**ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH**

**TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN**

**KHOA CÔNG NGHỆ THÔNG TIN**

****

**LAB 1 - Preprocessing**

**MÔN: KHAI THÁC DỮ LIỆU VÀ ỨNG DỤNG**

**| Giáo viên hướng dẫn |**

**Thầy: Dương Nguyễn Thái Bảo**

**Sinh viên thực hiện:**

**CAO TẤT CƯỜNG – 18120296**

**HÀ VĂN DUY – 18120339**

**Chuyên ngành: Khoa học máy tính**

Thành phố Hồ Chí Minh – 2020

# MỤC LỤC

[MỤC LỤC 2](#_Toc55475381)

[Phần I: Thông tin chung 3](#_Toc55475382)

[Phần II: Cài đặt Weka 4](#_Toc55475383)

[Phần 3: Làm quen với Weka 6](#_Toc55475384)

[2.1 Đọc dữ liệu vào Weka 6](#_Toc55475385)

[2.2 Khám phá tập dữ liệu Weather 11](#_Toc55475386)

[2.3 Khám phá tập dữ liệu Tín dụng Đức 14](#_Toc55475387)

[PHỤ LỤC 21](#_Toc55475388)

[TÀI LIỆU THAM KHẢO 22](#_Toc55475389)

# Phần I: Thông tin chung

Danh sách thành viên:

* Cao Tất Cường 18120296
* Hà Văn Duy 18120339

Mục tiêu của bài tập:

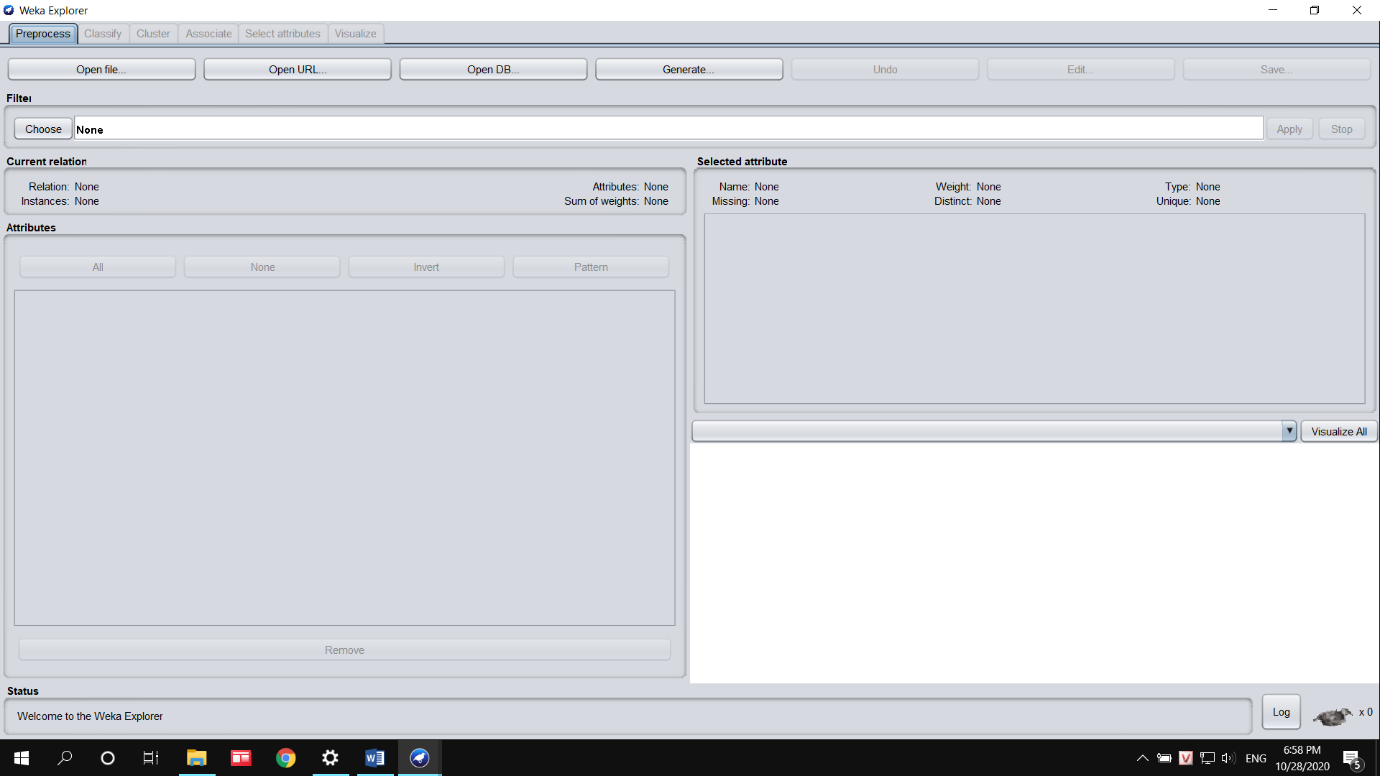
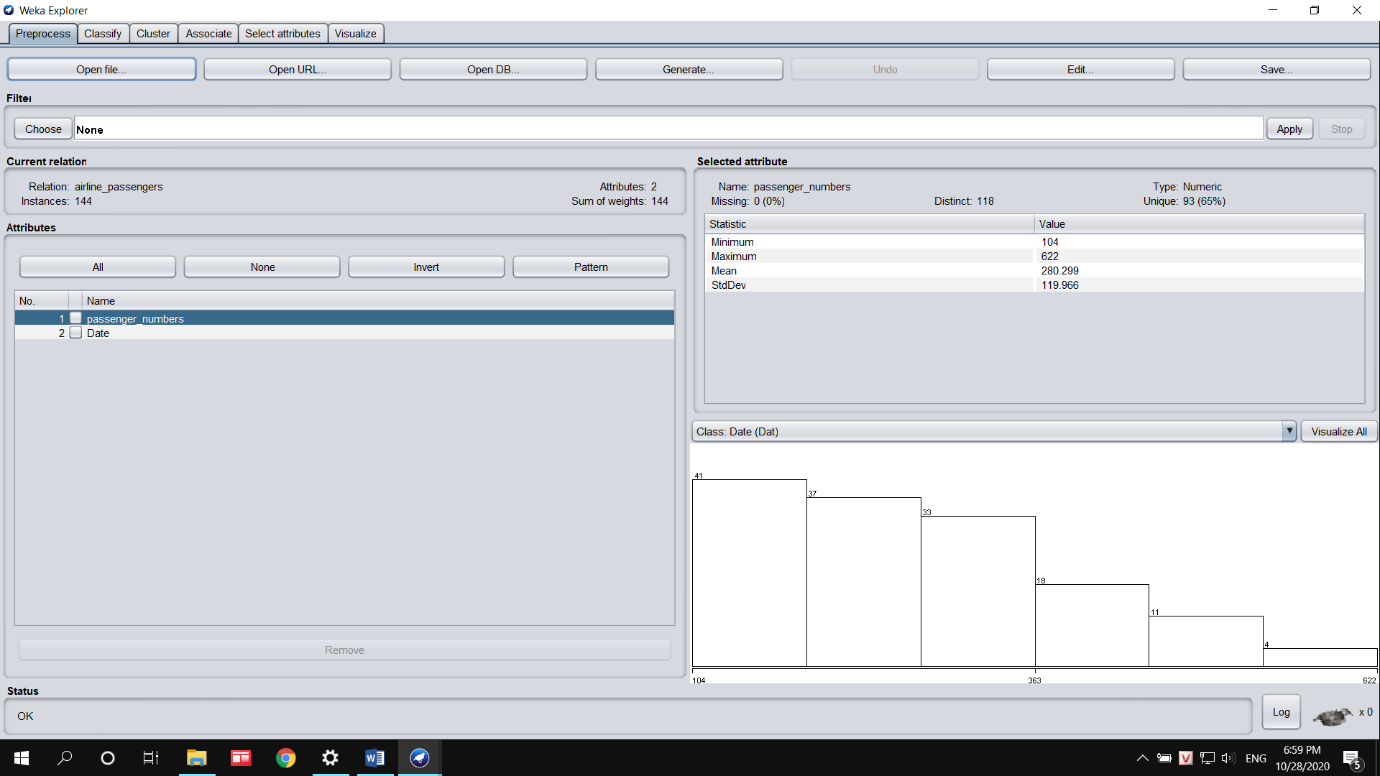
* Làm quen với các thao tác cơ bản trong tác vụ tiền xử lý dữ liệu thông qua việc áp dụng các công cụ hỗ trợ được cung cấp bởi phần mềm mã nguồn mở Weka.
* Phát huy kỹ năng lập trình để tự cài đặt các thủ tục tiền xử lý dữ liệu đơn giản.

Tỷ lệ thực hiện của mỗi thành viên

|  |  |
| --- | --- |
| Công việc | Người thực hiện |
| Làm việc với Weka | Cao Tất Cường |
| Cài đặt tiền xử lý dữ liệu | Hà Văn Duy |

Tự đánh giá mức độ hoàn thành đồ án: 100%

# Phần II: Cài đặt Weka

1. Sau khi cài đặt Weka ta có giao diện chức năng Explorer
2. Mở tập dữ liệu airline.arff ta có giao diện chức năng Explorer như sau:

* Ý nghĩa các nhóm điều khiển trong tab Preprocess

**Current relation**: Bảng này tóm tắt các thông tin chi tiết sau về các tập dữ liệu đang mở:

* Tên tập dữ liệu (quan hệ).
* Số lượng hàng (phiên bản).
* Số lượng cột (thuộc tính).

**Attributes**: Bảng này hiển thị các thuộc tính khác của tập dữ liệu. Khi nhấp vào một thuộc tính, Selected attribute sẽ liệt kê các thông tin chi tiết về thuộc tính đã chọn.

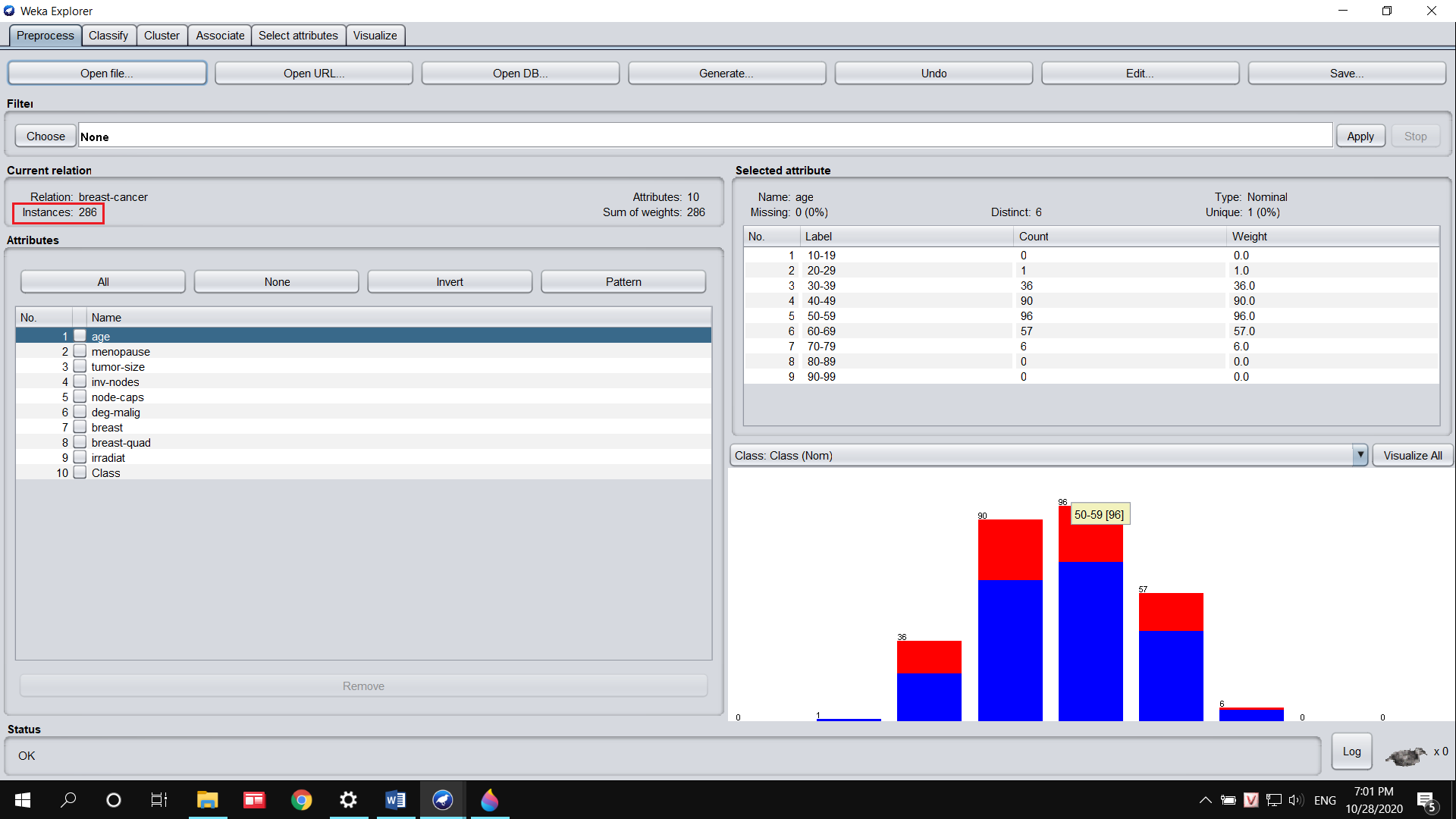
**Selected attribute**: Liệt kê rất nhiều thông tin về thuộc tính đã chọn ở Attributes, chẳng hạn như:

* Tên của thuộc tính.
* Số lượng giá trị bị thiếu và tỷ lệ giá trị bị thiếu trên toàn bộ tập dữ liệu.
* Số lượng các giá trị khác biệt.
* Kiểu dữ liệu.
* Giá trị nhỏ nhất.
* Gia trị lơn nhât.
* Giá trị trung bình.
* Độ lệch chuẩn.
* Ý nghĩa 5 tab còn lại trong giao diện Explorer
* **Classify**: Cung cấp một số thuật toán máy học để phân loại dữ liệu như Hồi quy tuyến tính, Hồi quy logistic, Máy vectơ hỗ trợ, Cây quyết định, RandomTree, RandomForest, NaiveBayes, v.v.
* **Cluster**: Dành cho việc đào tạo và đánh giá hiệu suất của các thuật toán gom nhóm không giám sát khác nhau liệu như SimpleKMeans, FilteredClusterer, HierarchicalClusterer, v.v trên tập dữ liệu chưa được gắn nhãn.
* **Associate**: Dùng để tự động tìm các liên kết trong tập dữ liệu. Yêu cầu dữ liệu trong đó tất cả các thuộc tính đều được phân loại.
* **Select attributes**: Dùng để thực hiện lựa chọn tính năng trên tập dữ liệu đã tải và xác định các tính năng đó có nhiều khả năng phù hợp nhất trong việc phát triển mô hình dự đoán.
* **Visualize**: Cho phép trực quan hóa dữ liệu đã xử lý để phân tích.

# Phần 3: Làm quen với Weka

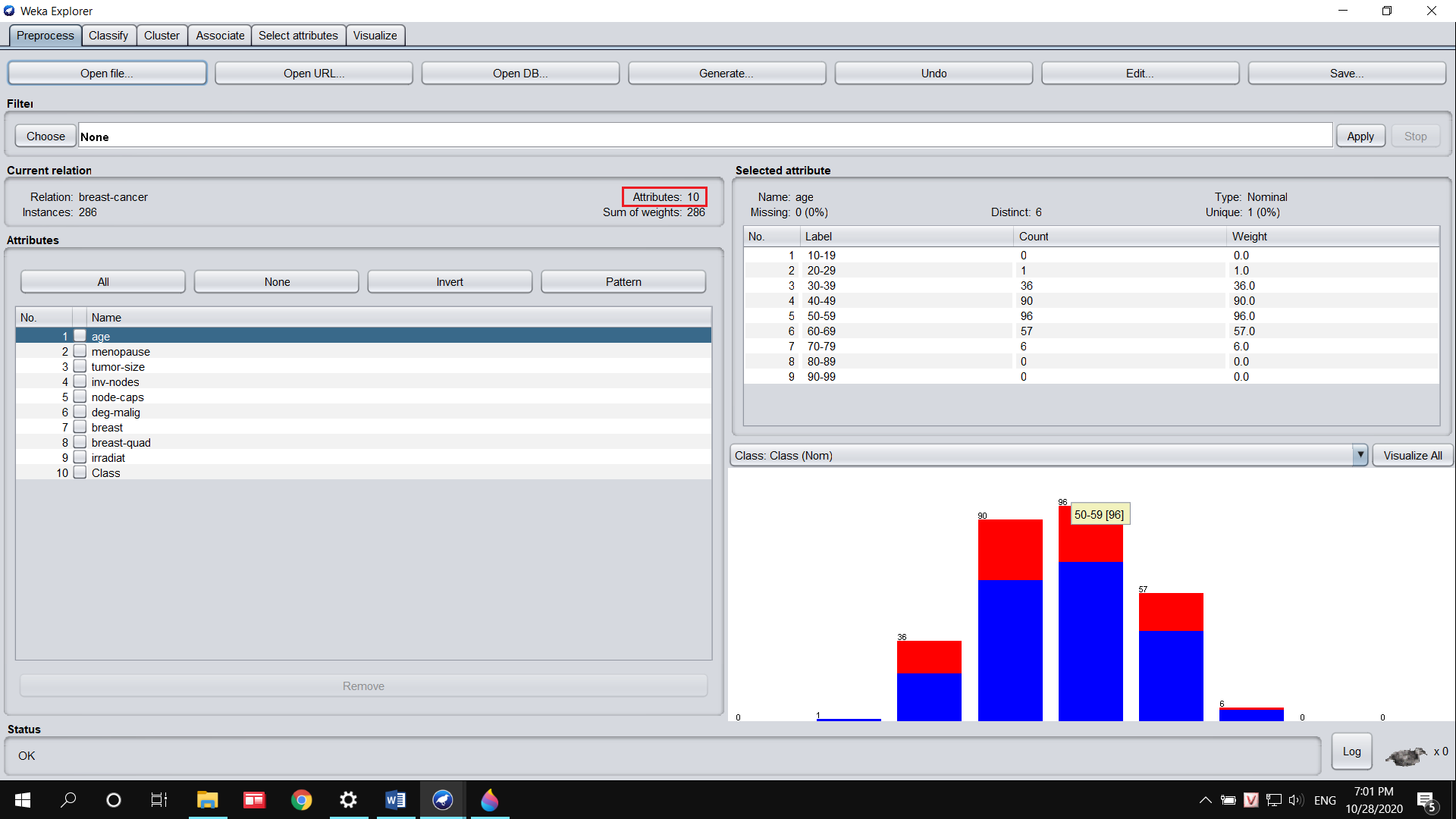
## Đọc dữ liệu vào Weka

1. *Tập dữ liệu có bao nhiêu mẫu (instances)?*

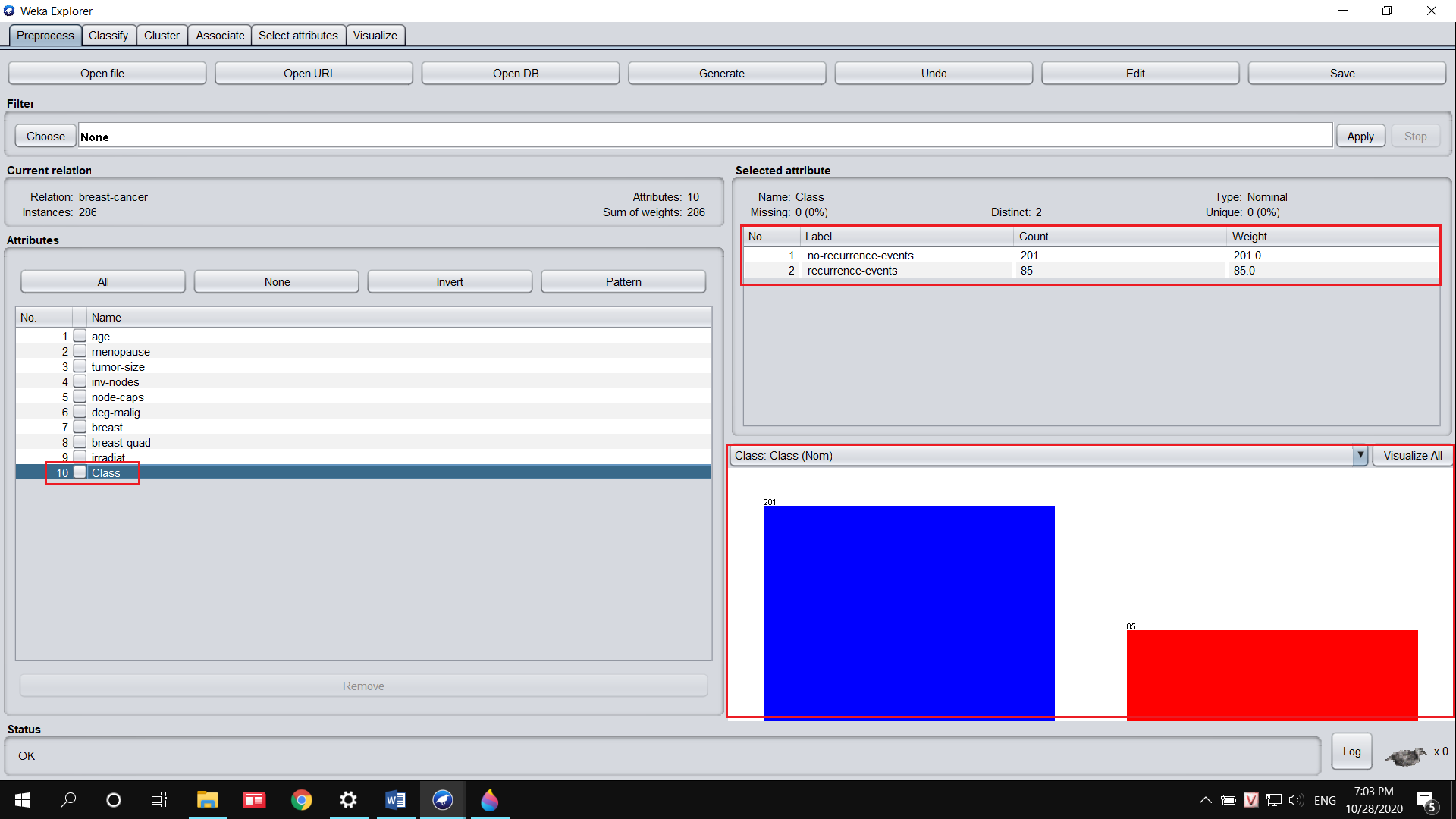
Tập dữ liệu có 286 mẫu.

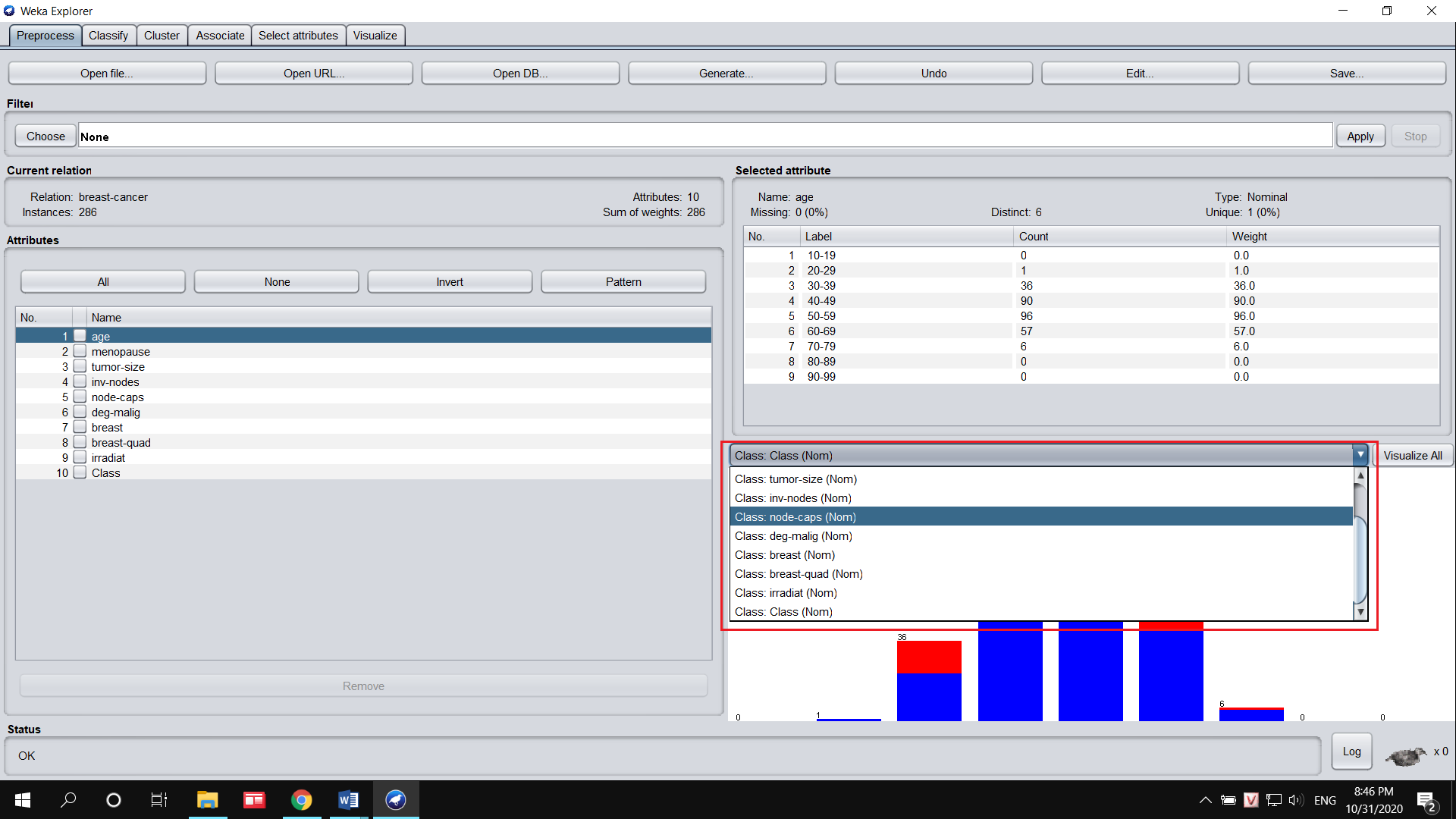
1. *Tập dữ liệu có bao nhiêu thuộc tính (attributes)?*

Tập dữ liệu có 10 thuộc tính.

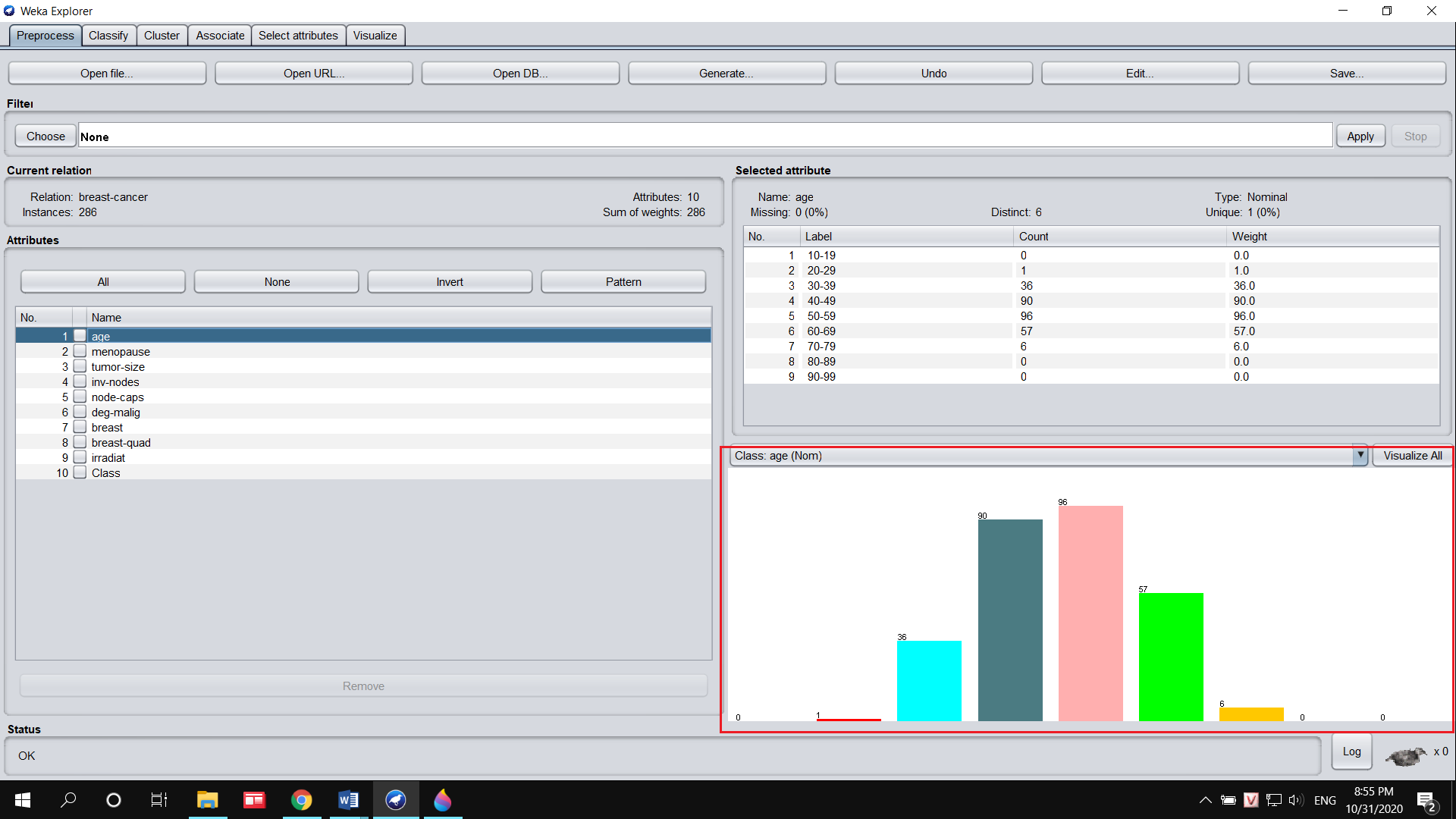


1. *Thuộc tính nào được dùng làm lớp (class)? Có thể thay đổi thuộc tính dùng làm lớp hay không? Nếu có thì bằng cách nào?*

* Thuộc tính Class {no-recurrence-events, recurrence-events} được dùng làm để lớp.
* Có thể thay đổi thuộc tính dùng làm lớp bằng cách chọn vào ô Class như hình bên dưới. Có thể chọn một thuộc tính bất kỳ trong dataset để làm thuộc tính lớp.



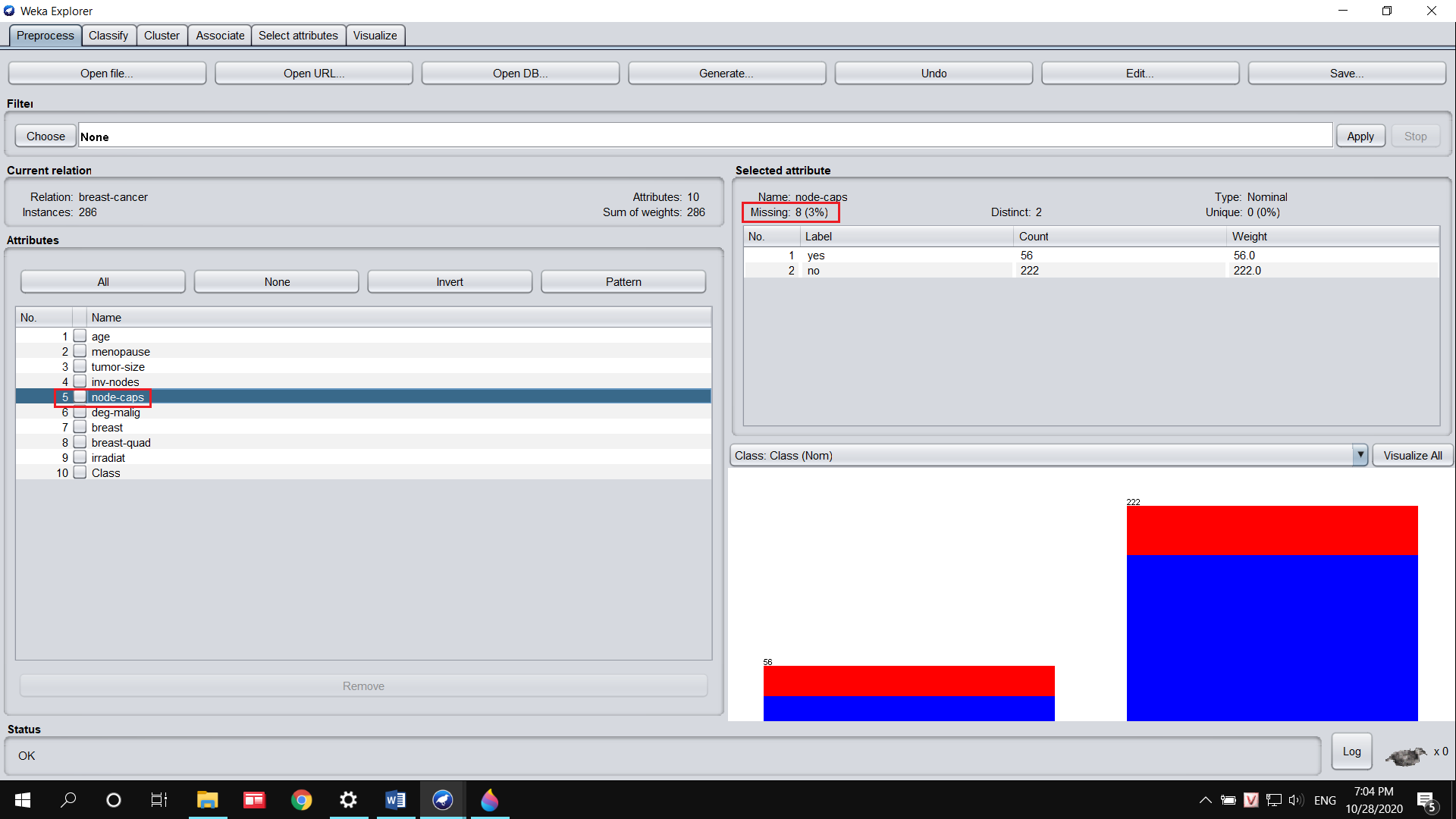
* Sau khi chọn một thuộc tính giả sử thuộc tính age thì đồ thị sẽ hiển thị như sau:

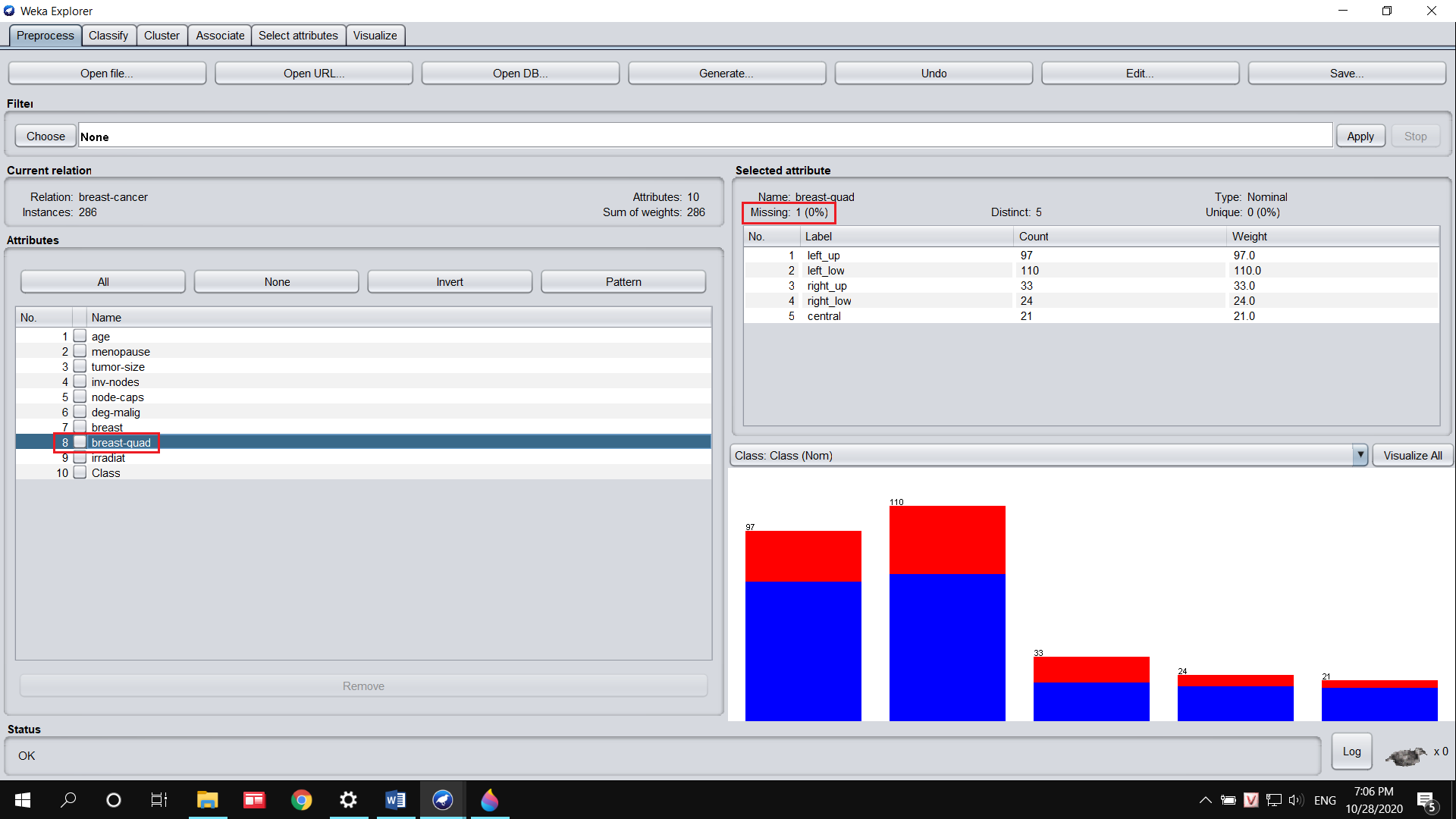


1. *Tìm hiểu chi tiết từng thuộc tính trong khung Attributes và cho biết: có bao nhiêu thuộc tính bị thiếu dữ liệu (missing values)? Thuộc tính nào thiếu dữ liệu ít nhất/nhiều nhất? Trình bày tổng quát các cách để giải quyết vấn đề missing values.*

* Có 2 thuộc tính bị thiếu dữ liệu: node-caps và breast-quad.

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

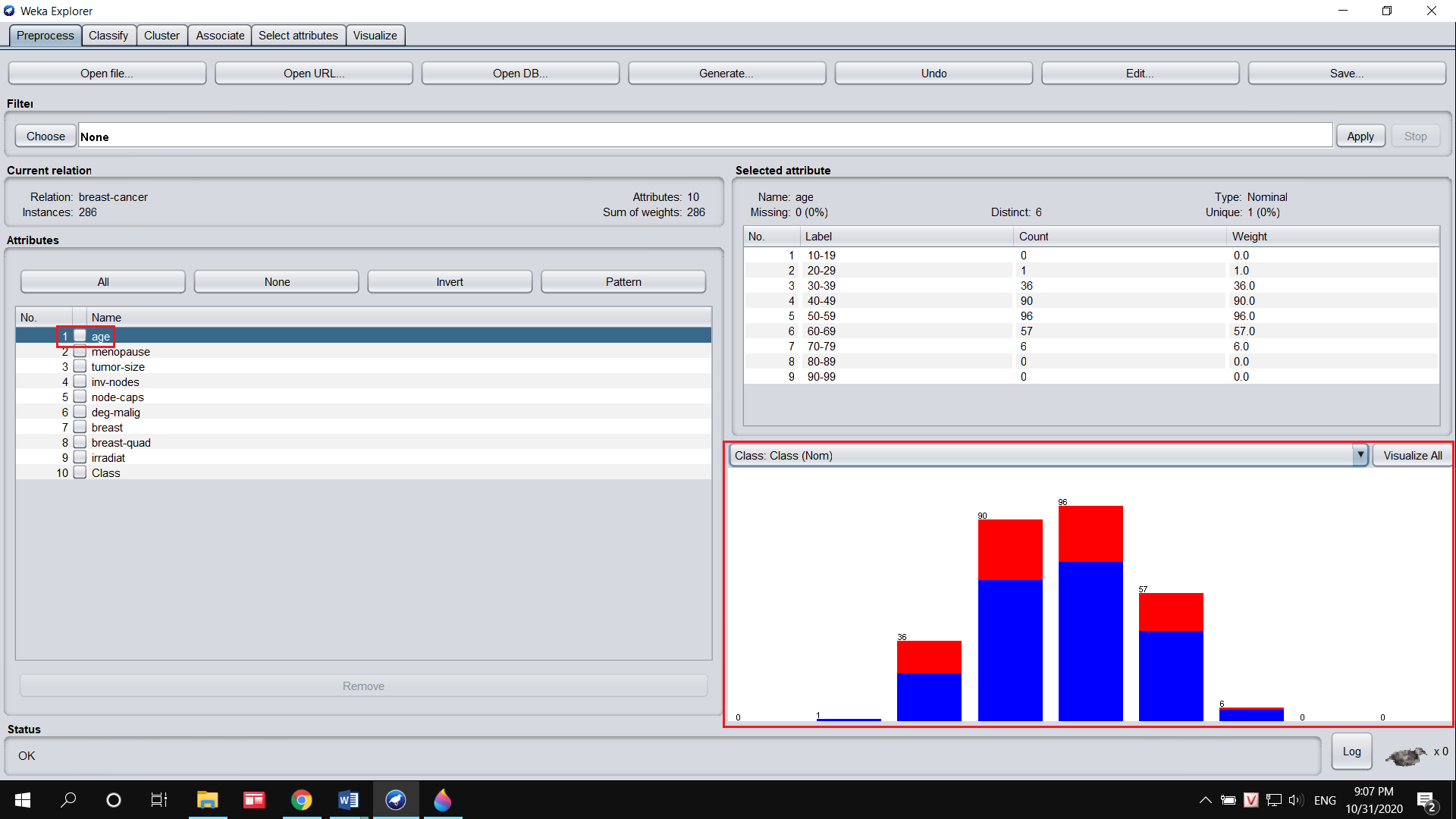
Thuộc tính breast-quad thiếu ít dữ liệu nhất: 1



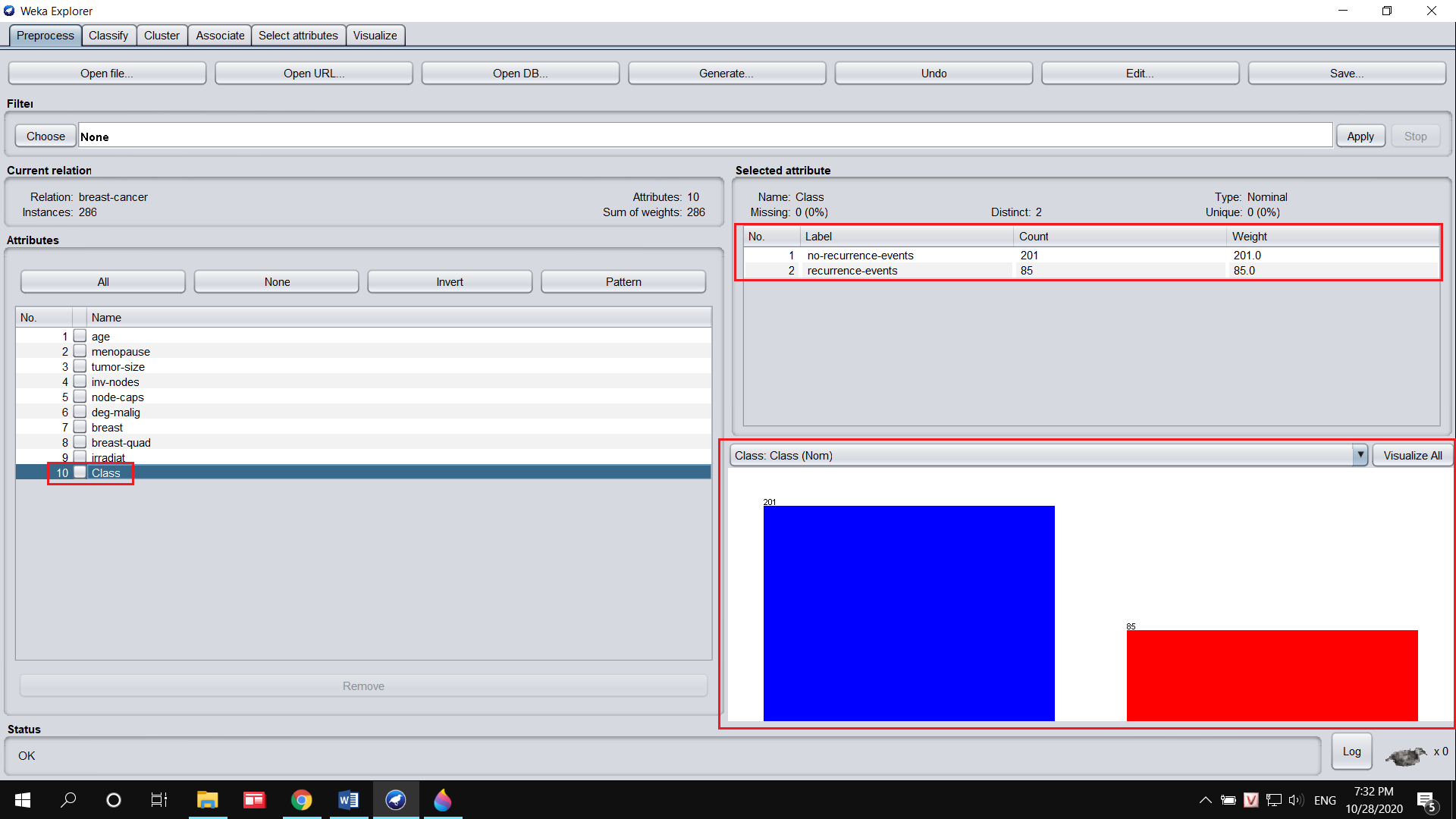
* Các cách giải quyết vấn đề thiếu dữ liệu:
* Sử dụng mean, median (cho thuộc tính numeric) và mode (cho thuộc tính categorical) để điền cho dữ liệu bị thiếu.
* Bỏ qua những bản ghi với dữ liệu thiếu.
* Lấp đầy dữ liệu thiếu một cách thủ công.
* Sử dụng hằng toàn cục để lấp đầy dữ liệu thiếu (NULL, unknown,....).
* Sử dụng giá trị thuộc tính trung bình để lấp đầy dữ liệu thiếu.
* Sử dụng giá trị thuộc tính trung bình của tất cả các mẫu thuộc cùng một lớp để lấp đầy dữ liệu thiếu.
* Suy diễn giá trị phù hợp nhất để lấp đầy dữ liệu thiếu.
* Có thể cần sử dụng những phương pháp như bộ phân lớp Bayesian hay cây quyết định để tự động suy diễn giá trị thiếu.

1. *Giải thích ý nghĩa của đồ thị trong cửa sổ Explorer. Bạn đặt tên cho đồ thị này là gì? Màu xanh và màu đỏ có nghĩa gì? Đồ thị này biểu diễn cho cái gì?*

* Đồ thị để thể hiện sự phân bố của thuộc tính hiện được chọn ở khung Attributes và được phân lớp theo thuộc tính được chọn ở box Class.
* Em đặt tên là histogram of the distribution attribute.
* Xét thuộc tính age, mỗi cột biểu diễn một khoảng tuổi, giá trị trên mỗi cột là số lượng mẫu có giá trị nằm trong khoảng đó.

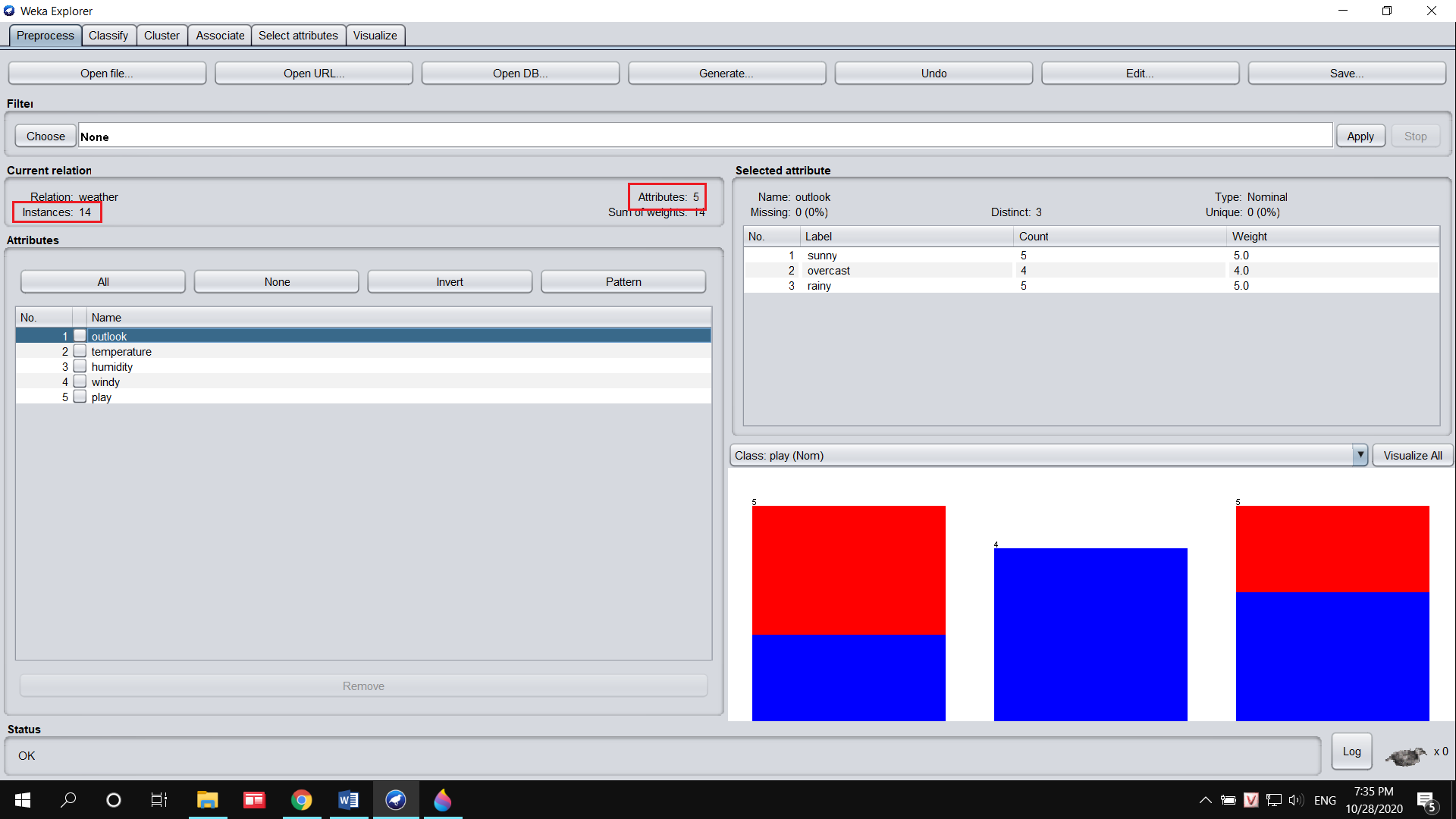


* Màu xanh ứng với Class có nhãn là no-recurrence-events.

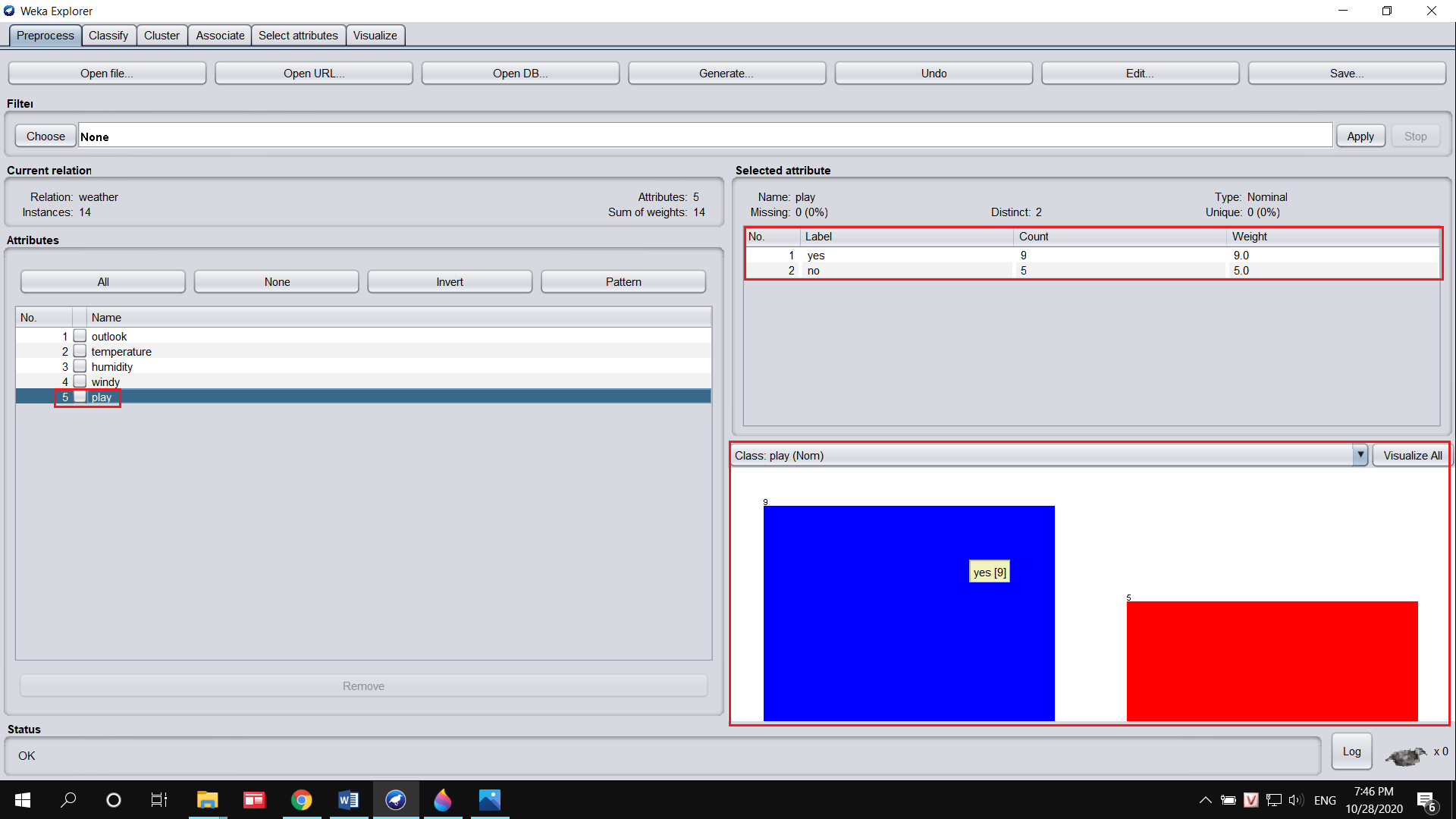
Màu đỏ ứng với Class có nhãn là recurrence-events.

## Khám phá tập dữ liệu Weather

1. *Tập dữ liệu có bao nhiêu thuộc tính? Bao nhiêu mẫu? Phân loại các thuộc tính theo kiểu dữ liệu (categorical/numeric). Thuộc tính nào là lớp?*

* Tập dữ liệu có 5 thuộc tính và 14 mẫu.
* Các thuộc tính có kiểu dữ liệu categorical: outlook, windy, play

Các thuộc tính có kiểu dữ liệu numeric: temperature, humidity

Thuộc tính lớp: play

1. *Liệt kê five-number summary của thuộc tính temperature và humidity. Weka có cung cấp những giá trị này không?*

* Five-number summary của thuộc tính temperature:

Minimum: 64.00

Lower quartile (Q1): 69.25

Median value (Q2): 72.00

Upper quartile (Q3): 78.75

Maximum: 85.00

* Five-number summary của thuộc tính humidity:

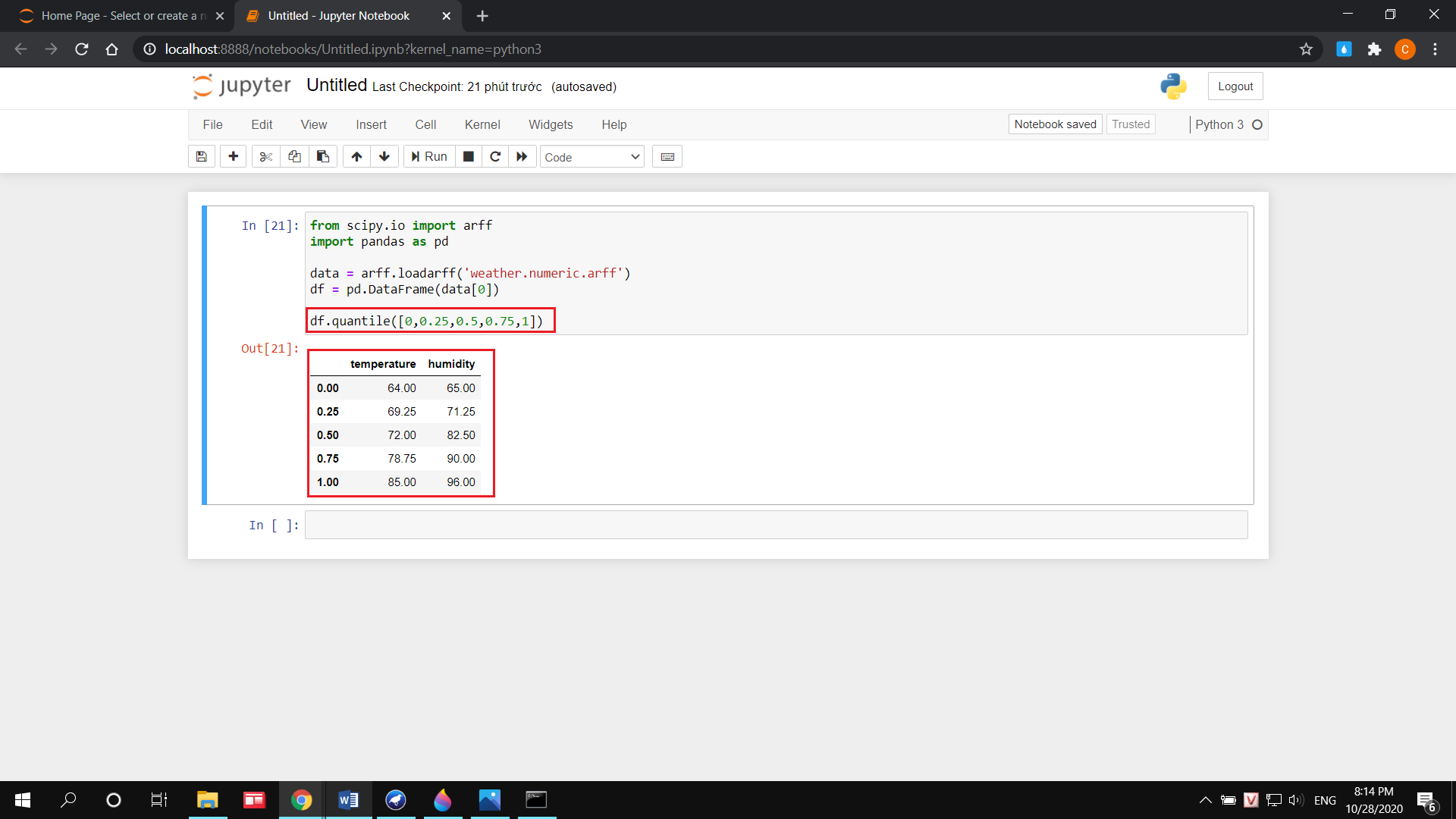
Minimum: 65.00

Lower quartile (Q1): 71.25

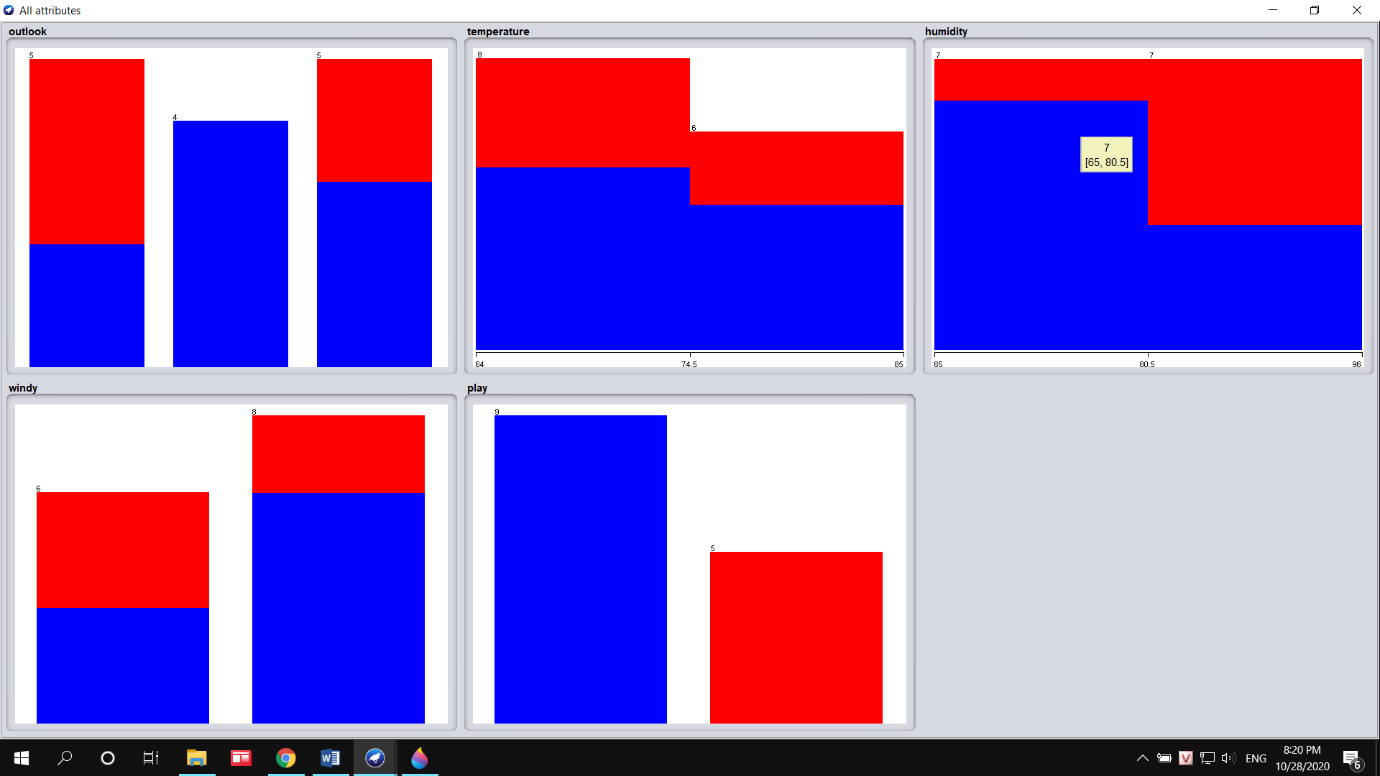
Median value (Q2): 82.50

Upper quartile (Q3): 90.00

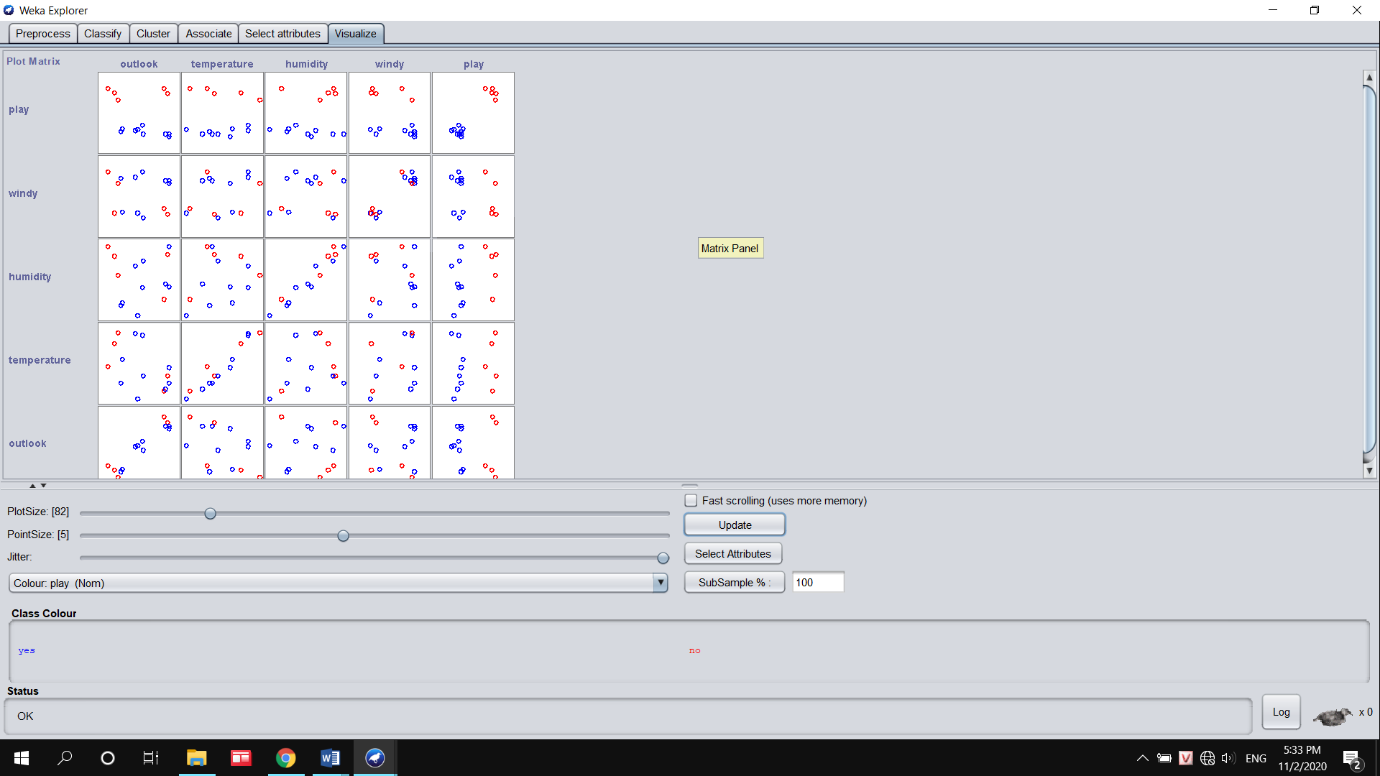
Maximum: 96.00

* Weka chỉ cung cấp chỉ cung cấp Minimum và Maximum, còn nhũng giá trị còn lại em tìm bằng cách sử dụng python như hình bên dưới:

1. *Lần lượt xem xét các thuộc tính khác của dataset dưới dạng đồ thị. Dán các ảnh chụp màn hình vào bài làm.*

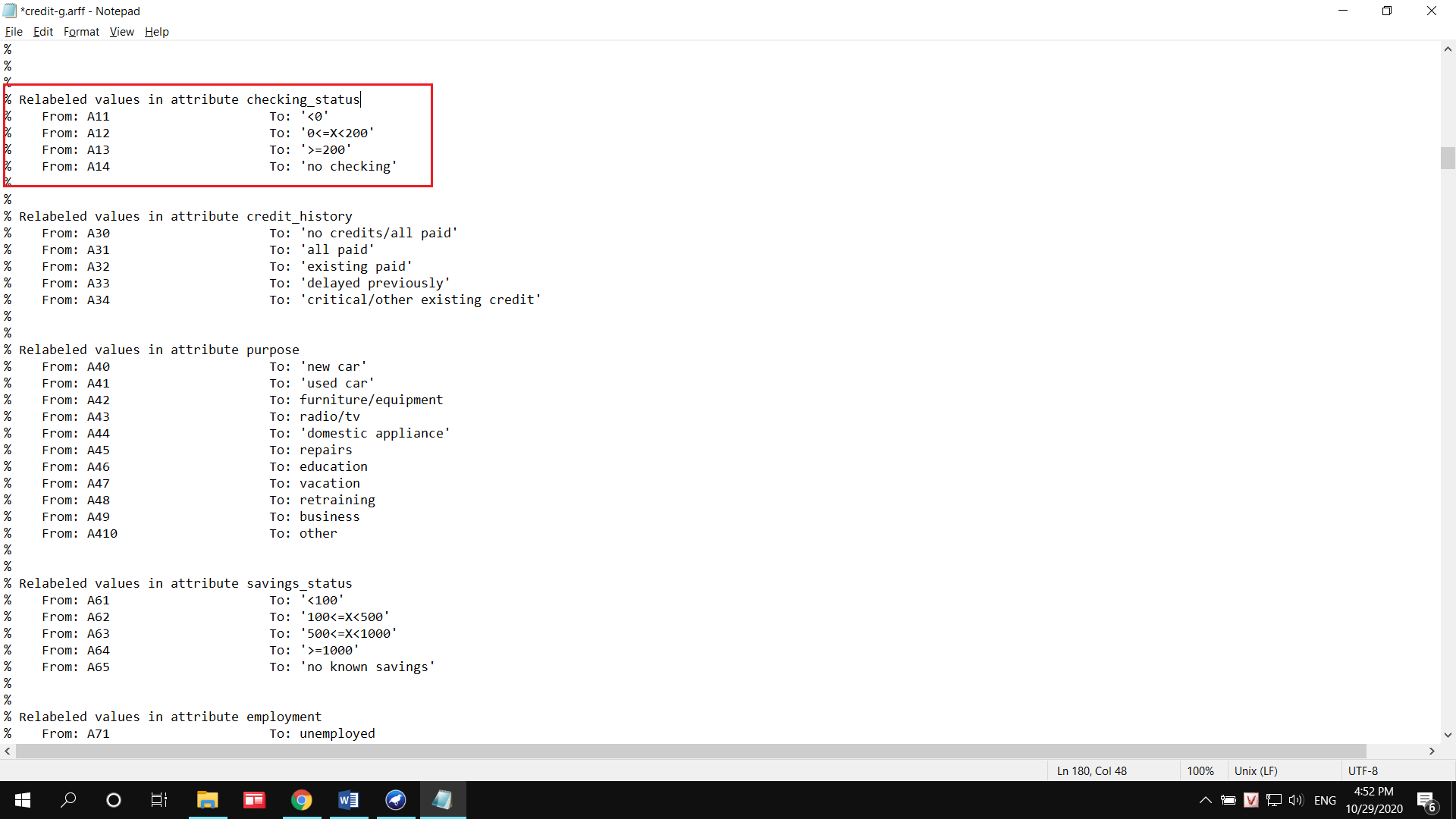
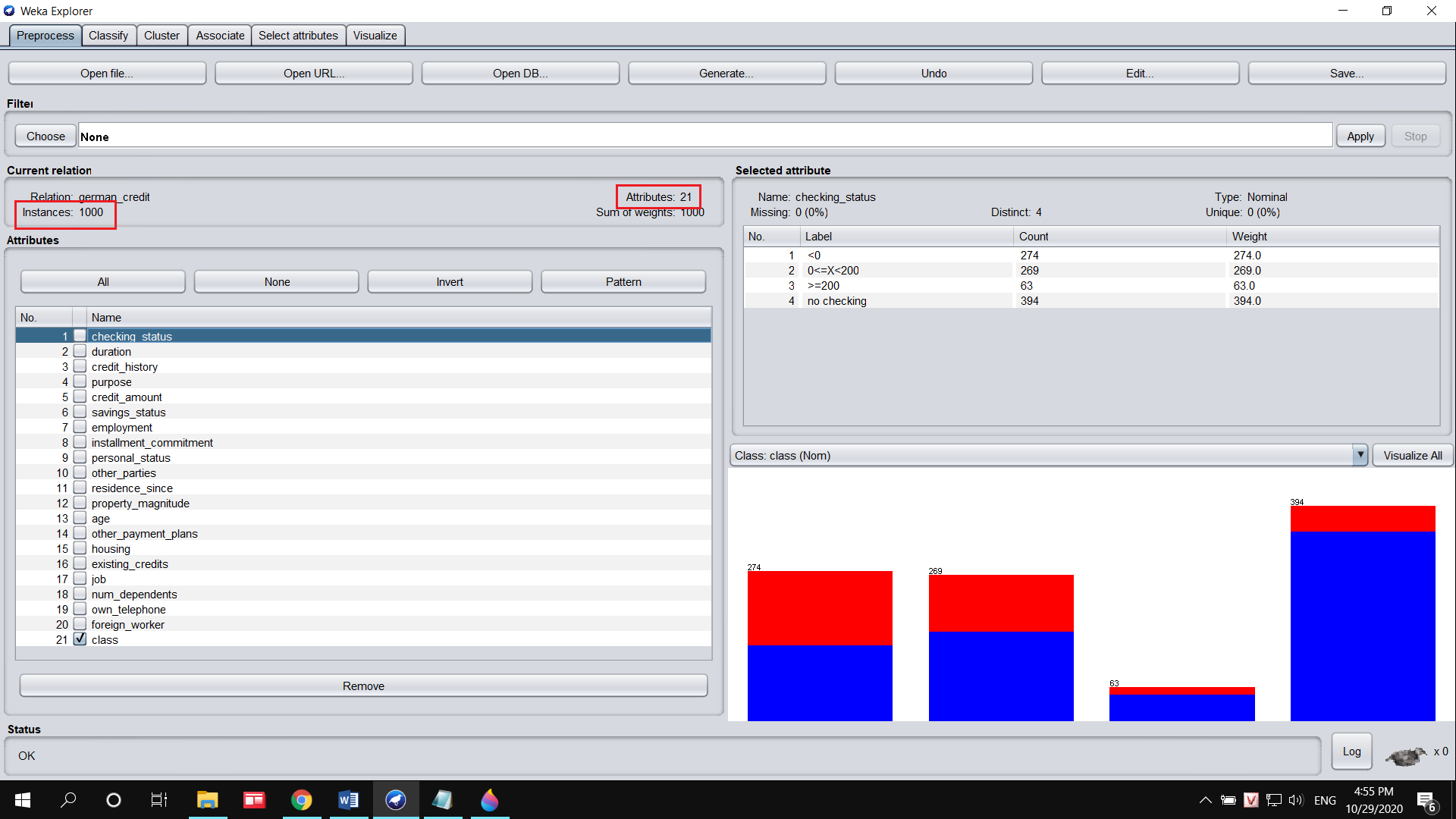
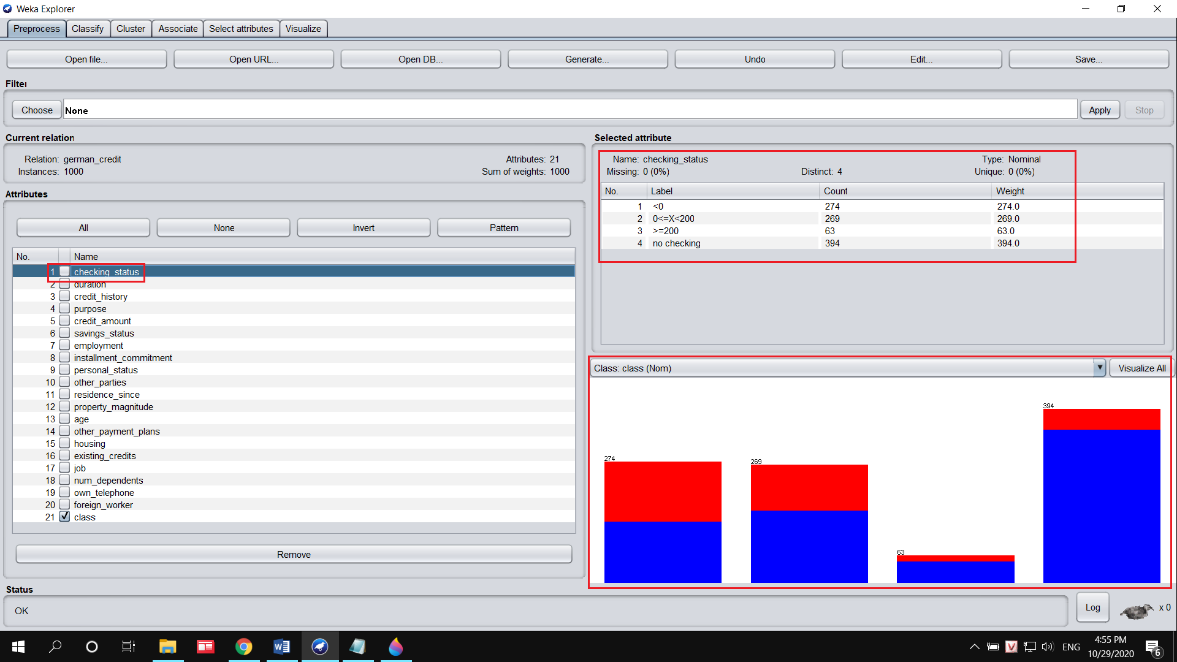
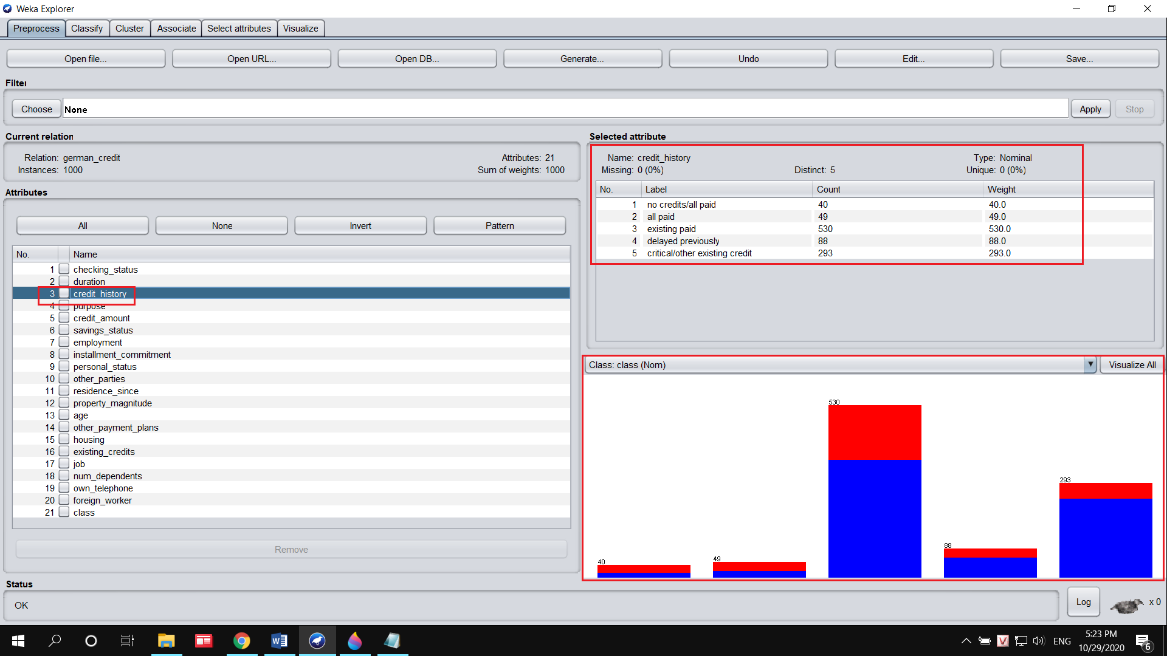
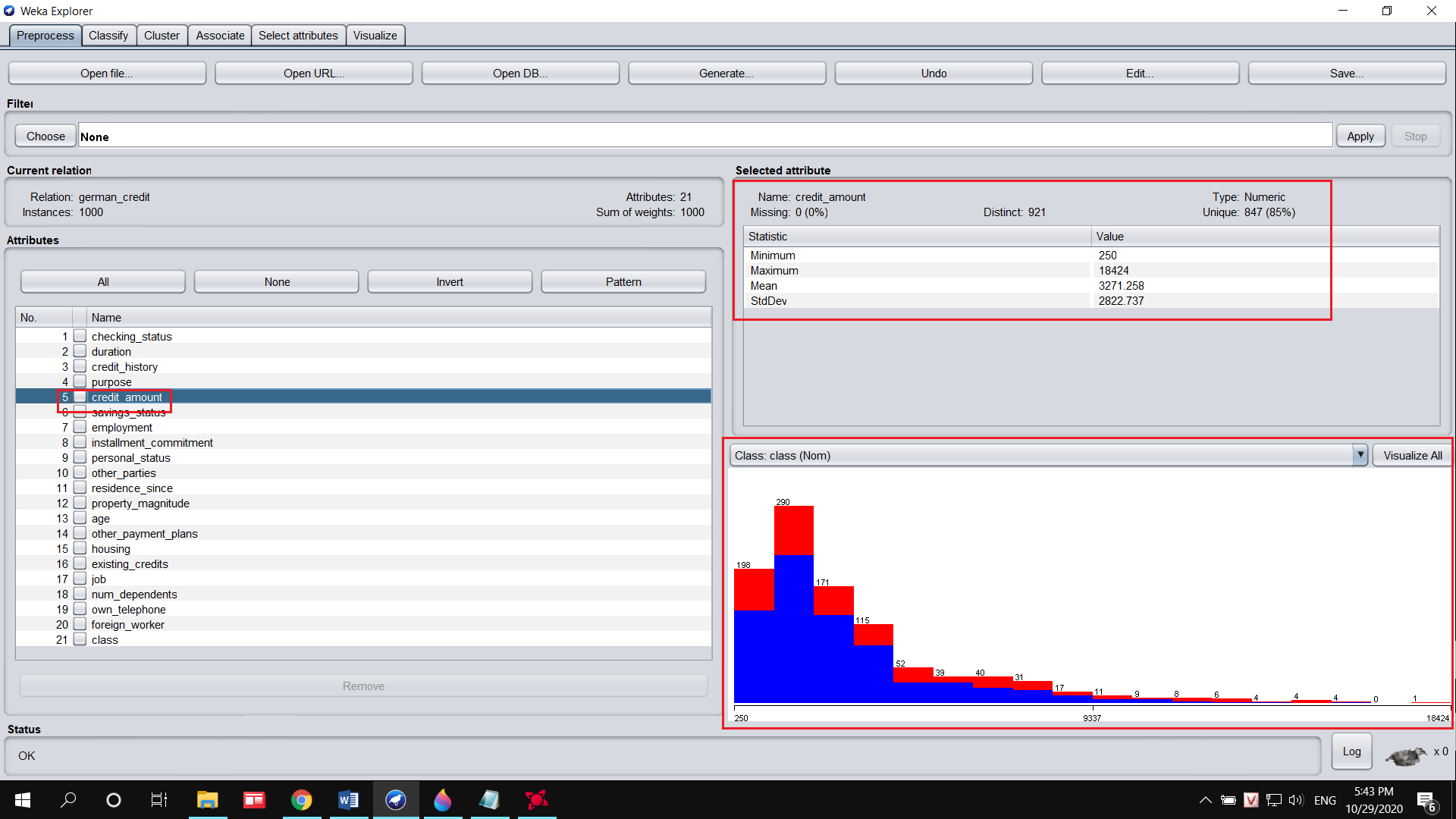
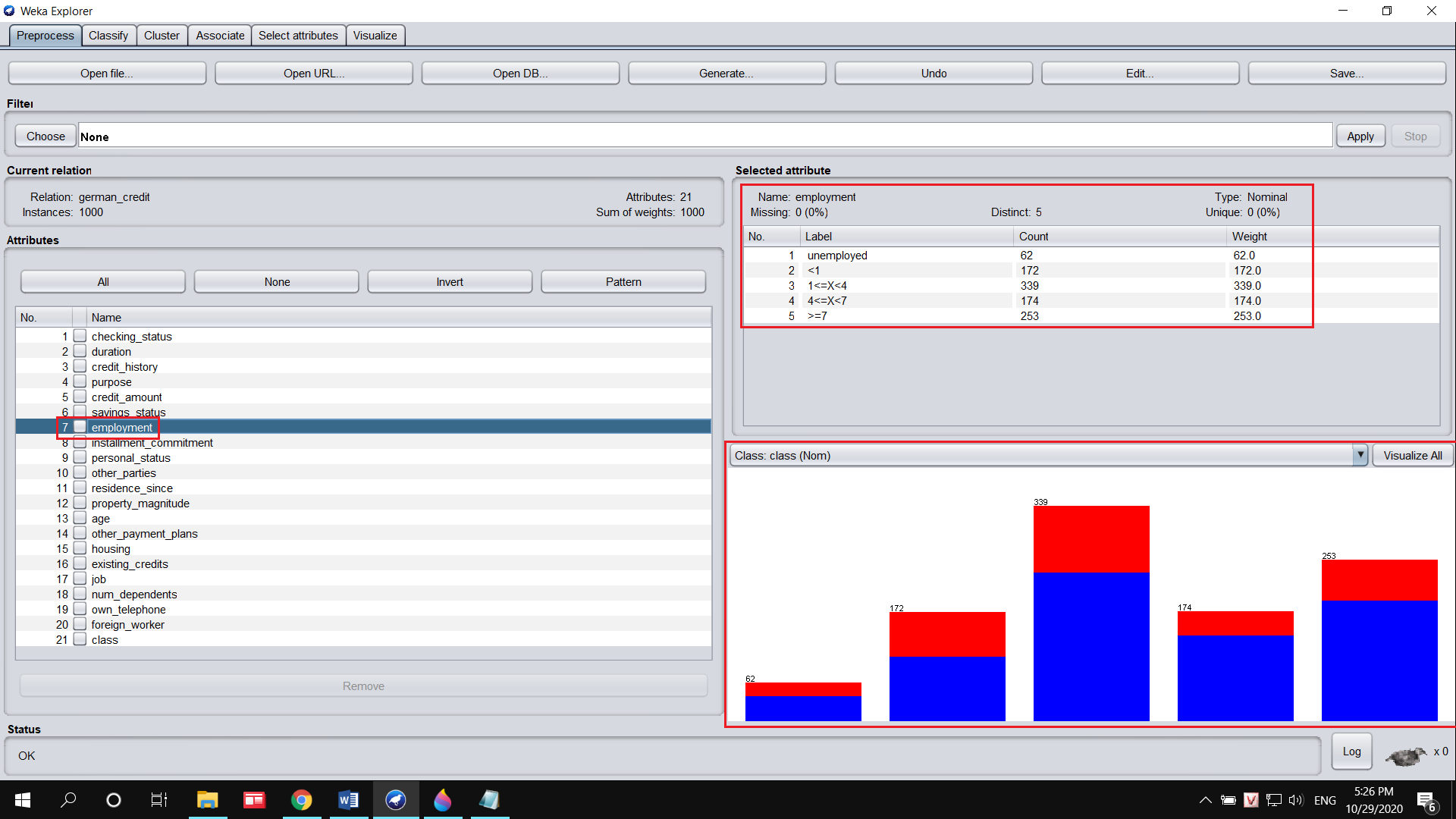
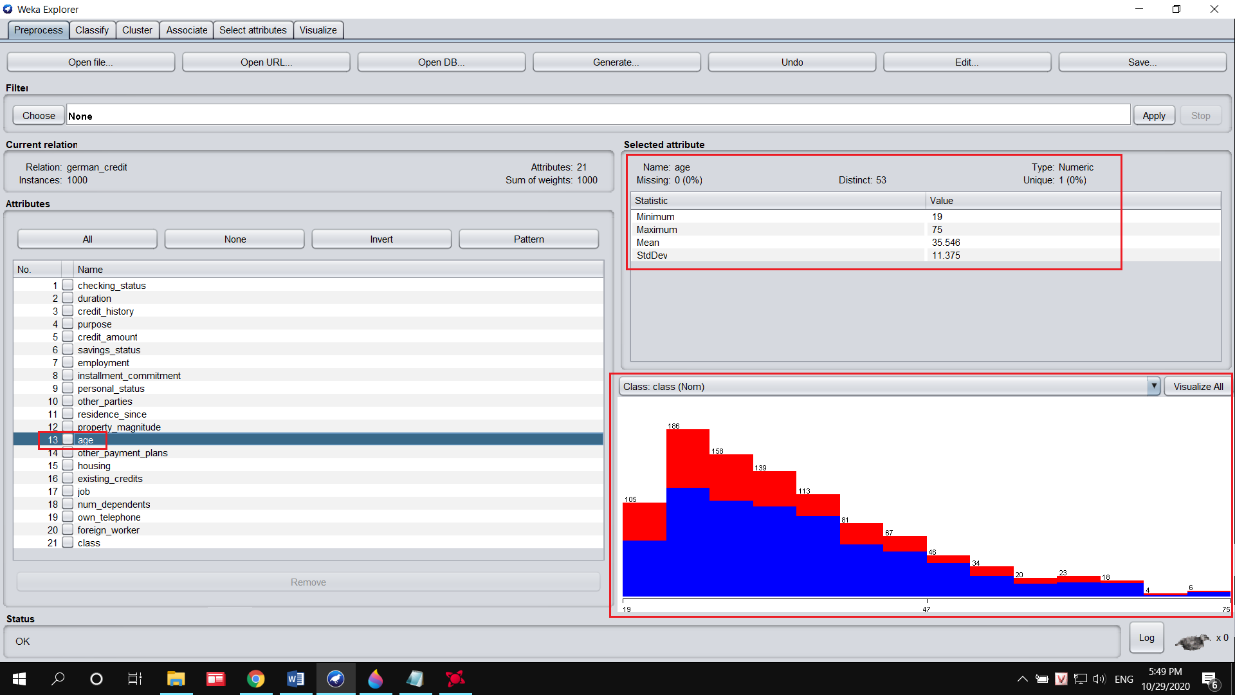
* Chỉ cần nhấn nút Visualize All ta có thể xem được tất cả các đồ thị.

1. *Chuyển sang tab Visualize. Thuật ngữ sử dụng trong textbook để đặt tên cho các đồ thị ở đây là gì? Chọn jitter tối đa để thấy tổng quan hơn về phân bố dữ liệu. Theo bạn có những cặp thuộc tính khác nhau nào có vẻ như tương quan với nhau không?*

* Thuật ngữ sử dụng trong textbook để đặt tên cho đồ thị này là Scatter plot matrix.
* Bằng trực quan em không thấy cặp thuộc tính nào có tương quan với nhau

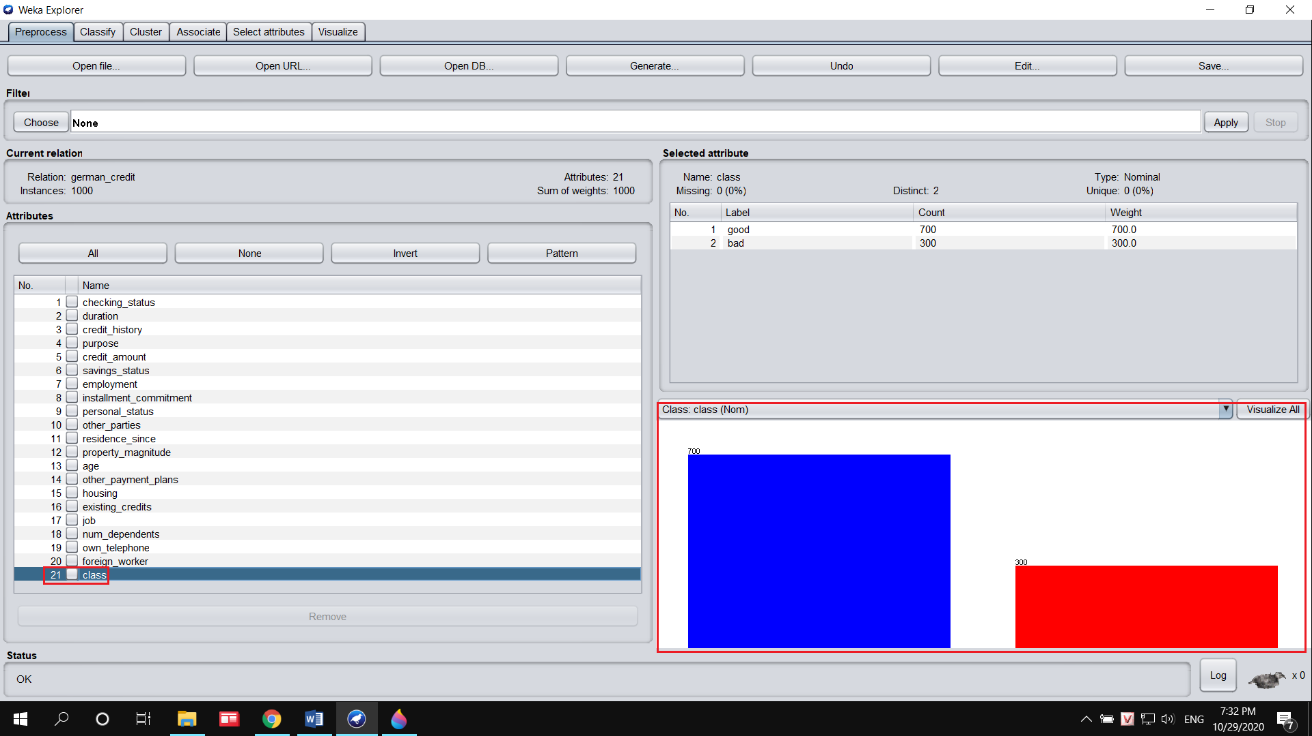
## Khám phá tập dữ liệu Tín dụng Đức

1. *Nội dung của phần ghi chú (comment) trong credit-g.arff (khi mở bằng 1 text editor bất kì) nói về điều gì? Tập dữ liệu có bao nhiêu mẫu? Bao nhiêu thuộc tính? Mô tả 5 thuộc tính bất kì (phải vừa có cả thuộc tính rời rạc và thuộc tính liên tục).*

* Phần ghi chú trong tập dữ liệu nói về các thông tin cơ bản của tập dữ liệu như: tiêu đề, nguồn gốc của tập dữ liệu, số lượng mẫu, mô tả thuộc tính, gán nhãn lại các giá trị trong từng thuộc tính categorical (như hình sau):
* Tập dữ liệu có 1000 mẫu và 21 thuộc tính
* Mô tả 5 thuộc tính:
* checking\_status
* Kiểu dữ liệu: nominal
* Dữ liệu thiếu: 0
* Nhãn = { ‘<0’, ‘0<=X<200’, ‘>=200’, ‘no checking’}
* ‘no checking’ và ‘>=200’ có tỷ lệ tín dụng tốt cao
* credit\_history
* Kiểu dữ liệu: nominal
* Dữ liệu thiếu: 0
* Nhãn = {‘no credits/all paid’, ‘all paid’, ‘existing paid’, ‘delayed previously’, ‘critical/other existing credit’}
* ‘existing paid’, ‘delayed previously’, ‘critical/other existing credit’ có tỷ lệ tín dụng tốt cao
* employment
* Kiểu dữ liệu: nominal
* Dữ liệu thiếu: 0
* Nhãn = {‘unemployed’, ‘<1’, ‘1<=X<4’, ‘4<=X<7’, ‘>=7’}
* ‘1<=X<4’, ‘4<=X<7’, ‘>=7’ có tỷ lệ tín dụng tốt cao
* credit\_amount
* Kiểu dữ liệu: numeric
* Dữ liệu thiếu: 0
* Dữ liệu thuộc [250, 18424]
* [250, 4288.667] có tỷ lệ tín dụng tốt cao
* age
* Kiểu dữ liệu: numeric
* Dữ liệu thiếu: 0
* Dữ liệu thuộc [19, 75]
* Độ tuổi thuộc (23 ,67] và (71, 75] có tỷ lệ tín dụng tốt cao

1. *Tên của thuộc tính lớp là gì? Đánh giá phân bố của các lớp, tức là cân bằng hay lệch về một lớp?*

* Thuộc tính lớp là class.
* Dựa vào đồ thị, ta thấy sự phân bố lệch về lớp ‘good’ (màu xanh) chiếm 70% số mẫu của tập dữ liệu cho thấy người Đức có tín dụng tốt có tỷ lệ khá cao.



1. *Sử dụng tab Select attributes. Liệt kê những lựa chọn khác nhau của Weka để chọn lọc thuộc tính, giải thích ngắn gọn từng phương pháp.*

* Các phương pháp của Weka để chọn lọc thuộc tính là:
* Các phương pháp đánh giá

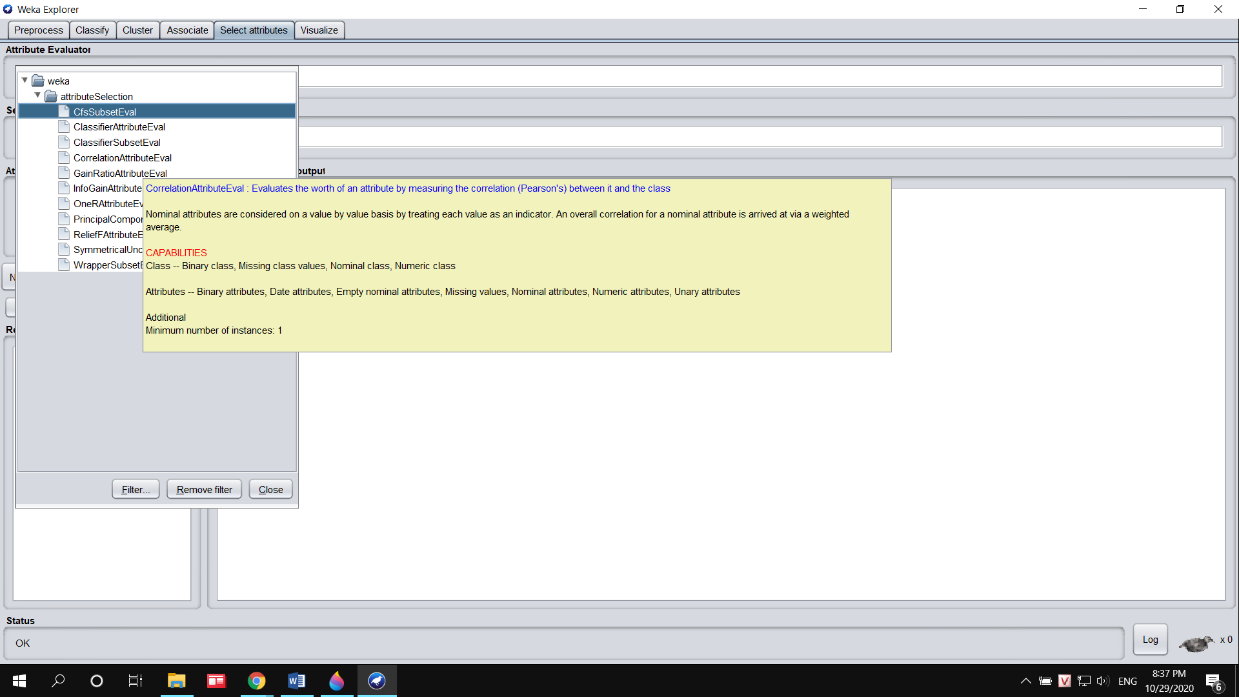
|  |  |  |
| --- | --- | --- |
| STT | Phương pháp | Mô tả |
| 1 | CfsSubsetEval | Đánh giá giá trị của một tập hợp con các thuộc tính bằng cách xem xét khả năng dự đoán riêng của từng tính năng cùng với mức độ dư thừa giữa chúng. |
| 2 | ClassifierAttributeEval | Đánh giá giá trị của thuộc tính bằng cách sử dụng trình phân loại do người dùng chỉ định. |
| 3 | ClassifierSubsetEval | Đánh giá tập hợp con thuộc tính trên dữ liệu đào tạo hoặc tập thử nghiệm tạm dừng riêng biệt. Sử dụng bộ phân loại để ước tính giá trị của một tập hợp các thuộc tính. |
| 4 | CorrelationAttributeEval | Đánh giá giá trị của một thuộc tính bằng cách đo lường mối tương quan giữa nó và lớp. |
| 5 | GainRatioAttributeEval | Đánh giá giá trị của một thuộc tính bằng cách đo tỷ lệ khuếch đại liên quan đến lớp. |
| 6 | InfoGainAttributeEval | Đánh giá giá trị của một thuộc tính bằng cách đo lường mức tăng thông tin liên quan đến lớp. |
| 7 | OneRAtrributeEval | Đánh giá giá trị của một thuộc tính bằng cách sử dụng bộ phân loại OneR. |
| 8 | PrincipalComponents | Thực hiện phân tích thành phần chính (chọn ra các thành phần chính) nhằm biến đổi dữ liệu. |
| 9 | ReliefFAttributeEval | Đánh giá giá trị của một thuộc tính bằng cách lấy mẫu liên tục một cá thể và xem xét giá trị của thuộc tính đã cho cho phiên bản gần nhất của cùng một lớp và khác lớp. Có thể hoạt động trên cả dữ liệu lớp rời rạc và liên tục. |
| 10 | SymmetricalUncertAttributeEval | Đánh giá giá trị của một thuộc tính bằng cách đo độ không đảm bảo đối xứng liên quan đến lớp. |
| 11 | WrapperSubsetEval | Đánh giá các tập thuộc tính bằng cách sử dụng lược đồ học tập. Xác thực chéo được sử dụng để ước tính độ chính xác của lược đồ học tập cho một tập hợp các thuộc tính. |

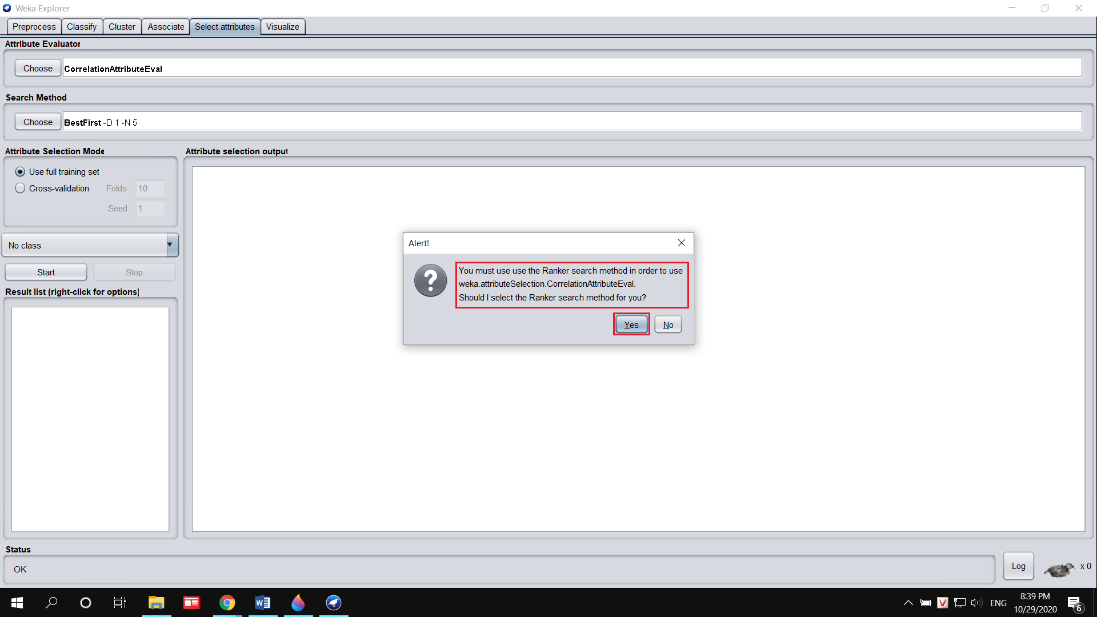
* Các phương pháp tìm kiếm

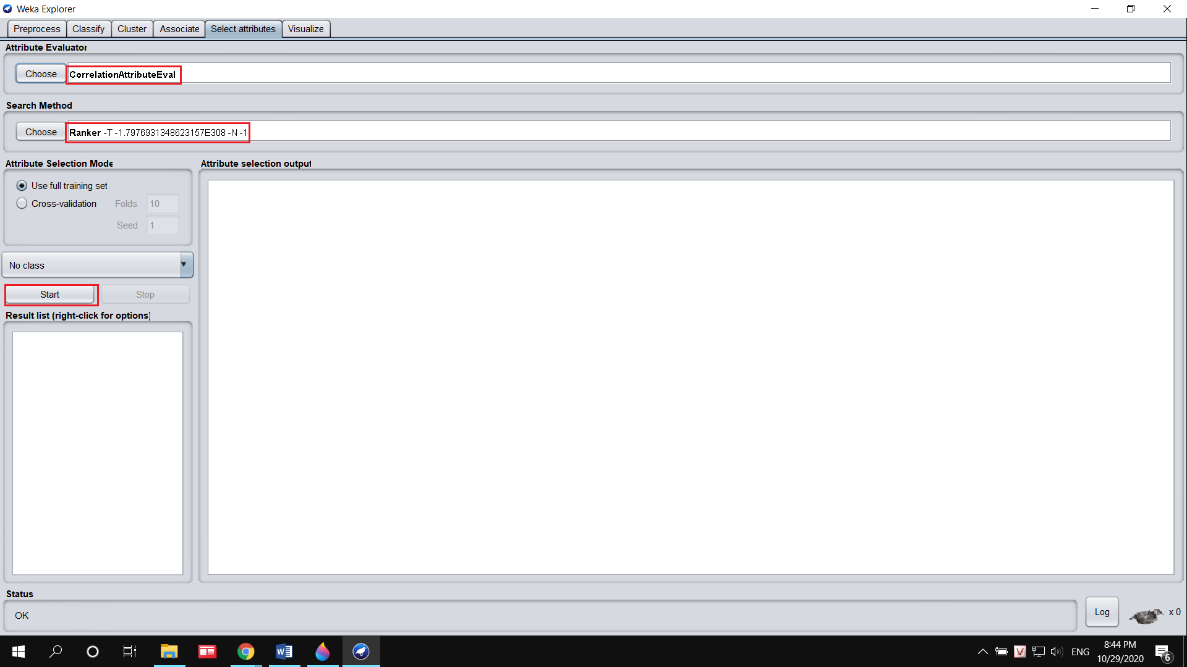
|  |  |  |
| --- | --- | --- |
| STT | Phương pháp | Mô tả |
| 1 | BestFirst | Tìm kiếm tập con không gian các thuộc tính bằng tham lam tăng cường với cơ sở quay lui. |
| 2 | GreedyStepwise | Thực hiện tìm kiếm tiến hoặc lùi tham lam thông qua không gian của các tập con thuộc tính. |
| 3 | Ranker | Xếp hạng thuộc tính theo đánh giá cá nhân của người dùng. Sử dụng kết hợp với các trình đánh giá thuộc tính (ReliefF, GainRatio, Entropy, v.v.). |

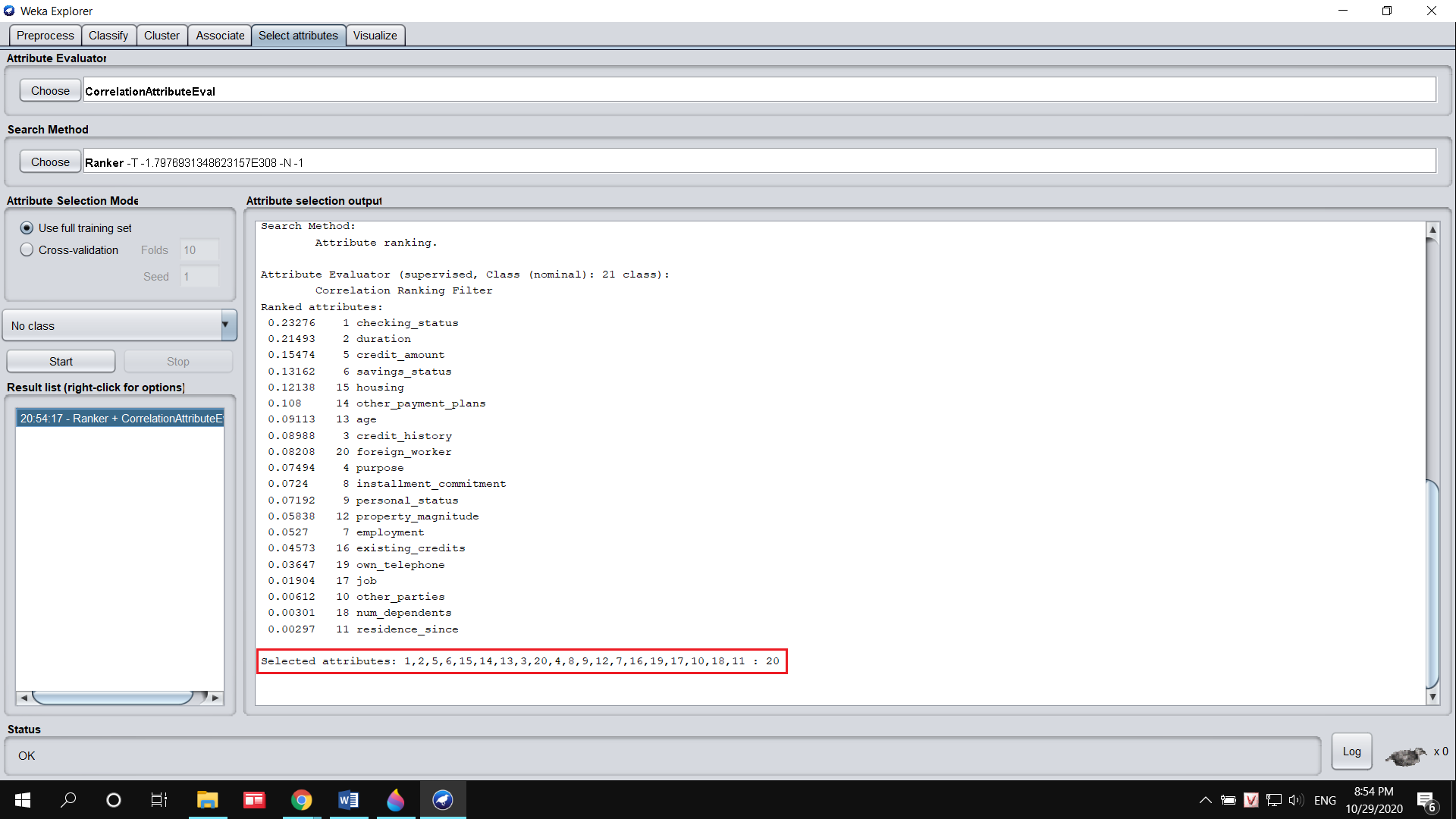
1. *Cần sử dụng bộ lọc nào để chọn ra 5 thuộc tính có tương quan cao nhất với thuộc tính lớp? Mô tả các bước làm, kèm theo hình chụp từng bước và kết quả cuối cùng.*

* Em sử dụng bộ lọc CorrelationAttributeEval để chọn ra 5 thuộc tính có tương quan cao nhất với thuộc tính lớp
* Các bước thực hiên

Bước 1: Chọn bộ lọc CorrelationAttributeEval

Bước 2: Khi chúng ta chọn bộ lọc CorrelationAttributeEval thì Weka buộc chúng ta phải sử dụng Ranker search nên chúng ta nhấn Yes.

Bước 3: Sau khi chọn xong bộ lọc thì nhấn nút Start

Bước 4: Sau khi thực hiện xong Weka xuất ra chuỗi thuộc tính có độ tương quan với thuộc tính lớp giảm dần.

Kết quả: 5 thuộc tính có tương quan với thuộc tính lớp nhất là: checking\_status, duration, credit\_amount, savings\_status, housing.

# PHỤ LỤC

# TÀI LIỆU THAM KHẢO

<https://www.tutorialspoint.com/weka/index.htm>

<https://machinelearningmastery.com/how-to-run-your-first-classifier-in-weka/>

<https://www.futurelearn.com/courses/data-mining-with-weka/0/steps/25377>