# 

# BÀI THỰC HÀNH SỐ 1

**1. MỤC TIÊU:**

Bài thực hành số 1 nhằm giúp sinh viên biết:

* Cài đặt weka
* Tiền xử lý dữ liệu

**2. YÊU CẦU:**

- Yêu cầu về điều kiện thực hành: Phòng máy thực hành có cài đặt phần mềm Weka, số lượng máy tương ứng với số sinh viên chia theo lớp thực hành.

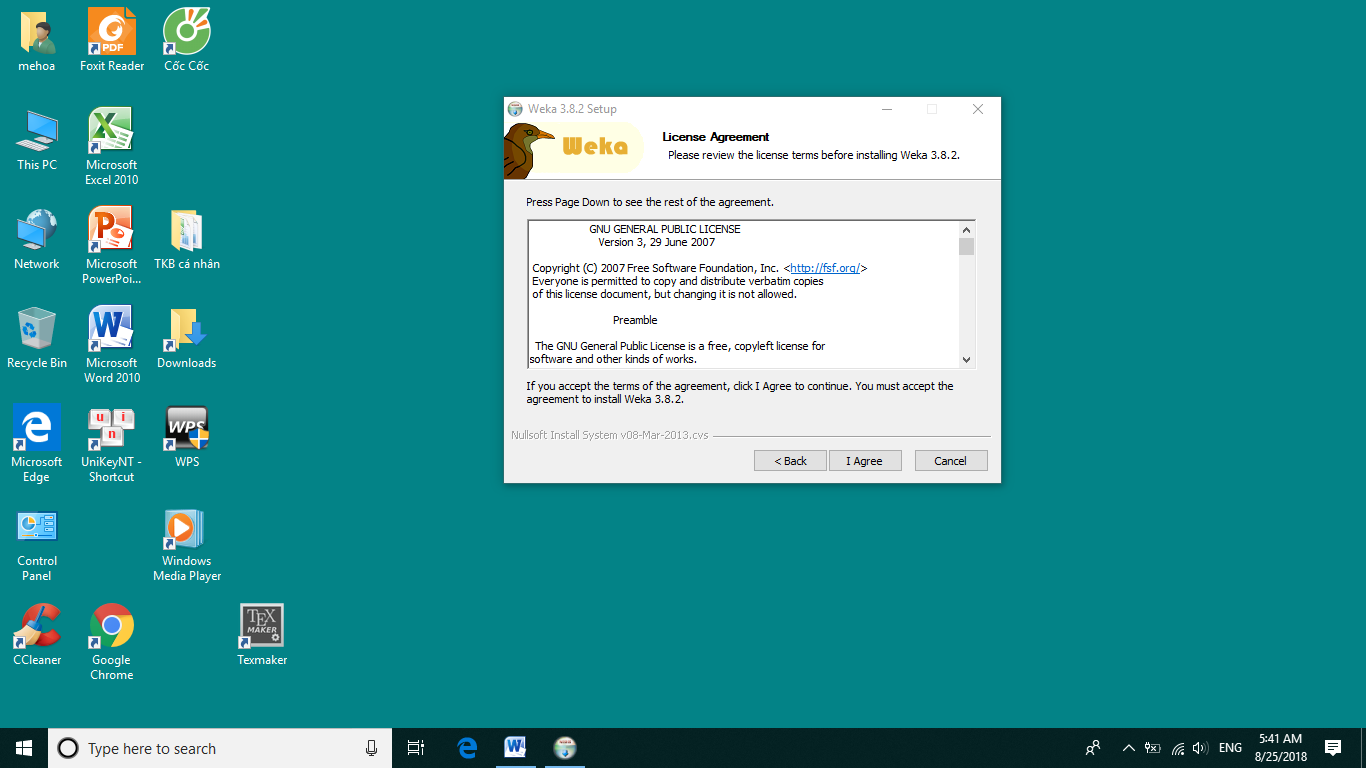
- Yêu cầu sinh viên: Chuẩn bị bài tập, tài liệu theo yêu cầu của đề cương môn học.

**3. NỘI DUNG BÀI THỰC HÀNH:**

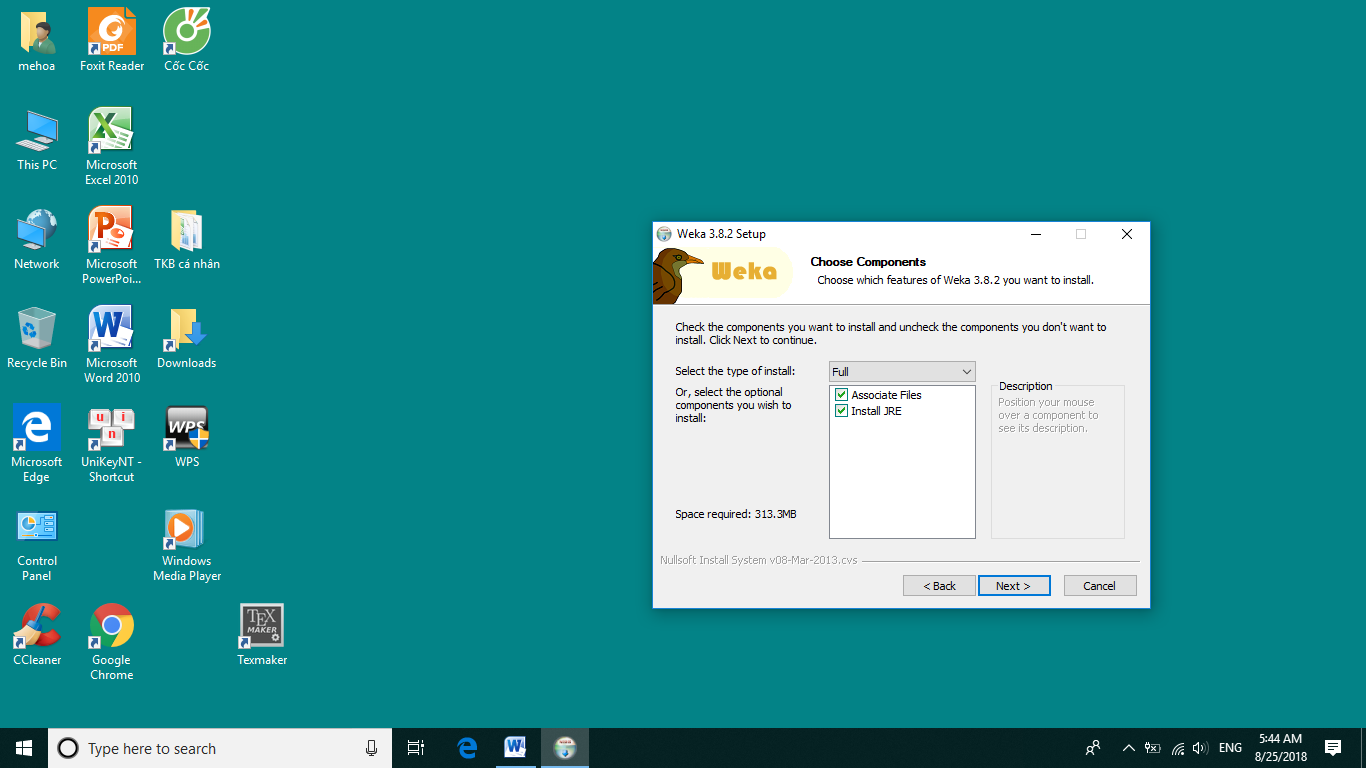
**3.1. Bài thực hành mẫu**

**\* Cài đặt wekwa**

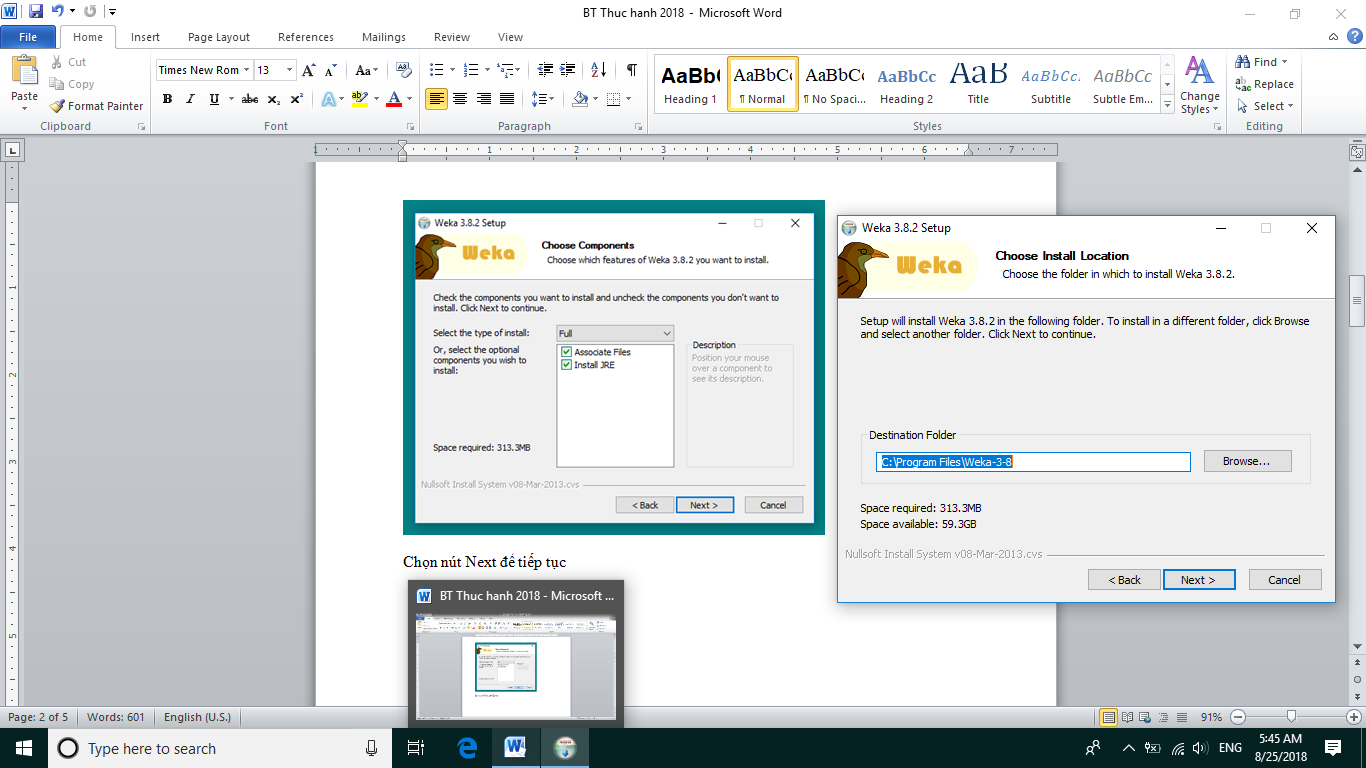
Chạy file cài weka



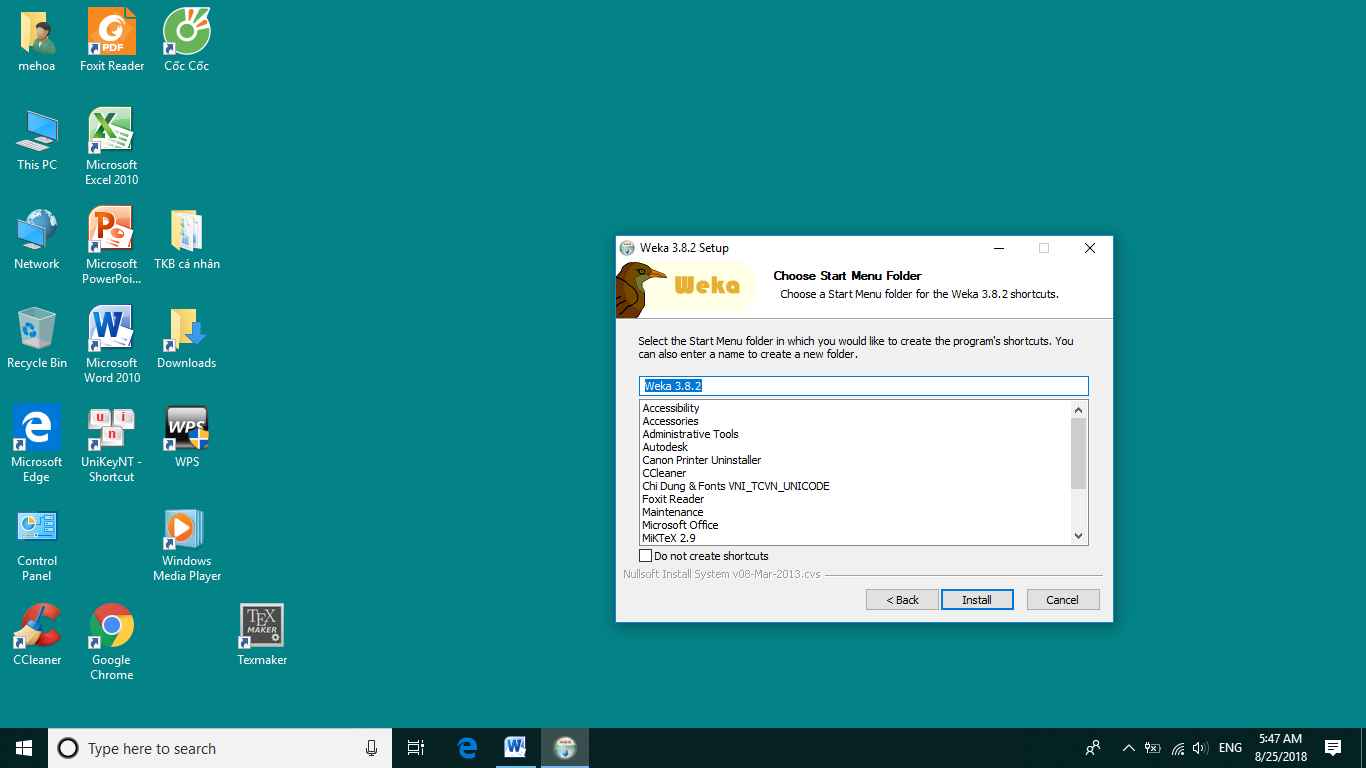
Chọn nút **I Agree** trên giao diện



Chọn nút **Next** để tiếp tục



Chỉ ra đường dẫn của thư mục chứa weka. Sau đó, chọn nút **Next**.



Chọn tên mục đặt ở phần **Start Menu**. Sau đó, chọn **Install.**

Chờ weka cài xong đưa ra thông báo thì bấm nút **Finish.**

**\*Tiền xử lý dữ liệu**

Các công việc cụ thể của tiền xử lý dữ liệu bao gồm những công việc như:

Filtering Attributes: Chọn các thuộc tính phù hợp với mô hình

Filtering samples: Lọc các mẫu (instances, patterns) dữ liệu cho mô hình

Clean data: Làm sạch dữ liệu như xóa bỏ các dữ liệu bất thường

Transformation: Chuyển đổi dữ liệu cho phù hợp với các mô hình như chuyển đổi dữ liệu từ numeric qua nomial hay ordinal

Discretization (rời rạc hóa dữ liệu): Nếu bạn có dữ liệu liên tục nhưng một vài mô hình chỉ áp dụng cho các dữ liệu rời rạc (như luật kết hợp chẳn hạn) thì bạn phải thực hiện việc rời rạc hóa dữ liệu.

**Tiền xử lý dữ liệu với Weka**

Dữ liệu được dùng là dữ liệu về khách hàng ngân hàng gồm 12 thuộc tính và 600 khách. Link đến tập dữ liệu: <http://bis.net.vn/files/storage/bank-data.rar>. Cấu trúc file bank-data.csv như sau:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **STT** | **Thuộc tính** | **Kiểu dữ liệu** | **Giá trị mẫu** | **Ý nghĩa** |
| **1** | **ID** | **Nominal** | **1,2,3,…** | **Mã KH** |
| **2** | **Age** | **Numeric** | **13** | **Tuổi** |
| **3** | **Sex** | **Nominal** | **Female, male** | **Giới** |
| **4** | **region** | **Nominal** | **City, town, rural, inner\_city, suburban** | **vùng** |
| **5** | **income** | **Numeric** | **34000** | **Thu nhập** |
| **6** | **Maried** | **Nominal** | **No,Yes** | **Tình trạng hôn nhân** |
| **7** | **children** | **Numeric** | **0,1,2,3** | **Số con** |
| **8** | **Car** | **Nominal** | **No,Yes** | **Có xe ko** |
| **9** | **save\_act** | **Nominal** | **No,Yes** | **Có gửi tiết kiệm ko** |
| **10** | **Current\_act** | **Nominal** | **No,Yes** | **Có tài khoản ko** |
| **11** | **Mortgage** | **Nominal** | **No,Yes** | **Có thể chấp ko** |
| **12** | **pep** | **Nominal** | **No,Yes** | **Có kế hoạch trả nợ ko** |

Vấn đề đặt ra là để thực hiện mô hình khai phá luật kết hợp trên file bank-data.csv ta cần phải làm các việc sau :

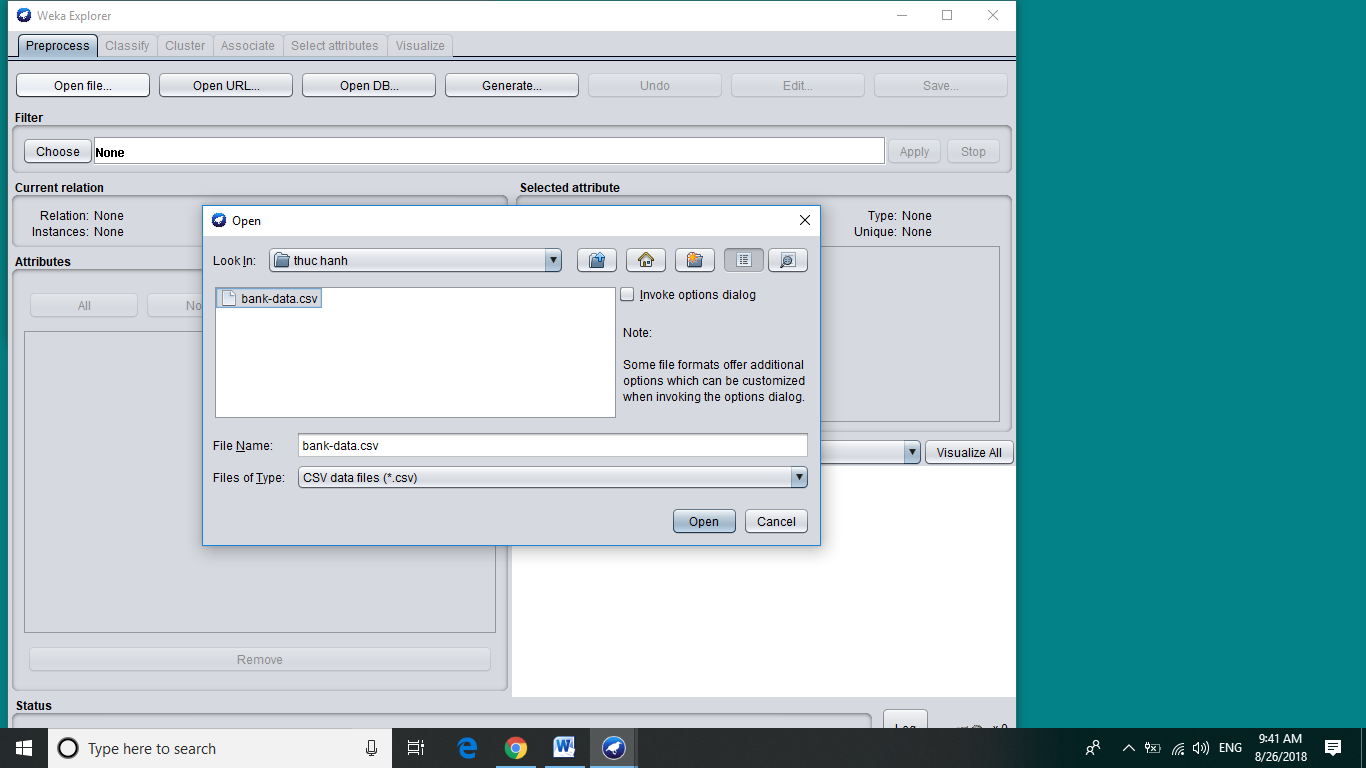
-        Loại bỏ thuộc tính id vì thuộc tính này không dùng trong mô hình. Quá trình này gọi là lọc thuộc tính (Filtering Attribute)

-        Rời rạc hóa giá trị của các trường liên tục là “children”,” age” và “income” vì mô hình khai phá luật kết hợp không làm việc với các kiểu dữ liệu liên tục. Quá trình này goi là rời rạc hóa dữ liệu (Discretization)

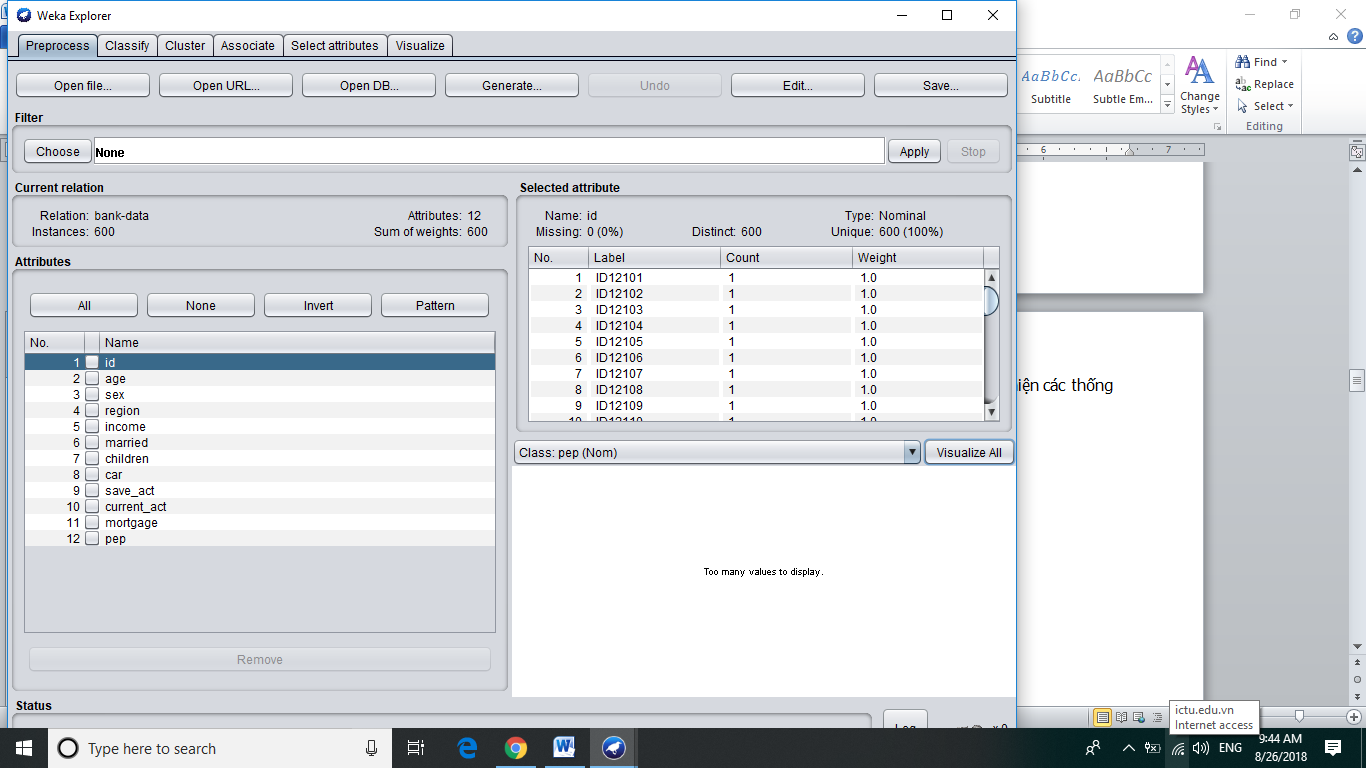
**Nạp dữ liệu (Loading the Data)**

Thông thường, định dạng chuẩn file dữ liệu của Weka la file ARFF (Attribute Relation File Format), tuy nhiên rất nhiều DBMS và Spreadsheet cho phép tổ chức file dữ liệu dưới dạng file .csv (comma-separated values) và một điều thuận lợi là Weka cho phép đọc dữ liệu từ file .csv.

 Nạp file dữ liệu data-bank.csv: Vào mục Weka Explorer, chọn tab “Preprocess” , chọn nút Open file, chỉ ra đường đẫn đến file, chọn file và bấm nút Open.



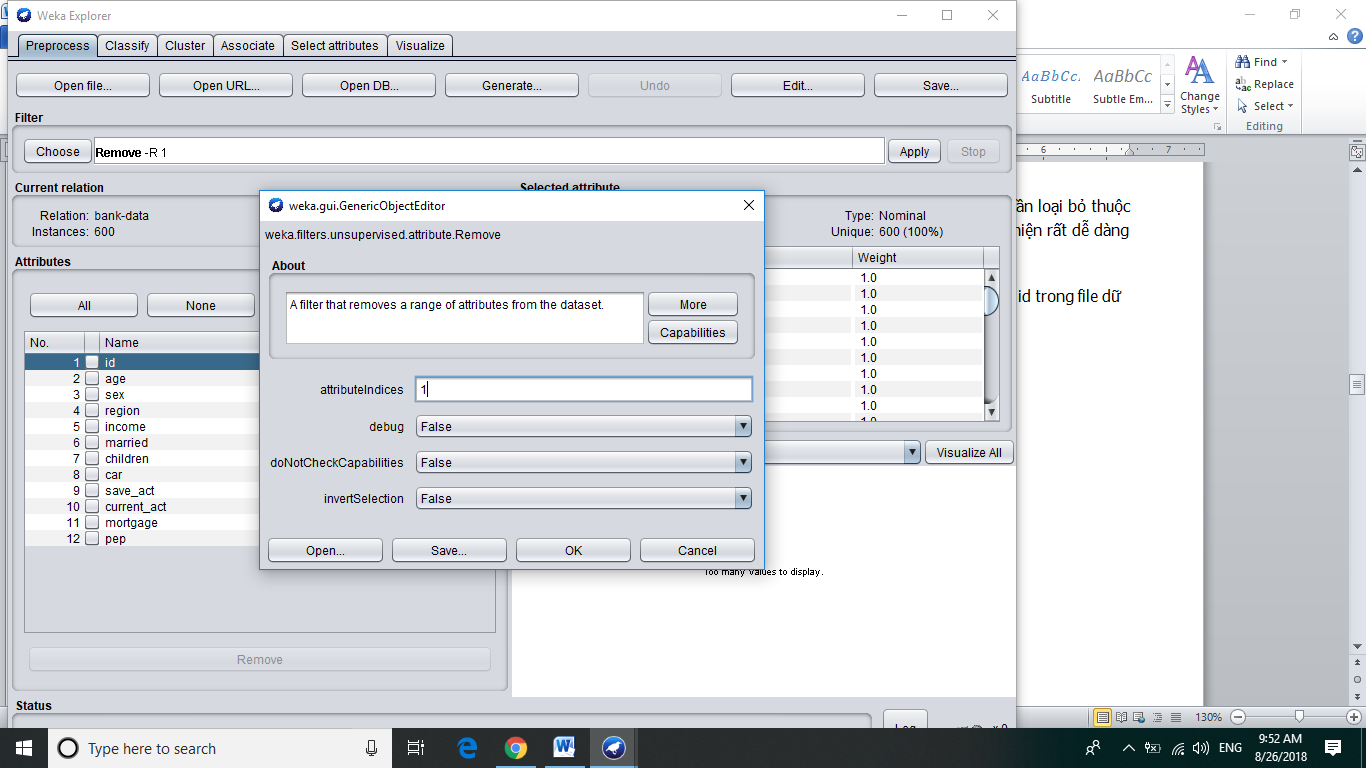
Sau khi dữ liệu được nạp lên, panel bên trái thể hiện các thuộc tính của file dữ liệu, panel bên phải thể hiện các thống kê tương ứng với thuộc tính bên trái. Chọn nút Visualize all để xem thống kê dữ liệu.



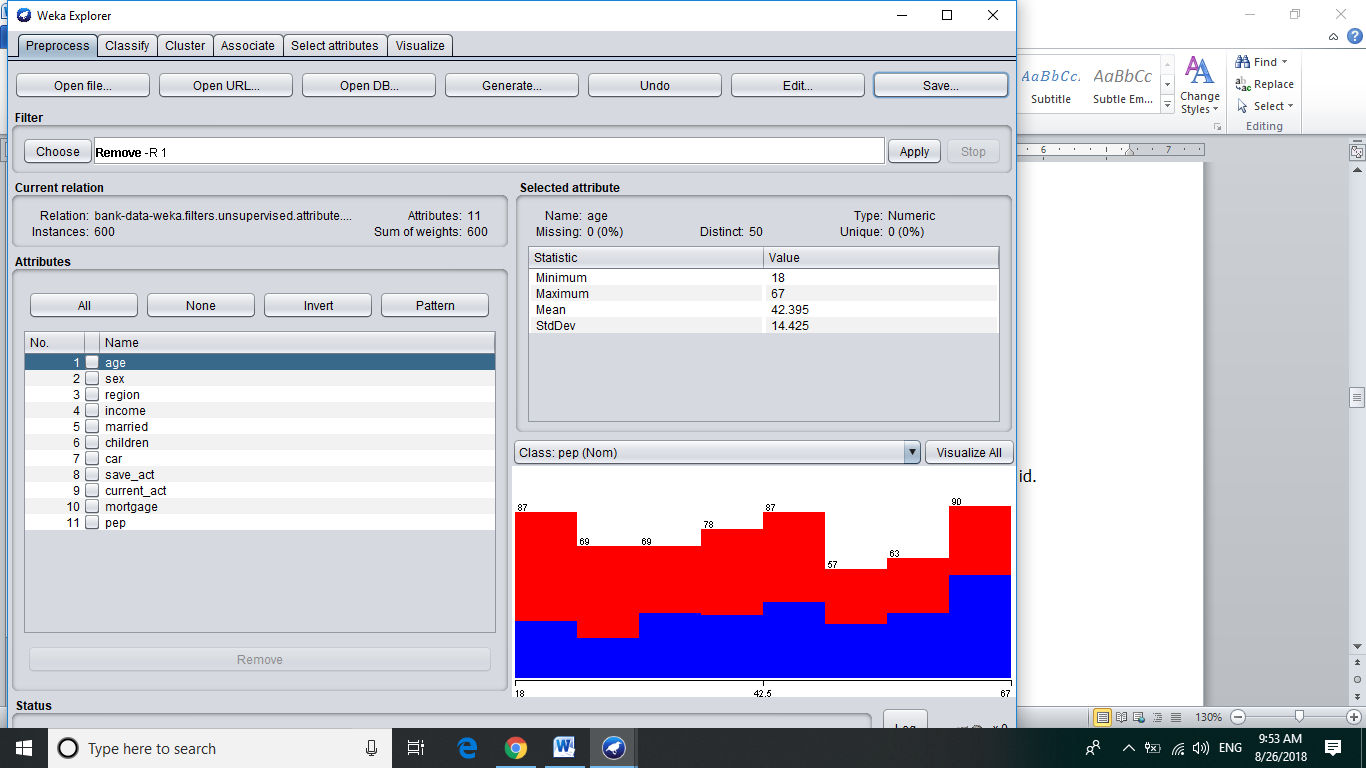
**Lọc thuộc tính (Filtering Attributes)**

Trong file dữ liệu bank-data.csv, mỗi khách hàng được xác định duy nhất bởi thuộc tính id. Chúng ta cần loại bỏ thuộc tính này trước khi thực hiện các bước tiếp theo trong Association Rule Mining. Công việc này được thực hiện rất dễ dàng trong Weka như sau: Trong panel Filter > Choose  > filters >unsupervised > attribute > Remove

Bước tiếp theo là bấm vào textbox ngay bên phải nút “Choose” và gỏ vào 1 (đây là index của thuộc tính id trong file dữ liệu). Chú ý rằng tùy chọn "invertSelection"  phải được thiết lập là false.



Bấm nút “Apply” sẽ tạo ra một liệu mới (working relation) với 11 thuộc tính sau khi đã loại bỏ thuộc tính id.



Bạn có thể thực hiện việc lọc nhiều thuộc tính khác nhau tùy theo ứng dụng của bạn theo cách trên.

Bây giờ ta lưu lại dữ liệu sau khi loại bỏ thuộc tính id dưới dạng file arff với tên là  "bank-data-R1.arff" .Để thực hiện việc rời rạc hóa dữ liệu ở bước tiếp theo.

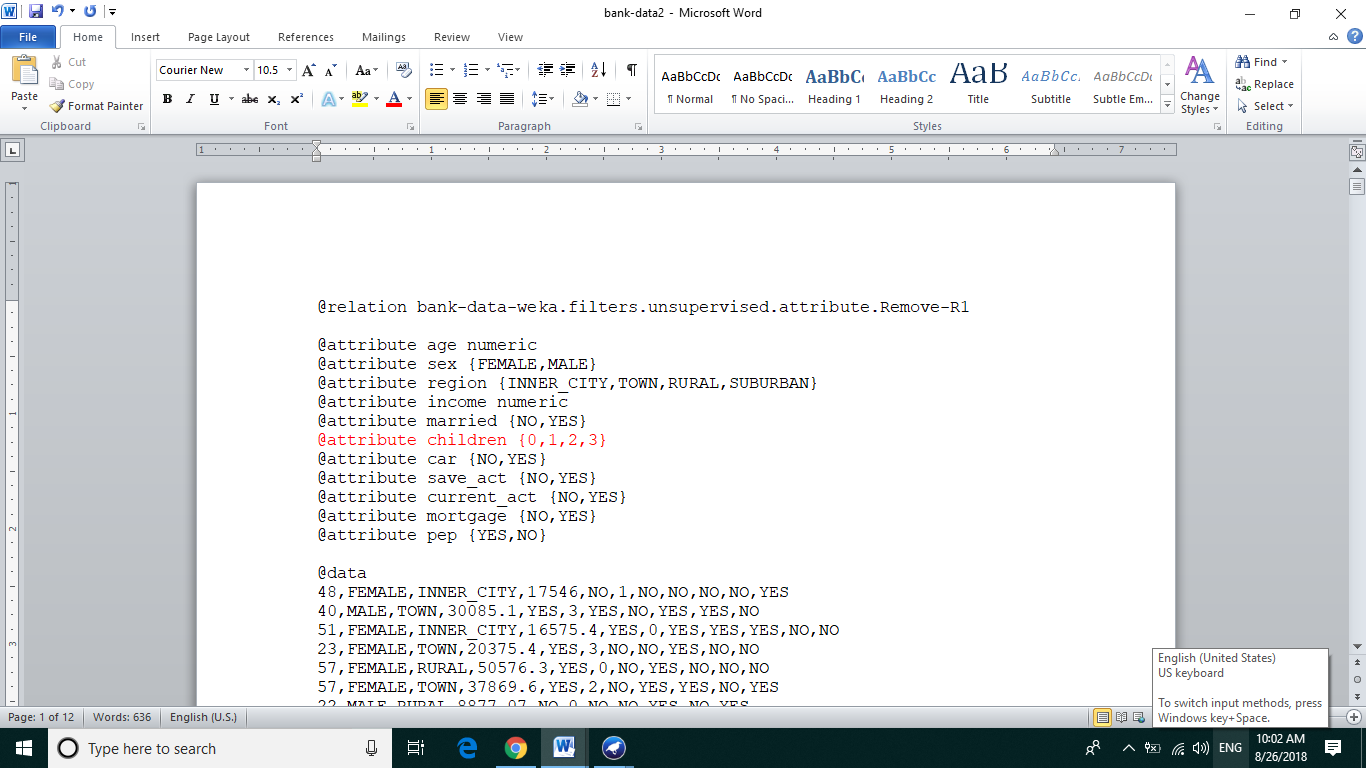
Chọn nút “save” và gỏ vào tên file "bank-data-R1.arff"

Chú ý rằng, khi thuộc tính “id” đã bị loại thì tất cả các giá trị của trường id trong các record cũng bị loại.

**Rời rạc hóa dữ liệu (Discretization)**

Trong Data Mining, một số kỹ thuật như khai phá luật kết hợp (association rule mining) chỉ có thể thực hiện trên các dữ liệu phân loại (categorical/ nominal data). Điều này yêu cầu phải thực hiện việc rời rạc hóa trên các thuộc tính có kiểu dữ liên tục (như kiểu numeric chẳn hạn) khi muốn áp dụng các kỹ thuật này. Trong file dữ liệu “bank-data-R1.arff” có 3 thuộc tính kiểu số, đó là "age", "income", và "children".

Đối với thuộc tính "children", vì phạm vi giá của nó chỉ có thể là 0,1,2 và 3 cho nên ta có thể giữ lại các giá trị của thuộc tính này. Việc rời rạc hóa cho thuộc tính “children” có thể thực hiện đơn giản bằng cách mở file “bank-data-R1.arff” bằng bất kỳ text editor nào (WordPad chẳn hạn) và thay từ khóa “numeric” bằng các giá trị rời rạc {0,1,2,3}. Sau đó lưu kết quả lại với tên file "bank-data2.arff".

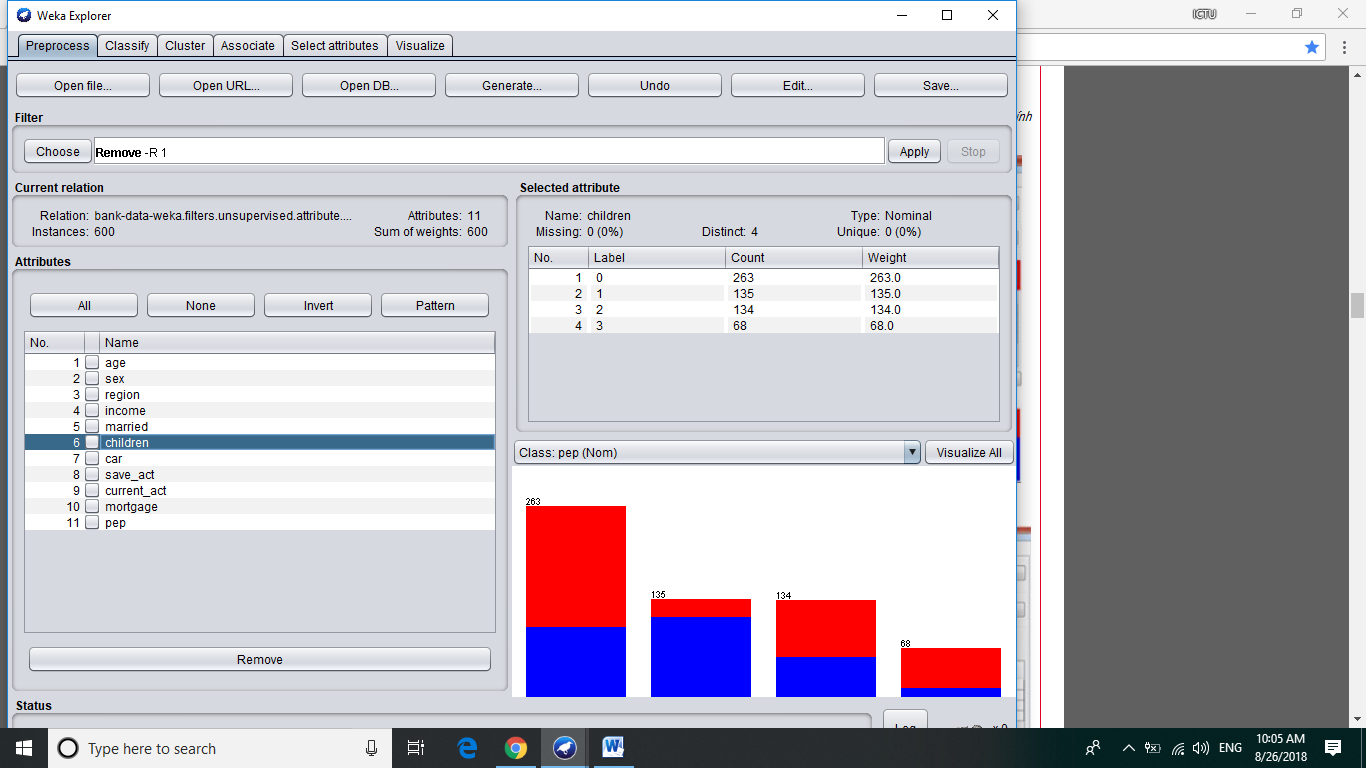


Bây giờ trên file “bank-data2.arff”, chúng ta sẽ thực hiện việc rời rạc hóa dữ liệu trên 2 thuộc tính là "age" và "income".

Trong ví dụ này, các dữ liệu liên tục trên 2 trường “age” và “income” sẽ được tự động chuyển vào 3 khoảng bins (intervals) như sau

**Nạp file dữ liệu “bank-data2.arff”**

Chú ý rằng thuộc tính “age” và “income” có kiểu dữ liệu là numeric. Và nếu ta chọn thuộc tính “children” ta sẽ thấy kiểu dữ liệu là nominal

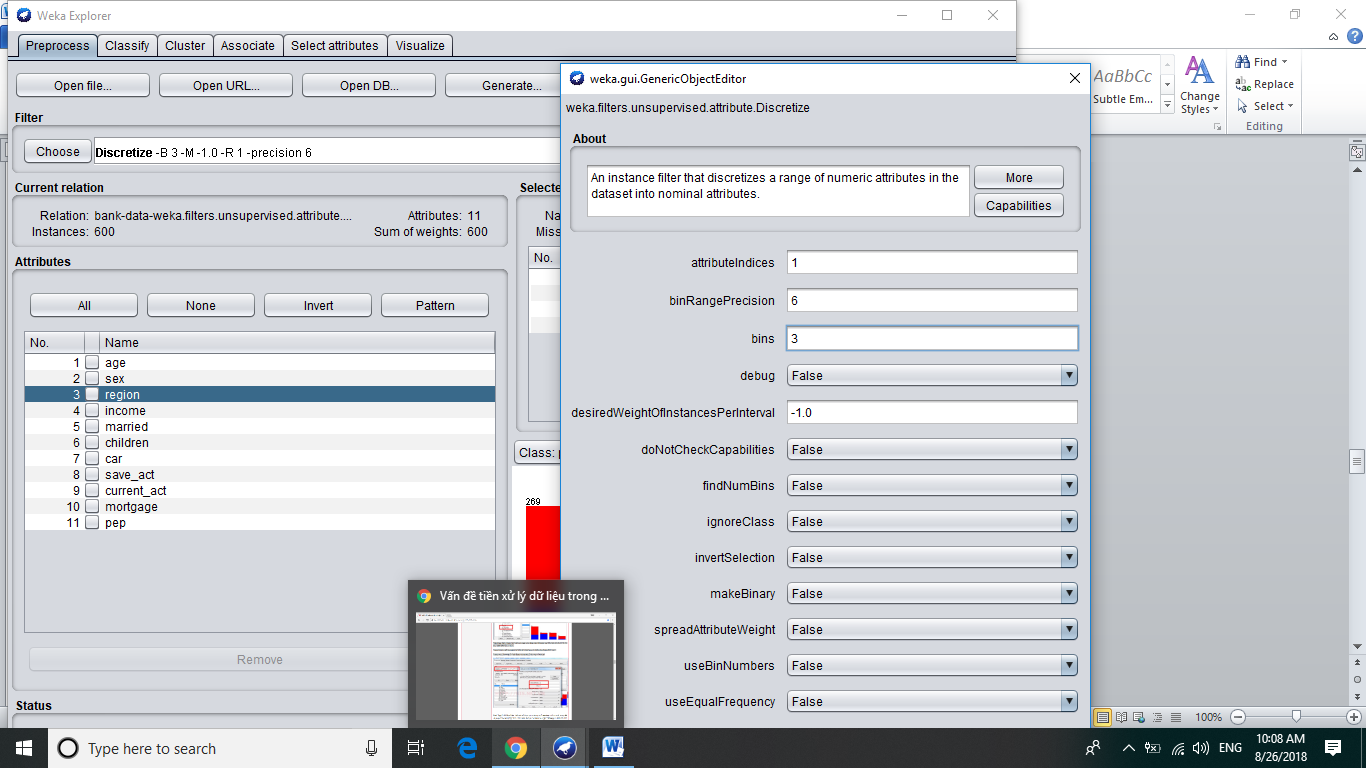


Bây giờ mở lại hộp thoại Filter và chọn: filters>unsupervised>attribute>Discretize

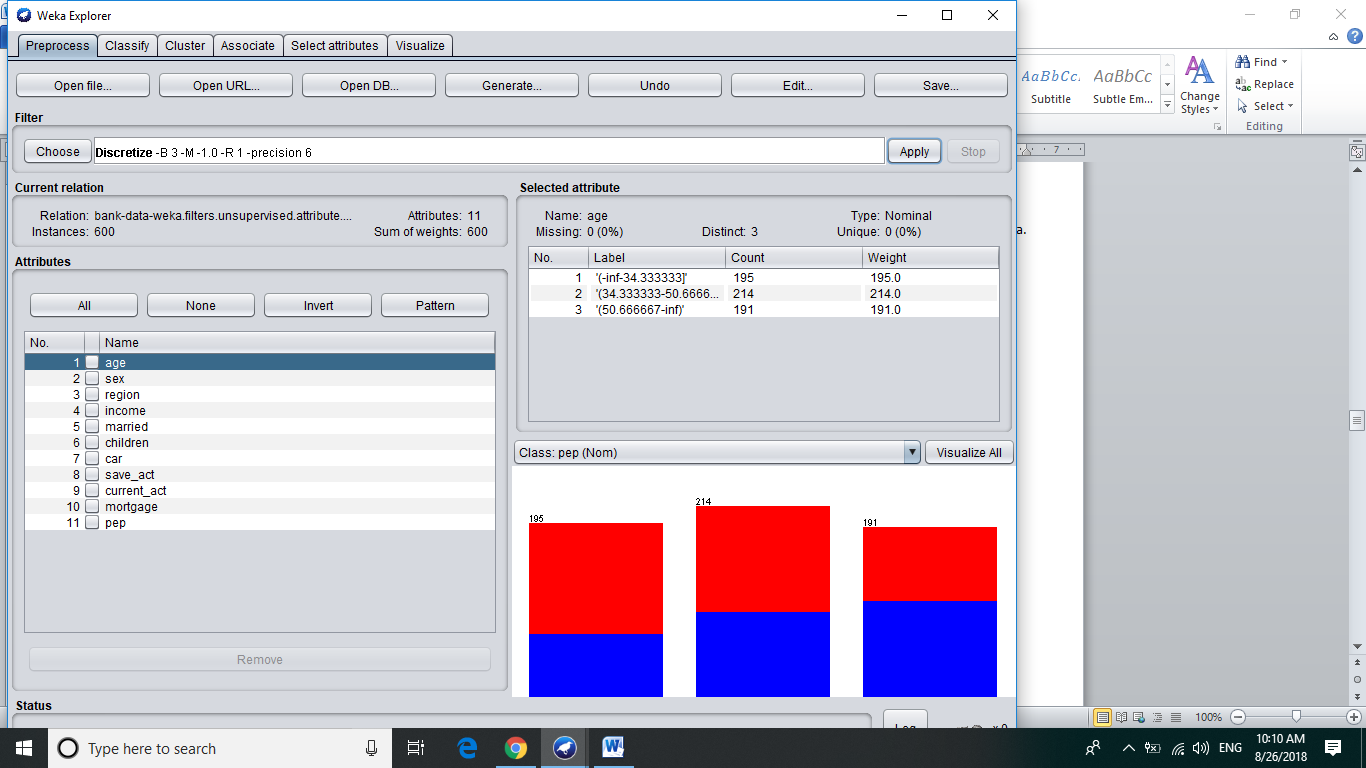
Tiếp theo, bấm chuột vào text box ngay bên phải nút “Choose” và thiết lập các tham số để thực hiện việc rời rạc hóa.

Trong textbox attributeIndices nhập 1 tương ứng với index của thuộc tính “age”

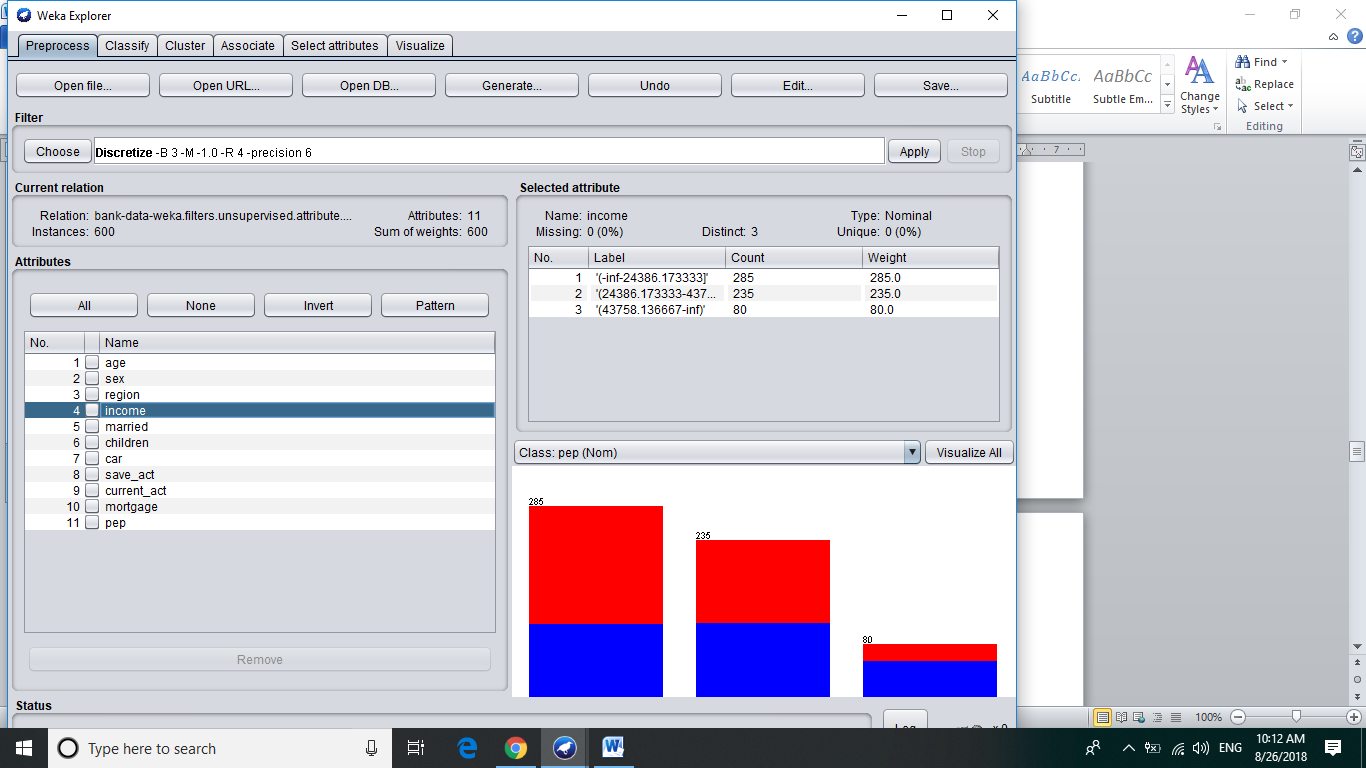
Trong mục bins nhập 3 (tuổi được chia thành 3 khoảng – interval)



Click "Apply" để thực hiện. Kết quả sẽ được tạo ra trong một working relation mới, trong đó các giá trị liên tục trong thuộc tính “age” sẽ được tự động chia vào 3 khoảng có nhãn lần lượt là "(-inf-34.333333]", "(34.333333-50.666667]" “(50.666667- inf)”



Tương tự, ta thực hiện việc rời rạc hóa thuộc tính “income” như sau (xem hình)và sau đó lưu lại file dữ liệu với tên “bank-data3.arff”



Chú ý rằng, kiểu dữ liệu cũng như giá trị của 2 thuộc tính “age” và “income” đã chuyển sang Nominal với 3 khoảng (bin, interval). Kiểm tra thuộc tính “age” ta sẽ thấy rằng có 3 độ tuổi có thể gọi là nhỏ, trung niên, già tương ứng với 3 khoảng là "(-inf-34.333333]", "(34.333333-50.666667]" “(50.666667- inf)”

Chúng ta có thể thay đổi các nhãn mà Weka tự động tạo ra bằng các nhãn dễ hiểu hơn. Chẳn hạn ta thay nhãn "(-inf-34.333333]" bằng 0\_34, nhãn "(34.333333-50.666667]" bằng 35\_51 và nhãn “(50.666667- inf)” bằng 52-max ta chỉ cần mở file dữ liệu ra bằng text editor có hỗ trợ find and replace (như wordpad chạn hạn) và thực hiện việc tìm kiếm và thay thế như trên

Đối với thuộc tính “income” ta thay thế các nhãn như sau:

(-inf-24386.173333] thay bằng 0\_24386

(24386.173333-43758.136667] thay bằng 24387\_43758

(43758.136667-inf)thay bằng 43759\_max

Sau khi đã thực hiện và thay thế các nhãn của 2 thuộc tính “age” và “income” ta lưu lại file dữ liệu cuối cùng có tên “bank-data-final.arff”.

Bây giờ, với file dữ liệu đã qua quá trình lọc thuộc tính id và rời rạc hóa dữ liệu trên 3 thuộc tính là “children”, “age” và “income”, bạn có thể thực hiện các kỹ thuật mining chỉ làm việc trên dữ liệu nominal (như association rule mining).

**3.2. Bài thực hành cơ bản**

Thực hiện chuẩn hóa dữ liệu của siêu thị cho nhiệm vụ tìm luật kết hợp

Link lấy dữ liệu: http://www.mediafire.com/file/pedsohybr24k99f/supermarket.arff

**3.3. Bài thực hành nâng cao**

Thực hiện chuẩn hóa dữ liệu về thời tiết cho nhiệm vụ tìm luật kết hợp. Link lấy dữ liệu: http://www.mediafire.com/download/0fd7pies72k7l79/weather.nominal.arff

# BÀI THỰC HÀNH SỐ 2

**1. MỤC TIÊU:**

Bài thực hành số 2 nhằm giúp sinh viên biết: Khai phá luật kết hợp bằng thuật toán Apriori

**2. YÊU CẦU:**

- Yêu cầu về điều kiện thực hành: Phòng máy thực hành có cài đặt phần mềm Weka, số lượng máy tương ứng với số sinh viên chia theo lớp thực hành.

- Yêu cầu sinh viên: Chuẩn bị bài tập, tài liệu theo yêu cầu của đề cương môn học.

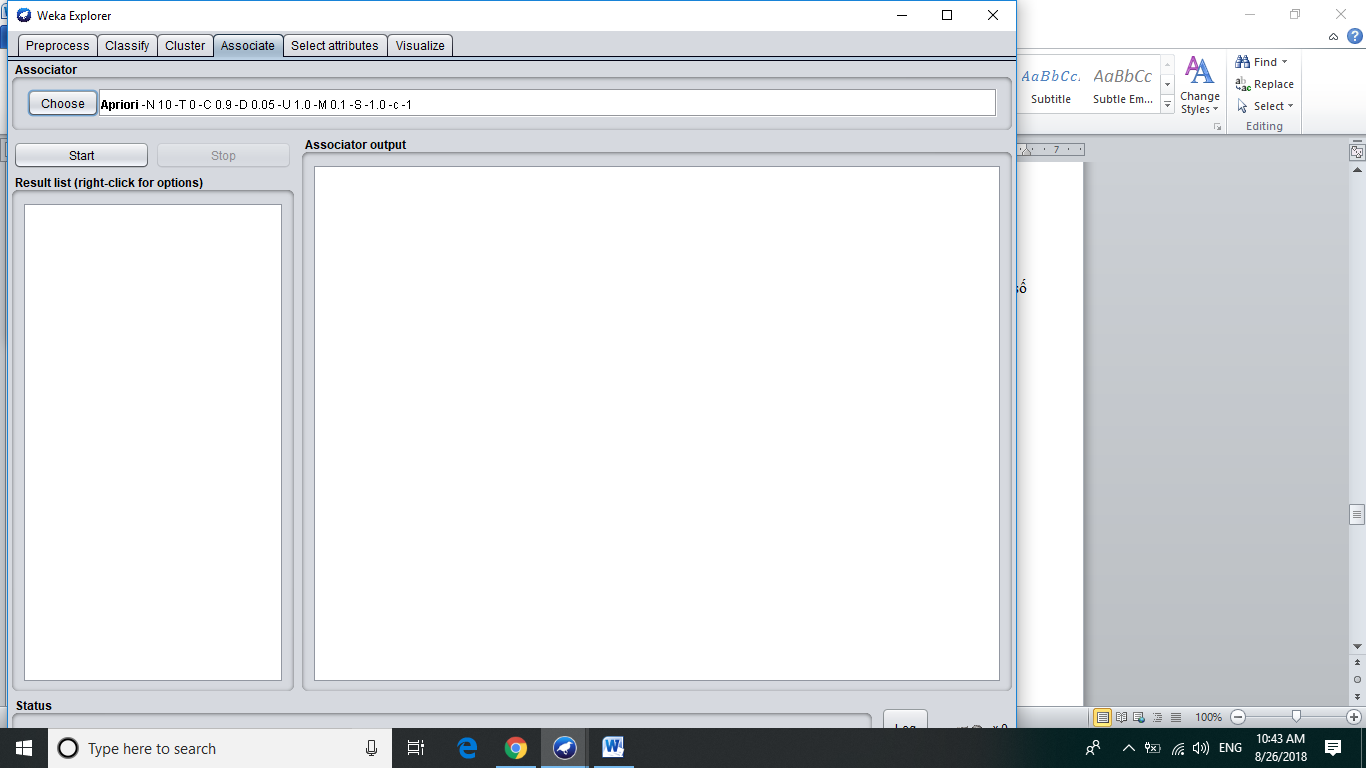
**3. NỘI DUNG BÀI THỰC HÀNH:**

**3.1. Bài thực hành mẫu**

Dữ liệu được dùng trong ví dụ này là dữ liệu về khách hàng ngân hàng (bank.arff) gồm 11 thuộc tính và 600 khách hàng mà đã được tiền xử lý trong bài thực hành số 1.

**Nạp dữ liệu (nạp file bank.arff)**

**Chọn tab Association và chọn thuật toán Apriori**



**Thiết lập các tham số (numRules, Support, Confidence,…)**

**Một số tham số chính của thuật toán Apriori sinh luật kết hợp**

*lowerBoundMinSupport:* Cận dưới của minimum support.

*metricType:* Có 4 loại metricType là Confidence, Lift , Leverage và Conviction.

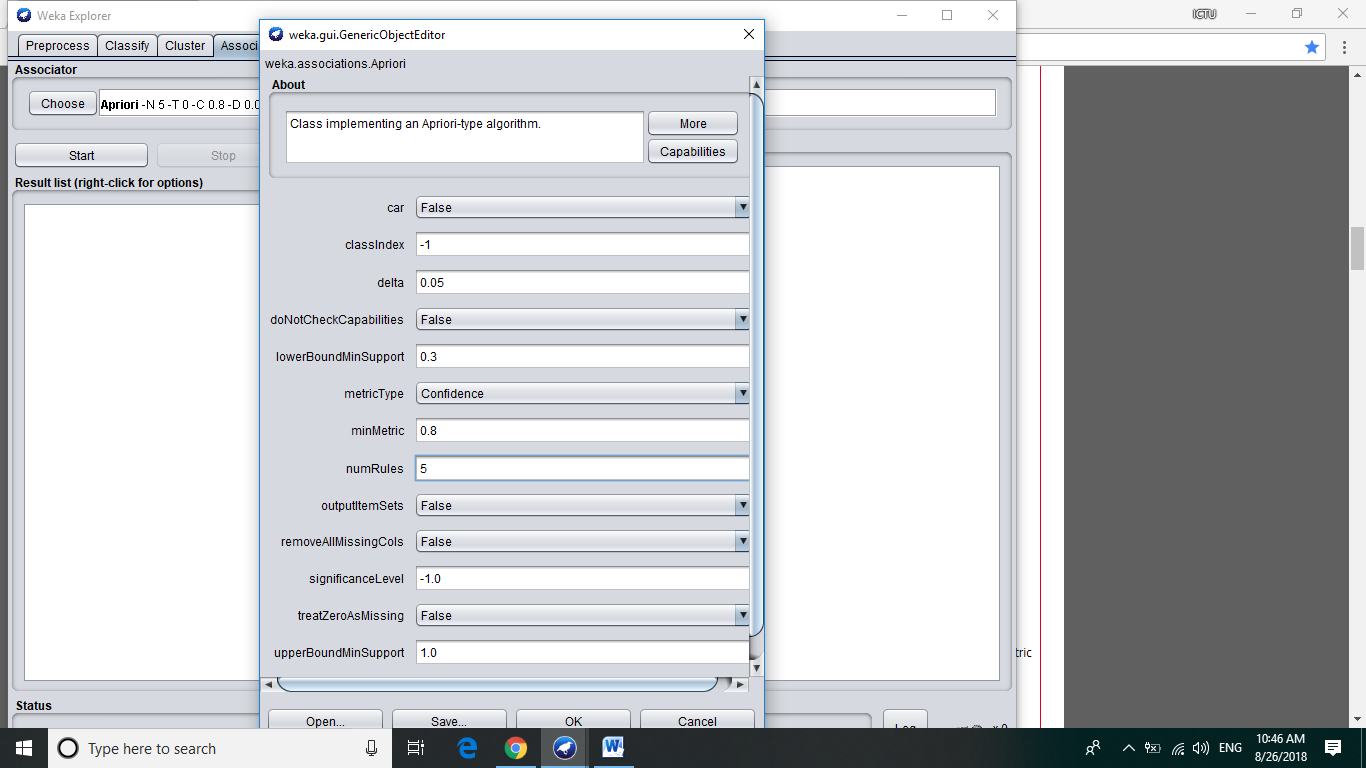
*Minimum metric score*: Chỉ quan quan đến các luật có metric score cao hơn giá trị này.

*numRules:* Số luật muốn tìm (các luật sẽ được sắp xếp theo thứ tự giảm dần của metric score.

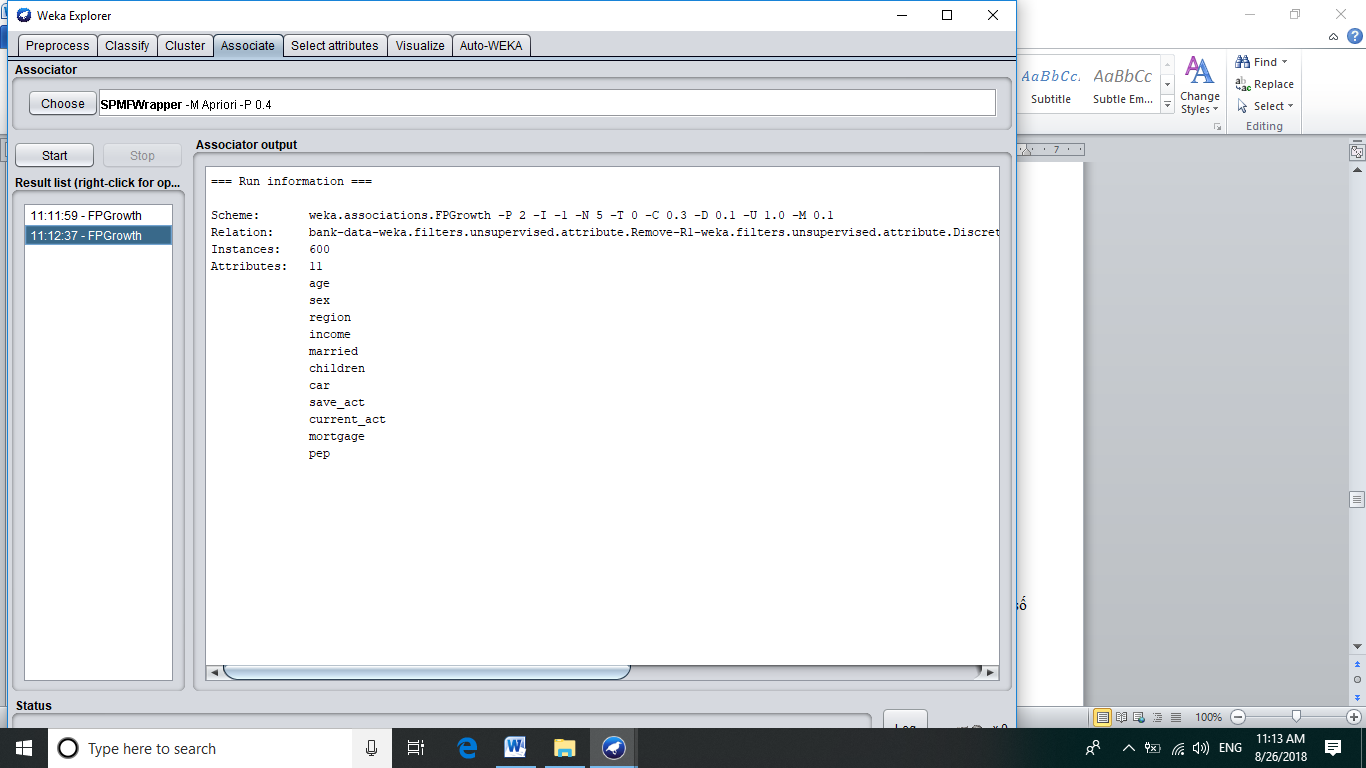
*significanceLevel:* Mức ý nghĩa (chỉ dùng khi metricType là confidence).

upperBoundMinSupport: Cận trên của minimum support (bắt đầu lặp lại việc giảm minimum support từ upperBoundMinSupport đến lowerBoundMinSupport).

Các tham số đặt (số luật=10, min-supp=0.1, min\_conf=0.6)



**Bấm nút Start để xem kết quả**



**3.2. Bài thực hành cơ bản**

a. Nhập dữ liệu về các đăng kí môn học của 50 sinh viên. Tìm luật kết hợp thể hiện sự liên kết giữa các môn học trong việc đăng kí.

b. Nhập dữ liệu về các bộ phim đã xem của 50 sinh viên. Tìm luật kết hợp thể hiện sự liên kết giữa các bộ phim đã xem.

**3.3. Bài thực hành nâng cao**

a. Thực hiện tìm luật kết hợp từ dữ liệu của siêu thị.

b. Thực hiện tìm luật kết hợp từ dữ liệu về thời tiết .

# BÀI THỰC HÀNH SỐ 3

**1. MỤC TIÊU:**

Bài thực hành số 3 nhằm giúp sinh viên biết: Khai phá luật kết hợp bằng thuật toán Fp-growth

**2. YÊU CẦU:**

- Yêu cầu về điều kiện thực hành: Phòng máy thực hành có cài đặt phần mềm Weka, số lượng máy tương ứng với số sinh viên chia theo lớp thực hành.

- Yêu cầu sinh viên: Chuẩn bị bài tập, tài liệu theo yêu cầu của đề cương môn học.

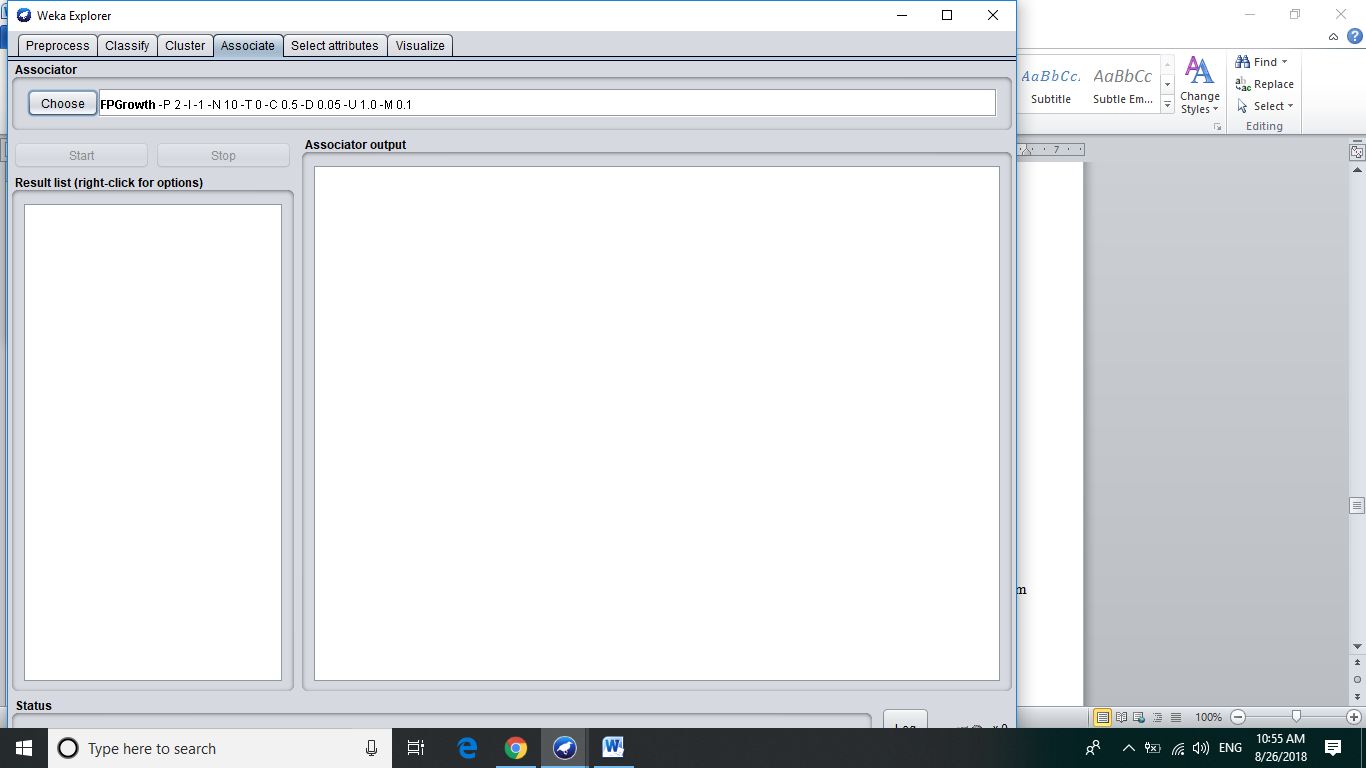
**3. NỘI DUNG BÀI THỰC HÀNH:**

**3.1. Bài thực hành mẫu**

Dữ liệu được dùng trong ví dụ này là dữ liệu về khách hàng ngân hàng (bank.arff) gồm 11 thuộc tính và 600 khách hàng mà đã được tiền xử lý trong bài thực hành số 1.

**Nạp dữ liệu (nạp file bank.arff)**

**Chọn tab Association và chọn thuật toán FP-growth**



**Thiết lập các tham số (numRules, Support, Confidence,…)**

**Một số tham số chính của thuật toán FPgrowth sinh luật kết hợp**

*lowerBoundMinSupport:* Cận dưới của minimum support.

*metricType:* Có 4 loại metricType là Confidence, Lift , Leverage và Conviction.

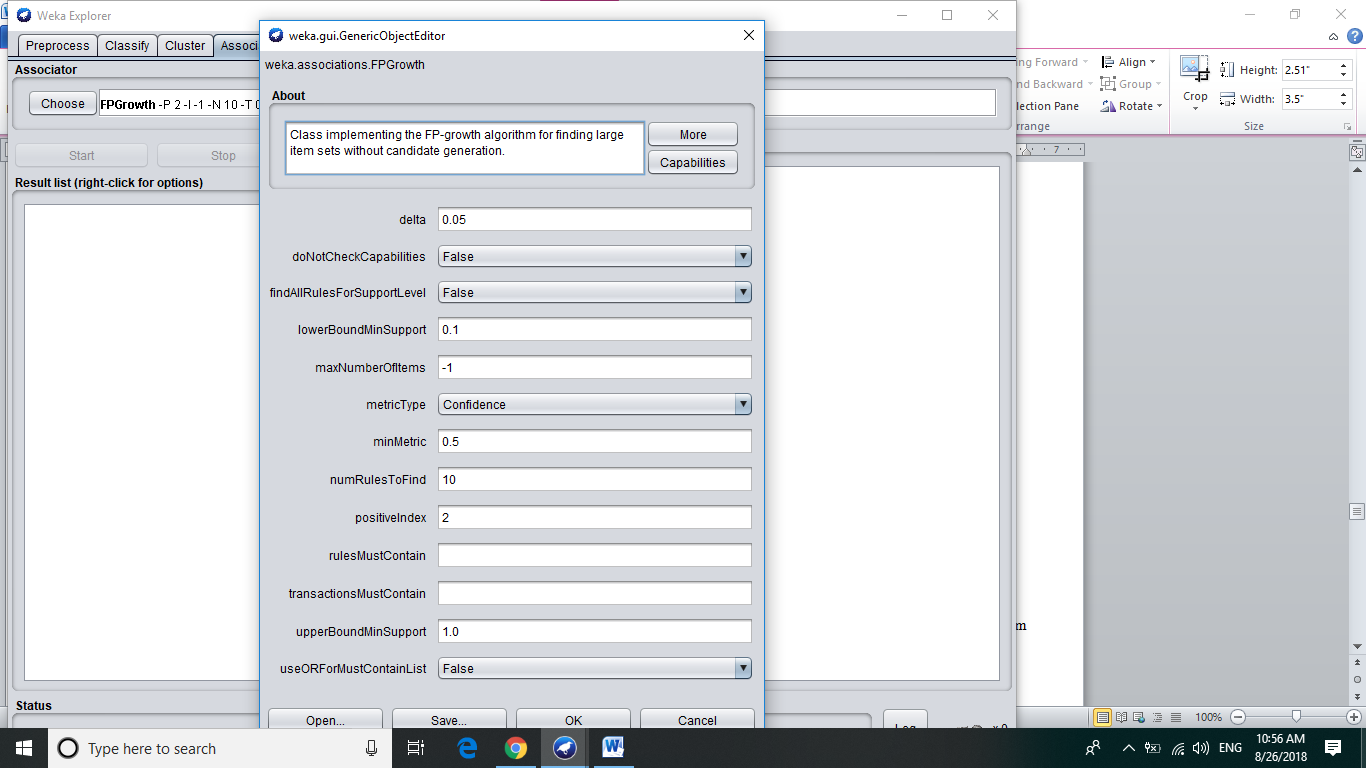
*Minimum metric score*: Chỉ quan quan đến các luật có metric score cao hơn giá trị này.

*numRules:* Số luật muốn tìm (các luật sẽ được sắp xếp theo thứ tự giảm dần của metric score.

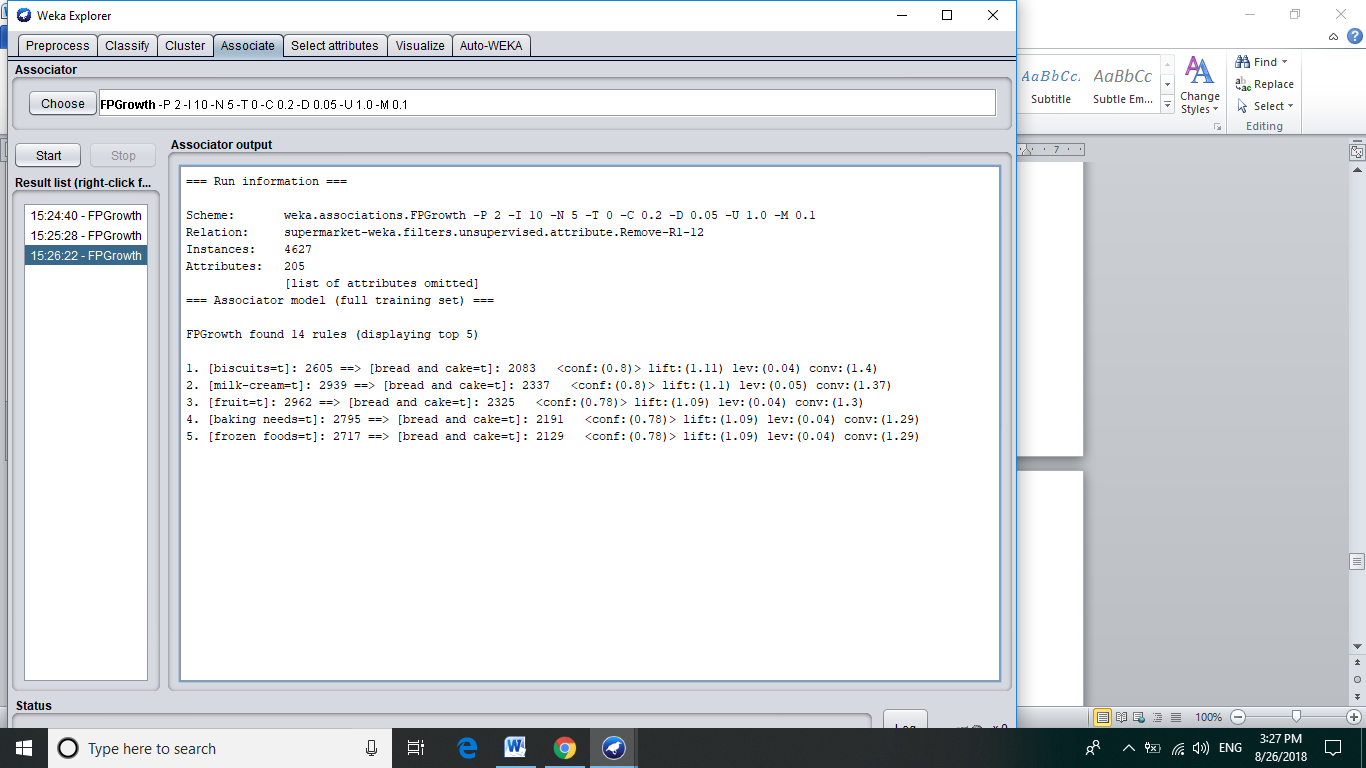
*significanceLevel:* Mức ý nghĩa (chỉ dùng khi metricType là confidence).

upperBoundMinSupport: Cận trên của minimum support (bắt đầu lặp lại việc giảm minimum support từ upperBoundMinSupport đến lowerBoundMinSupport).

Các tham số đặt (số luật=10, min-supp=0.1, min\_conf=0.6)



**Bấm nút Start để xem kết quả**



**3.2. Bài thực hành cơ bản**

a. Nhập dữ liệu về các đăng kí môn học của 50 sinh viên. Tìm luật kết hợp thể hiện sự liên kết giữa các môn học trong việc đăng kí.

b. Nhập dữ liệu về các bộ phim đã xem của 50 sinh viên. Tìm luật kết hợp thể hiện sự liên kết giữa các bộ phim đã xem.

**3.3. Bài thực hành nâng cao**

a. Thực hiện tìm luật kết hợp từ dữ liệu của siêu thị.

b. Thực hiện tìm luật kết hợp từ dữ liệu về thời tiết .

# BÀI THỰC HÀNH SỐ 4

**1. MỤC TIÊU:**

Bài thực hành số 4 nhằm giúp sinh viên biết:

Chia tập dữ liệu thành tập huấn luyện (training set) và tập kiểm tra (testing set)

Phân lớp dữ liệu bằng cây quyết định

**2. YÊU CẦU:**

- Yêu cầu về điều kiện thực hành: Phòng máy thực hành có cài đặt phần mềm Weka, số lượng máy tương ứng với số sinh viên chia theo lớp thực hành.

- Yêu cầu sinh viên: Chuẩn bị bài tập, tài liệu theo yêu cầu của đề cương môn học.

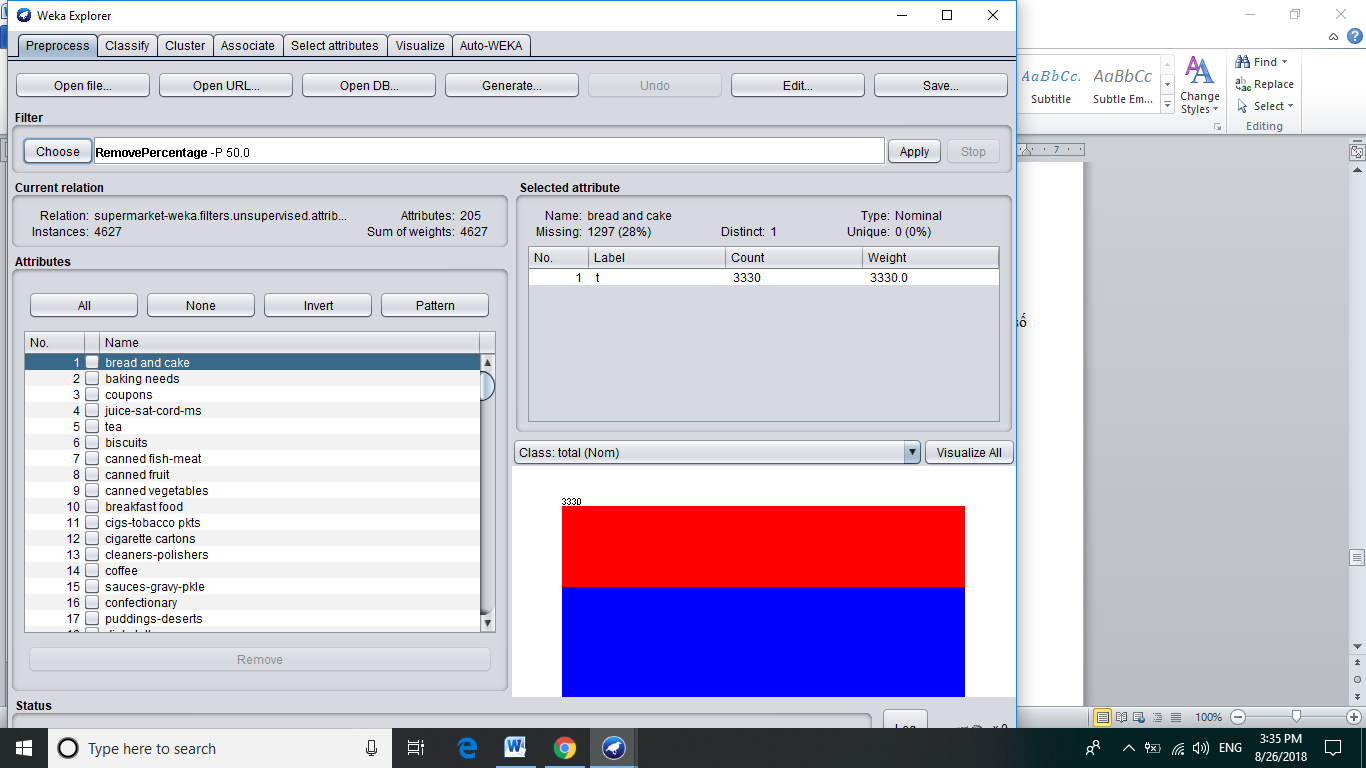
**3. NỘI DUNG BÀI THỰC HÀNH:**

**3.1. Bài thực hành mẫu**

\***Chia đôi tập dữ liệu**

Mở file chứa dữ liệu (supermarket1.arff)

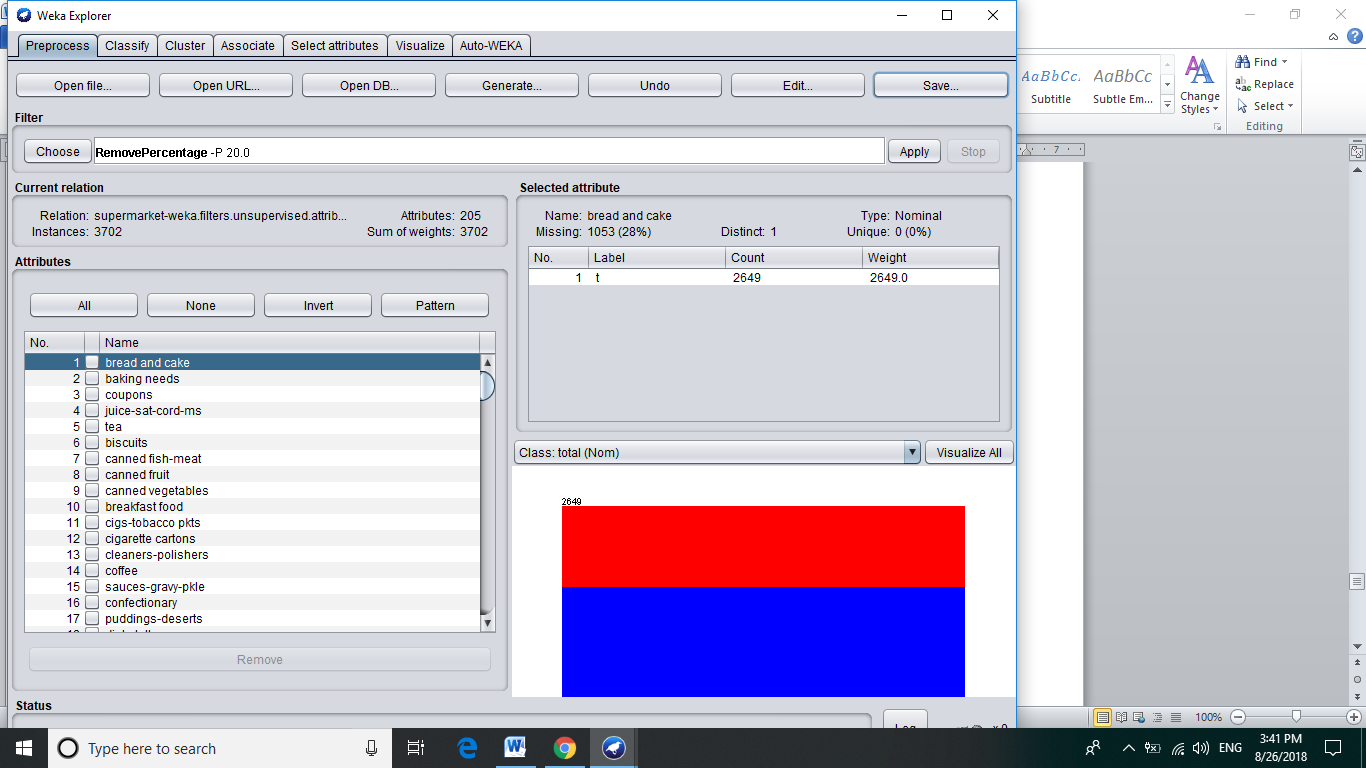
Chọn Filters>Unsupervised>Instance>RemovePercentage



Đặt các tham số cho việc chia dữ liệu

**Percentage: tỷ lệ % số mẫu bị loại bỏ**

Đặt Percentage=20, chọn OK



Chọn Apply và Save vào file supermarket\_train.arff

Tiếp tục để tạo tập testing.

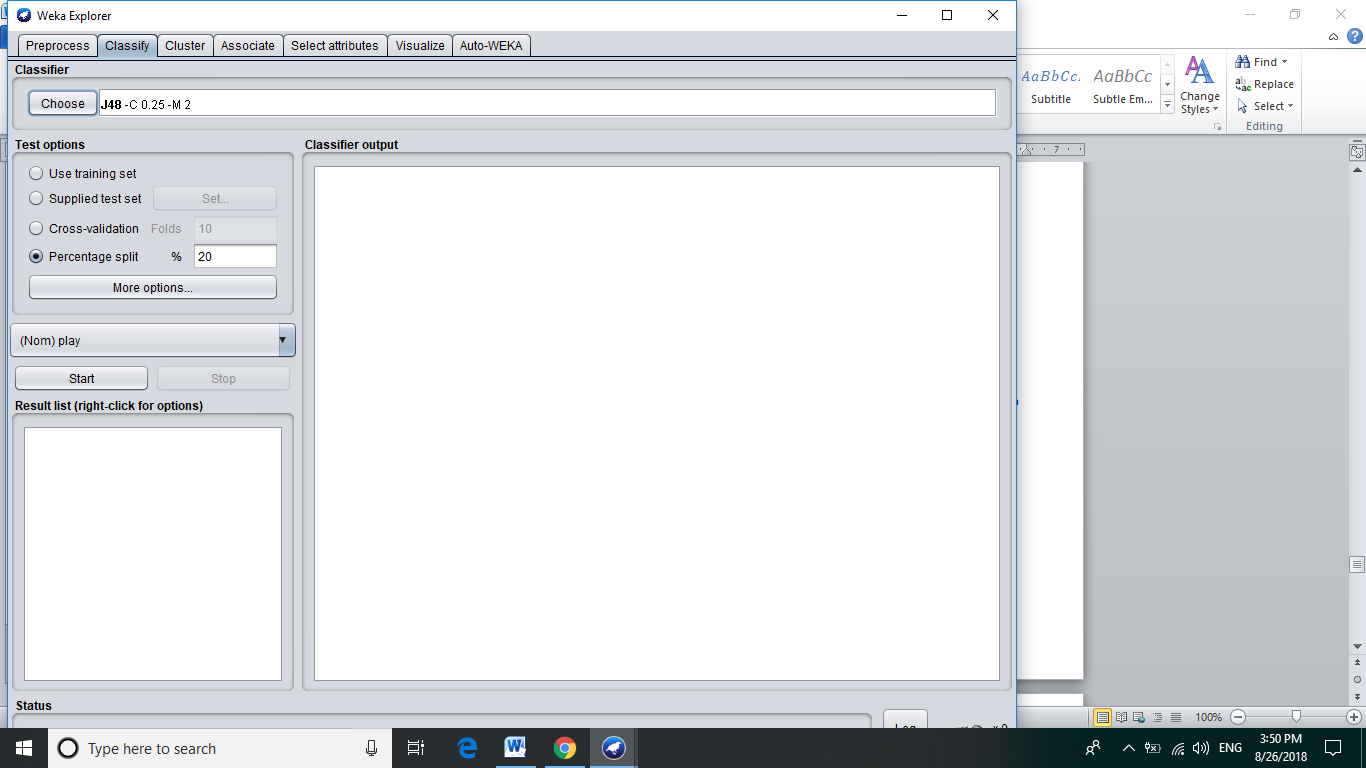
Đặt tham số InvertSelection=true, chọn OK

Chọn Apply và Save vào file supermarket\_test.arff

**\*Xây dựng cây quyết định từ tập training**

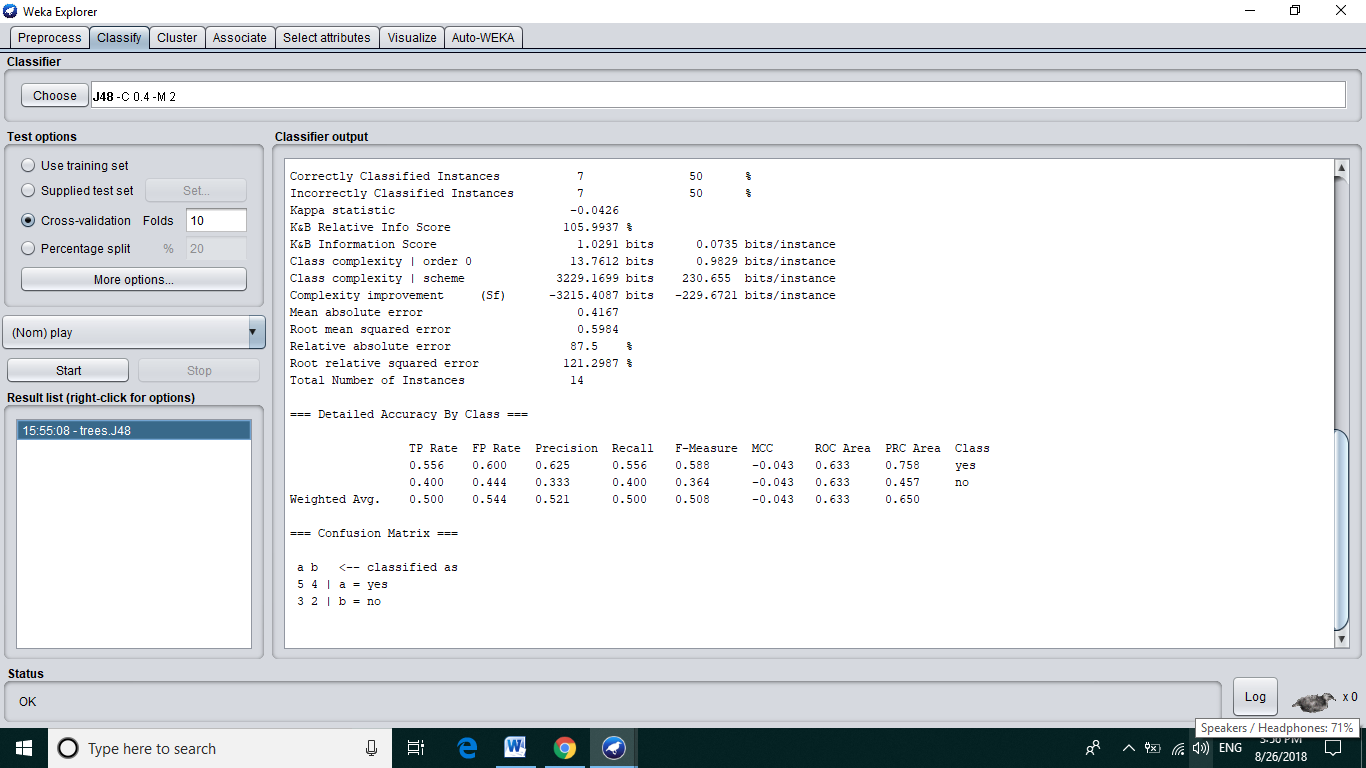
Mở file chứa dữ liệu (weather.nominal.arff)

Chọn tab **Classify**, chọn **tree**và chọn thuật toán **J48** (có thể chọn các thuật toán khác để so sánh hiệu quả phân lớp của các thuật toán khác nhau như ID3, SimpleCart,….)

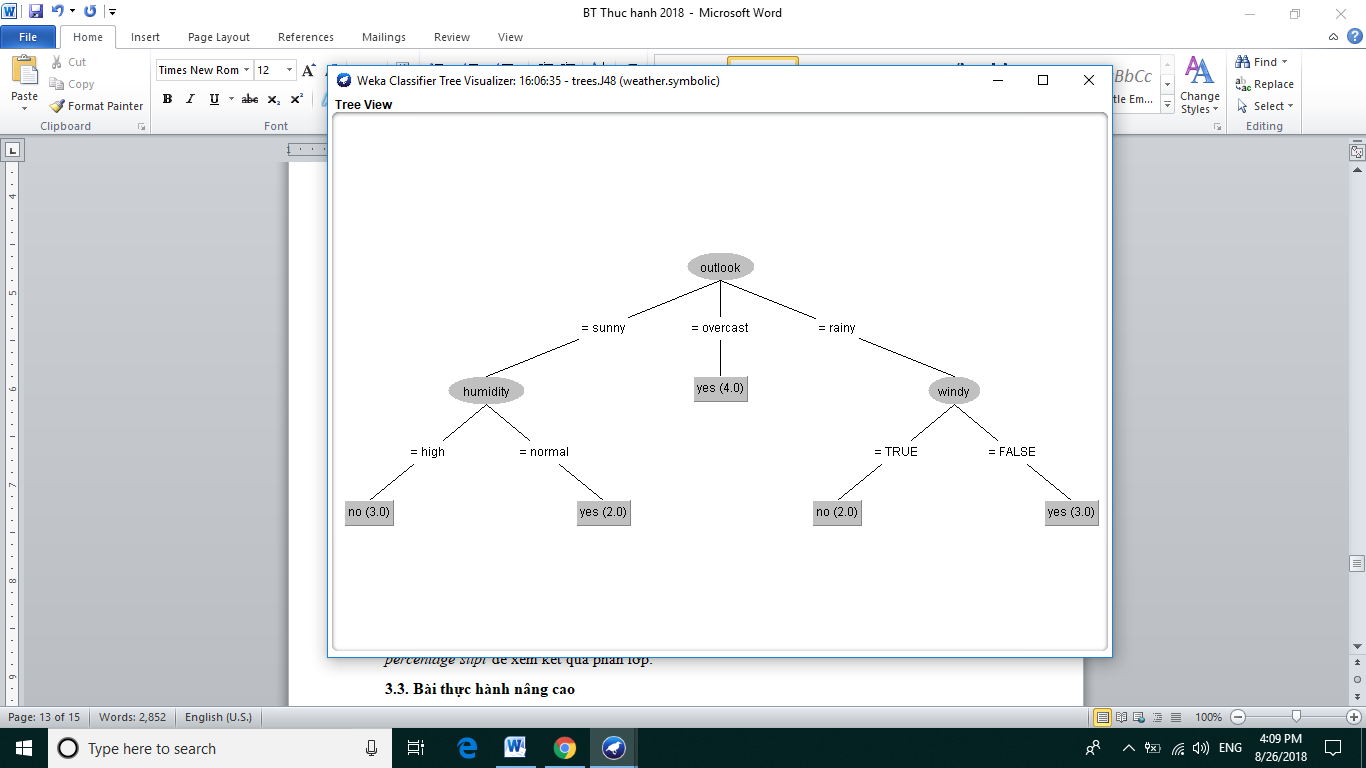


Thiết đặt các lựa chọn ở Test Option và đặt tham số cho bộ phân lớp (nếu cần)

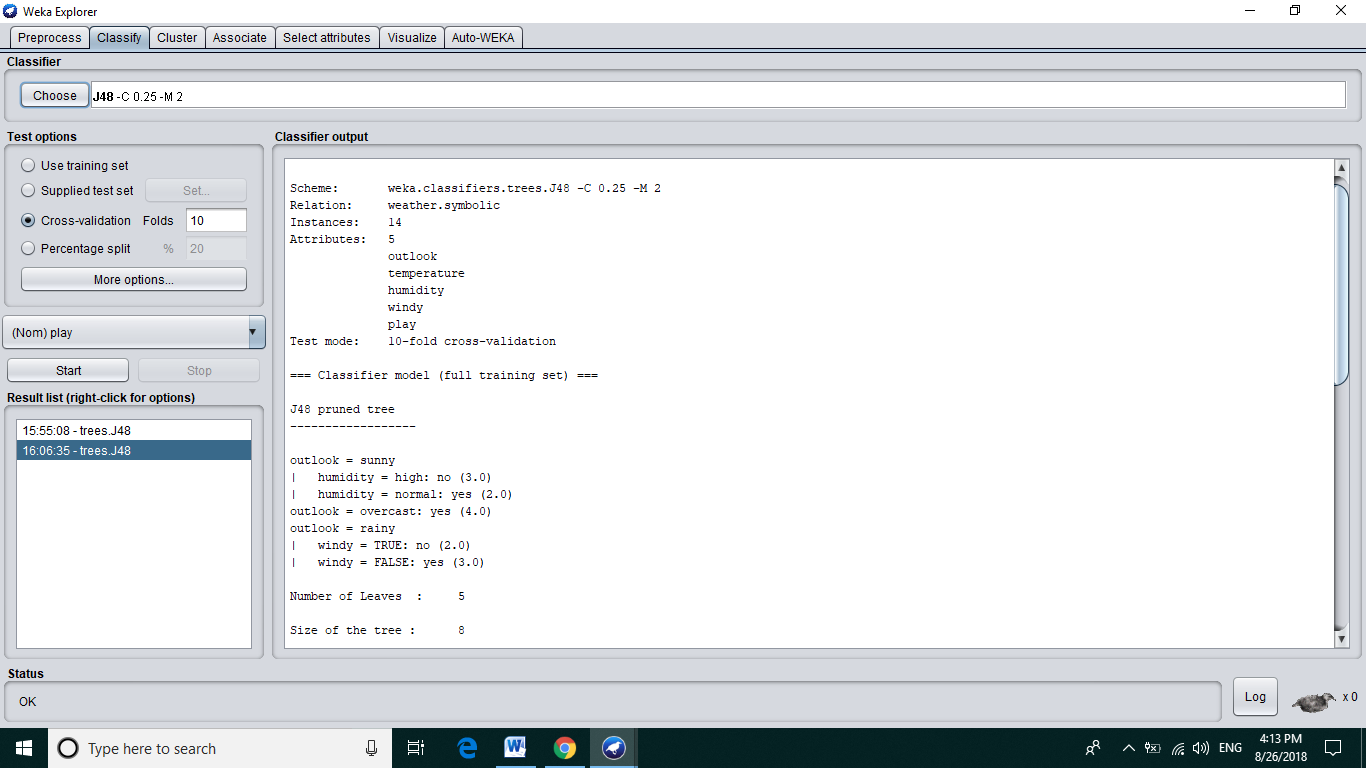
Bấm nút Start để xem kết quả



Xem cây quyết định kết quả: ở mục Result list, chuột phải vào tên thuật toán về cây quyết định vừa chạy và chọn Visualize tree.



Tập các luật được liệt kê



**3.2. Bài thực hành cơ bản**

a. Chia tập dữ liệu về thời tiết trong weather.nominal.arff thành 2 tập theo tỷ lệ 70% cho training và 30% cho testing.

b. Dùng bộ phân lớp **J48 với từng lựa chọn *use traning set, supplied test set, cross validation và percentage slipt* để xem kết quả phân lớp.**

**Làm tương tự với tập dữ liệu labor.arff.**

**3.3. Bài thực hành nâng cao**

a. Chia tập dữ liệu về thời tiết trong weather.nominal.arff thành 2 tập theo tỷ lệ 60% cho training và 40% cho testing.

b. Dùng các bộ phân lớp **J48, Randomtree, Random Forest, REPtree với lựa chọn *use traning set* để so sánh hiệu quả của các bộ phân lớp.**

**c.** Dùng các bộ phân lớp **J48, Randomtree, Random Forest, REPtree với lựa chọn *cross validation* với giá trị là 10để so sánh hiệu quả của các bộ phân lớp.**

**Làm tương tự với tập dữ liệu labor.arff.**

# BÀI THỰC HÀNH SỐ 5

**1. MỤC TIÊU:**

Bài thực hành số 5 nhằm giúp sinh viên biết: phân lớp với mô hình Naïve Bayes.

**2. YÊU CẦU:**

- Yêu cầu về điều kiện thực hành: Phòng máy thực hành có cài đặt phần mềm Weka, số lượng máy tương ứng với số sinh viên chia theo lớp thực hành.

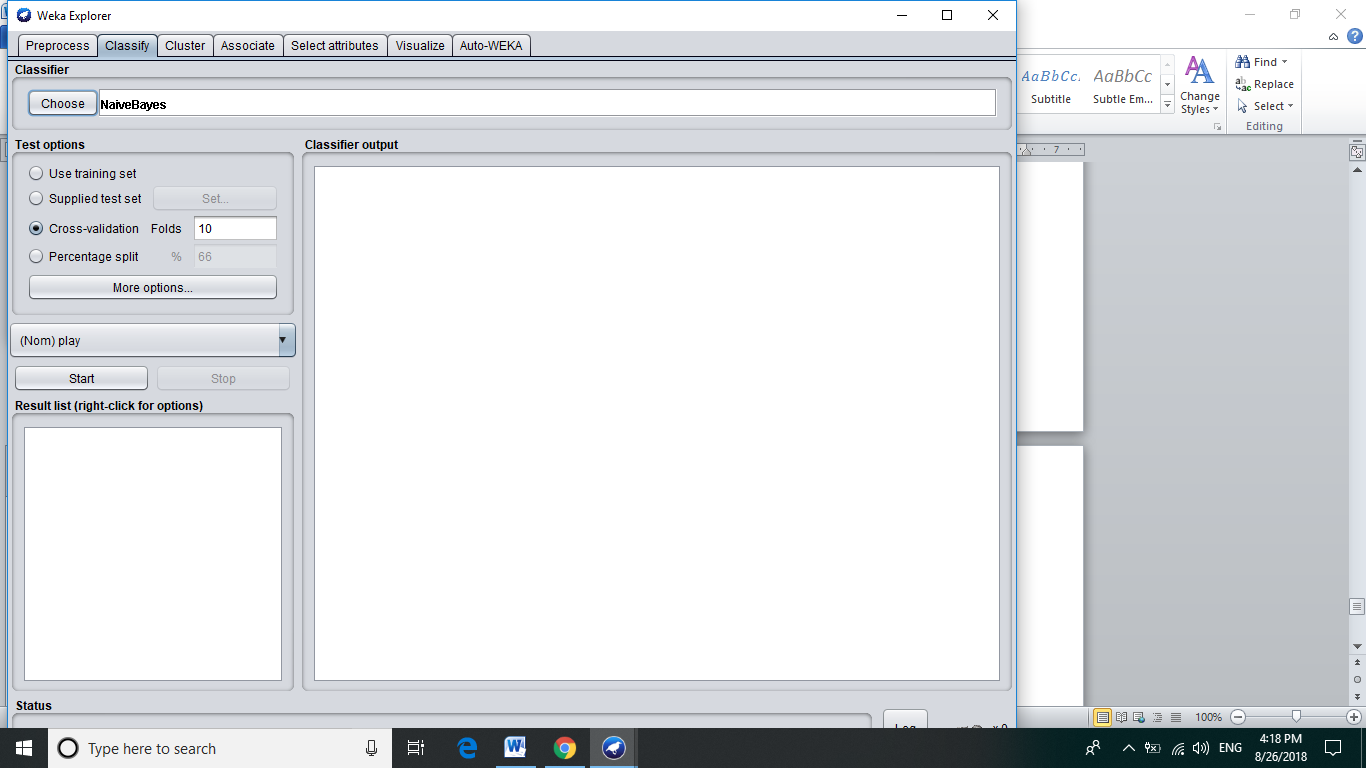
- Yêu cầu sinh viên: Chuẩn bị bài tập, tài liệu theo yêu cầu của đề cương môn học.

**3. NỘI DUNG BÀI THỰC HÀNH:**

**3.1. Bài thực hành mẫu**

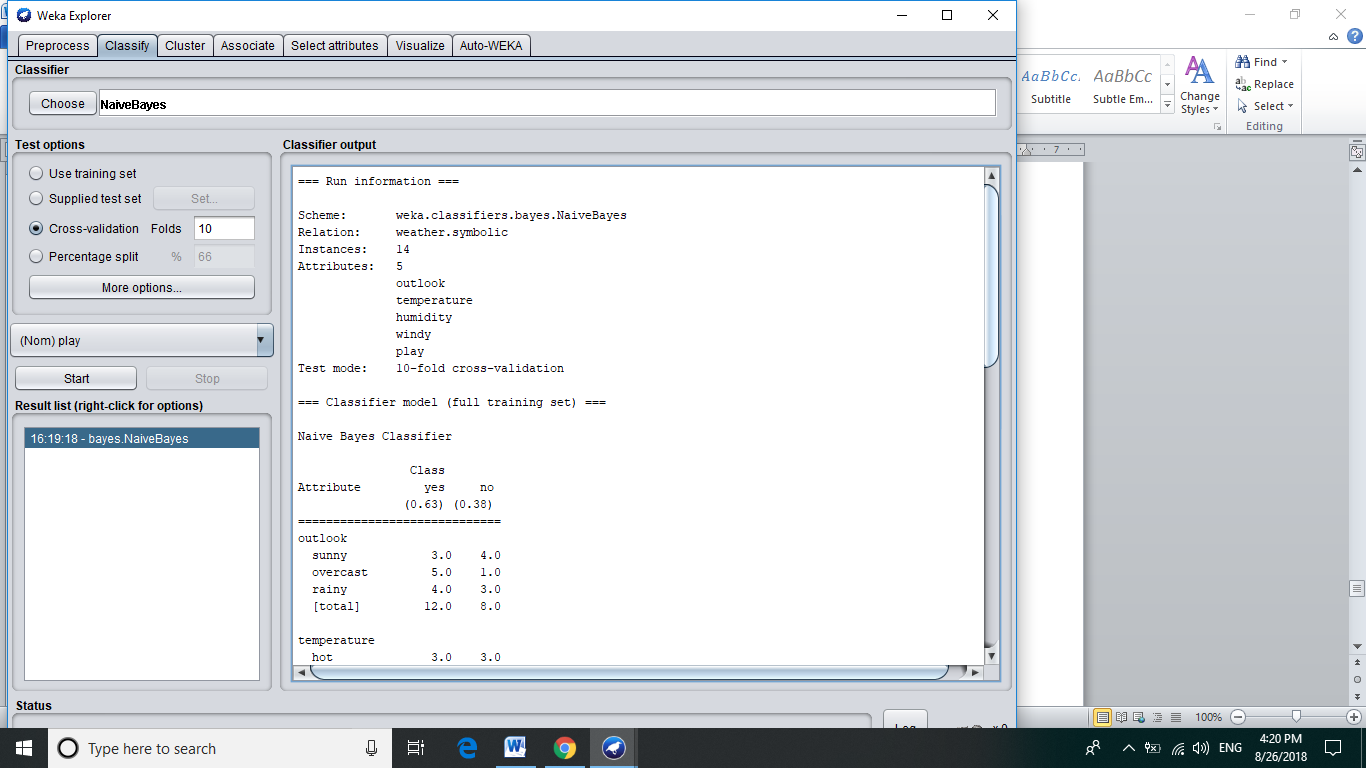
Mở file chứa dữ liệu (weather.nominal.arff)

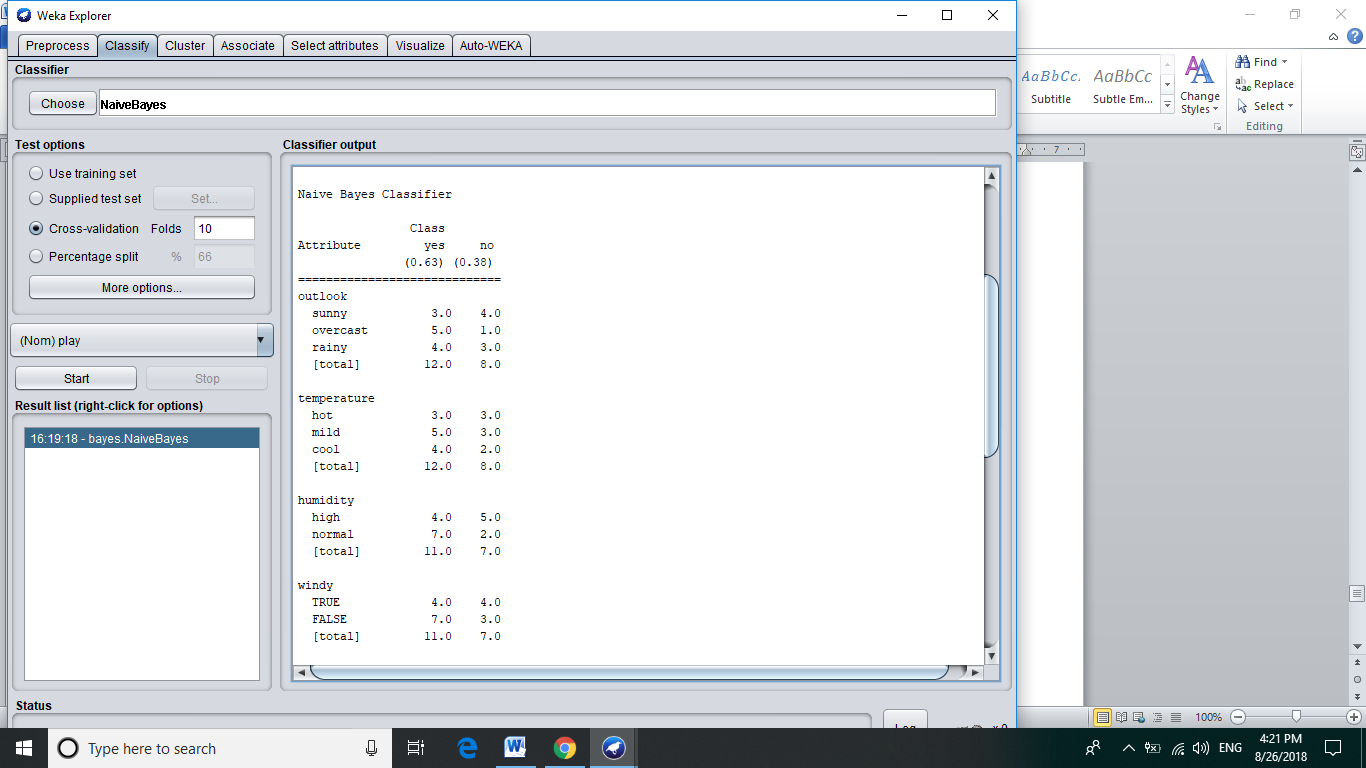
Chọn tab **Classify**, chọn **Bayes** và chọn **Naïve Bayes**.

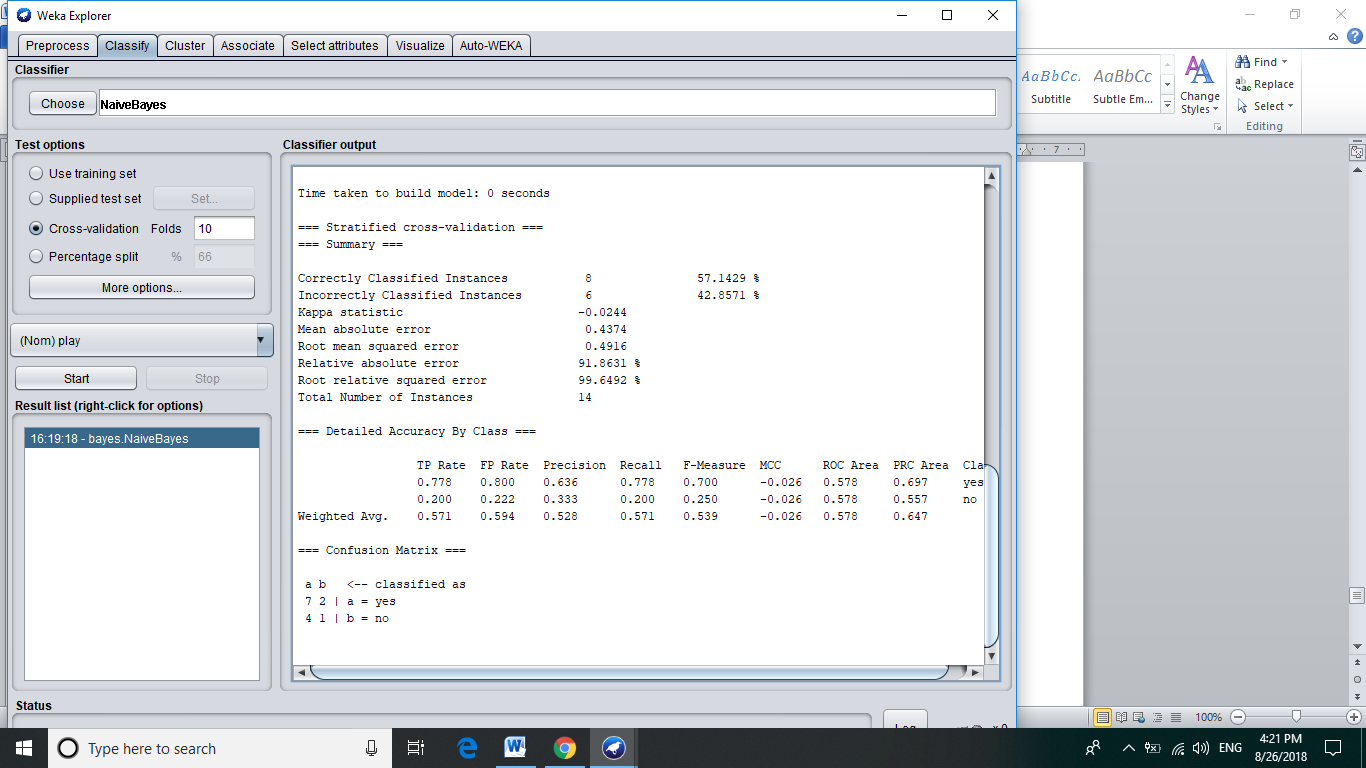


Thiết đặt các lựa chọn ở Test Option và đặt tham số cho bộ phân lớp (nếu cần)

Bấm nút Start để xem kết quả







**3.2. Bài thực hành cơ bản**

a. Chia tập dữ liệu về thời tiết trong weather.nominal.arff thành 2 tập theo tỷ lệ 50% cho training và 50% cho testing.

b. Dùng bộ phân lớp **Naïve Bayes với từng lựa chọn *use traning set, supplied test set, cross validation và percentage slipt* để xem kết quả phân lớp.**

**Làm tương tự với tập dữ liệu labor.arff.**

**3.3. Bài thực hành nâng cao**

a. Chia tập dữ liệu về thời tiết trong weather.nominal.arff thành 2 tập theo tỷ lệ 40% cho training và 60% cho testing.

b. Dùng các bộ phân lớp **Naïve Bayes, BayesNet với lựa chọn *use traning set* để so sánh hiệu quả của các bộ phân lớp.**

**c.** Dùng các bộ phân lớp **Naïve Bayes, BayesNet với lựa chọn *cross validation* với giá trị là 5để so sánh hiệu quả của các bộ phân lớp.**

**Làm tương tự với tập dữ liệu labor.arff.**

# BÀI THỰC HÀNH SỐ 6

**1. MỤC TIÊU:**

Bài thực hành số 6 nhằm giúp sinh viên biết: phân cụm với K-means

**2. YÊU CẦU:**

- Yêu cầu về điều kiện thực hành: Phòng máy thực hành có cài đặt phần mềm Weka, số lượng máy tương ứng với số sinh viên chia theo lớp thực hành.

- Yêu cầu sinh viên: Chuẩn bị bài tập, tài liệu theo yêu cầu của đề cương môn học.

**3. NỘI DUNG BÀI THỰC HÀNH:**

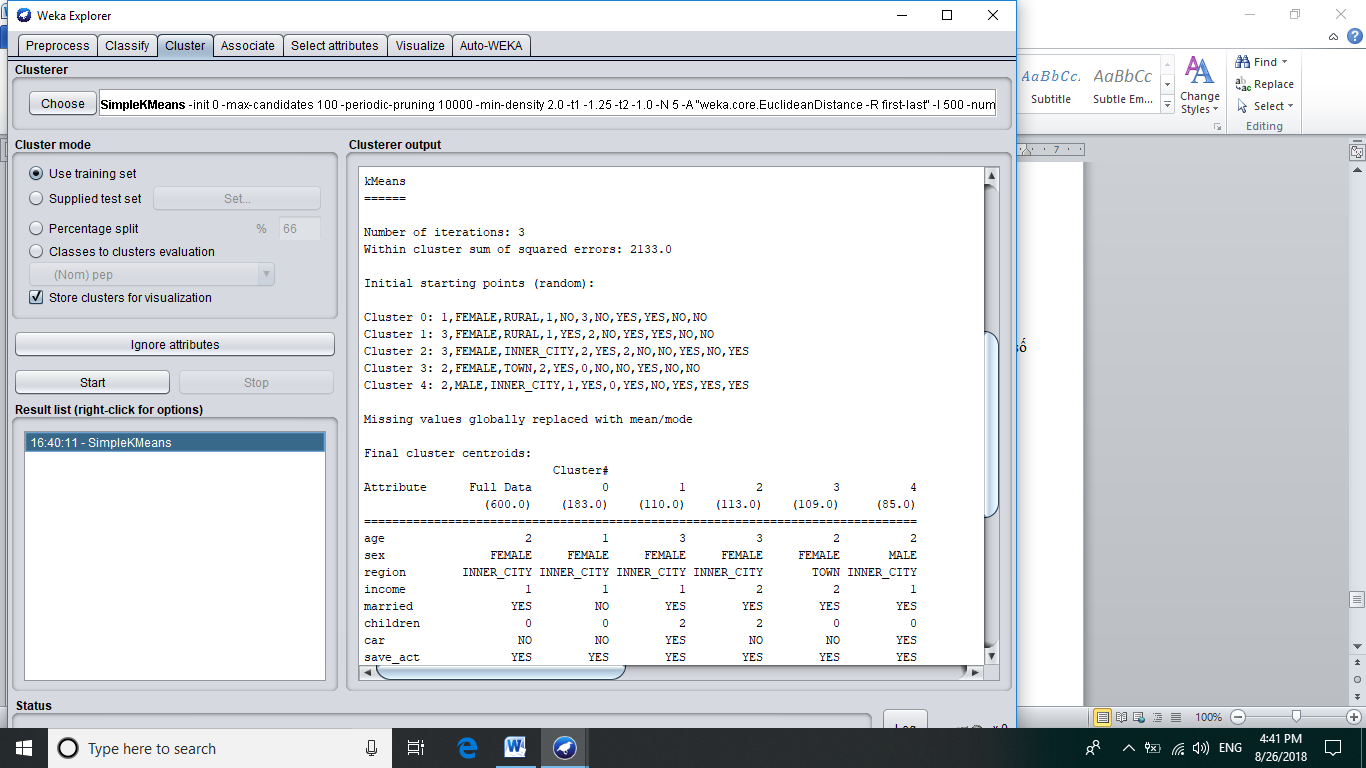
**3.1. Bài thực hành mẫu**

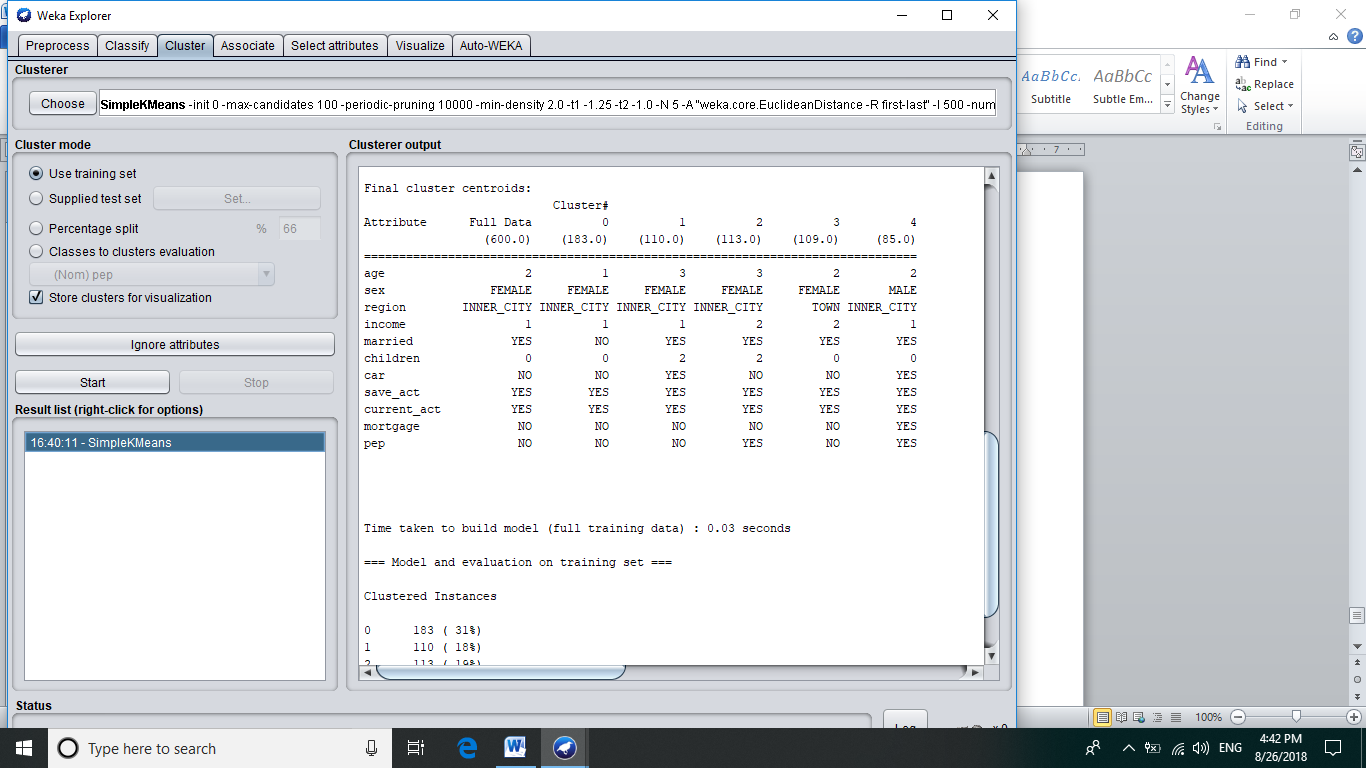
Mở file dữ liệu bank-data.arff

Chọn Tab Cluster, chọn Clusters>SimpleKmeans

Đặt tham số numCluster=5, chọn Ok.

Chọn Start để xem kết quả





**3.2. Bài thực hành cơ bản**

Mở file **diabetes.arff** ở link sau: http://www.mediafire.com/file/6achyh69qxbxcqs/diabetes.arff

Phân cụm K-means với số cụm lần lượt là 3, 4, 5.

**3.3. Bài thực hành nâng cao**

Mở file **credit.arff** ở đường link sau: http://www.mediafire.com/file/nf59zz8r6ipkrct/credit-g.arff

Phân cụm K-means với số cụm lần lượt là 2,5,7.

**Tham khảo:** http://cuongndh.blogspot.com/p/khai-pha-du-lieu.html