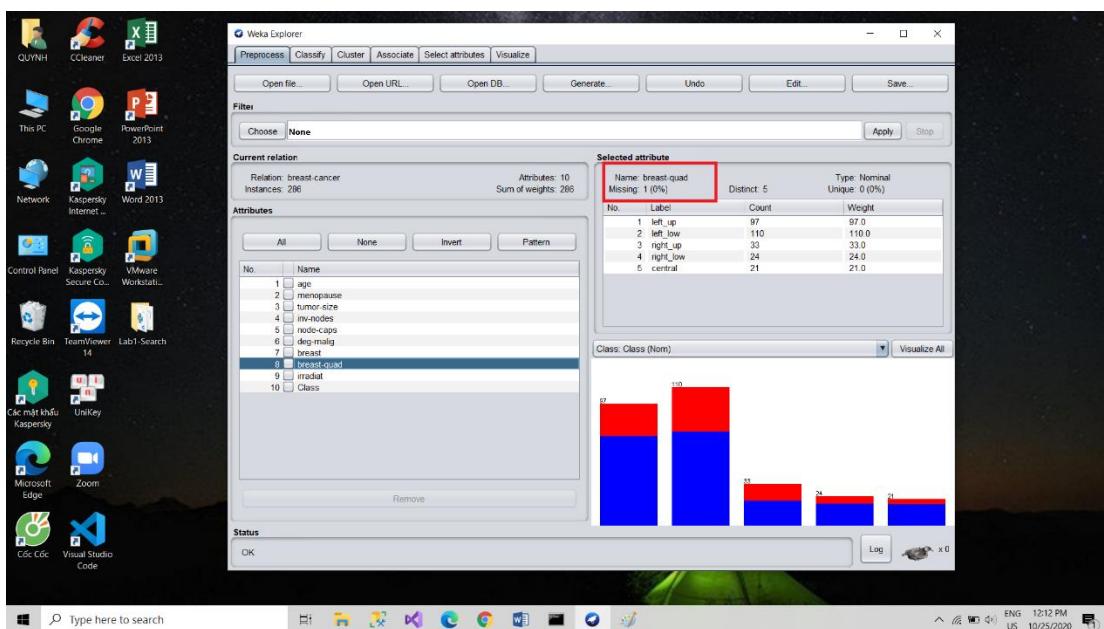
### 2.1.4.

- Các thuộc tính thiếu dữ liệu:

+ node-caps: mất 8 dữ liệu



+ breast-quad: mất 1 dữ liệu



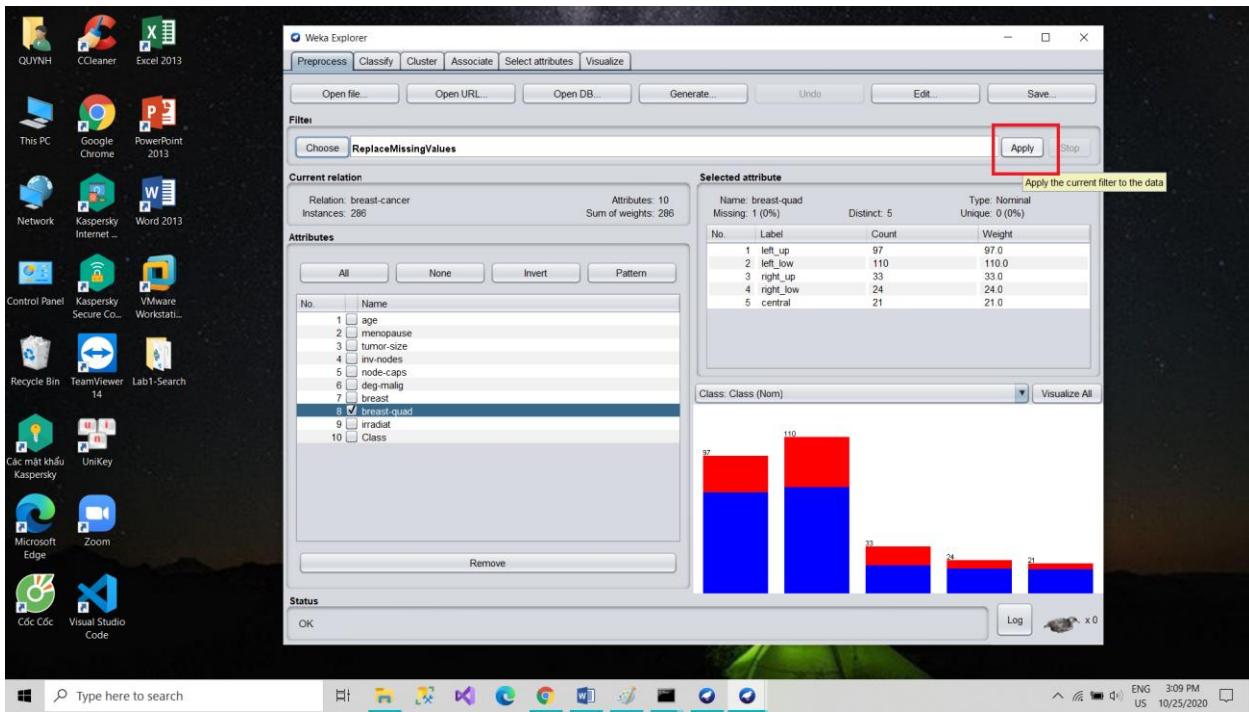
- Thuộc tính mất nhiều dữ liệu nhất: node-caps (mất 8)

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

- Cách xử lí khi bị mất dữ liệu:

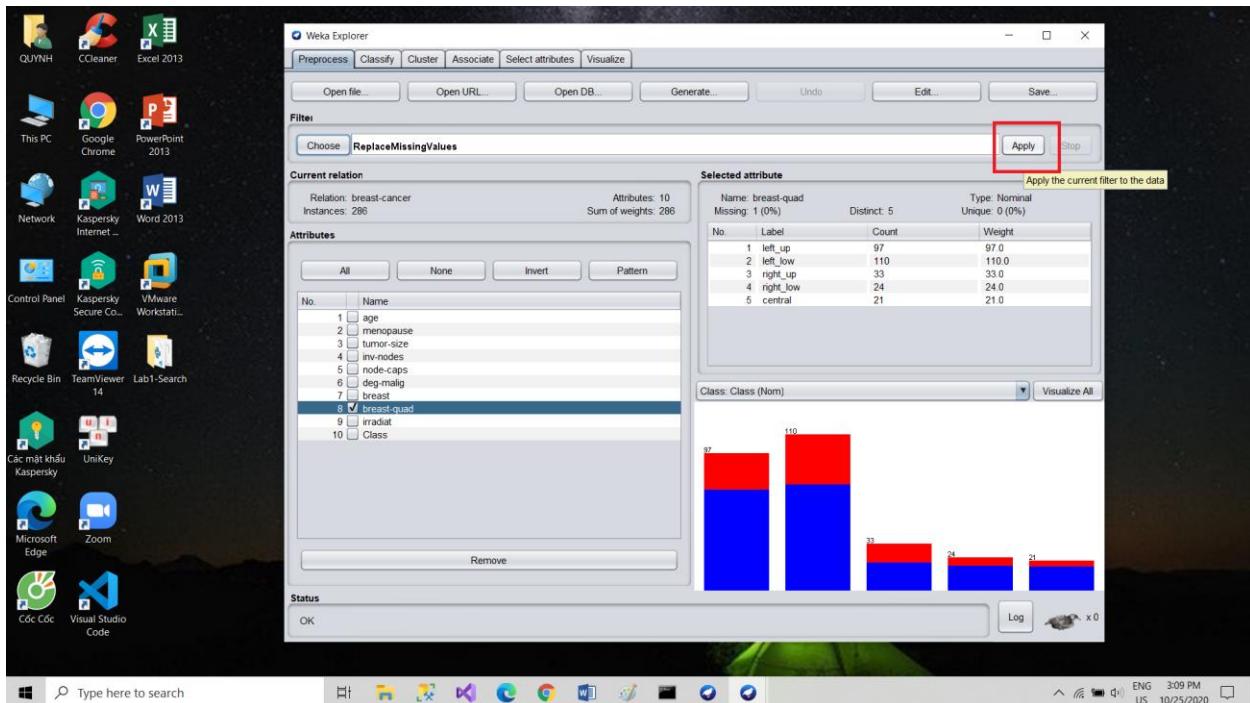
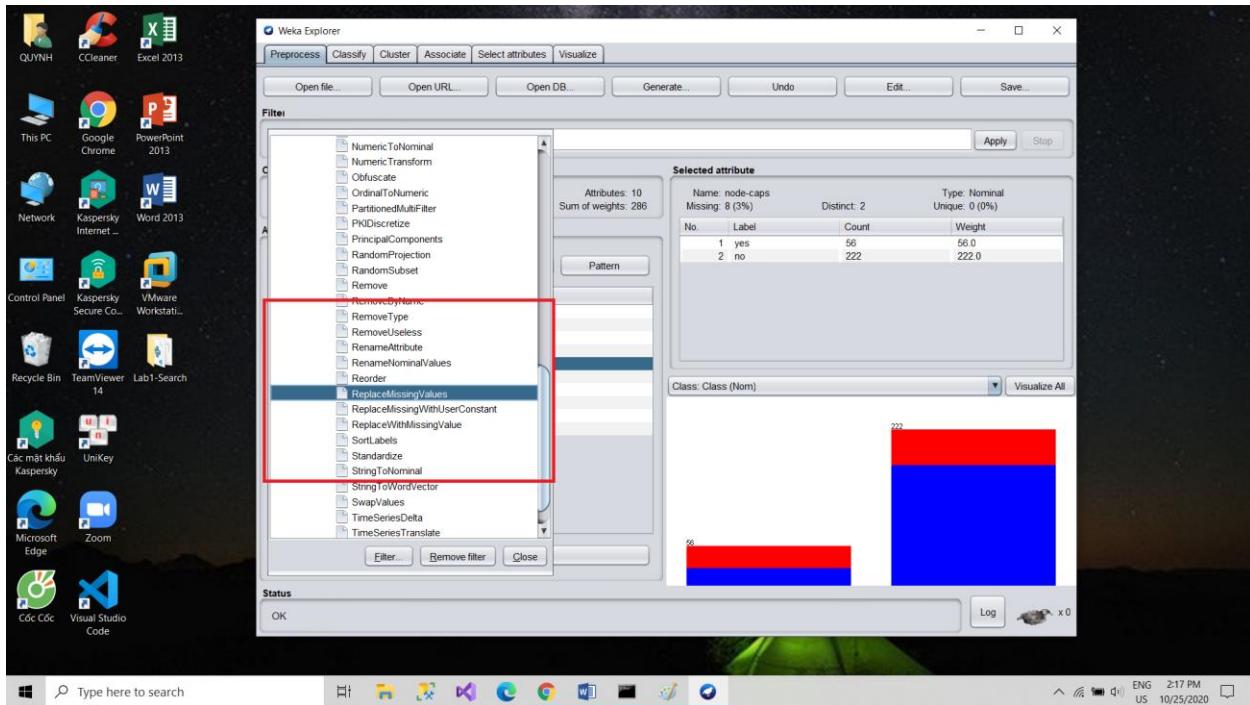
Cách 1:

- + Thay các giá trị thiếu bằng giá trị trung bình (mean) đối với các thuộc tính kiểu dữ liệu số (numeric)
- + Thay bằng mode đối với thuộc tính định danh (nominal)



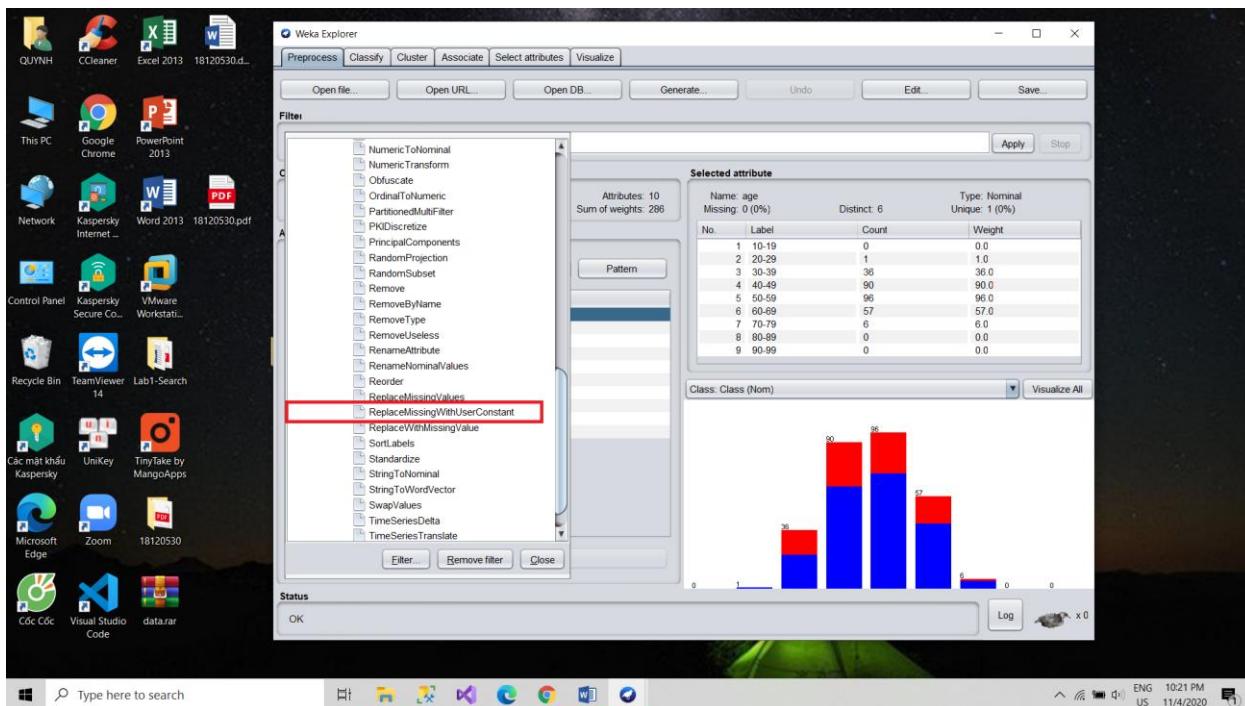
+ Bước làm:

Trong tab: Filter -> bấm nút choose -> chọn filters -> unsupervised -> attribute -> replaceMissingValues -> nhấn apply

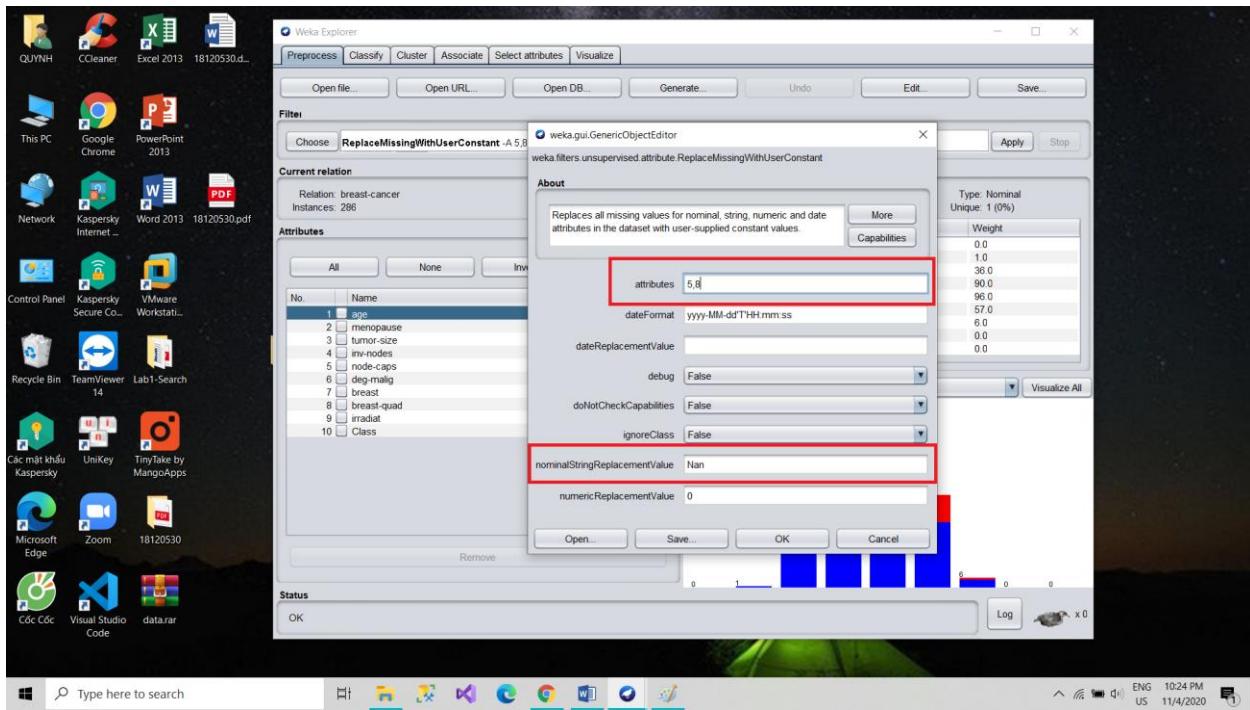


Cách 2:

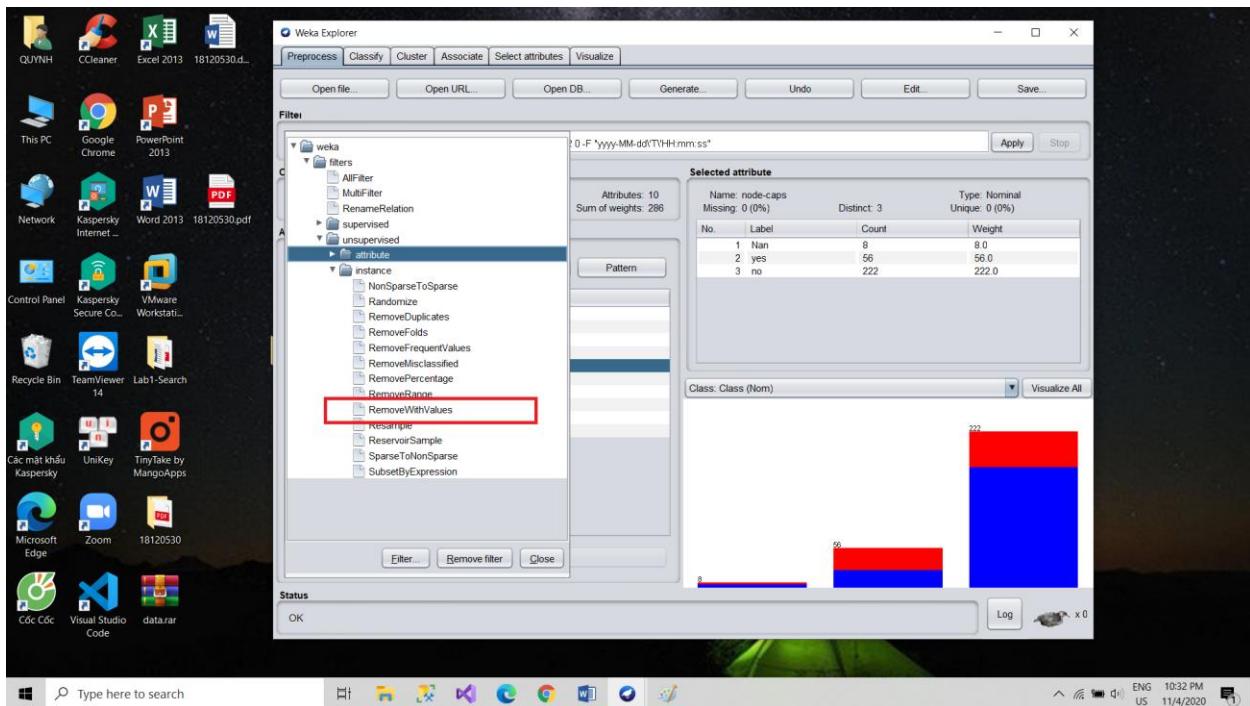
- Thay thế các giá trị bị thiếu bằng ‘Nan’
- Xóa những dòng dữ liệu có giá trị Nan
- Bước làm:
  - + Trong tab: Filter -> bấm nút choose -> chọn filters -> unsupervised -> attribute -> replaceMissingValuesWithUserConstant.



- + Click chuột vào Filter.
- + Attributes: là thứ tự thuộc tính bị mất giá trị. Cụ thể trong bài ta điền 5,8
- + NominalStringRepalcementValue: điền Nan.
- + Chọn Ok.
- + Chọn Apply

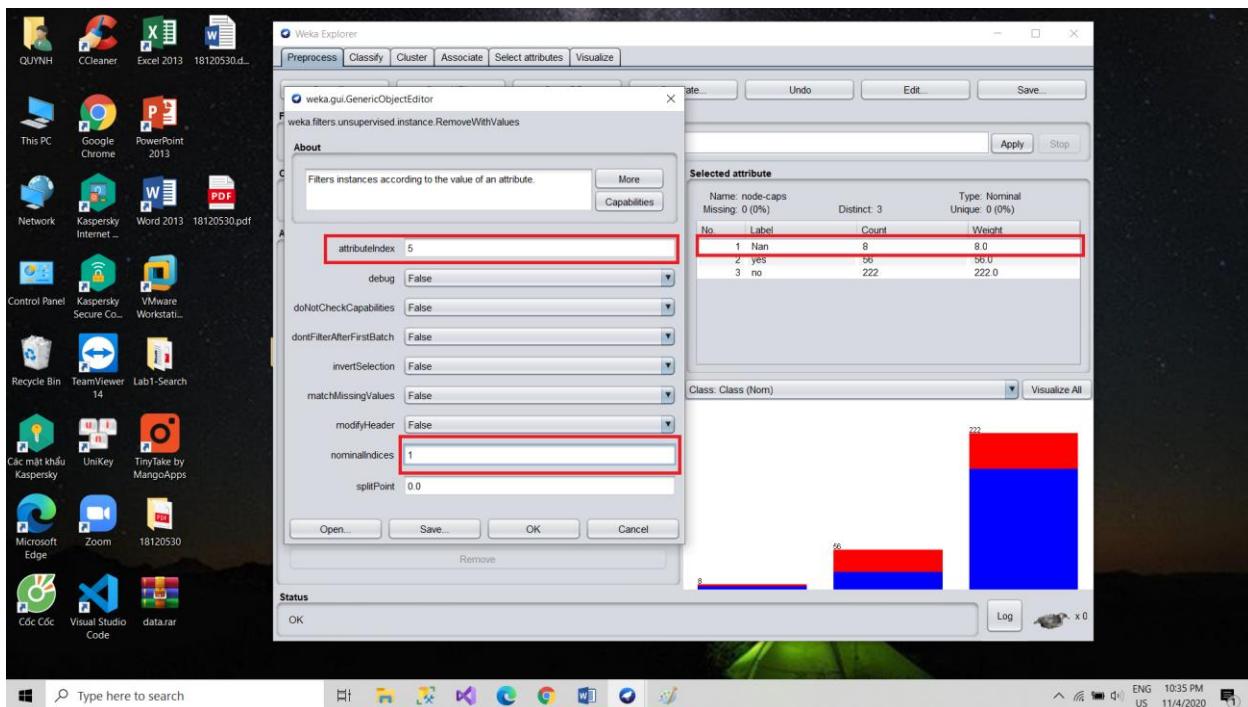


- Trong tab: Filter -> bấm nút choose -> chọn filters -> unsupervised -> instance -> RemoveWithValues



- Chọn Filter:

- + AttributeIndex: thứ tự thuộc tính từng bị mất giá trị, ở đây ta nhập 5
- + NominalIndices: 1 (thuộc tính Nan)
- + Chọn OK
- + Nhấn Apply
- + Làm tương tự với thuộc tính thứ 8 (breast-quad).

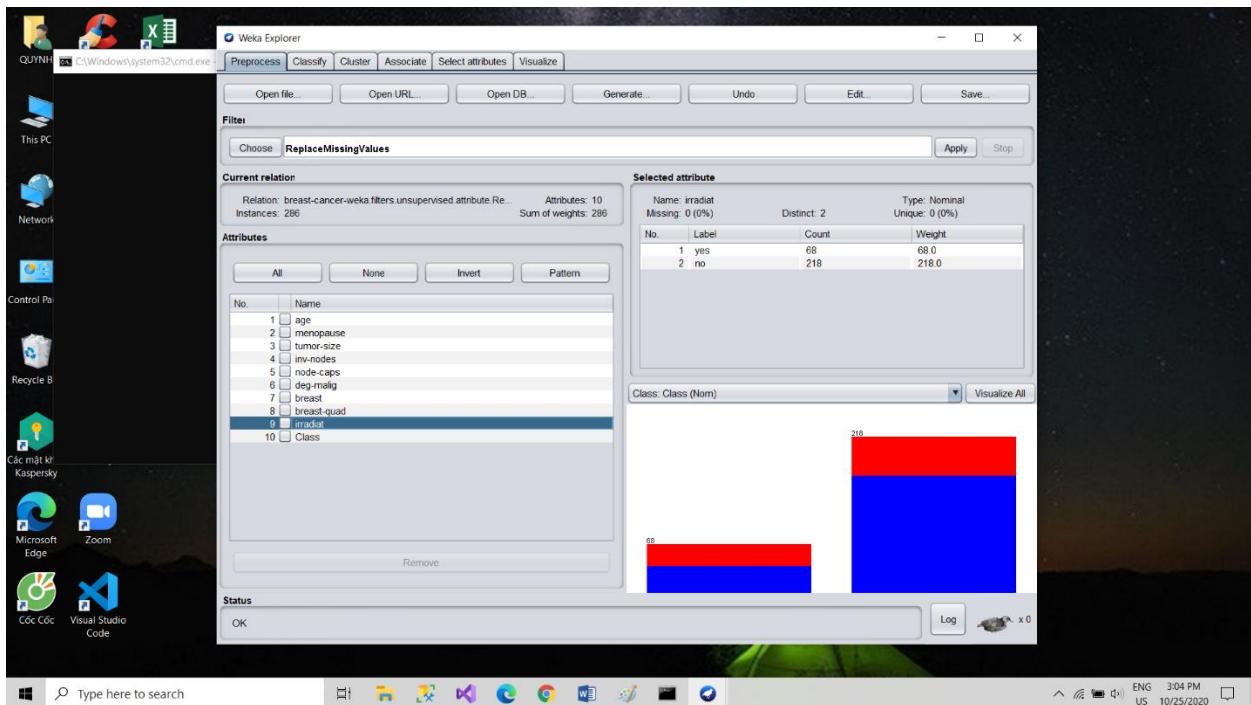


### 2.1.5.

Chọn một thuộc tính bên attributes thì tại selected attribute và biểu đồ sẽ thể hiện cho thuộc tính đó.

Ý nghĩa đồ thị là biểu diễn trực quan từng thuộc tính. Cho biết thuộc tính gồm các đặc tính nào (số cột), mỗi màu trên cột biểu diễn một lớp.

Khi chọn attributes là irradiat, thì cột bên trái thể hiện đặc tính yes, cột bên phải thể hiện đặc tính no. Trên cột bên trái màu đỏ thể hiện lớp recurrence-events, màu xanh thể hiện lớp no-recurrence-events.



Đặt tên: Đồ thị phân bố giá trị của thuộc tính theo nhãn

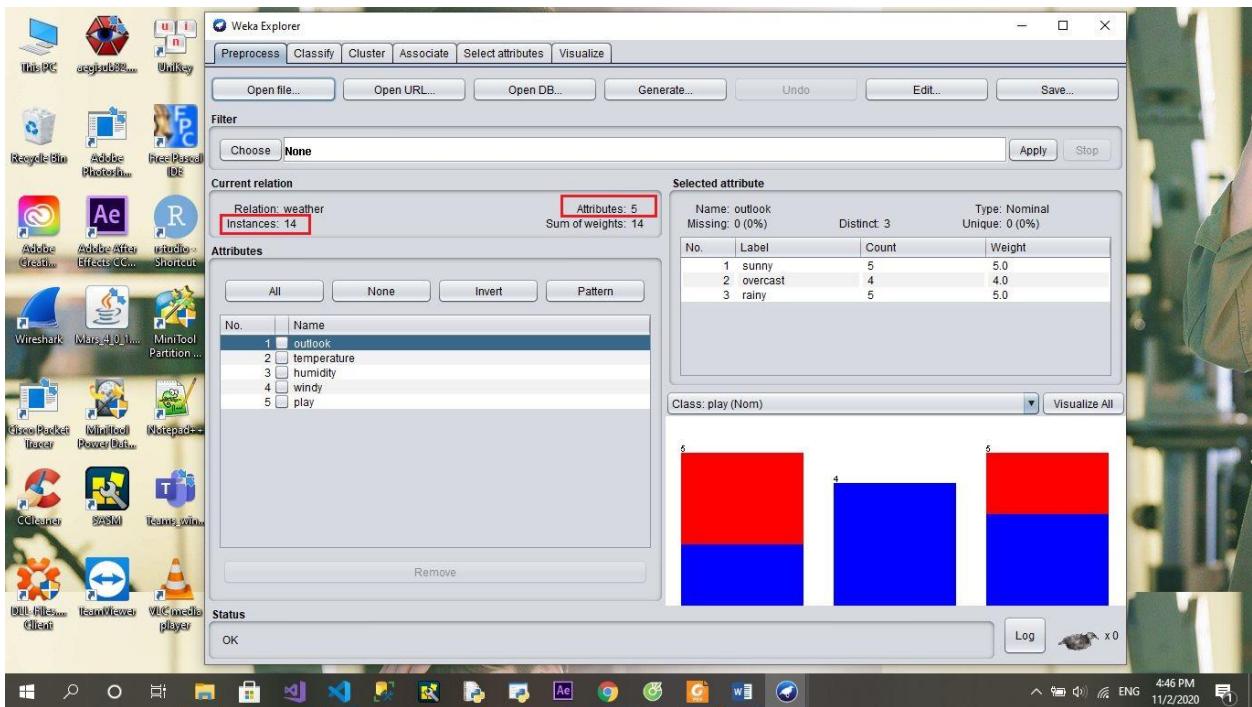
Màu đỏ thể hiện lớp recurrence-events

Màu xanh thể hiện lớp no-recurrence-events

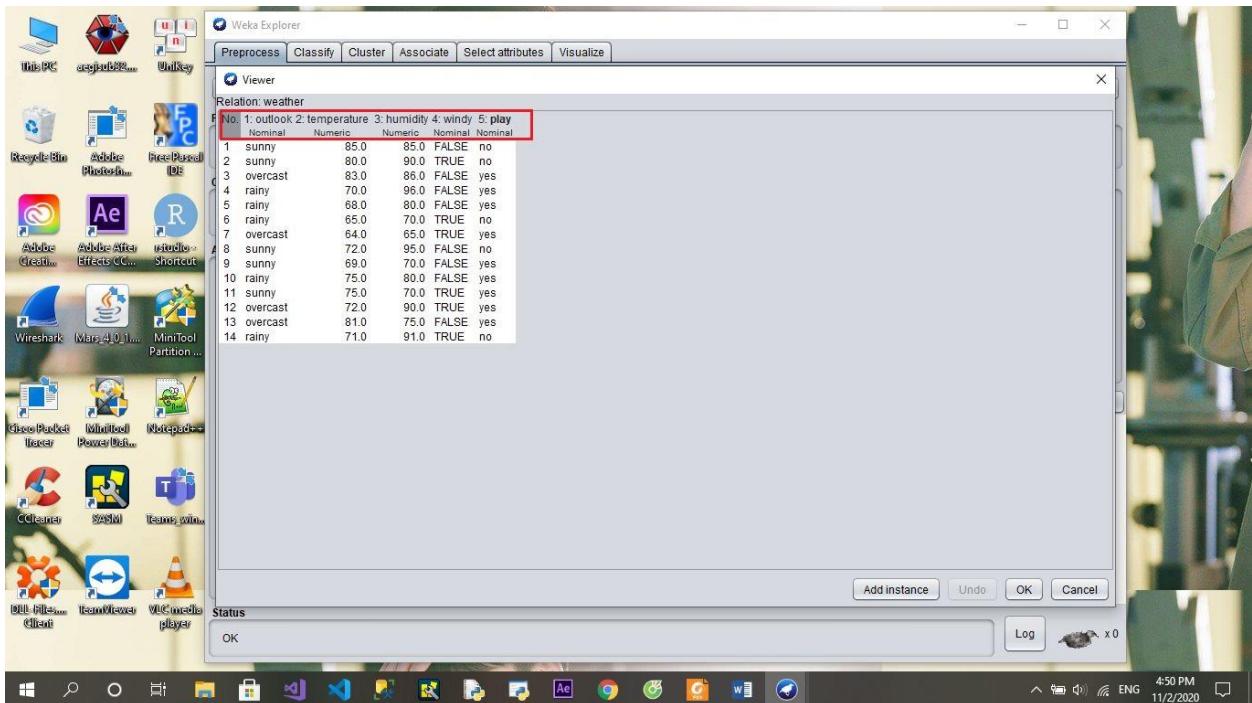
## 2.2. Khám phá tập dữ liệu Weather

### 2.2.1.

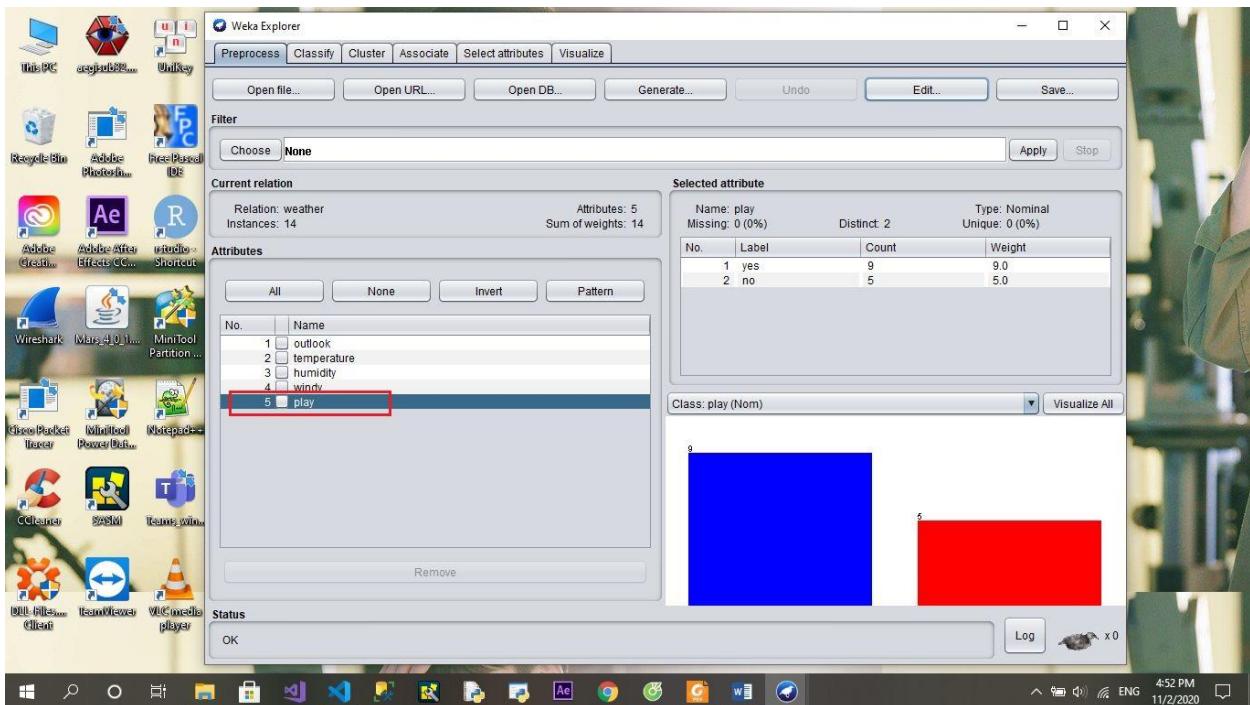
- Tập dữ liệu có 5 thuộc tính, 14 mẫu.



- Phân loại thuộc tính:
- + Categorical: outlook, windy, play
- + Numeric: temperature, humidity



- Thuộc tính play là lớp (mặc định là thuộc tính cuối cùng)

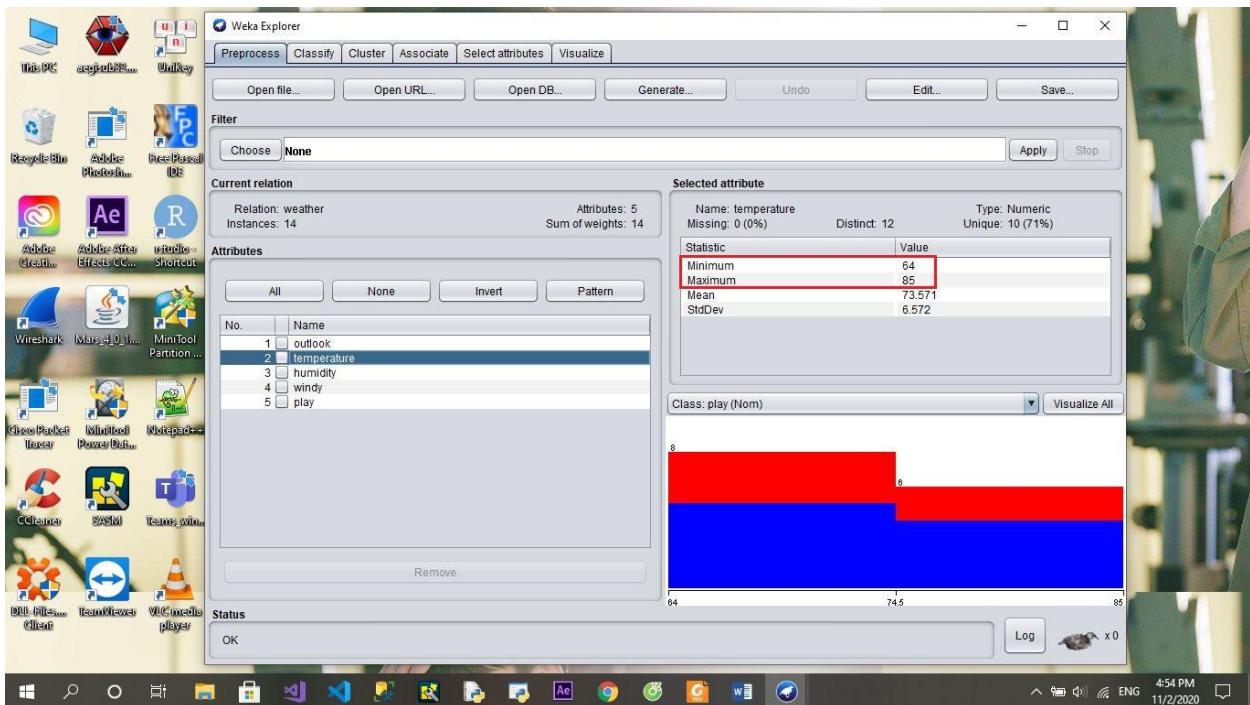


## 2.2.2.

- Five-number summary:

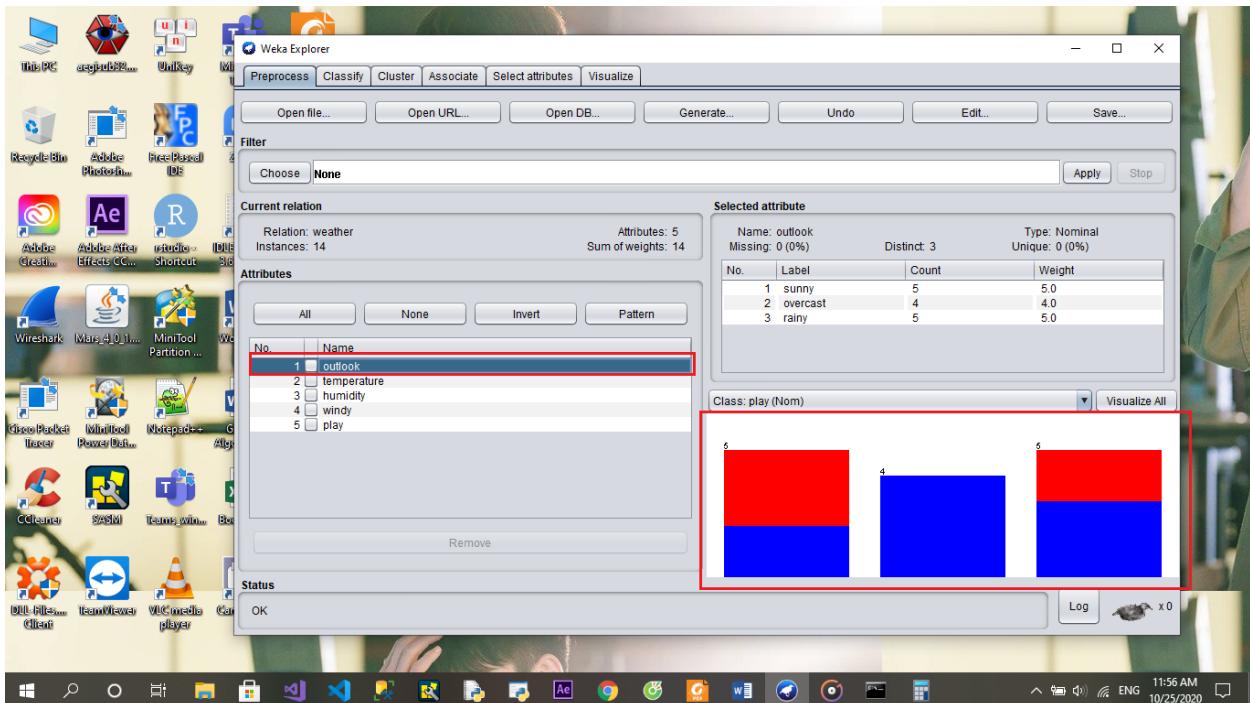
	<b>Minimum</b>	<b>Q1</b>	<b>Median</b>	<b>Q3</b>	<b>Maximum</b>
<b>Temparature</b>	64	69	72	80	85
<b>Humidity</b>	65	70	92.5	90	96

- Weka chỉ cung cấp giá trị Minimum và Maximum còn các giá trị còn lại thì phải tự tính.

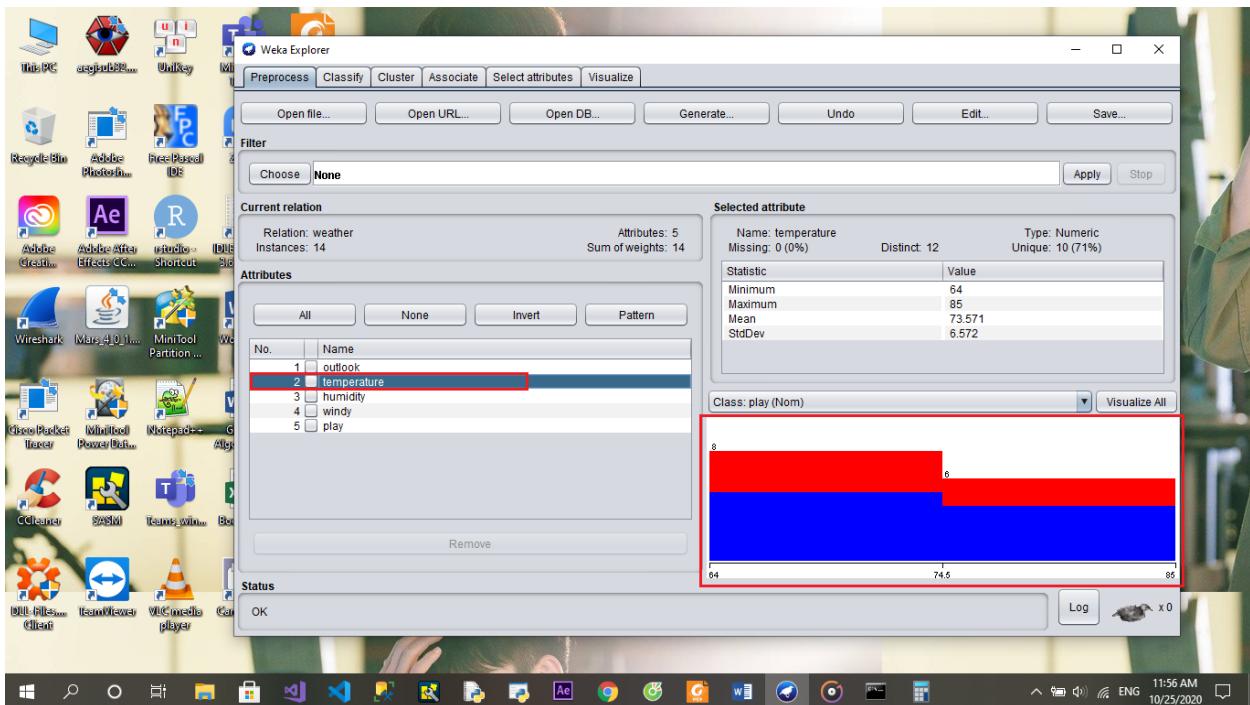


### 2.2.3. Xem xét các thuộc tính khác của dataset dưới dạng đồ thị

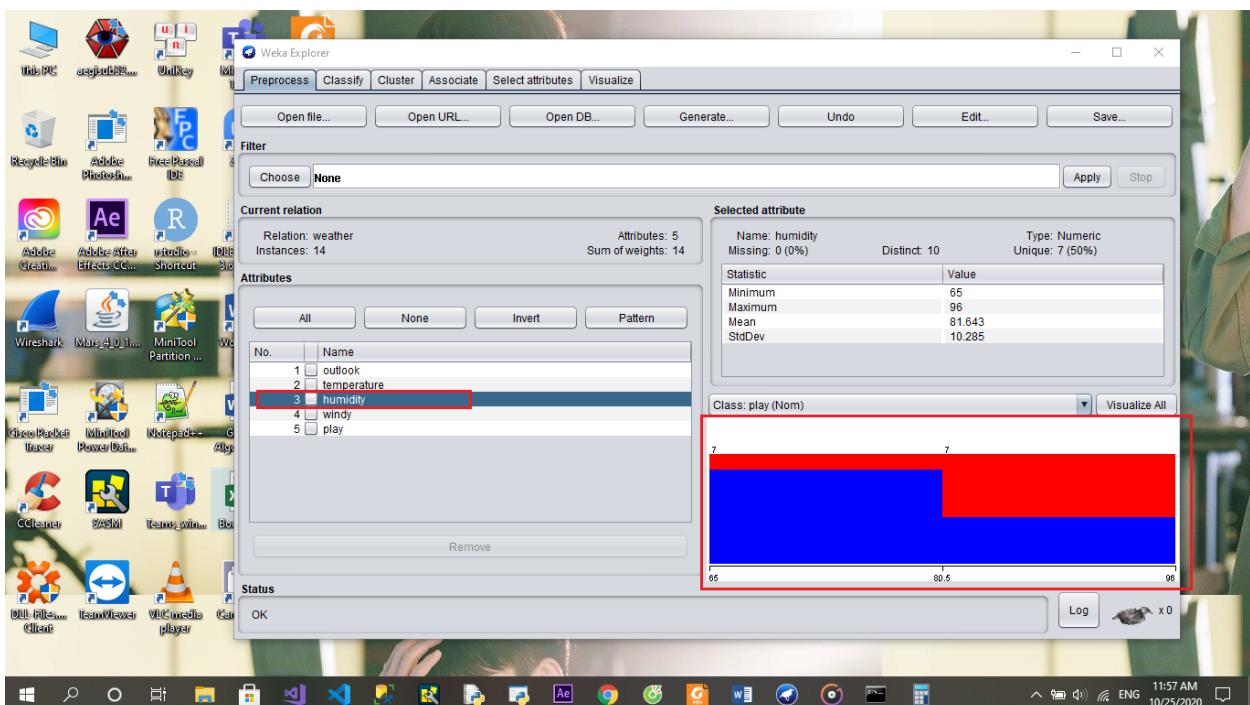
- Thuộc tính outlook:



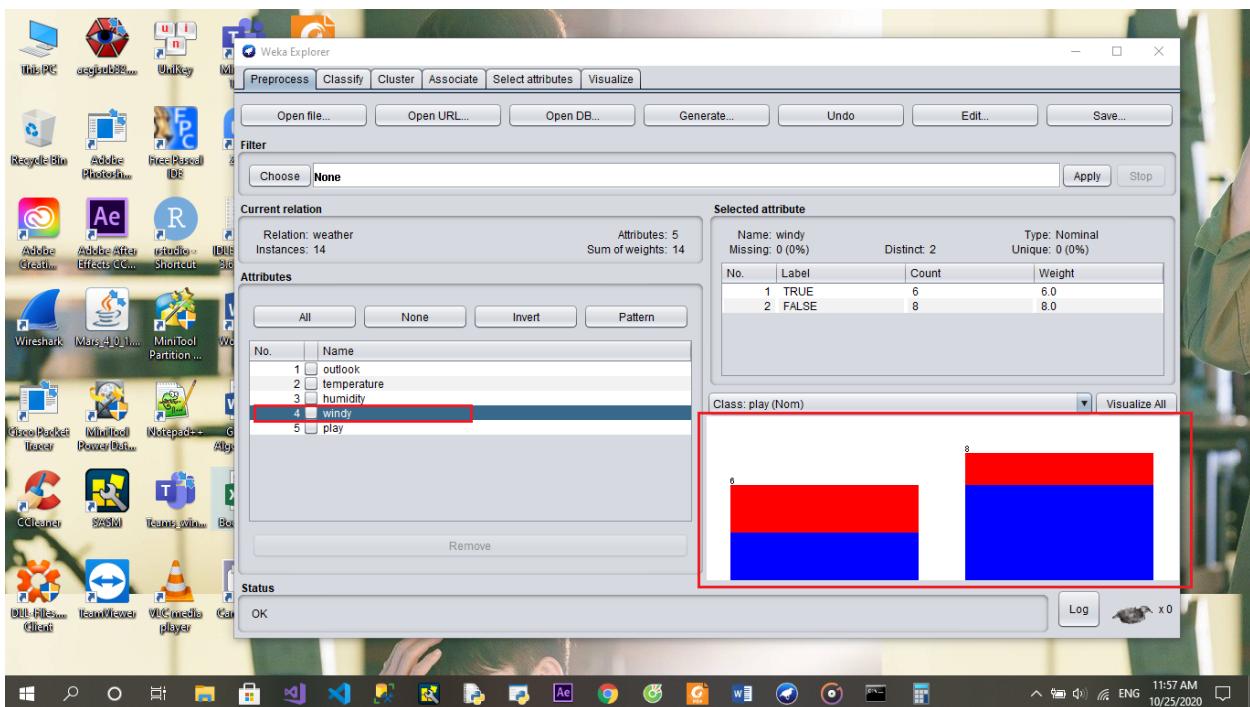
- Thuộc tính temperature:



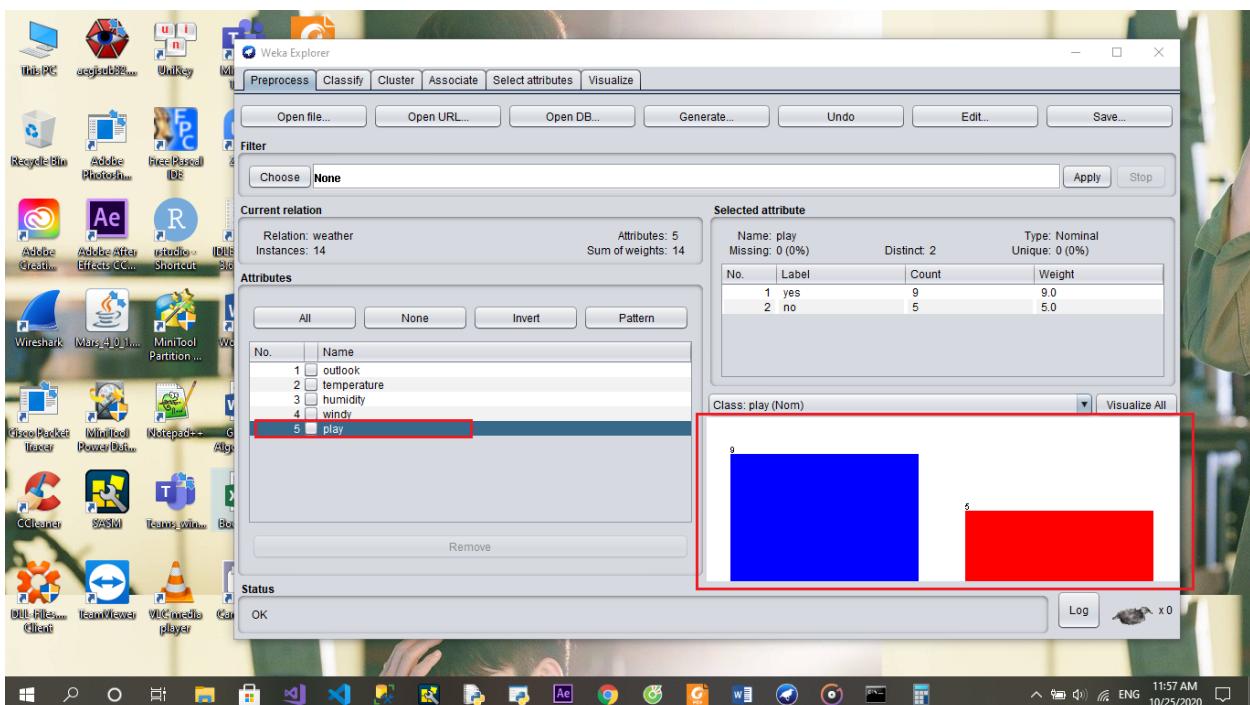
- Thuộc tính humidity:



- Thuộc tính windy:



- Thuộc tính play:



## 2.2.4.

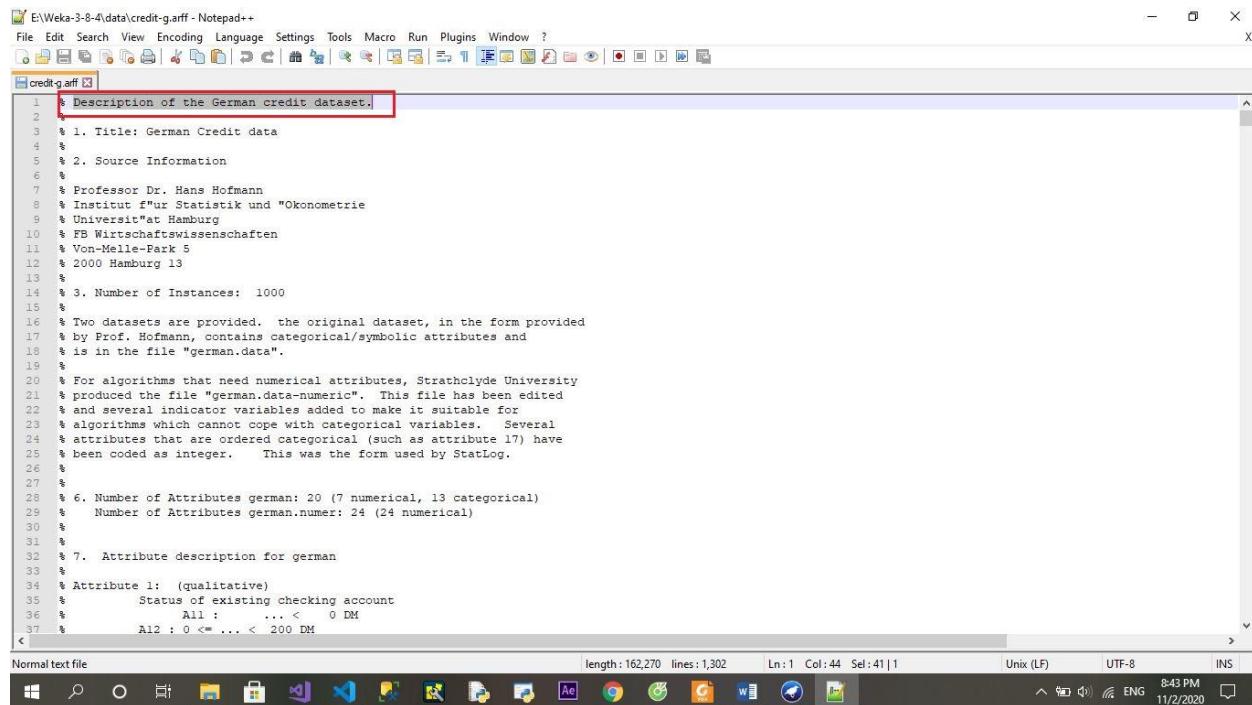
- Thuật ngữ sử dụng trong textbook để đặt tên cho các đồ thị là: Scatter Plot

- Theo em những cặp thuộc tính khác nhau có vẻ tương quan với nhau là: (play, humidity), (windy, temperature), (outlook, temperature).

### 2.3. Khám phá tập dữ liệu Tín dụng Đức

#### 2.3.1.

- Nội dung của phần ghi chú (comment) trong credit-g.arff (khi mở bằng 1 text editor bất kỳ) nói về: Mô tả bộ dữ liệu Tín dụng Đức bao gồm tiêu đề, nguồn thông tin, các thuộc tính,...

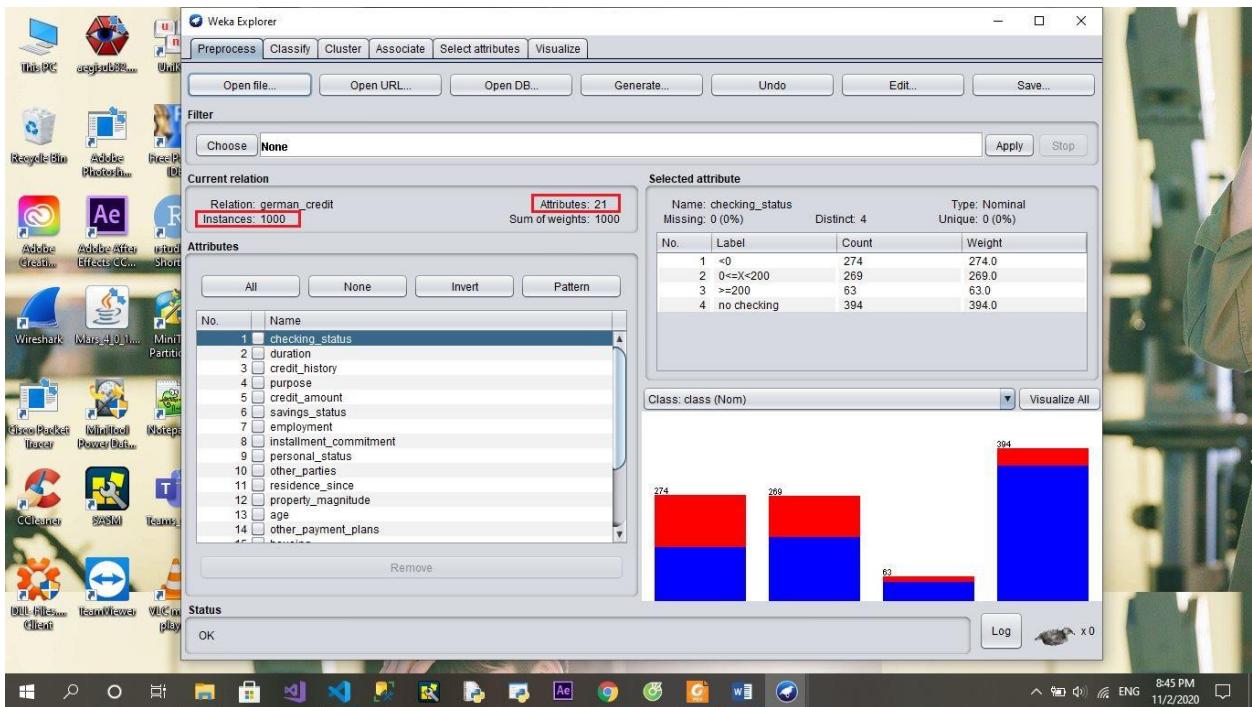


```

1 % Description of the German credit dataset.
2 %
3 % 1. Title: German Credit data
4 %
5 % 2. Source Information
6 %
7 % Professor Dr. Hans Hofmann
8 % Institut f"ur Statistik und "Okonometrie
9 % Universit"at Hamburg
10 % FB Wirtschaftswissenschaften
11 % Von-Melle-Park 5
12 % 2000 Hamburg 19
13 %
14 % 3. Number of Instances: 1000
15 %
16 % Two datasets are provided. the original dataset, in the form provided
17 % by Prof. Hofmann, contains categorical/symbolic attributes and
18 % is in the file "german.data".
19 %
20 % For algorithms that need numerical attributes, Strathclyde University
21 % produced the file "german.data-numeric". This file has been edited
22 % and several indicator variables added to make it suitable for
23 % algorithms which cannot cope with categorical variables. Several
24 % attributes that are ordered categorical (such as attribute 17) have
25 % been coded as integer. This was the form used by StatLog.
26 %
27 %
28 % 6. Number of Attributes german: 20 (7 numerical, 13 categorical)
29 %    Number of Attributes german.numer: 24 (24 numerical)
30 %
31 %
32 % 7. Attribute description for german
33 %
34 % Attribute 1: (qualitative)
35 %      Status of existing checking account
36 %      All : ... < 0 DM
37 %      A12 : 0 <= ... < 200 DM

```

- Tập dữ liệu có 1000 mẫu và 21 thuộc tính.

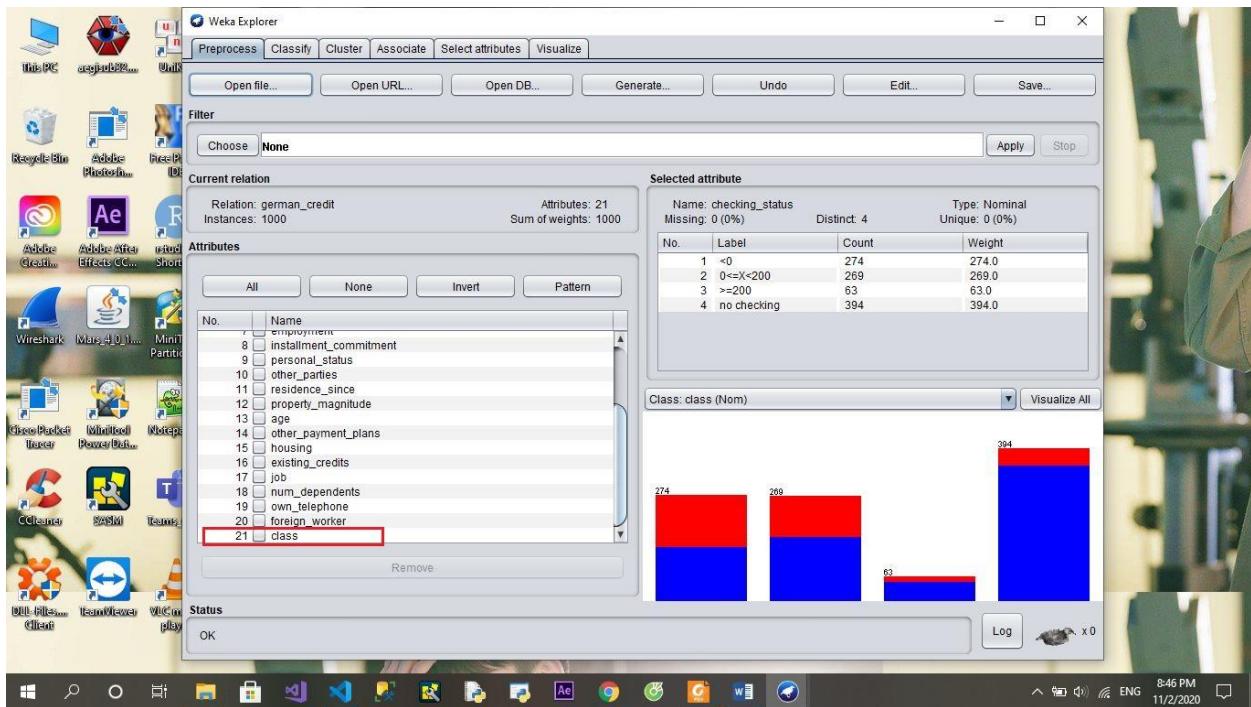


- Mô tả 5 thuộc tính:

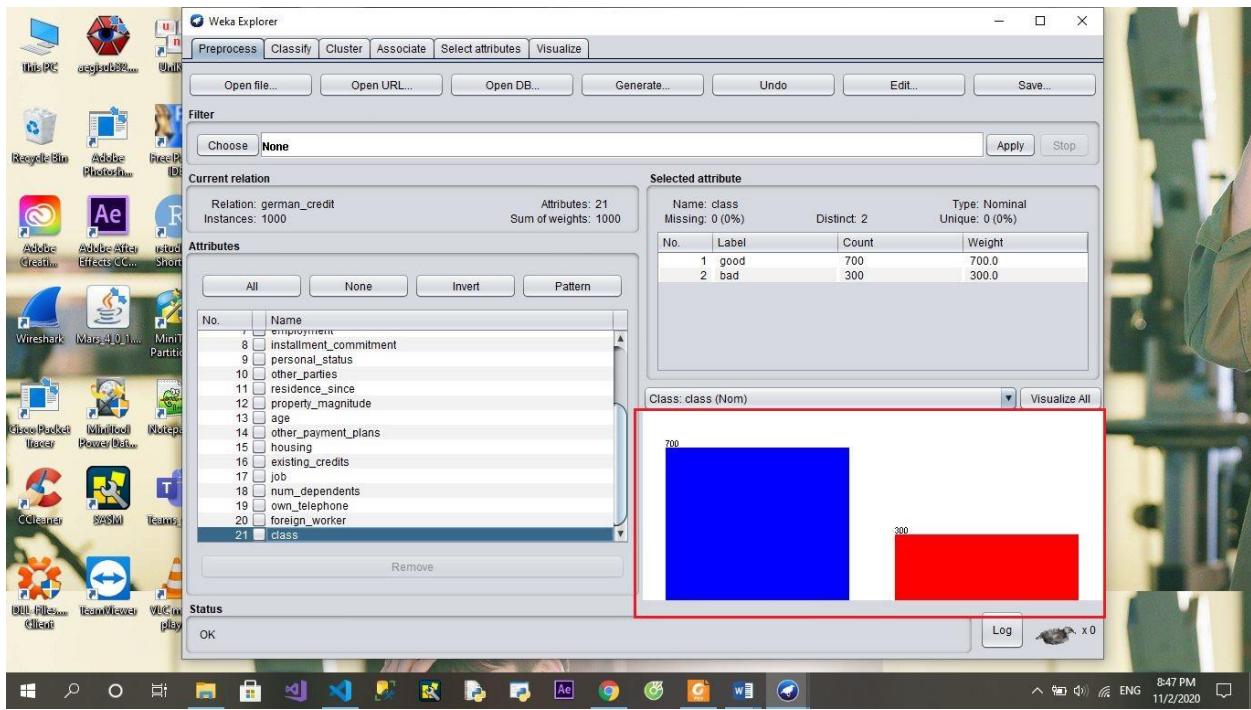
- + checking\_status: thuộc tính này có 4 giá trị ( $X < 0$ ,  $0 \leq X < 200$ ,  $X \geq 200$ , no checking) và số lượng tương ứng của 4 giá trị này là (274, 269, 63, 394)
- + credit\_history: thuộc tính này có 5 giá trị (no credits/all paid, all paid, existing paid, delayed previously, critical/other existing credit) và số lượng tương ứng của 5 giá trị này là (40, 49, 530, 88, 293)
- + purpose: thuộc tính này có 6 giá trị (new car, used car, furniture/equipment, radio/tv, domestic appliance, repairs) và số lượng tương ứng của 6 giá trị này là (234, 103, 181, 280, 12, 22)
- + foreign\_worker: thuộc tính này có 2 giá trị (yes, no) và số lượng tương ứng của 2 giá trị này là (963, 37)
- + age: thuộc tính này gồm các giá trị số thực với giá trị nhỏ nhất là 19, lớn nhất là 75, giá trị trung bình là 35.546 và độ lệch chuẩn là 11.375.

### 2.3.2.

- Tên của thuộc tính lớp: class (mặc định là thuộc tính cuối cùng)



- Đánh giá phân bố của các lớp: lệch về 1 phía

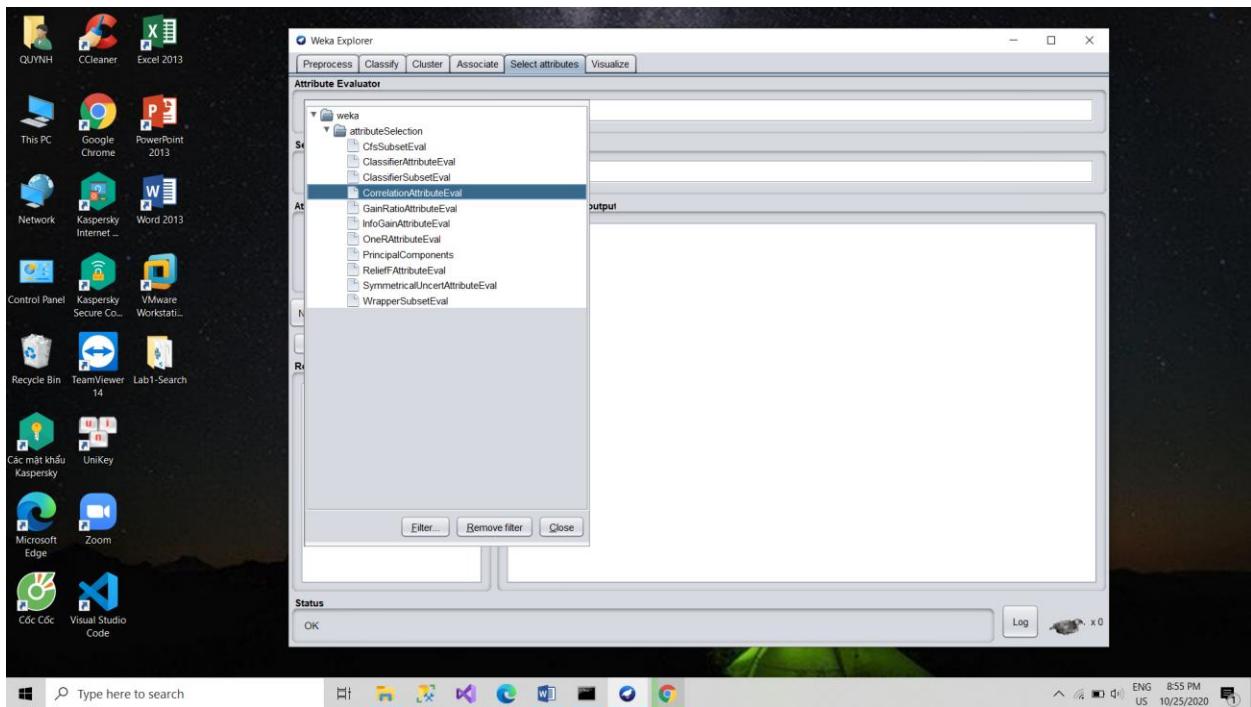


### 2.3.3.

Trong Weka, một phương pháp lựa chọn thuộc tính (attribute selection) bao gồm 2 phần:

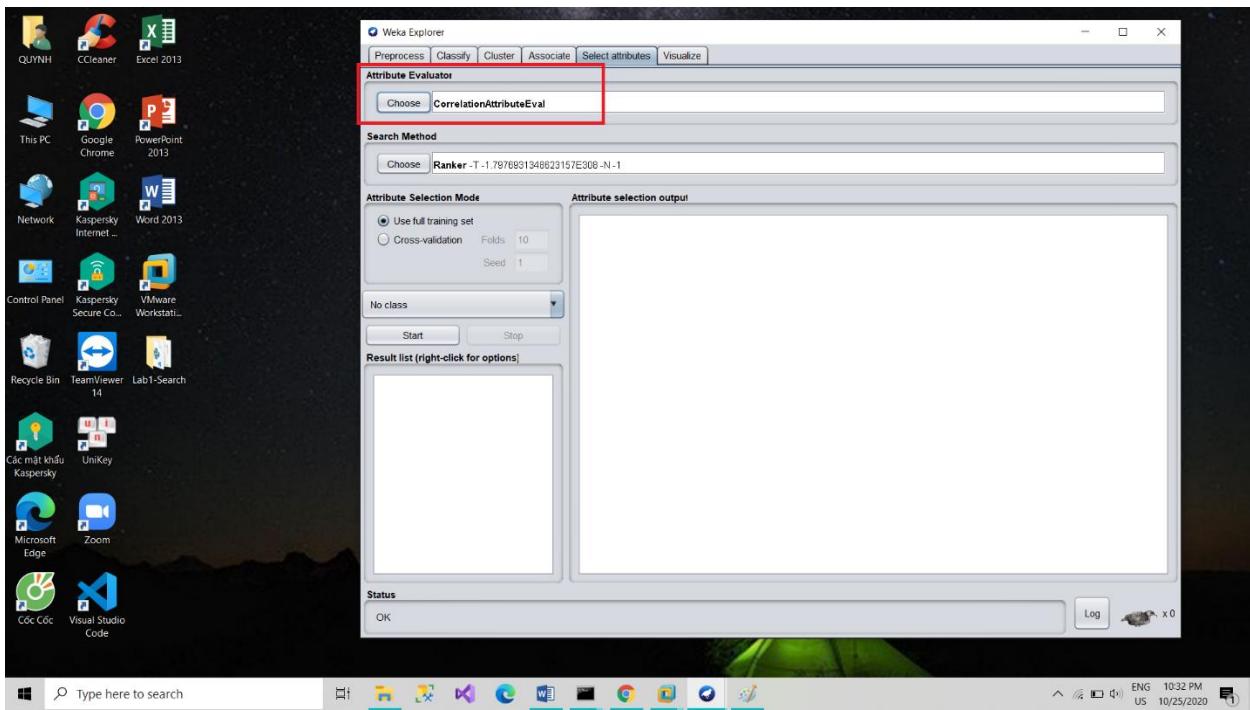
- Attribute Evaluator: Weka cung cấp 9 phương pháp đánh giá thuộc tính, gồm:

- + CfsSubsetEval: Đánh giá tập thuộc tính bằng cách xem xét khả năng dự đoán của từng thuộc tính riêng lẻ và mức độ dự thừa giữa chúng.
  - + 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.
- Search Method: Để xác định phương pháp tìm kiếm được thực hiện. Weka cung cấp 3 phương thức tìm kiếm, gồm:
- + BestFirst: Tiến hành kỹ thuật leo đồi tham lam kết hợp với quay lui.
  - + GreedyStepwise: Thực hiện tìm kiếm tham lam về phía trước hoặc phía sau thông qua không gian các tập con thuộc tính.
  - + Ranker: Xếp hạng các thuộc tính theo đánh giá trọng số của từng thuộc tính. Sử dụng kết hợp với các bộ đánh giá thuộc tính (ReliefF, GainRatio,...).

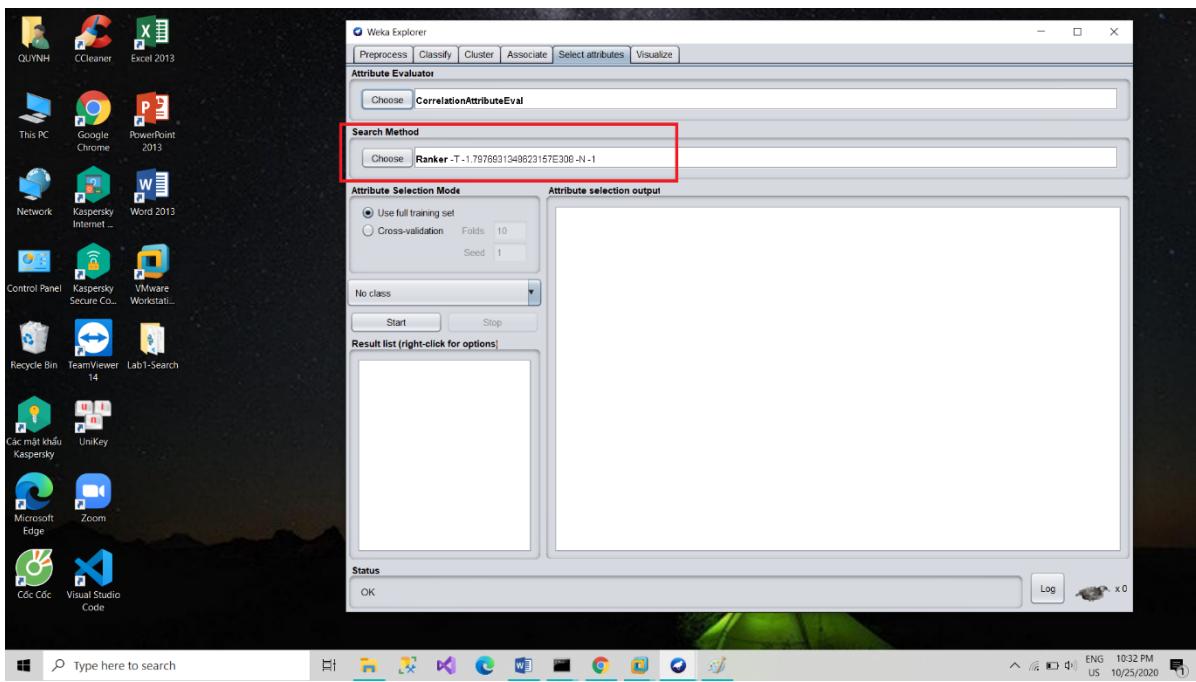


### 2.3.4.

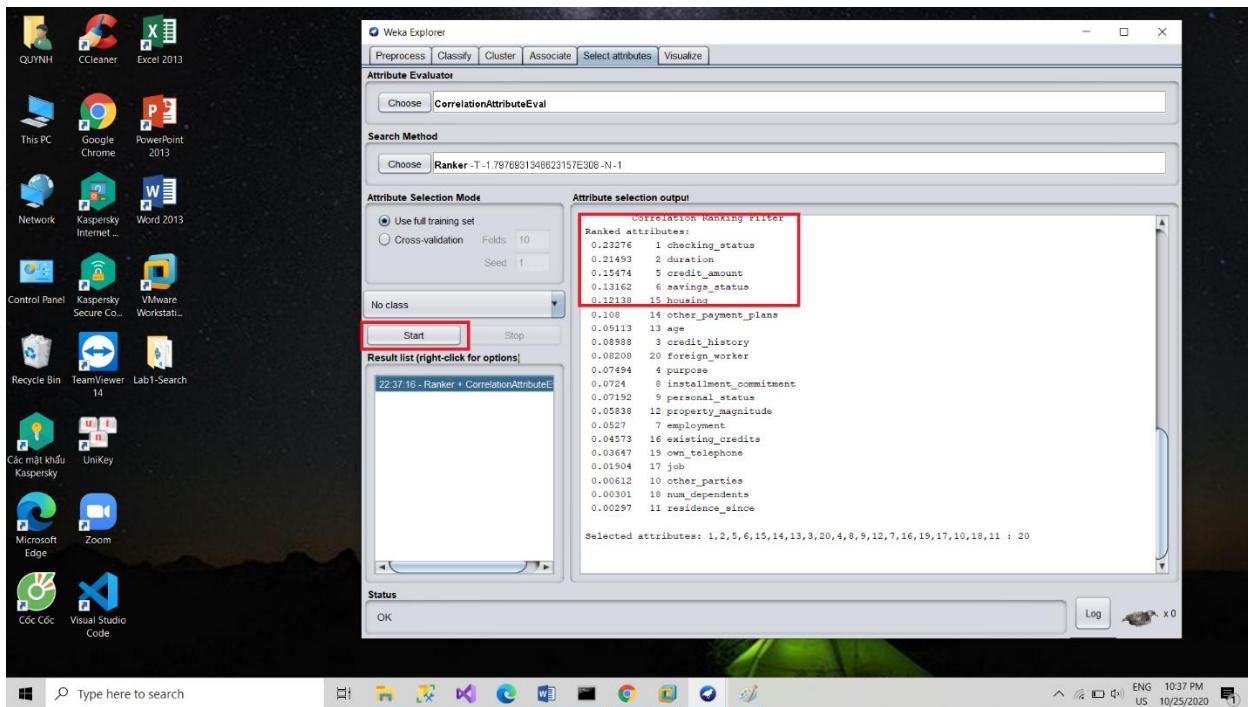
- Bộ lọc để chọn ra 5 thuộc tính có tương quan cao nhất với thuộc tính lớp là:
    - + Attribute Evaluator: CorrelationAttributeEval
    - + Search Method: Ranker
  - Các bước thực hiện:
    - + Bước 1: Chọn Attribute evaluator CorrelationAttributeEval



+ Bước 2: Chọn search method: ranker



+Bước 3: Bấm start, chọn 5 thuộc tính đầu tiên trong attribute select output



Kết quả cuối cùng: các thuộc tính được chọn: checking\_status, duration, credit\_amount, savings\_status, housing.

### III. CÀI ĐẶT TIỀN XỬ LÝ DỮ LIỆU

Chạy các chức năng với bộ dữ liệu “house-prices.csv”:

Chức năng	Cú pháp tham số dòng lệnh	Kết quả
Chức năng 1	python bai1.py house-prices.csv	Các cột bị thiếu dữ liệu: LotFrontage, Alley, MasVnrType, MasVnrArea, BsmtQual, BsmtCond, BsmtExposure, BsmtFinType1, BsmtFinType2, FireplaceQu, GarageType, GarageYrBlt, GarageFinish, GarageQual, GarageCond, PoolQC, Fence, MiscFeature
Chức năng 2	python bai2.py house-prices.csv	Số dòng bị thiếu dữ liệu: 1000
Chức năng 3	python bai3.py house-prices.csv mean bai3_mean.csv	Kết quả lưu ở file bai3_mean.csv trong thư mục Result

	python bai3.py house-prices.csv median bai3_median.csv	Kết quả lưu ở file bai3_median.csv trong thư mục Result
<b>Chức năng 4</b>	python bai4.py house-prices.csv 0.05 bai4.py	Kết quả lưu ở file bai4.csv trong thư mục Result
	python bai4.py house-prices.csv 0 bai4_01.csv	Kết quả lưu ở file bai4_01.csv trong thư mục Result
<b>Chức năng 5</b>	python bai5.py 50 house-prices.csv bai5_50.csv	Kết quả lưu ở file bai5_50.csv trong thư mục Result
	python bai5.py 0 house-prices.csv bai5_0.csv	Kết quả lưu ở file bai5_0.csv trong thư mục Result
<b>Chức năng 6</b>	python bai6.py house-prices.csv bai6.csv	Kết quả lưu ở file bai6.csv trong thư mục Result
<b>Chức năng 7</b>	python bai7.py house-prices.csv Id min-max bai7_Id.csv	Kết quả lưu ở file bai7_Id.csv trong thư mục Result
	python bai7.py house-prices.csv LotFrontage z-score bai7_LotFrontage.csv	Kết quả lưu ở file bai7_LotFrontage.csv trong thư mục Result
<b>Chức năng 8</b>	python bai8.py house-prices.csv EnclosedPorch * 3SsnPorch bai8_01.csv	Kết quả lưu ở file bai8_01.csv trong thư mục Result
	python bai8.py house-prices.csv EnclosedPorch / 3SsnPorch bai8_02.csv	Kết quả lưu ở file bai8_02.csv trong thư mục Result
	python bai8.py house-prices.csv EnclosedPorch + 3SsnPorch bai8_03.csv	Kết quả lưu ở file bai8_03.csv trong thư mục Result
	python bai8.py house-prices.csv EnclosedPorch - 3SsnPorch bai8_04.csv	Kết quả lưu ở file bai8_04.csv trong thư mục Result
	python bai8.py house-prices.csv EnclosedPorch * 3SsnPorch * MoSold bai8_05.csv	Kết quả lưu ở file bai8_05.csv trong thư mục Result

**IV. TÀI LIỆU THAM KHẢO.**

[http://data.uet.vnu.edu.vn:8080/xmlui/bitstream/handle/123456789/1062/NinhHoaiAnh\\_LuanVan.pdf?sequence=1](http://data.uet.vnu.edu.vn:8080/xmlui/bitstream/handle/123456789/1062/NinhHoaiAnh_LuanVan.pdf?sequence=1)

<https://www.slideshare.net/HoQuangThanh/la-chn-thuc-tnh-v-khai-ph-lut-kt-hp-trn-weka>

[https://github.com/vltanh/hcmus-DataMining/blob/master/Lab01/submit/1612838\\_1612849.pdf](https://github.com/vltanh/hcmus-DataMining/blob/master/Lab01/submit/1612838_1612849.pdf)

[http://ccs.hnue.edu.vn/hungtd/DM2012/NhatQuang/L2-Gioi\\_thieu\\_WEKA.pdf](http://ccs.hnue.edu.vn/hungtd/DM2012/NhatQuang/L2-Gioi_thieu_WEKA.pdf)

[https://www.academia.edu/8629014/Le\\_Thu\\_H%C6%B0%C6%A1ng\\_CNPM](https://www.academia.edu/8629014/Le_Thu_H%C6%B0%C6%A1ng_CNPM)

<https://ongxuanhong.wordpress.com/2015/08/25/ap-dung-cac-phuong-phap-phan-lop-classification-tren-tap-du-lieu-mushroom/>

Sách J. Han and M. Kamber: Data Mining, Concepts and Techniques, Second Edition