**ĐỀ BÀI TẬP CHO CÁC NHÓM**

1. **Tiền xử lý dữ liệu**

Trình bày đầy đủ các bước tiền xử lý dữ liệu bằng WEKA

* Sử dụng 1 tập dữ liệu của UCI có dữ liệu thiếu, dữ liệu liên tục, dữ liệu rời rạc.
* Khuyến khích: Tìm hiểu thêm các chức năng WEKA.

1. **Xây dựng cây quyết định**

* Làm bài tập 4.1, 4.2, 4.3 trong tài liệu [1]. (các bước tính toán)
* Trình bày ứng dụng Naïve Bayes trong WEKA ( [2], trang 139) cho một tập dữ liệu đã tiền xử lý. (tập dữ liệu khác Weather.norminal.arff)
* Trình bày các bước tạo training set và testing set trong WEKA ( [2] trang 150)

1. **Luật kết hợp**

* Trình bày thuật toán Naïve (Naïve Algorithm) tìm luật kết hợp ([2],240)
* Làm bài tập 12.1, 12.2 trong tài liệu [1]
* Sử dụng WEKA hoặc R tìm luật kết hợp của bài tập 12.1, 12.2 trong tài liệu [1]
* So sánh kết quả khai phá luật kết hợp và các luật phân lớp ([2], trang 360)
* Liệt kê tên một số thuật toán khai phá luật kết hợp. (nếu được chỉ ra ý tưởng chính của các thuật toán)

1. **Phân cụm**

* Thử nghiệm phân cụm dữ liệu bằng WEKA/R cho các thuật toán:

K-means, HCA (Hierarchical Clustering Algorithms).

* Trình bày thuật toán DBSCAN ([2], trang 203).

**Tài liệu tham khảo**

[1] Data Mining, Theories, Algorithms, and Examples – Nong Ye, CRC Press

[2] Data Ming and Data Warehousing, Parteek Bhatia

Câu 1:

**Tổng hợp dữ liệu**

* Chọn bộ dữ liệu Horse Colic trên UCI
* Dữ liệu ban đầu có dạng:

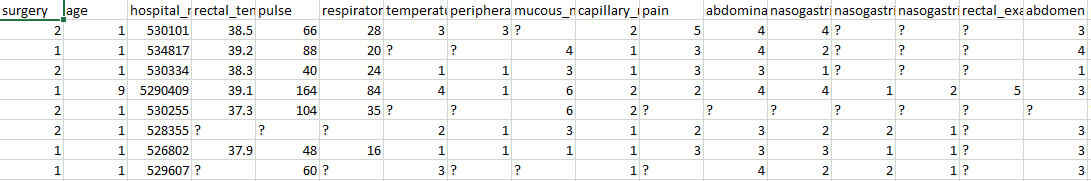


* Lấy tên thuộc tính là dòng đầu tiên lấy trong file horse-colic.names lưu vào dòng đầu tiên có dạng:



**Làm sạch và chuyển đổi dữ liệu**

* Tiến hành mở file vừa tao được bằng Excel các cột phân chia với nhau bởi dấu “ “ chọn save as file với đuôi arff.



* Tiến hành mở file vừa tạo bằng phần mềm WEKA, do là bài toán học không dám sát nên.

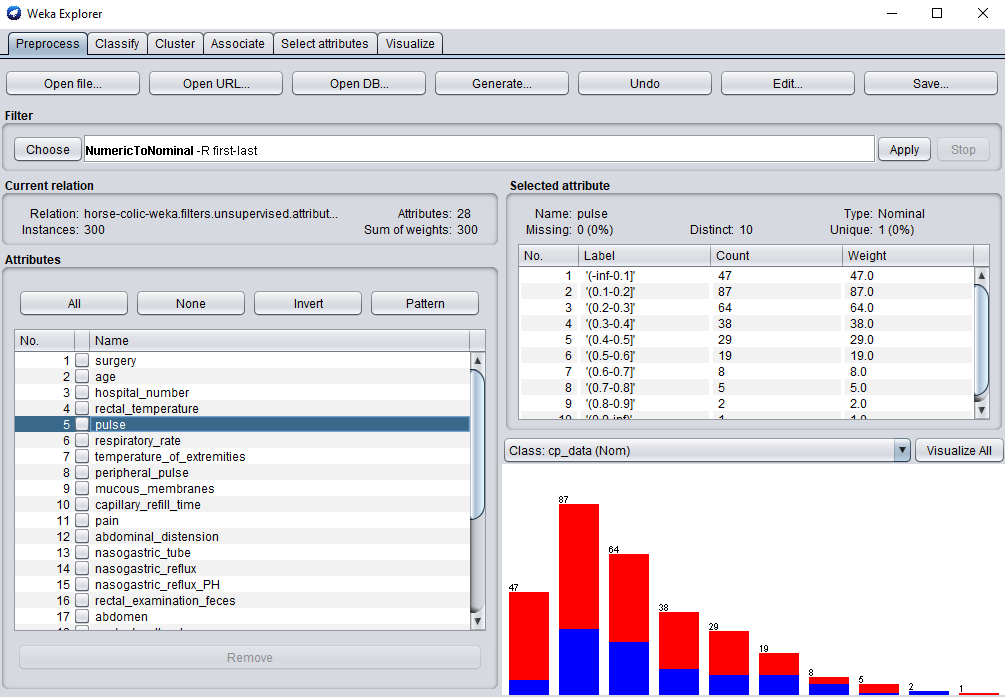
Làm sạch dữ liệu: nằm trong mục Unspervised/Filter

Chuẩn hóa các thuộc tính số về đoạn [0, 1] bằng bộ lọc Normalize  click Apply.

Dùng bộ lọc ReplaceMissingValue để thay thế các giá trị thiếu thành giá trị trung bình của hệ thống  click Apply.

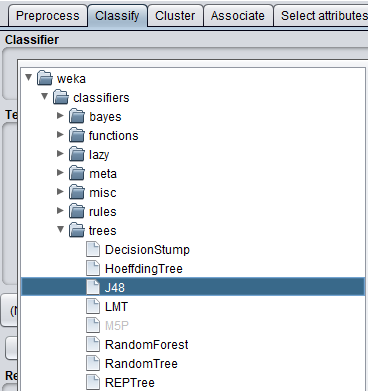
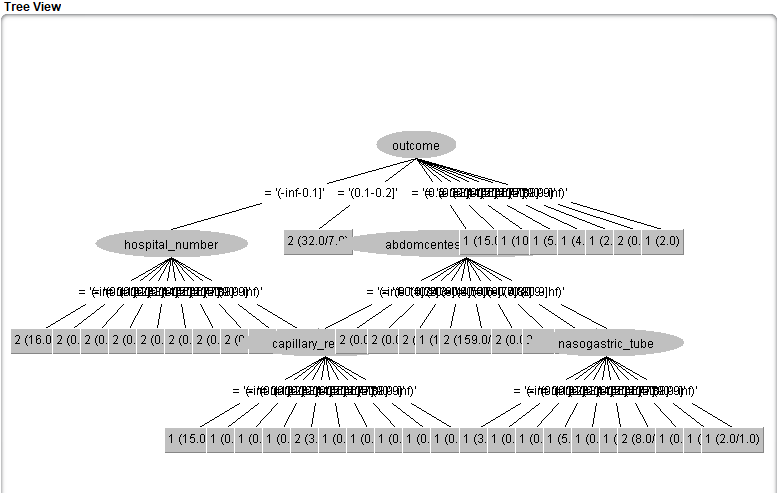
Dùng bộ lọc Discretize để rời rạc hóa thuộc tính numberic về Nominal  click Apply.

Dùng bộ lọc NumbericToNominal để chuyển các thuộc tính numberic thành nominal  click Apply.

Cuối cùng sẽ có kết quả như hình:

**Rút gọn dữ liệu:**

Trong tab classify chọn classifier – tree – J48. Sau đó ta thấy chỉ thuộc tính Hospital number, capillary refill time, adominal distention, adominosentensis appearance, adomcentesis total protein là xuất hiện trên cây do đó loại bỏ các thuộc tính thừa.

Bài 2:

**4.1 Tài liệu [1]**

|S| = 16, m = 2 : C1 = True, C2 = False, |C1| = 7, |C2| = 7

Entropy hệ thống I(S) =

|  |  |  |  |
| --- | --- | --- | --- |
| Color | Yellow | Purple |  |
| T | 5 | 2 | 7 |
| F | 3 | 6 | 9 |
|  | 8 | 8 | 16 |

Entropy(SYellow)

Entropy(Purple)

|  |  |  |  |
| --- | --- | --- | --- |
| Size | Small | Large |  |
| T | 5 | 2 | 7 |
| F | 3 | 6 | 9 |
|  | 8 | 8 | 16 |

Entropy(SSmall)

Entropy(SLarge)

|  |  |  |  |
| --- | --- | --- | --- |
| Act | Stretch | Dip |  |
| T | 5 | 2 | 7 |
| F | 3 | 6 | 9 |
|  | 8 | 8 | 16 |

Entropy(SStretch)

Entropy(SDip)

|  |  |  |  |
| --- | --- | --- | --- |
| Age | Adult | Child |  |
| T | 5 | 2 | 7 |
| F | 3 | 6 | 9 |
|  | 8 | 8 | 16 |

Entropy(SAdult)

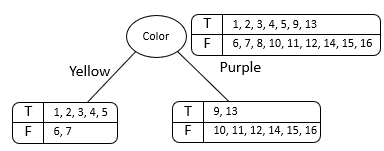
Entropy(SChild)

Gain(Color, S) = I(S) – I(SColor) = 0.9887 – I(SYellow) – I(SPurple) = 0.10585

Gain(Size, S) = I(S) – I(SSize) = 0.9887 – I(SSmall) – I(SLarge) = 0.10585

Gain(Act, S) = I(S) – I(SAct) = 0.9887 – I(SStretch) – I(SDip) = 0.10585

Gain(Age, S) = I(S) – I(SAge) = 0.9887 – I(SAdult) – I(SChild) = 0.10585



, 8

**Xét sự phân lớp với nhánh Yellow ứng với Size, Act, Age**

Entropy hệ thống I(S1) =

|  |  |  |  |
| --- | --- | --- | --- |
| Size | Small | Large |  |
| T | 4 | 1 | 5 |
| F | 0 | 3 | 3 |
|  | 4 | 4 | 8 |

Entropy(SSmall)

Entropy(SLarge )

|  |  |  |  |
| --- | --- | --- | --- |
| Act | Stretch | Dip |  |
| T | 3 | 2 | 5 |
| F | 1 | 2 | 2 |
|  | 4 | 4 | 8 |

Entropy(SStretch)

Entropy(SDip)

|  |  |  |  |
| --- | --- | --- | --- |
| Age | Adult | Child |  |
| T | 3 | 2 | 5 |
| F | 1 | 2 | 3 |
|  | 4 | 4 | 8 |

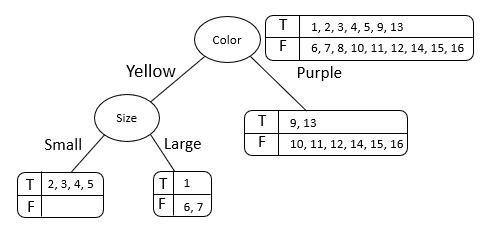
Entropy(SAdult)

Entropy(SChild)

Gain(Size, S) = I(S1) – I(SSize) = 0.86132 – I(SLarge) – I(SSmall) = 0.455684

Gain(Act, S) = I(S1) – I(SAct) = 0. 86132 – I(SStretch) – I(SDip) = -0.044315

Gain(Age, S) = I(S1) – I(SAge) = 0. 86132 – I(SAdult) – I(SChild) = -0. 044315



, 8

5

1,

**Xét sự phân lớp với nhánh Large ứng với Act, Age**

Entropy hệ thống I(S2) =

|  |  |  |  |
| --- | --- | --- | --- |
| Act | Stretch | Dip |  |
| T | 1 | 0 | 1 |
| F | 1 | 1 | 2 |
|  | 2 | 1 | 3 |

Entropy(SStretch)

Entropy(SDip)

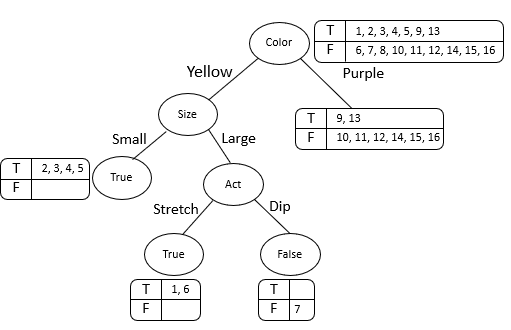
|  |  |  |  |
| --- | --- | --- | --- |
| Age | Adult | Child |  |
| T | 1 | 0 | 1 |
| F | 1 | 1 | 2 |
|  | 2 | 1 | 3 |

Entropy(SAdult)

Entropy(SChild)

Gain(Act, S) = I(S2) – I(SAct) = 0.91829 – I(SStretch) – I(SDip) = 0.25162

Gain(Age, S) = I(S2) – I(SAge) = 0.91829 – I(SAdult) – I(SChild) = 0.25162



1,

Child

Adult

False

True

Age

5

6

, 8

**Xét sự phân lớp với nhánh Purple ứng với Size, Act, Age**

Entropy hệ thống I(S3) =

|  |  |  |  |
| --- | --- | --- | --- |
| Size | Small | Large |  |
| T | 1 | 1 | 2 |
| F | 3 | 3 | 6 |
|  | 4 | 4 | 8 |

Entropy(SSmall)

Entropy(SLarge )

|  |  |  |  |
| --- | --- | --- | --- |
| Act | Stretch | Dip |  |
| T | 2 | 0 | 2 |
| F | 2 | 4 | 6 |
|  | 4 | 4 | 8 |

Entropy(SStretch)

Entropy(SDip)

|  |  |  |  |
| --- | --- | --- | --- |
| Age | Adult | Child |  |
| T | 2 | 0 | 2 |
| F | 2 | 4 | 6 |
|  | 4 | 4 | 8 |

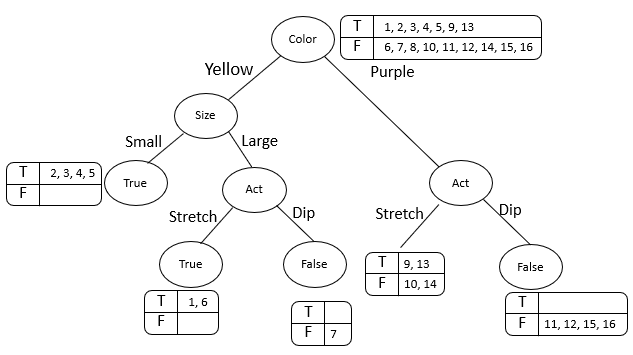
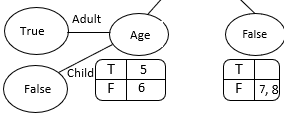
Entropy(SAdult)

Entropy(SChild)

Gain(Size, S) = I(S3) – I(SSize) = 0.81127 – I(SLarge) – I(SSmall) = 0

Gain(Act, S) = I(S3) – I(SAct) = 0.81127 – I(SStretch) – I(SDip) = 0.31127

Gain(Age, S) = I(S3) – I(SAge) = 0.81127 – I(SAdult) – I(SChild) = 0.31127



**Xét sự phân lớp với nhánh Stretch ứng với Size, Age**

Entropy hệ thống I(S4) =

|  |  |  |  |
| --- | --- | --- | --- |
| Size | Small | Large |  |
| T | 1 | 1 | 2 |
| F | 1 | 1 | 2 |
|  | 2 | 2 | 4 |

Entropy(SSmall)

Entropy(SLarge )

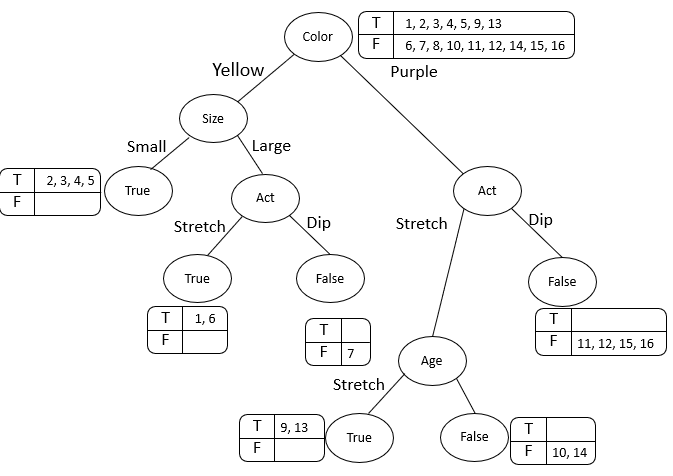
|  |  |  |  |
| --- | --- | --- | --- |
| Age | Adult | Child |  |
| T | 2 | 0 | 2 |
| F | 0 | 2 | 2 |
|  | 2 | 2 | 4 |

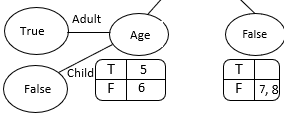
Entropy(SAdult)

Entropy(SChild)

Gain(Size, S) = I(S4 – I(SSize) = 1– I(SLarge) – I(SSmall) = 0

Gain(Age, S) = I(S4) – I(SAge) = 1– I(SAdult) – I(SChild) = 1





**4.2**

|S| = 24, m = 3   
C1 = Noncontact, C2 = Soft contact, C3 = Hard contact  
|C1| = 15, |C2| = 5, |C3| = 4

Entropy hệ thống I(S) =

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Age | Young | Pre-presbyopic | Presbyopic |  |
| Noncontact | 4 | 5 | 6 | 15 |
| Softcontact | 2 | 2 | 1 | 5 |
| Hardcontact | 2 | 1 | 1 | 4 |
|  | 8 | 8 | 8 | 24 |

Entropy(SYoung)

Entropy(Pre-presbyopic)

Entropy(Presbyopic)

|  |  |  |  |
| --- | --- | --- | --- |
| SP | Myope | Hypermetrope |  |
| Noncontact | 7 | 8 | 15 |
| Softcontact | 2 | 3 | 5 |
| Hardcontact | 3 | 1 | 4 |
|  | 12 | 12 | 24 |

Entropy(SMyope)

Entropy(SHypermetrope)

|  |  |  |  |
| --- | --- | --- | --- |
| Astigmatic | No | Yes |  |
| Noncontact | 7 | 8 | 15 |
| Softcontact | 5 | 0 | 5 |
| Hardcontact | 0 | 4 | 4 |
|  | 12 | 12 | 24 |

Entropy(SNo)

0.98

Entropy(SYes)

|  |  |  |  |
| --- | --- | --- | --- |
| TPR | Reduced | Normal |  |
| Noncontact | 12 | 3 | 15 |
| Softcontact | 0 | 5 | 5 |
| Hardcontact | 0 | 4 | 4 |
|  | 12 | 12 | 24 |

Entropy(SReduced)

Entropy(SNormal)

Gain(Age, S) = I(S) – I(SAge) = 1.326– I(SYoung) – I(SPre-presbyopic) – I(SPresbyopic) = 0.0143

Gain(SP, S) = I(S) – I(SSP) = 1.326 – I(SMyope) – I(SHypermetrope) = 0.041

Gain(Astigmatic, S) = I(S) – I(SAstigmatic) = 1.326 – I(SNo) – I(SYes) = 0.376

Gain(TPR, S) = I(S) – I(STPR) = 1.326 – I(SReduced) – I(SNormal) = 0.551

|  |  |
| --- | --- |
| Noncontact | 1,3,5,7,9,11,13,15,16,17,18,19,21,23,24 |
| Softcontact | 2,6,10,14,22 |
| Hardcontact | 4,8,12,20 |

Reduced

Normal

|  |  |
| --- | --- |
| Noncontact | 1,3,5,7,9,11,13,15,17,19,21,23 |
| Softcontact |  |
| Hardcontact |  |

|  |  |
| --- | --- |
| Noncontact | 16,18,24 |
| Softcontact | 2,6,10,14,22 |
| Hardcontact | 4,8,12,20 |

**Xét sự phân lớp với nhánh Normal ứng với Age, Spectacle Prescription, Astigmatic.**

|S| = 12, m = 3   
C1 = Noncontact, C2 = Soft contact, C3 = Hard contact  
|C1| = 3, |C2| = 5, |C3| = 4

Entropy hệ thống I(S) =

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Age | Young | Pre-presbyopic | Presbyopic |  |
| Noncontact | 0 | 1 | 2 | 3 |
| Softcontact | 2 | 2 | 1 | 5 |
| Hardcontact | 2 | 1 | 1 | 4 |
|  | 4 | 4 | 4 | 12 |

Entropy(SYoung)

Entropy(Pre-presbyopic)

Entropy(Presbyopic)

|  |  |  |  |
| --- | --- | --- | --- |
| SP | Myope | Hypermetrope |  |
| Noncontact | 1 | 2 | 3 |
| Softcontact | 2 | 3 | 5 |
| Hardcontact | 3 | 1 | 4 |
|  | 6 | 6 | 12 |

Entropy(SMyope)

1.46

Entropy(SHypermetrope)

|  |  |  |  |
| --- | --- | --- | --- |
| Astigmatic | No | Yes |  |
| Noncontact | 1 | 2 | 3 |
| Softcontact | 5 | 0 | 5 |
| Hardcontact | 0 | 4 | 4 |
|  | 6 | 6 | 12 |

Entropy(SNo)

0.65

Entropy(SYes)

Gain(Age, S) = I(S) – I(SAge) = 1.55– I(SYoung) – I(SPre-presbyopic) – I(SPresbyopic) = 0.2383

Gain(SP, S) = I(S) – I(SSP) = 1.55 – I(SMyope) – I(SHypermetrope) = 0.09

Gain(Astigmatic, S) = I(S) – I(SAstigmatic) = 1.55 – I(SNo) – I(SYes) = 0.765

No

Yes

|  |  |
| --- | --- |
| Noncontact | 18 |
| Softcontact | 2,6,10,14,22 |
| Hardcontact |  |

|  |  |
| --- | --- |
| Noncontact | 16,24 |
| Softcontact |  |
| Hardcontact | 4,8,12,20 |

**Xét sự phân lớp của nhánh No ứng với các thuộc tính Age, Spectacle Prescription**

|S| = 6, m = 3   
C1 = Noncontact, C2 = Soft contact, C3 = Hard contact  
|C1| = 1, |C2| = 5, |C3| = 0

Entropy hệ thống I(S) =

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Age | Young | Pre-presbyopic | Presbyopic |  |
| Noncontact | 0 | 0 | 1 | 1 |
| Softcontact | 2 | 2 | 1 | 5 |
| Hardcontact | 0 | 0 | 0 | 0 |
|  | 2 | 2 | 2 | 6 |

Entropy(SYoung)

Entropy(Pre-presbyopic)

Entropy(Presbyopic)

|  |  |  |  |
| --- | --- | --- | --- |
| SP | Myope | Hypermetrope |  |
| Noncontact | 1 | 0 | 1 |
| Softcontact | 2 | 3 | 5 |
| Hardcontact | 0 | 0 | 0 |
|  | 3 | 3 | 6 |

Entropy(SMyope)

0.92

Entropy(SHypermetrope)

Gain(Age, S) = I(S) – I(SAge) = 0.65– I(SYoung) – I(SPre-presbyopic) – I(SPresbyopic) = 0.316

Gain(SP, S) = I(S) – I(SSP) = 0.65 – I(SMyope) – I(SHypermetrope) = 0.19

**Xét sự phân lớp của nhánh Yes ứng với các thuộc tính Age, Spectacle Prescription**

|S| = 6, m = 3   
C1 = Noncontact, C2 = Soft contact, C3 = Hard contact  
|C1| = 2, |C2| = 0, |C3| = 4

Entropy hệ thống I(S) =

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Age | Young | Pre-presbyopic | Presbyopic |  |
| Noncontact | 0 | 1 | 1 | 2 |
| Softcontact | 0 | 0 | 0 | 0 |
| Hardcontact | 2 | 1 | 1 | 4 |
|  | 2 | 2 | 2 | 6 |

Entropy(SYoung)

Entropy(Pre-presbyopic)

Entropy(Presbyopic)

|  |  |  |  |
| --- | --- | --- | --- |
| SP | Myope | Hypermetrope |  |
| Noncontact | 0 | 2 | 2 |
| Softcontact | 0 | 0 | 0 |
| Hardcontact | 3 | 1 | 4 |
|  | 3 | 3 | 6 |

Entropy(SMyope)

0

Entropy(SHypermetrope)

.92

Gain(Age, S) = I(S) – I(SAge) = 0.92 – I(SYoung) – I(SPre-presbyopic) – I(SPresbyopic) = 0.253

Gain(SP, S) = I(S) – I(SSP) = 0.92 – I(SMyope) – I(SHypermetrope) = 0.46

|  |  |
| --- | --- |
| Noncontact | 16,24 |
| Softcontact |  |
| Hardcontact | 8 |

No

Yes

Young

Pre-presbyopic

Presbyopic

Myope

Hypermetrope

Presbyopic

Pre-presbyopic

Young

|  |  |
| --- | --- |
| Noncontact | 18 |
| Softcontact | 2,6,10,14,22 |
| Hardcontact |  |

|  |  |
| --- | --- |
| Noncontact | 16,24 |
| Softcontact |  |
| Hardcontact | 4,8,12,20 |

|  |  |
| --- | --- |
| Noncontact |  |
| Softcontact | 2,6 |
| Hardcontact |  |

|  |  |
| --- | --- |
| Noncontact |  |
| Softcontact |  |
| Hardcontact | 4,12,20 |

|  |  |
| --- | --- |
| Noncontact |  |
| Softcontact |  |
| Hardcontact | 8 |

|  |  |
| --- | --- |
| Noncontact | 16 |
| Softcontact |  |
| Hardcontact |  |

|  |  |
| --- | --- |
| Noncontact | 24 |
| Softcontact |  |
| Hardcontact |  |

|  |  |
| --- | --- |
| Noncontact |  |
| Softcontact | 10,14 |
| Hardcontact |  |

|  |  |
| --- | --- |
| Noncontact | 18 |
| Softcontact | 22 |
| Hardcontact |  |

|  |  |
| --- | --- |
| Noncontact | 18 |
| Softcontact |  |
| Hardcontact |  |

|  |  |
| --- | --- |
| Noncontact |  |
| Softcontact | 22 |
| Hardcontact |  |

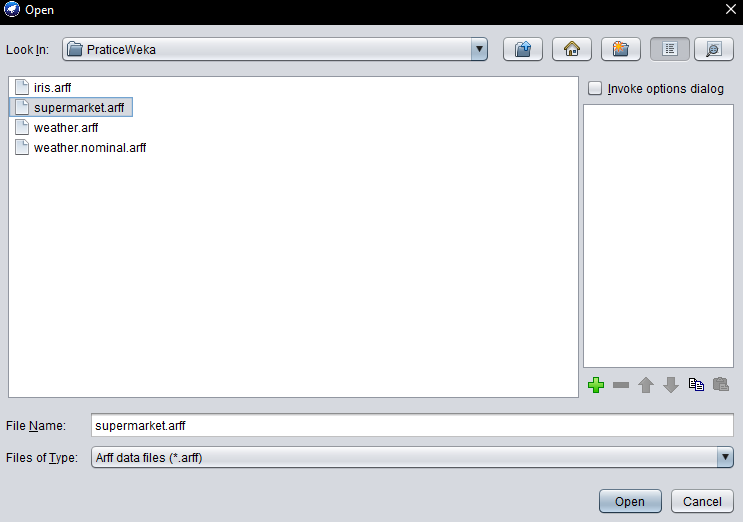
**Ứng dụng Naïve Bayes trong WEKA**

Để áp dụng Naïve Bayes trên bất kỳ tập dữ liệu nhất định nào, cần thực hiện các bước sau:

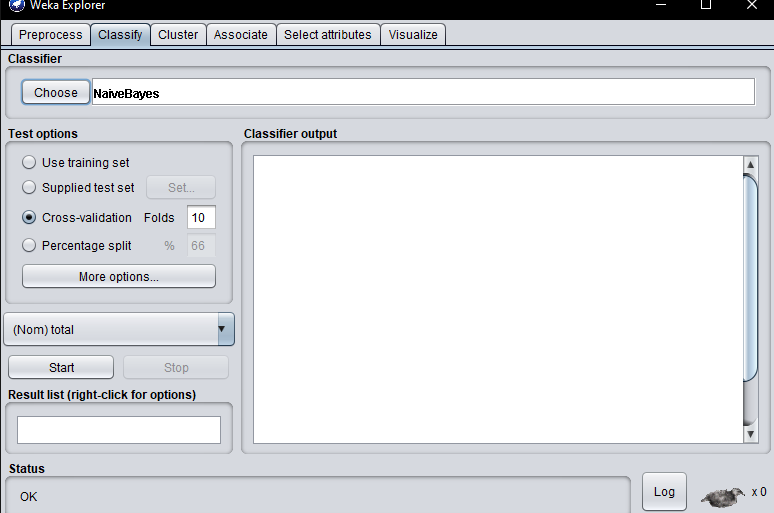
1. Khởi động Weka Explorer và chuyển đến tab Preprocess.



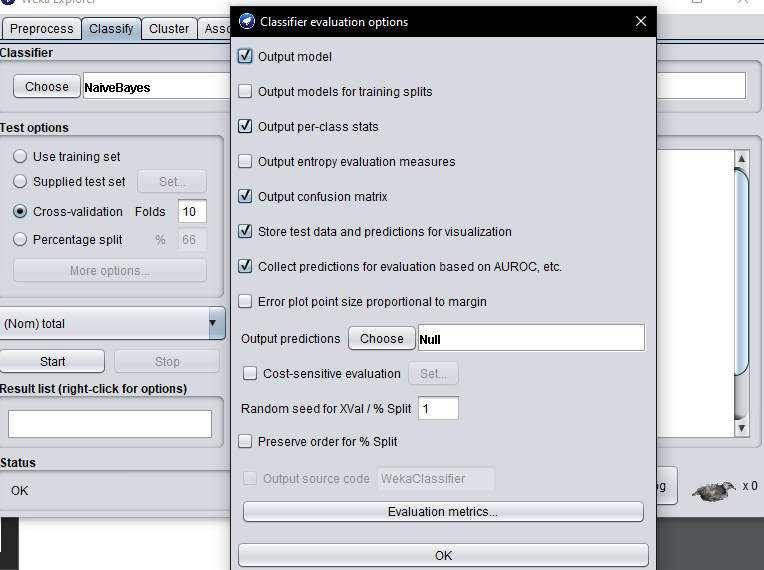
1. Bây giờ chọn tab Open file và chọn supermarket.arff



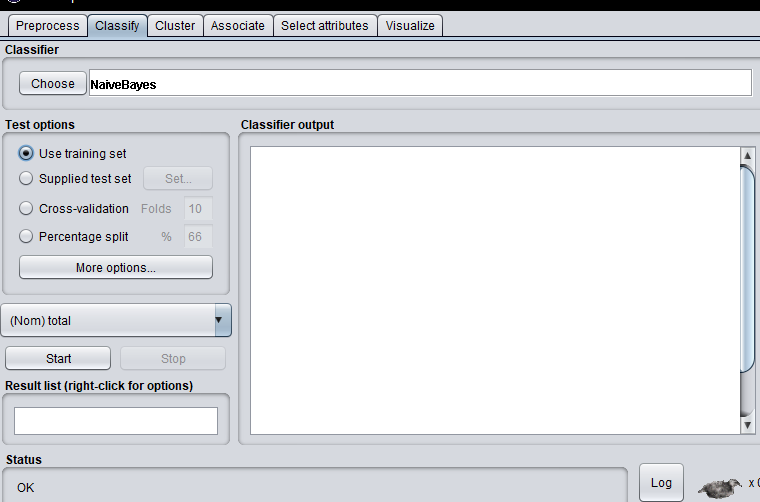
1. Bây giờ chọn tab Classify và chọn Naivebayes



1. Nhấp vào nút More Options và thiết lập các tùy chọn trong classifier evaluation options

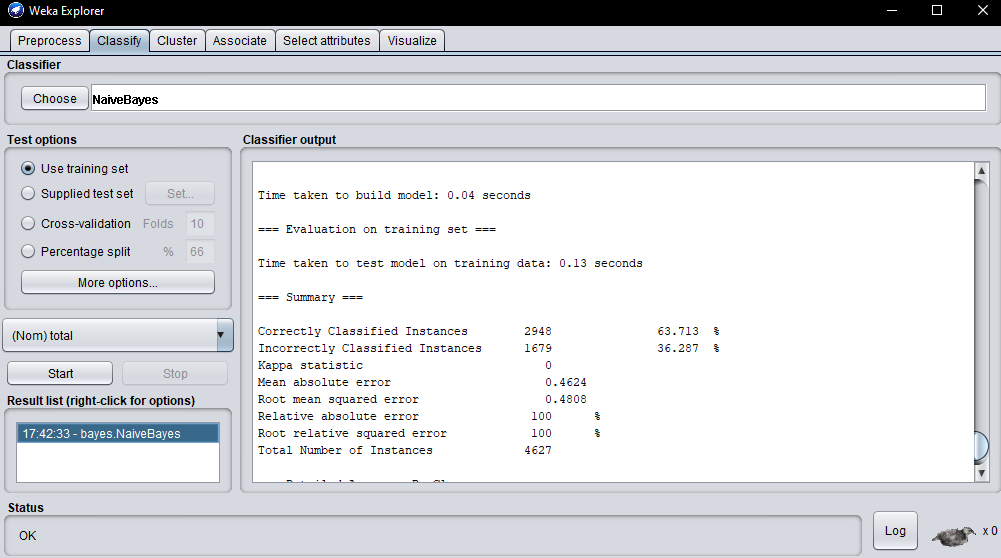


1. Chọn tùy chọn Use Training Set

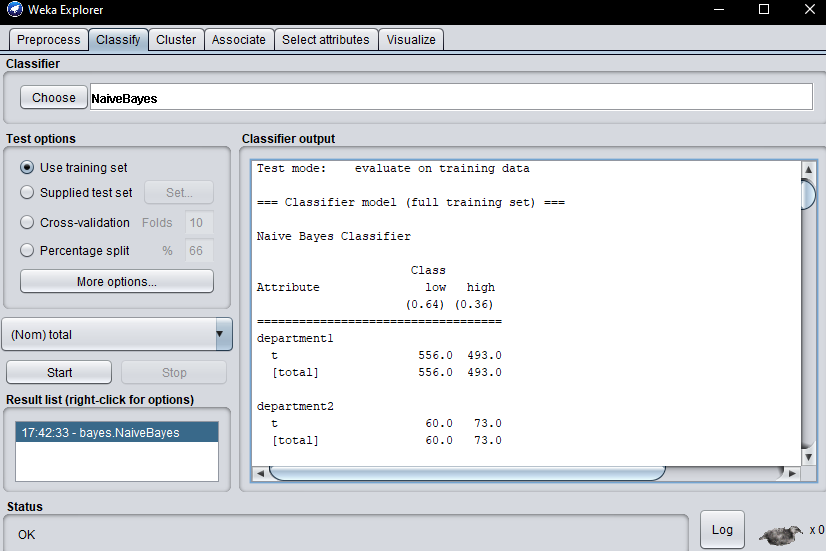


1. Bấm nút Start

Sau khi xử lý, kết quả sẽ xuất hiện như trong Hình dưới, trong đó nó cho thấy các trường hợp được phân loại chính xác là 2948 và trường hợp không chính xác là 1679 và kết quả tổng thể là 63.713 %.



Kết quả này cũng cho thấy bộ phân loại Naïve Bayes của Weka đã tự động áp dụng bộ ước lượng Laplace như được hiển thị trong số lượng đã sửa đổi.



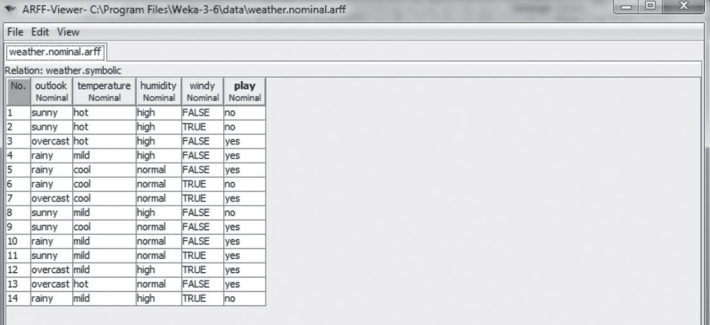
**Các bước tạo testing set trong WEKA**

Để kiểm tra trình phân loại trên các phiên bản không xác định, chúng ta phải tạo dữ liệu kiểm tra. Trong phần này, một quy trình đã được minh họa để tạo một tập dữ liệu thử nghiệm từ một tập dữ liệu hiện có. Để tạo các phiên bản thử nghiệm, hãy làm theo các bước sau.

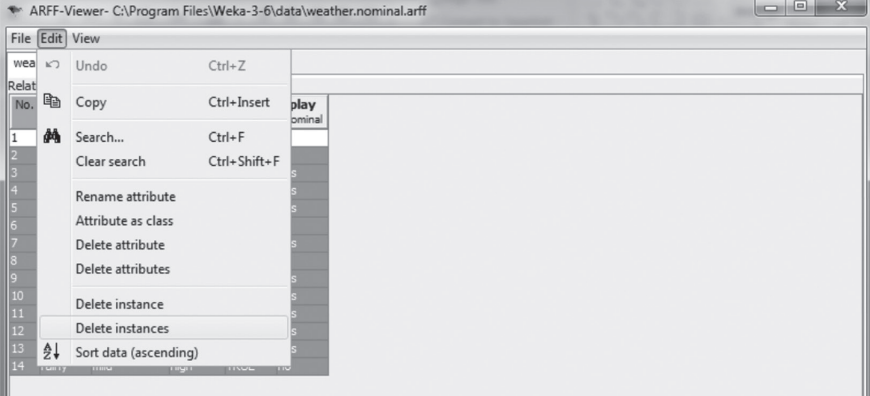
1. Nhấp vào tùy chọn Tool trong tab menu và chọn ArffViewer



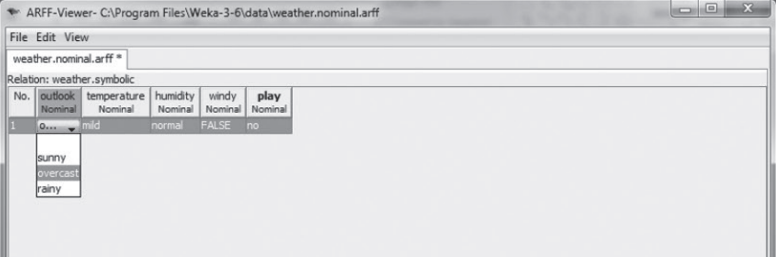
1. Chọn File → Open → chọn weather.nominal.arf

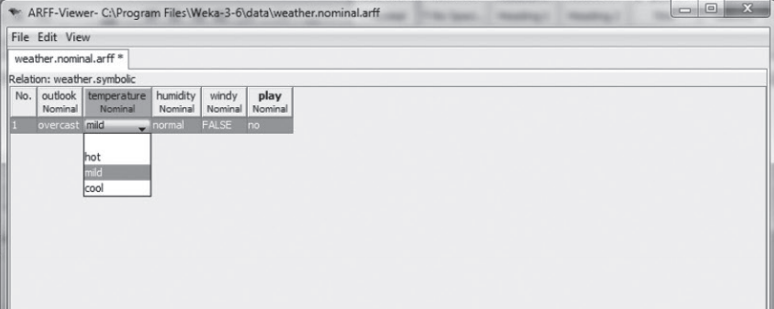


1. Chọn tất cả các bản ghi trong ArffViewer ngoại trừ một bản ghi vì chúng tôi muốn giữ bản ghi đó như một trường hợp bản ghi thử nghiệm. Trong trường hợp bạn muốn có n bản ghi làm tập dữ liệu thử nghiệm thì hãy chọn tất cả các bản ghi ngoại trừ n bản ghi đầu tiên
2. Chọn Edit → Delete Instances

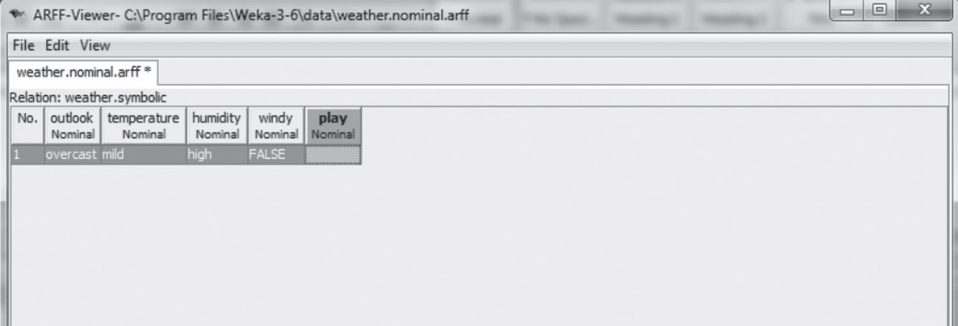


1. Bây giờ, hãy thay đổi các giá trị của một bản ghi và đặt nó theo các giá trị mà chúng ta muốn có cho bản ghi thử nghiệm bằng cách sử dụng tùy chọn thả xuống để đặt giá trị của các thuộc tính

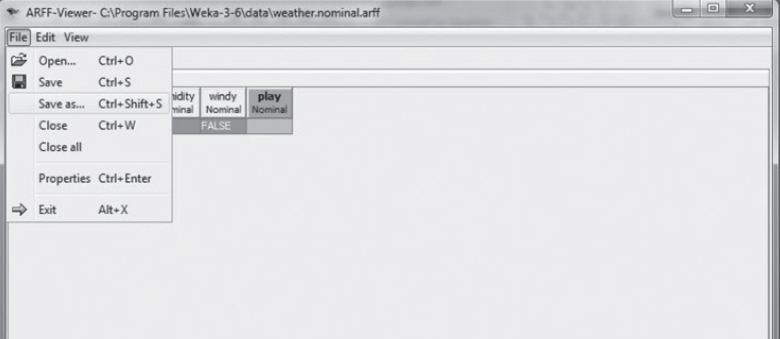


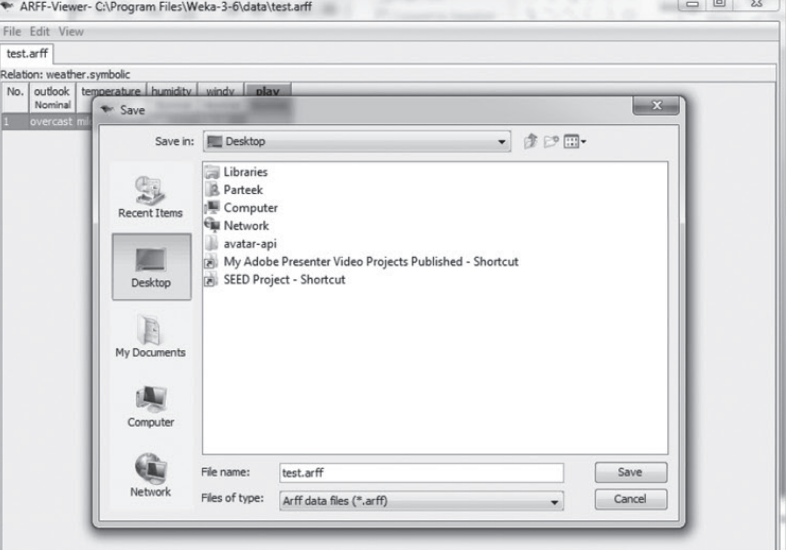


1. Giữ thuộc tính lớp trống vì chúng tôi muốn dự đoán các giá trị của nó bằng bộ phân loại

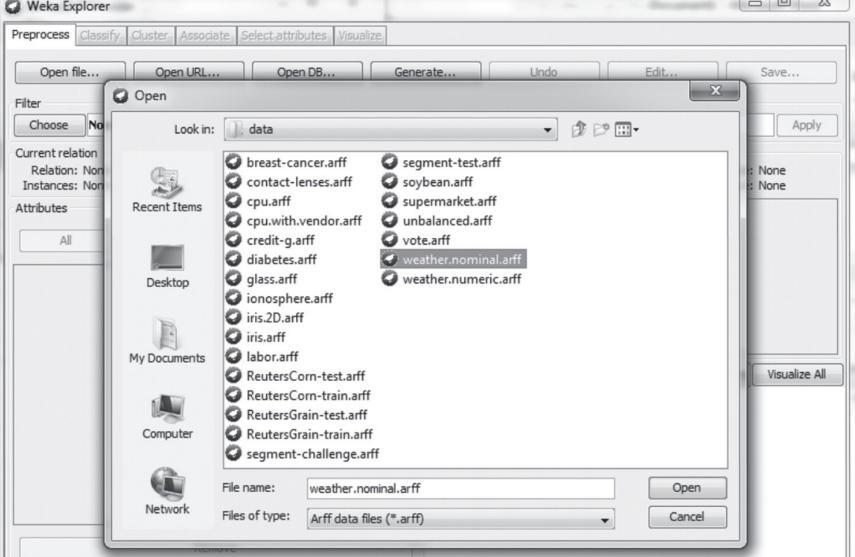


1. Sau khi thay đổi giá trị của các hàng, Lưu tệp dưới dạng test.arff

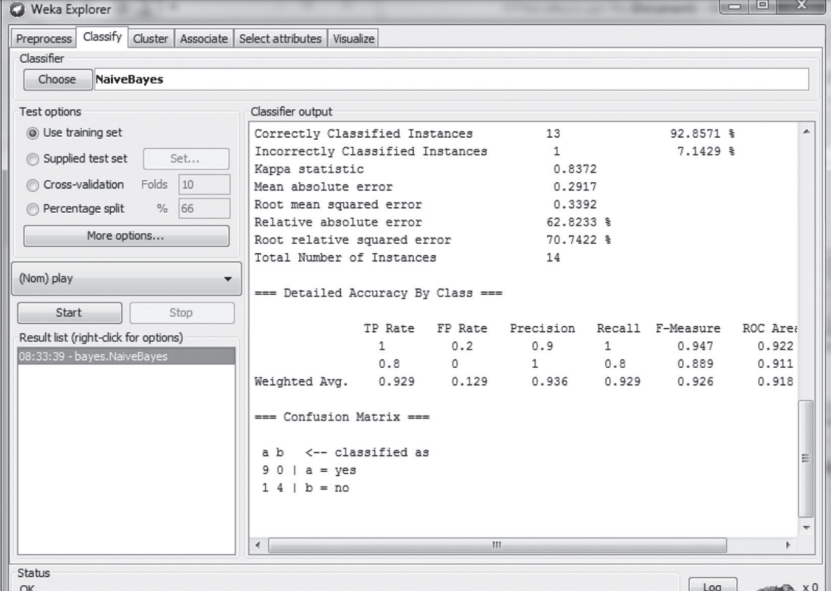




1. Bây giờ chuyển đến tab Preprocess và nhấp vào tùy chọn Open file và chọn weather.nominal.arf

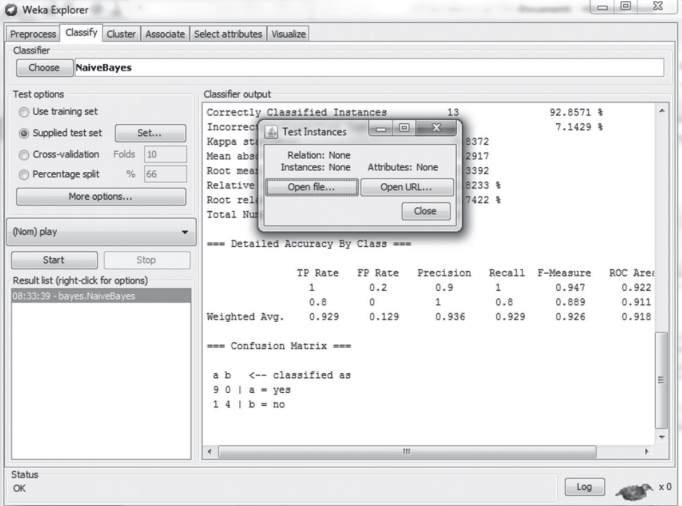


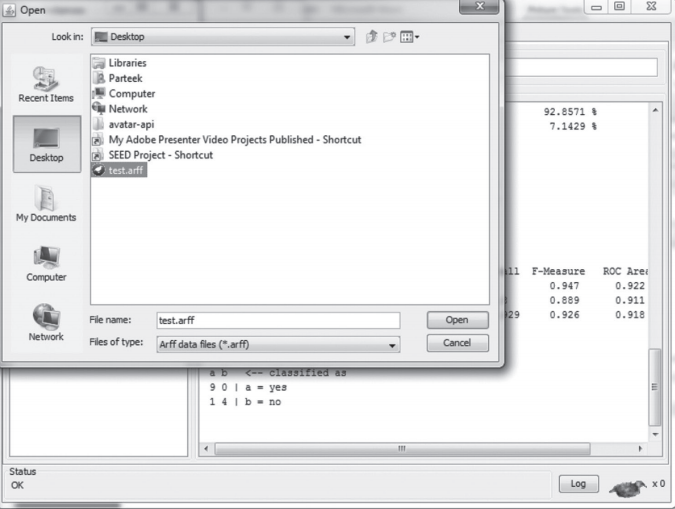
1. Chọn tab Classify và chọn NaiveBayes Classifier.
2. Nhấp vào nút Bắt đầu, nó sẽ xây dựng bộ phân loại

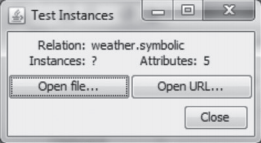


Chọn tùy chọn Supplied test set và nhấp vào nút Set

1. Sau đó Open file → chọn tệp test.arff. Khi tệp thử nghiệm đã được chọn, nó sẽ hiển thị chi tiết về tệp, tức là tên của mối quan hệ, số lượng thuộc tính và cá thể, v.v.

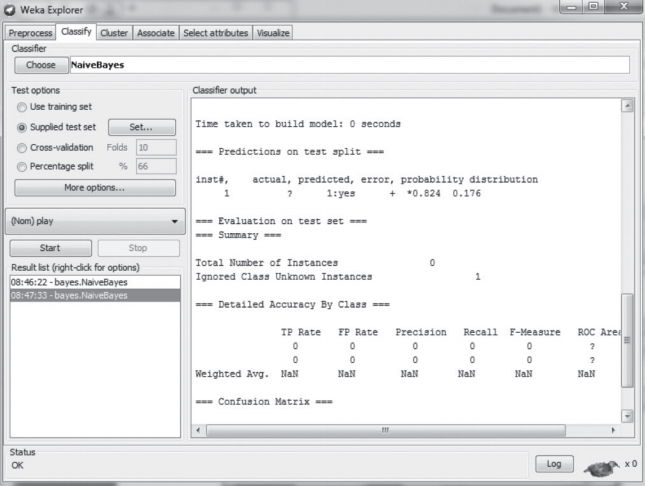






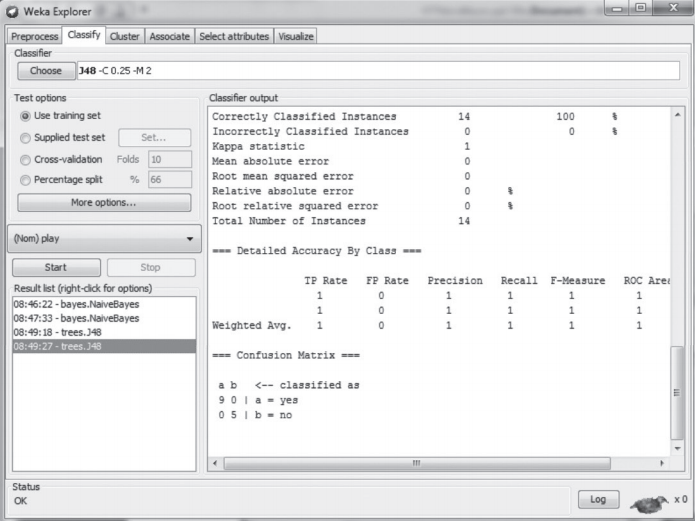
1. Nhấp vào nút Start và áp dụng bộ phân loại trên các trường hợp thử nghiệm. Trong trường hợp này, trình phân loại đã dự đoán một trường hợp chưa xác định là Play: Yes

Trong trường hợp của bạn, giá trị dự đoán có thể khác vì nó phụ thuộc vào giá trị của các thuộc tính được đặt trong quá trình tạo tập dữ liệu thử nghiệm.



Bây giờ, chúng ta hãy áp dụng bộ phân loại J48 trên tập dữ liệu thử nghiệm này để xác minh xem nó có hiển thị cùng một dự đoán hay không bằng cách thực hiện các bước sau.

* 1. Thay đổi trình phân loại thành J48 bằng cách sử dụng các bước sau. Nhấp vào Choose → Weka → classifiers  → trees  → J48
  2. Chọn tùy chọn Use Training Set và nhấp vào nút Bắt đầu để xây dựng bộ phân loại J48



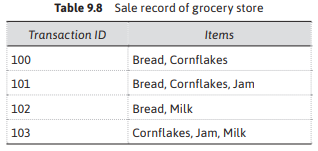
* 1. Bây giờ chọn tùy chọn supplied test set được cung cấp và chọn bộ dữ liệu kiểm tra là test.arff và sau đó nhấp vào nút bắt đầu để áp dụng J48 trên bộ dữ liệu kiểm tra. Bạn có thể xác minh rằng J48 cũng đã dự đoán giá trị của Play là Yes, giống như trong Bộ phân loại Naïve Bayes. Tuy nhiên, J48 chắc chắn hơn về dự đoán của nó so với Naïve Bayes vì ​​nó đã dự đoán với xác suất 1,0 trong khi Naïve Bayes dự đoán tương tự với xác suất 0,824. Bạn có thể tìm thấy các giá trị này trong đầu ra

**Câu 3:**

**Thuật toán Naïve (Naïve Algorithm) tìm luật kết hợp.**

Cho bộ dữ liệu như bảng 9.8 thuật toán Naïve được sử dụng để xác định mối liên hệ giữa các mặt hàng được bán trong bảng.

Giả sử min\_sup = 0.5 và min\_conf = 0.75

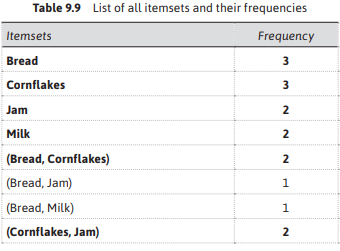
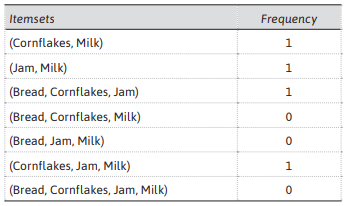


* Cách thức hoạt động

Liệt kê tất cả các sự kết hợp của các mặt hàng sau đó xác định tần xuất xuất hiện của mối kết hợp, hoặc sự kết hợp có nhiều hơn hoặc bằng với độ hỗ trợ xác định.

Sử dụng cách kết hợp này thì những luật kết hợp có độ tin cậy lớn hơn hoặc bằng giới hạn ngưỡng được xác định.

Với 4 item trong bảng sẽ có khả năng tạo ra 24 sự kết hợp bao gồm cả sự kết hợp rỗng. Nếu không xét mối kết hợp rỗng thì ta có bảng sau:



min\_sub = 0.5 => sup = 2 từ đây thì ta xác định các itemsets xuất hiện trong ít nhất 2 giao dịch. Theo như trong bảng ta thấy 4 items Bread, Cornflakes, Jam, Milk là phổ biến và độ phổ biến giảm dần khi có sự kết hợp từ 2 items trở đi ngoại trừ (Bread, Cornflakes), (Conrnflakes, Jam) không có tập 3 itemset phổ biến nào.

Luật kết hợp chỉ có ý nghĩa khi có ít nhất 1 cặp items trở lên, vì thế chúng ta bỏ qua các itemset chỉ có 1 item và chỉ quan tâm đến các cặp (Bread, Cornflakes), (Conrnflakes, Jam).

*Ghi chú: 1 luật kết hợp với 2-itemset (A, B) có thể dẫn đến các luật A → B, B → A nếu cả 2 đều thỏa mãn yêu cầu về độ tin cậy.*

Với yêu cầu độ tin cậy nhỏ nhất là 0.75:

Bread→Cornflakes

Confidence = Support of (Bread, Cornflakes) / Support of (Bread) = 2/3 = 67% Cornflakes→Bread

Confidence = Support of (Cornflakes, Bread) / Support of (Cornflakes) = 2/3 = 67% Cornflakes→Jam

Confidence = Support of (Cornflakes, Jam) / Support of (Cornflakes) = 2/3 = 67% Jam→Cornflakes

Confidence = Support of (Jam, Cornflakes) / Support of (Jam) = 2/2 = 100%

Kết luận chỉ có luật Jam→Cornflakes thỏa mãn độ tin cậy tối thiểu 0.75

**Hạn chế của thuật toán Naïve Bayes**

Chỉ hoạt động tốt với bộ dữ liệu có 4 itemsets, khó có thể xử lý với bộ dữ liệu lớn càng nhiều itemset thì các mối kết hợp gia tăng theo cấp số mũ

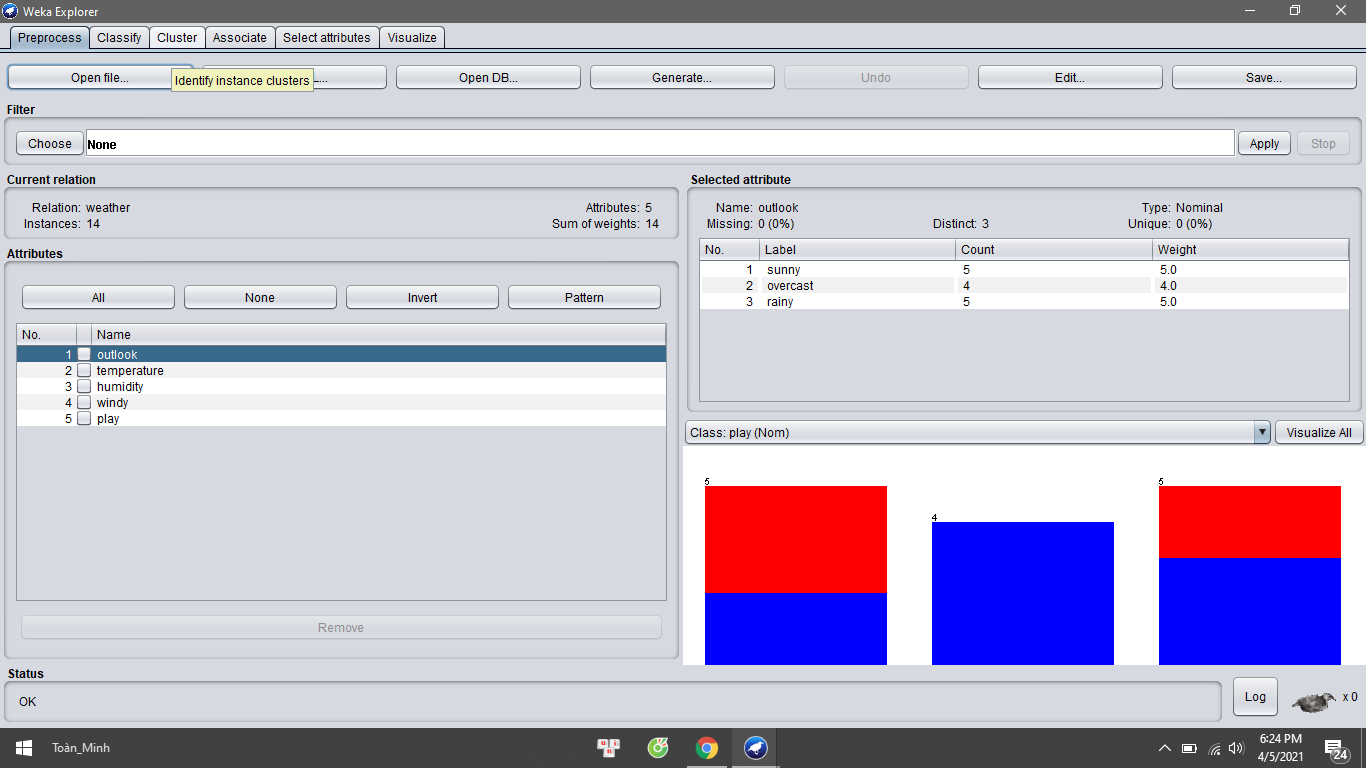
|  |  |
| --- | --- |
| TIDs | Items |
| 1 | x1, x2, x4, x5, x7, x8, x9 |
| 2 | x2, x3, x4, x6, x7, x8 |
| 3 | x1, x3, x4, x5, x6, x7, x8, x9 |
| 4 | x1, x4, x5, x7, x8, x9 |
| 5 | x1, x5, x6, x7, x9 |
| 6 | x2, x4, x5, x6, x7, x8 |
| 7 | x2, x4, x5, x7, x8 |
| 8 | x3, x4, x5, x6, x7, x8, x9 |
| 9 | x4, x7, x8 |
| 10 | x5, x7, x8 |
| 11 | x3, x4, x6, x7, x8, x9 |
| 12 | x1, x5, x7, x8, x9 |
| 13 | x1, x2, x3, x4, x5, x6, x7, x8, x9 |
| 14 | x2, x3, x4, x5, x6, x7, x8, x9 |
| 15 | x2, x3, x4, x6, x7, x8, x9 |
| 16 | x1, x5, x6, x7, x8, x9 |

|  |  |
| --- | --- |
| Itemset | Frequency |
| x1 | 7 |
| x2 | 7 |
| x3 | 6 |
| x4 | 14 |
| x5 | 12 |
| x6 | 10 |
| x7 | 16 |
| x8 | 15 |
| x9 | 11 |
| x1, x2 | 2 |
| x1, x3 | 2 |
| x1, x4 | 3 |
| x1, x5 | 7 |
| x1, x6 | 4 |
| x1, x7 | 7 |
| x1, x8 | 6 |
| x1, x9 | 6 |
| x2, x3 | 3 |
| x2, x4 | 7 |
| x2, x5 | 4 |
| x2, x6 | 5 |
| x2, x7 | 6 |
| x2, x8 | 7 |
| x2, x9 | 4 |
| x3, x4 | 7 |
| x3, x5 | 4 |
| x3, x6 | 7 |
| x3, x7 | 7 |
| x3, x8 | 7 |
| x3, x9 | 6 |
| x4, x5 | 8 |
| x5, x6 | 7 |
| x5, x7 | 12 |
| x5, x8 | 11 |
| x5, x9 | 9 |
| x6, x7 | 10 |
| x6, x8 | 9 |
| x6, x9 | 8 |
| x7, x8 | 15 |
| x7, x9 | 11 |
| x8, x9 | 10 |
| x1, x2, x3 | 1 |
| x1, x2, x4 | 2 |
| x1, x2, x5 | 2 |
| x1, x2, x6 | 1 |
| x1, x2, x7 | 2 |
| x1, x2, x8 | 2 |
| x1, x2, x9 | 2 |
| x2, x3, x4 | 4 |
| x2, x3, x5 | 2 |
| x2, x3, x6 | 4 |
| x2, x3, x7 | 4 |
| x2, x3, x8 | 4 |
| x2, x3, x9 | 3 |
| x3, x4, x5 | 4 |
| x3, x4, x6 | 7 |
| x3, x4, x7 | 7 |
| x3, x4, x8 | 7 |
| x3, x4, x9 | 6 |
| x4, x5, x6 | 5 |
| x4, x5, x7 | 8 |
| x4, x5, x8 | 8 |
| x4, x5, x9 | 6 |
| x5, x6, x7 | 7 |
| x5, x6, x8 | 6 |
| x5, x6, x9 | 6 |
| x6, x7, x8 | 9 |
| x6, x7, x9 | 8 |
| x7, x8, x9 | 10 |
| x1, x2, x3, x4 | 1 |
| x1, x2, x3, x5 | 1 |
| x1, x2, x3, x6 | 1 |
| x1, x2, x3, x7 | 1 |
| x1, x2, x3, x8 | 1 |
| x1, x2, x3, x9 | 1 |
| x1, x2, x3, x4, x5 | 1 |
| x1, x2, x3, x4, x6 | 1 |
| x1, x2, x3, x4, x7 | 1 |
| x1, x2, x3, x4, x8 | 1 |
| x1, x2, x3, x4, x9 | 1 |
| x1, x2, x3, x4, x5, x6 | 1 |
| x1, x2, x3, x4, x5, x7 | 1 |
| x1, x2, x3, x4, x5, x8 | 1 |
| x1, x2, x3, x4, x5, x9 | 1 |
| x1, x2, x3, x4, x5, x6, x7 | 1 |
| x1, x2, x3, x4, x5, x6, x8 | 1 |
| x1, x2, x3, x4, x5, x6, x9 | 1 |
| x1, x2, x3, x4, x5, x6, x7, x8 | 1 |
| x1, x2, x3, x4, x5, x6, x7, x9 | 1 |
| x1, x2, x3, x4, x5, x6, x7, x8, x9 | 1 |

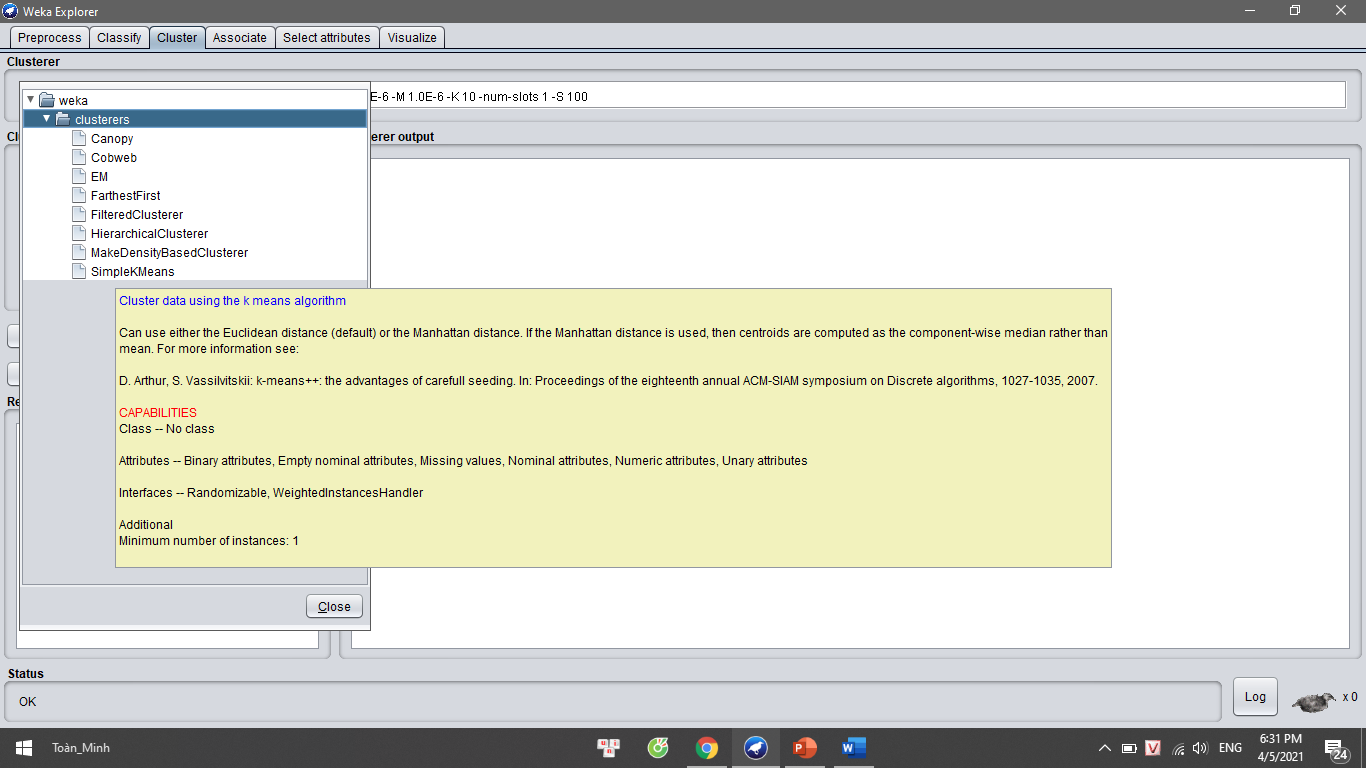
Câu 4:

**Thuật toán: K-means**

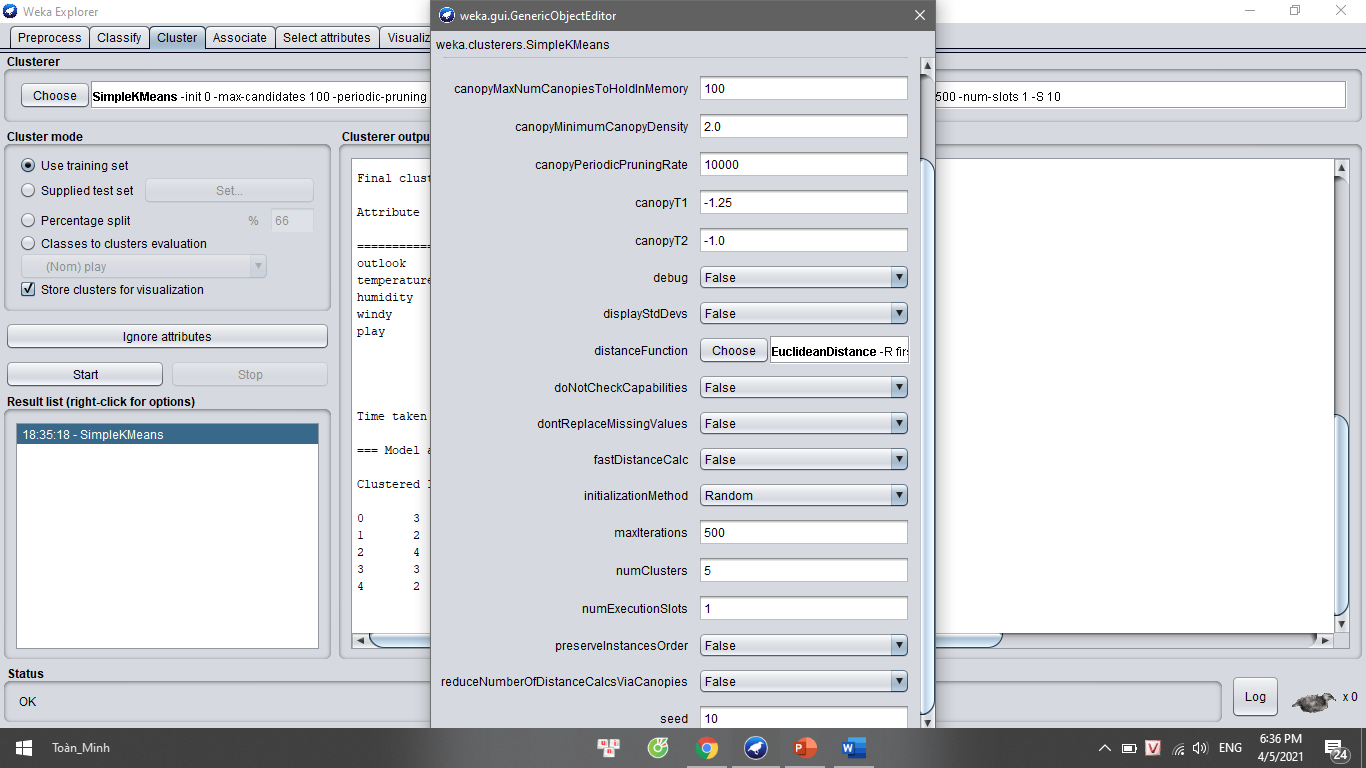
Sau khi mở file .arff. Chọn vào thẻ *Cluster* rồi chọn *Choose* .



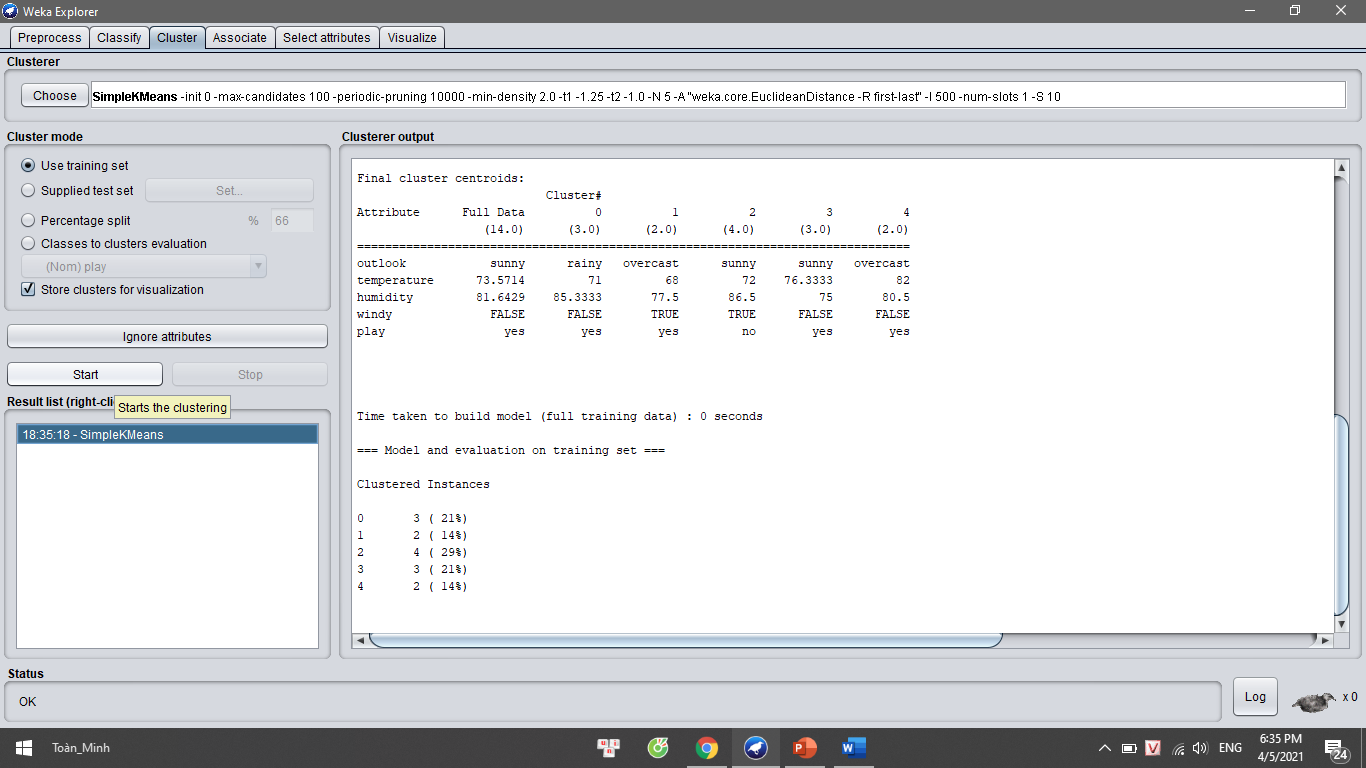
Tại đây trong folder Clusterers, chọn vào *SimpleKMeans* .



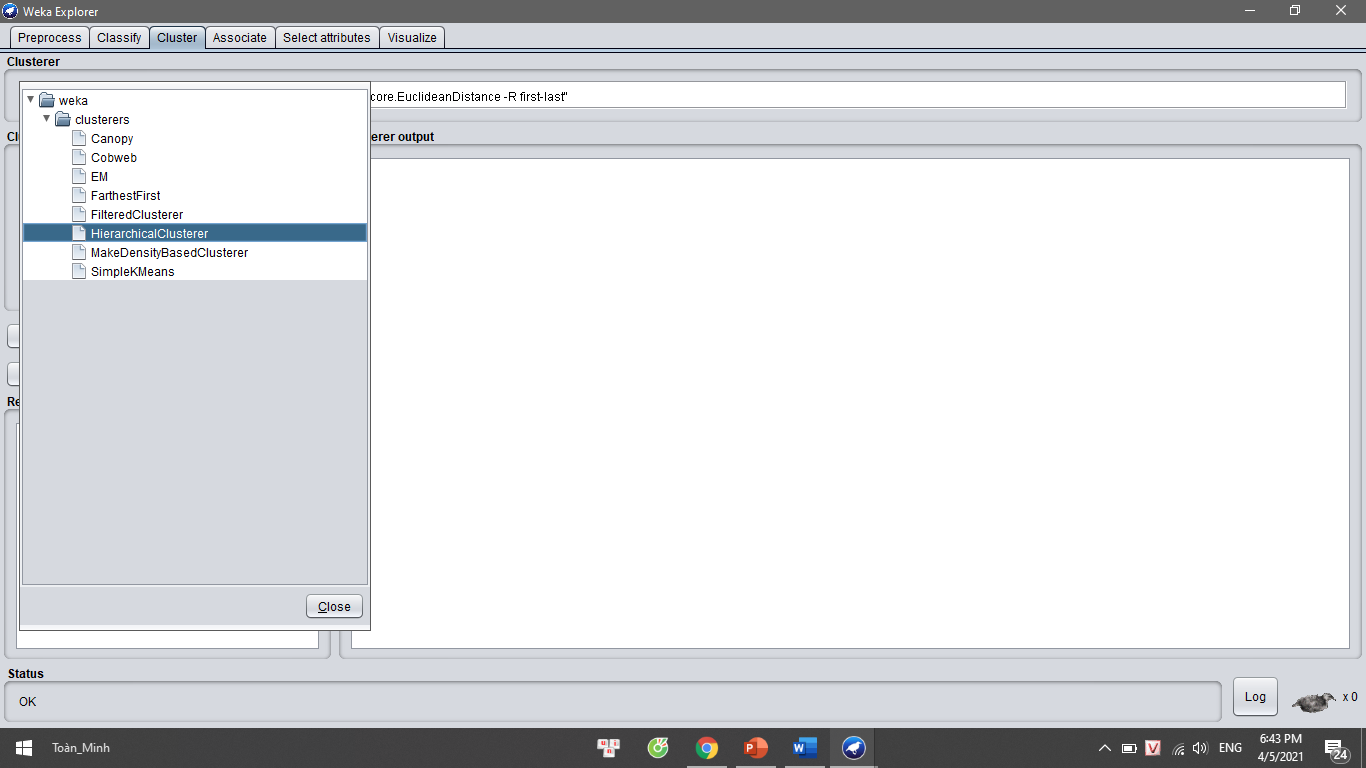
Nhấp vào thanh lệnh rồi kéo xuống thay đổi thuộc tính *numClusters* theo số cụm mong muốn( ví dụ chọn 5).



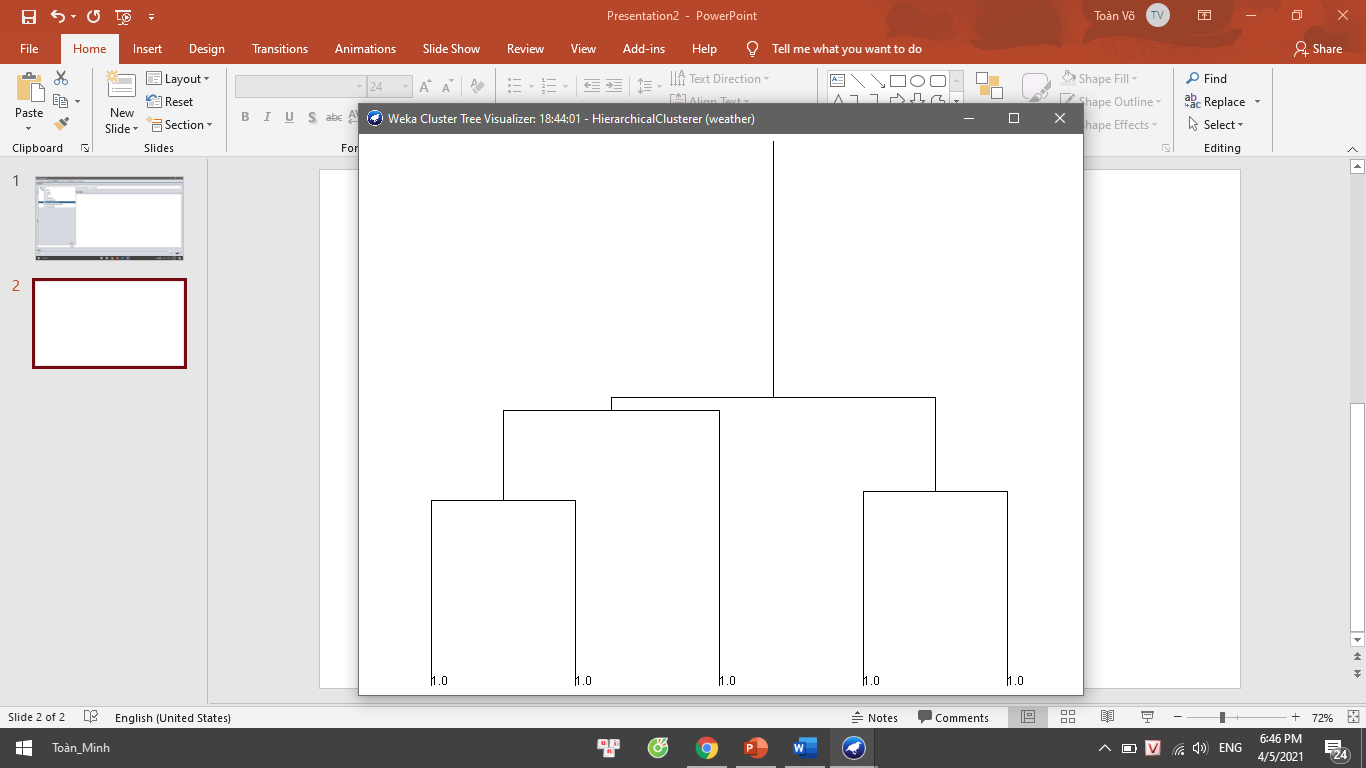
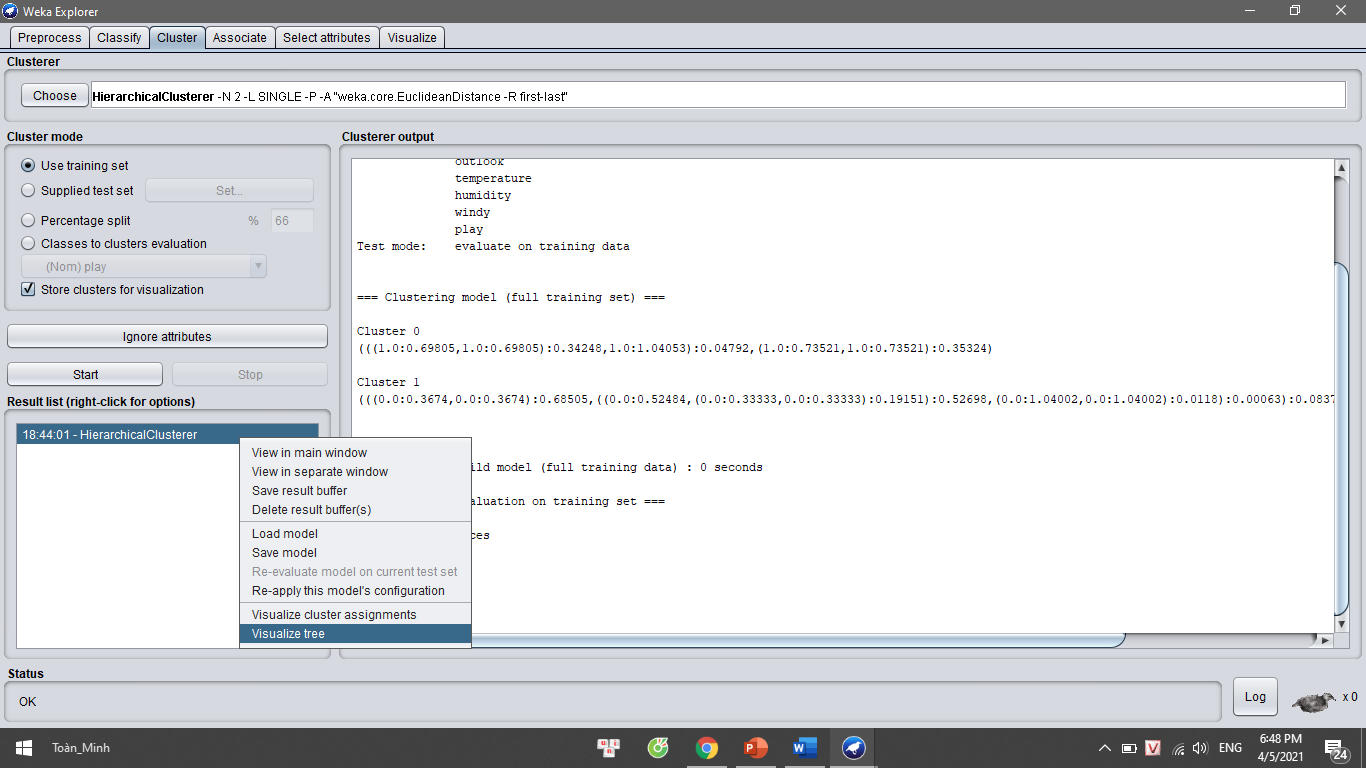
Nhấn *OK* rồi nhấn *Start* .



**Thuật toán: HCA(Hierarchical Clustering Algorithms)**

Cũng với bộ dữ liệu trên, trong thư mục Clusterers chọn vào *HierarchicalClusterers*.

Chọn *Start*. Nhấp chuột phải vào ô *Result list* và chọn *Visualize tree* để xem cây được tạo.



Ý tưởng chính đằng sau DBSCAN là một điểm thuộc về một cụm nếu nó gần với nhiều điểm từ cụm đó.

Có hai tham số chính của DBSCAN:

* **eps** : Khoảng cách chỉ định các vùng lân cận. Hai điểm được coi là hàng xóm nếu khoảng cách giữa chúng nhỏ hơn hoặc bằng eps.
* **minPts:** Số điểm dữ liệu tối thiểu để xác định một cụm.
* **Điểm cốt lõi:** Một điểm là điểm cốt lõi nếu có ít nhất số điểm minPts (bao gồm cả chính điểm đó) trong khu vực xung quanh của nó với bán kính eps.
* **Điểm biên giới:** Một điểm là điểm biên giới nếu nó có thể tiếp cận được từ một điểm cốt lõi và có số điểm nhỏ hơn minPts trong khu vực xung quanh của nó.

**Điểm ngoại lệ** **:** Điểm là điểm ngoại lệ nếu nó không phải là điểm cốt lõi và không thể đạt được từ bất kỳ điểm cốt lõi nào.

Hoạt động của thuật toán:

minPts và eps được xác định.

**Điểm bắt đầu:** được chọn ngẫu nhiên tại khu vực lân cận của nó được xác định bằng cách sử dụng bán kính eps. Nếu có số điểm ít nhất minPts trong vùng lân cận, điểm đó được đánh dấu là điểm cốt lõi và bắt đầu hình thành cụm. Nếu không, điểm được đánh dấu là nhiễu. Khi bắt đầu hình thành cụm (giả sử cụm A), tất cả các điểm trong vùng lân cận của điểm ban đầu sẽ trở thành một phần của cụm A. Nếu các điểm mới này cũng là điểm cốt lõi, thì các điểm nằm trong vùng lân cận của chúng cũng được thêm vào cụm A.

Bước tiếp theo là chọn ngẫu nhiên một điểm khác trong số các điểm chưa được thăm ở các bước trước. Sau đó, quy trình tương tự được áp dụng.

Quá trình này kết thúc khi tất cả các điểm được truy cập.

**Ưu điểm:**

* Có thể xác định được cụm có hình dáng bất kỳ.
* Có thể thay đổi quy mô (hiệu quả khi làm việc với cơ sở dữ liệu lớn)
* Có khả năng xử lý nhiễu tốt.

**Nhược điểm:**

* Nếu các cụm rất khác nhau về mật độ trong cụm, thì DBSCAN không phù hợp để xác định các cụm.
* Trong một số trường hợp, xác định khoảng cách thích hợp của vùng lân cận (eps) là không dễ dàng