TRƯỜNG HÈ HỌC MÁY THỐNG KÊ

THỰC HÀNH BUỔI 1: HỆ HỖ TRỢ QUYẾT ĐỊNH VÀ PHƯƠNG PHÁP DỰ ĐOÁN

(Decision Tree and regression method)

I. Nội dung chính:

- Giới thiệu tổng quan về công cụ weka
- Định dạng của dữ liệu
- Phân tích đánh giá đặc trưng dựa trên tần suất.
- Phương pháp hiển thị dữ liệu
- Phương pháp trích chọn đặc trưng
- Phương pháp J84: sử dụng, chọn tham số cho mô hình, đánh giá kết quả cho dữ liệu kiểm tra
- Phương pháp regression: sử dụng, chọn tham số cho mô hình, đánh giá kết quả cho dữ liệu kiểm tra

II. Giới thiệu tổng quan về công cụ weka:

+ Weka là một phần mềm được viết bằng ngôn ngữ JAVA do Witten và Frank xây dựng.

+ Weka bao gồm các phương pháp học máy cơ bản phục vụ cho các mục tiêu sau:

- Tiền xử lý dữ liệu, các phương pháp học máy cơ bản và các phương pháp đánh giá mô hình
- Sử dụng đồ họa để biểu diễn dữ liệu
- Là một môi trường được dùng để so sánh các thuật toán học.

+ Weka website: http://www.cs.waikato.ac.nz/ml/weka/

+ Tài liệu hướng dẫn sử dụng weka:

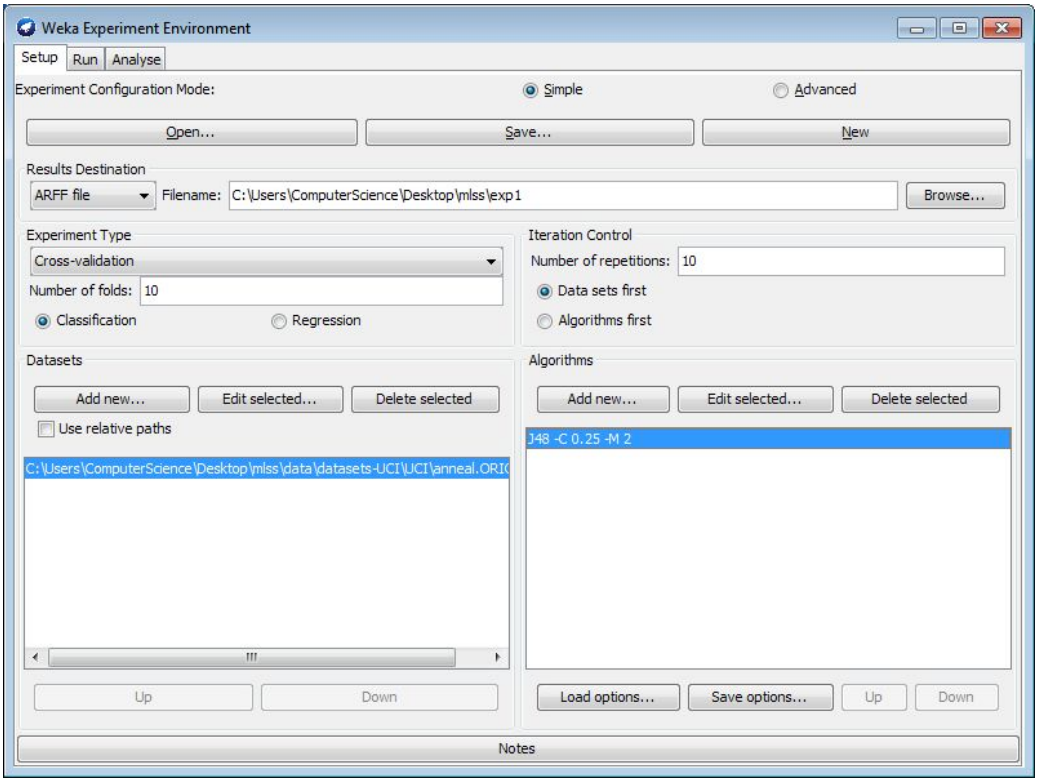
http://www.cs.waikato.ac.nz/ml/weka/documentation.html

+ Cơ sở dữ liệu: http://www.cs.waikato.ac.nz/ml/weka/datasets.html

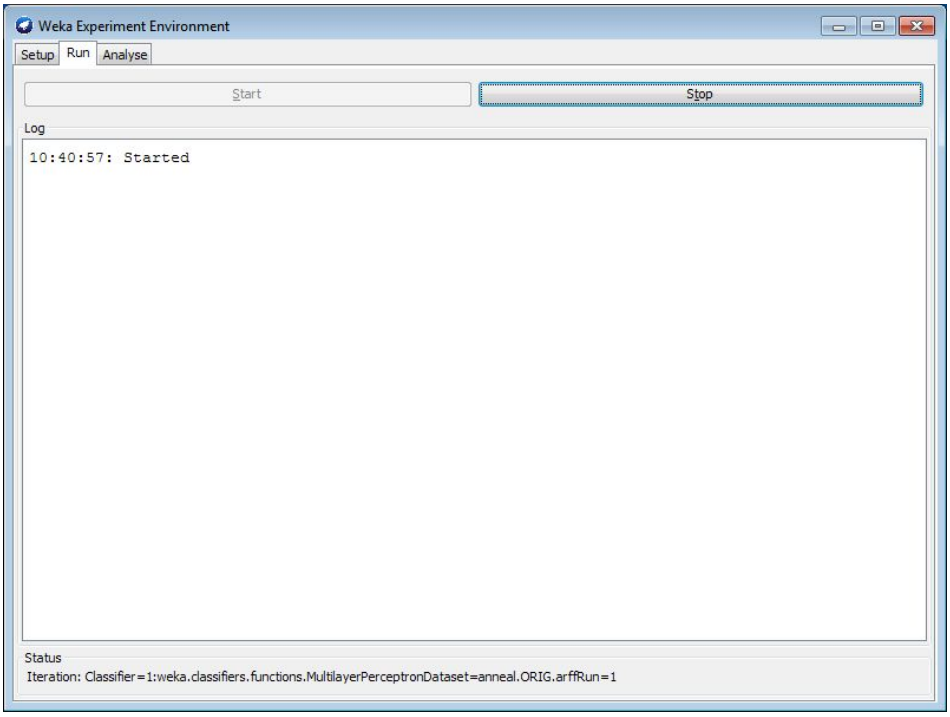
+ Giao diện chính:

Hình 1: Giao diện chính của Weka

- **Experimenter:** Dùng để thiết kế mô hình, thay đổi mô hình, thực thi mô hình và đánh giá mô hình của người sử dụng.

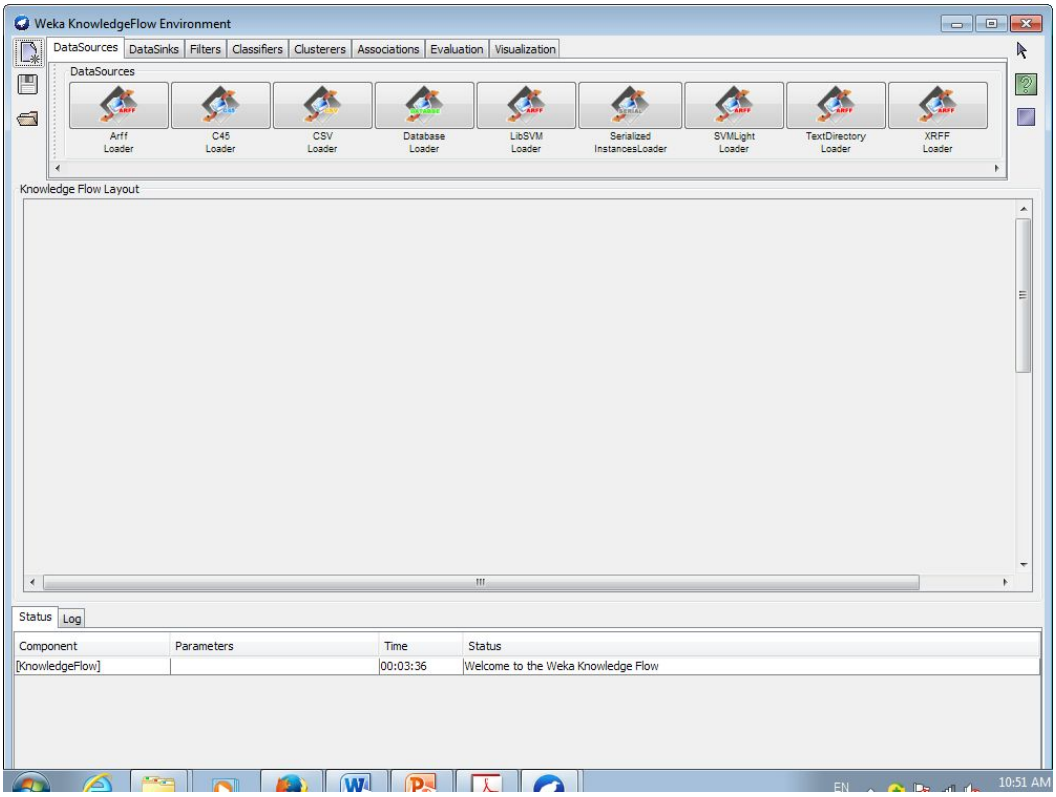


Hình 2. Thiết lập thực nghiệm trong phần Experimenter



Hình 3. Chạy thực nghiệm trong phần Experimenter

- **KnowledgeFlow:** Cung cấp một giao diện khác cho Explorer.



Hình 4. Giao diện chính của KnowledgeFlow

- **Explorer:**

Các chức năng chính:

+ Preprocess (tiền xử lý dữ liệu): Chọn và thay đổi dữ liệu.

+ Phân lớp hoặc hồi quy (Classification or regression): Học và kiểm tra các mô hình cho bài toán phân lớp hoặc bài toán dự đoán.

+ Phân cụm (Cluster). Học các cụm từ dữ liệu.

+ Luật kết hợp (Associate): Học các luật kết hợp từ dữ liệu

+ Lựa chọn thuộc tính (Select attributes): Lựa chọn các thuộc tính "hữu ích" để biểu diễn dữ liệu.

+ Trực quan hóa (Visualize). Hiển thị biểu đồ 2D.

III. Định dạng dữ liệu (ARFF)

ARFF: Attribute-Relation File Format

a. Head file:

- Comments: bắt đầu bằng dấu %
- Relations: @relation <relation-name>

<relation-name>: là một xâu cho biết tên của dữ liệu. Nếu là một xâu chứa dấu các thì xâu phải nằm trong dấu '' hoặc "".

- Khai báo kiểu dữ liệu (data declaration): @attribute <tên thuộc tính/đặc trưng> <kiểu dữ liệu>

- Tên thuộc tính/đặc trưng: là xâu bắt đầu bằng chữ cái. Nếu tồn tại dấu cách trong xâu thì xâu phải nằm trong dấu ‘’ hoặc “”.
- Kiểu dữ liệu:
  + integer: số nguyên
  + real: số thực

  + numeric: bao gồm số nguyên (integer) và số thực (real)

  + string

  + Nominal (categorical data):

  + Date: @ATTRIBUTE timestamp DATE "yyyy-MM-dd HH:mm:ss"

  + Relational

```
@attribute <name> relational
        <further attribute definitions>
@end <name>
```

- Ví dụ minh họa:

```
% 1. Title: Iris Plants Database
%
% 2. Sources:
% (a) Creator: R.A. Fisher
% (b) Donor: Michael Marshall (MARSHALL%PLU@io.arc.nasa.gov)
% (c) Date: July, 1988
%
@RELATION iris
@ATTRIBUTE sepallength NUMERIC
@ATTRIBUTE sepalwidth NUMERIC
@ATTRIBUTE petallength NUMERIC
@ATTRIBUTE petalwidth NUMERIC
@ATTRIBUTE class {Iris-setosa,Iris-versicolor,Iris-virginica}
```

b. Dữ liệu: @data

```
5.1,3.5,1.4,0.2,Iris-setosa
4.9,3.0,1.4,0.2,Iris-setosa
4.7,3.2,1.3,0.2,Iris-setosa
4.6,3.1,1.5,0.2,Iris-setosa
5.0,3.6,1.4,0.2,Iris-setosa
5.4,3.9,1.7,0.4,Iris-setosa
```

- Biểu diễn dữ liệu rời rạc:

```
@data
        0, X, 0, Y, "class A"
        0, 0, W, 0, "class B"
```

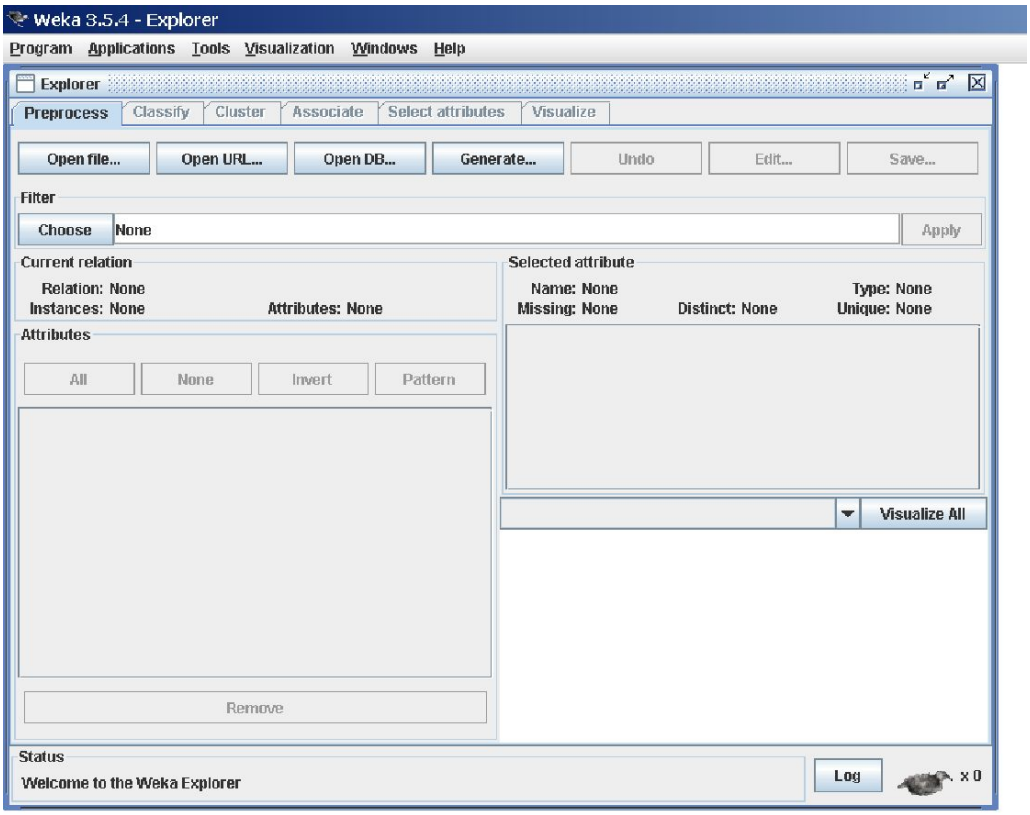
Có thể biểu diễn như sau:

@data

{1 X, 3 Y, 4 "class A"}
{2 W, 4 "class B"}

IV. Explorer

a. Preprocess



Hình 5. Giao diện của tiền xử lý dữ liệu

- **Tải dữ liệu:**

  + Open file....: cho phép đọc file dữ liệu từ hệ thống file cục bộ. Các kiểu file có thể đọc: ARFF, CSV, C4.5, binary.

  + Open URL....: Cho phép tải dữ liệu từ một tài nguyên chuẩn.

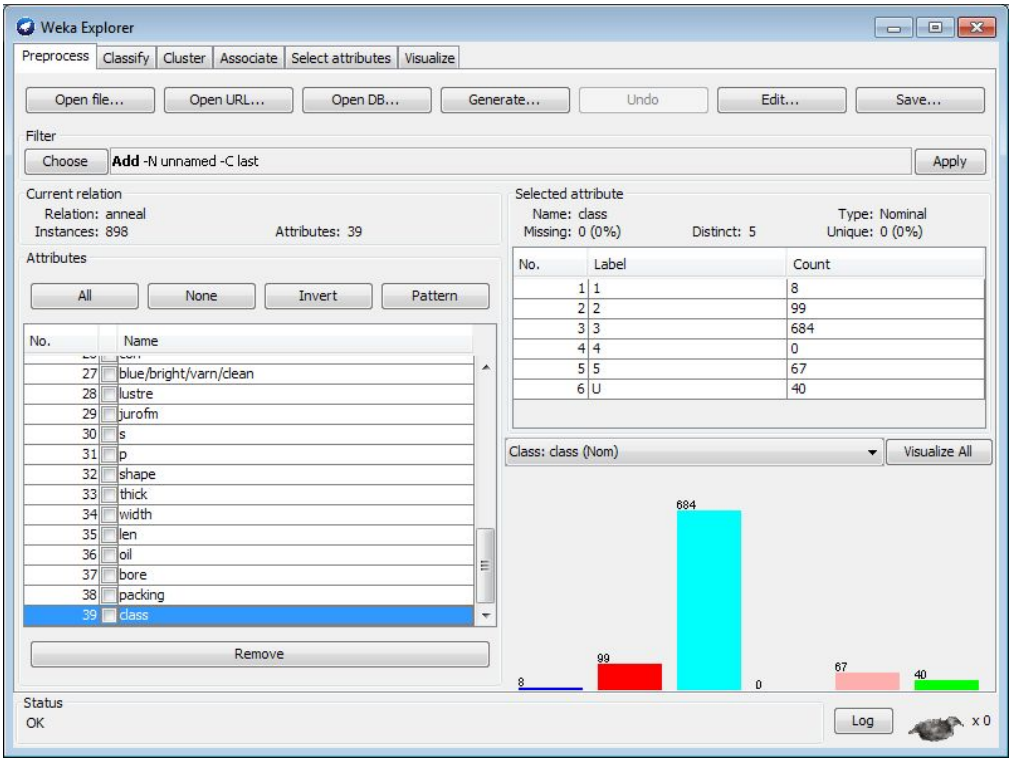
  + Open DB....: Đọc dữ liệu từ cơ sở dữ liệu

  + Generate....: Tạo dữ liệu nhân tạo bằng các thuật toán đã được xây dựng trong weka.

- **Trạng thải dữ liệu (current relation):** mô tả tên dữ liệu (relation), số lượng đối tượng (Instances), và số lượng thuộc tính (Attributes)
- **Lọc dữ liệu:** Có nhiều bộ lọc được cung cấp để lọc dữ liệu theo nhiều cách khác nhau.
- **Làm việc với các thuộc tính:** Cho phép lựa chọn thuộc tính, loại bỏ thuộc tính, hiển thị tên thuộc tính, kiểu dữ liệu của thuộc tính, số lượng giá trị mà thuộc tính nhận được, xác định

thuộc tính có phải là định danh, hiển thị tần suất xuất hiện các giá trị của thuộc tính, biểu diễn đồ thị mỗi quan hệ của thuộc tính với thuộc tính lớp (class label)

- **Ví dụ minh họa:**

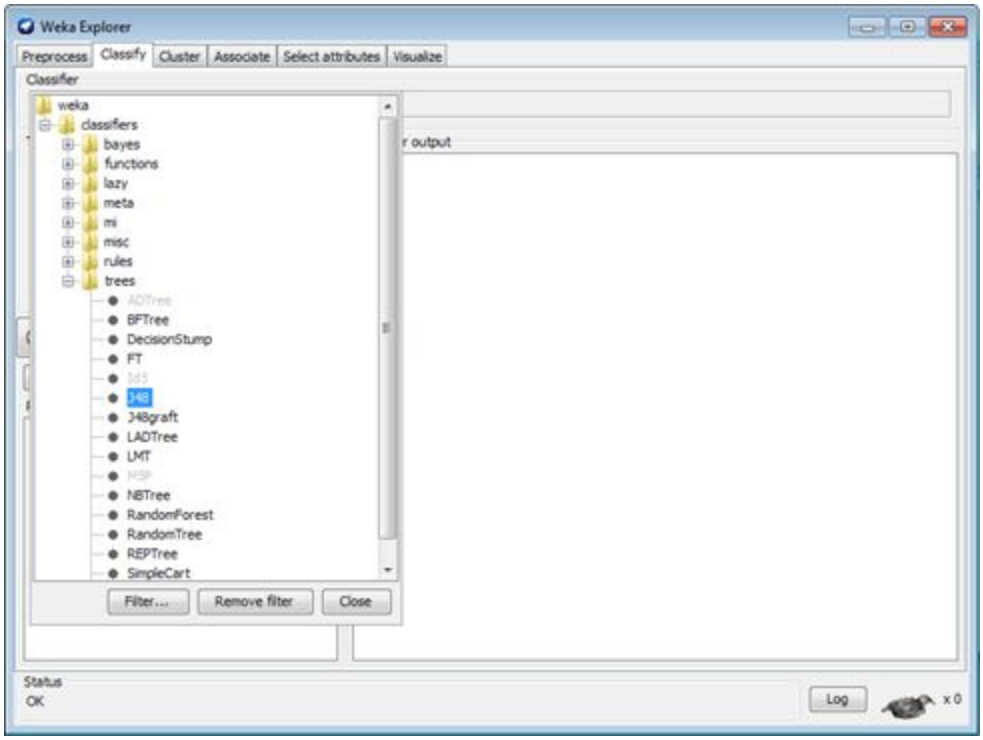


Hình 6. Ví dụ minh họa về tiền xử lý dữ liệu

Trong ví dụ này dữ liệu được sử dụng có tên là “anneal”, số lượng đối tượng là 898, số lượng thuộc tính là 39, các tên thuộc tính được liệt kê như ở hình 6. Khi cho thuộc tính class, các thông tin về thuộc tính được hiển thị như tên của thuộc tính, số lượng đối tượng nhiều (missing value), số lượng các giá trị của thuộc tính mà dữ liệu “anneal” chứa, kiểu dữ liệu là Nominal, ….. Ngoài ra, lược đồ histogram biểu diễn phân bố của các giá trị của thuộc tính class cũng được thể hiện như hình trên hình 6.

b. Phân lớp (classify tab)

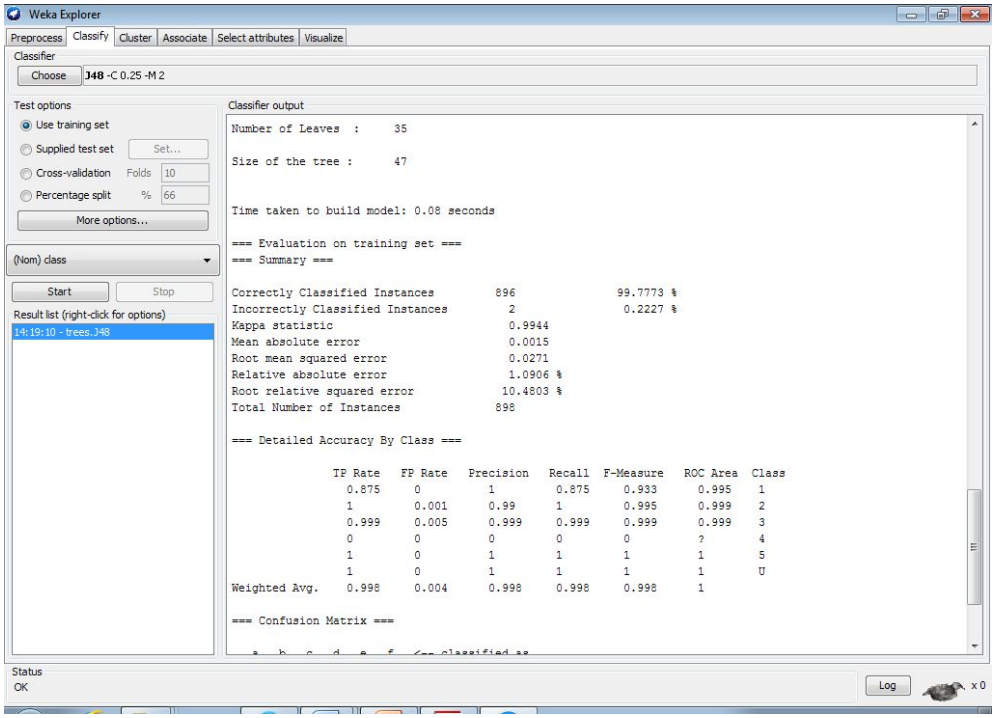
- **Lựa chọn bộ phân lớp (classifier)**: chọn Choose hộp thoại classifier sẽ hiện thị một danh sách các danh sách thuật toán phân lớp như hình sau:



Hình 7. Lựa chọn các thuật toán phân lớp

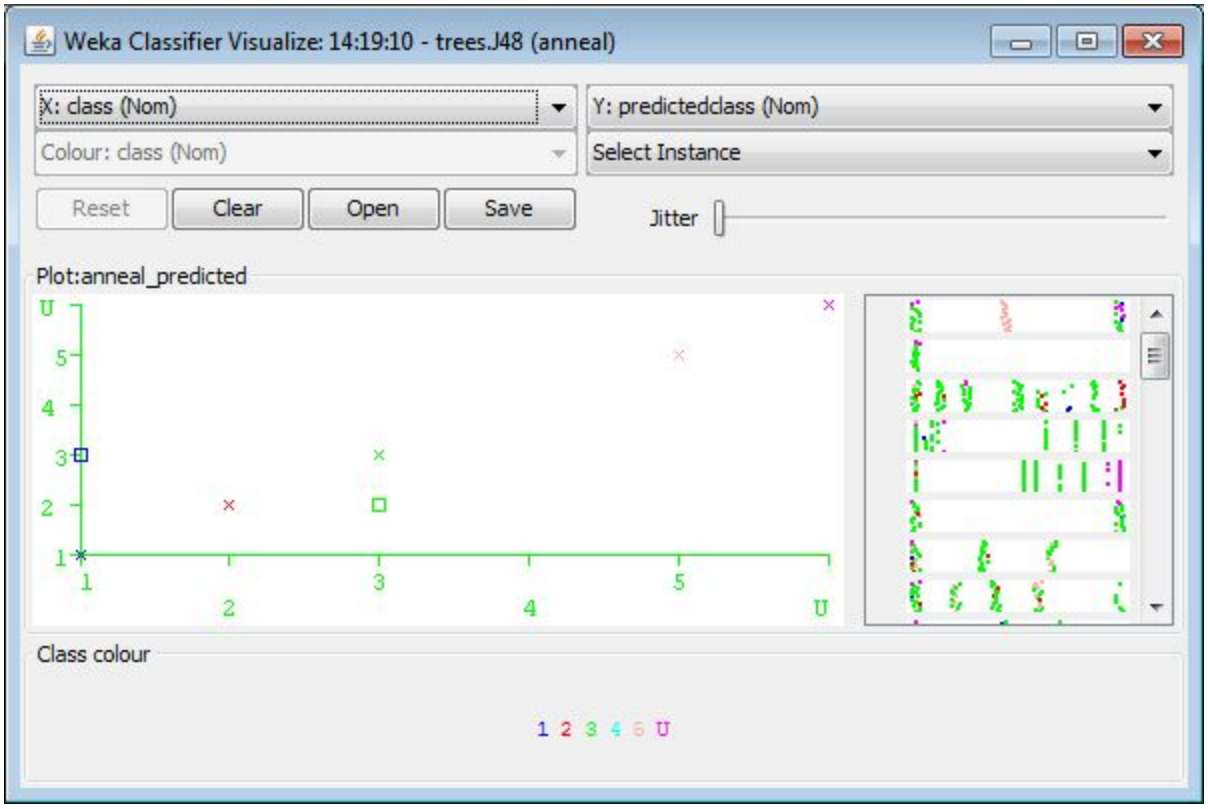
Click lựa chọn thuật toán phù hợp cho bài toán phân lớp cụ thể.

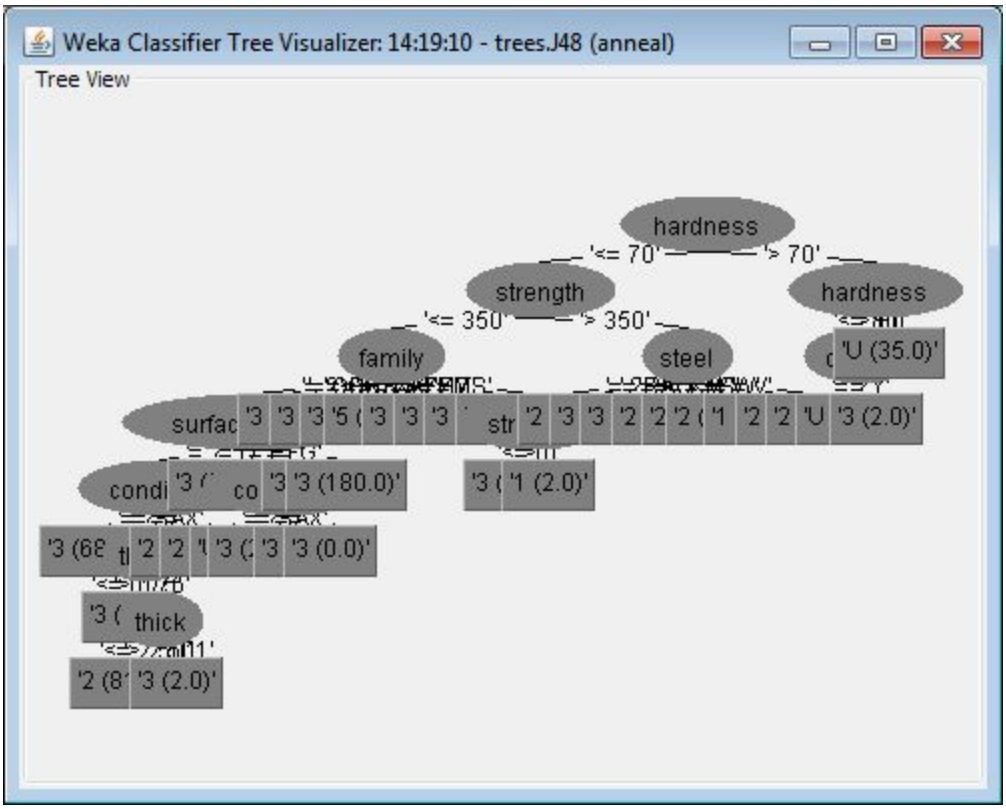
- **Lựa chọn cách học mô hình (test options):**
  - Use training set: Mô hình được học sẽ được kiểm tra trên dữ liệu được sử dụng để học.
  - Supplied test set: sử dụng một bộ dữ liệu nào đó để kiểm tra mô hình vừa được học.
  - Cross-validation: sử dụng kiểm tra chéo để kiểm tra chất lượng của mô hình được học.
  - Percentage split: chia dữ liệu thành hai phần, một phần dùng để học, một phần dùng để kiểm tra độ chất lượng của mô hình.
- **Thuộc tính class:** là thuộc tính được nhằm mục đích cho việc dự đoán.
- **Huấn luyện mô hình và kiểm tra:** chọn nút start để thực hiện việc học mô hình theo thuật toán lựa chọn và kiểm tra khả năng của mô hình được học dựa vào dữ liệu kiểm tra (testing data)
- **Kết quả phân lớp:**
  - **Thông tin chung:** Hiển thị các thông tin liên quan tới thuật toán, tên dữ liệu, số lượng đối tương, các thuộc tính, cách thức kiểm tra
  - **Mô hình phân lớp:** hiển thị mô hình phân lớp
  - **Mô tả thống kê:** liệt kê các thống kê về độ chính xác của mô hình trên dữ liệu kiểm tra.
  - **Mô tả chi tiết:** liệt kê các độ đo để đánh giá mô hình dựa trên dữ liệu kiểm tra.
  - **Confusion Matrix:** hiển thị số lượng đối tượng được kết gán tới mỗi lớp.



Hiển thị kết quả: phải chuột vào result list để chọn hiển thị các kết quả về: View in separate window, Save result buffer, Load model, Save model, Re-evaluate model on current test set, Visualize classifier errors, Visualize tree or Visualize graph, Visualize margin curve, Visualize threshold curve, Visualize cost curve.

Hình 8. Visualize classifier errors





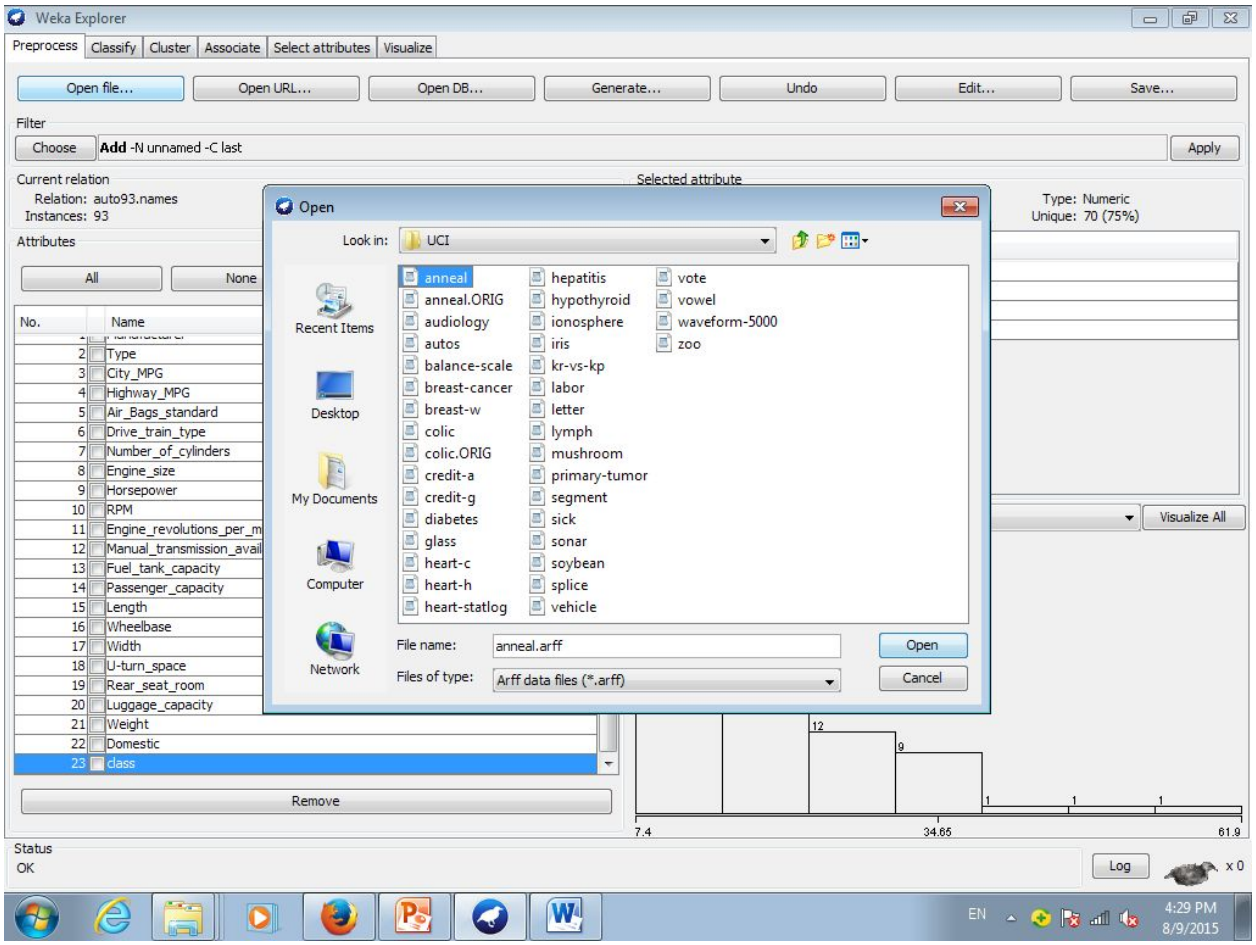
Hình 9. Visualize tree or Visualize graph

III. **Hướng dẫn thực hành sử dụng thuật toán J48**

- Dữ liệu đầu vào: anneal.arff
- Thuật toán xây dựng cây: J48
- Kết quả đầu ra: mô hình (cây) được học, các thông tin chi tiết về độ chính xác của mô hình

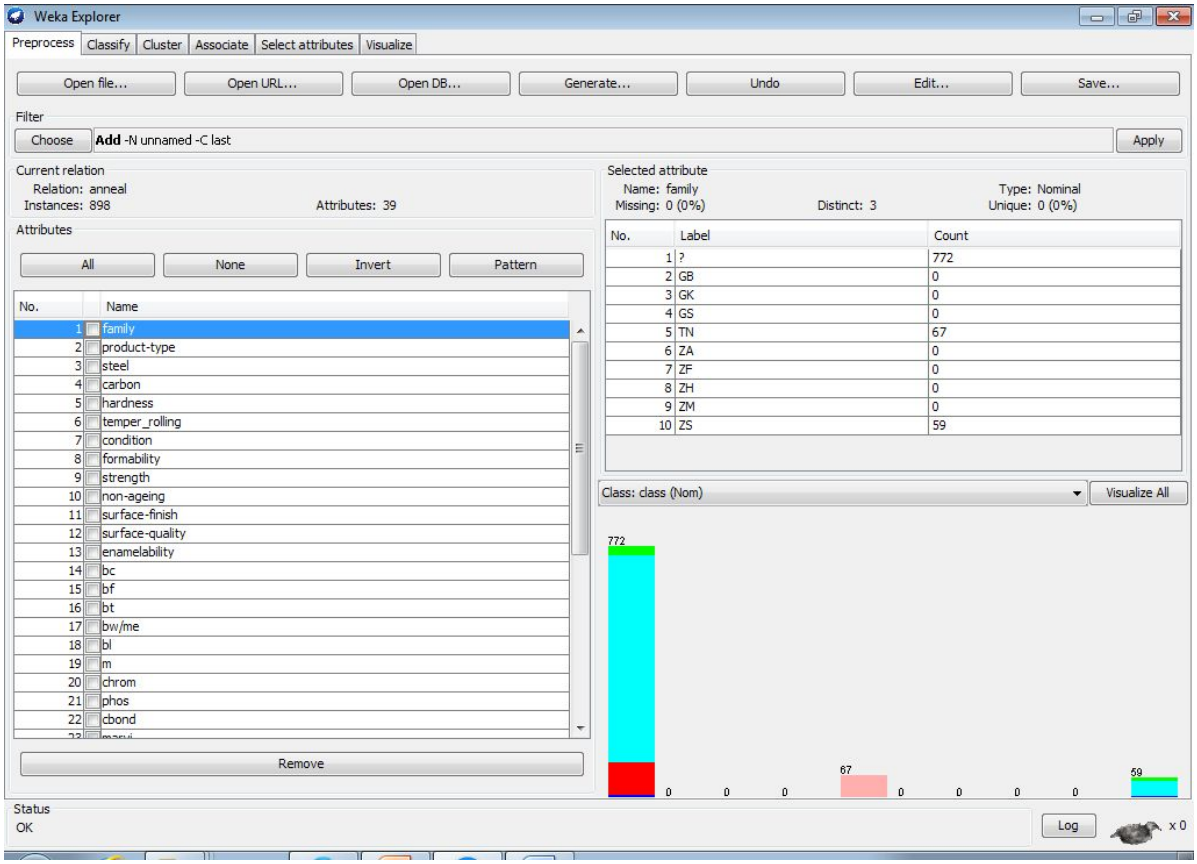
Các bước thực hiện:

Bước 1. Từ tab “preprocess” chọn “open file…” và lựa chọn file “anneal.arff”



Hình 10. Lựa chọn file “anneal.arff” để xây dựng cây quyết định

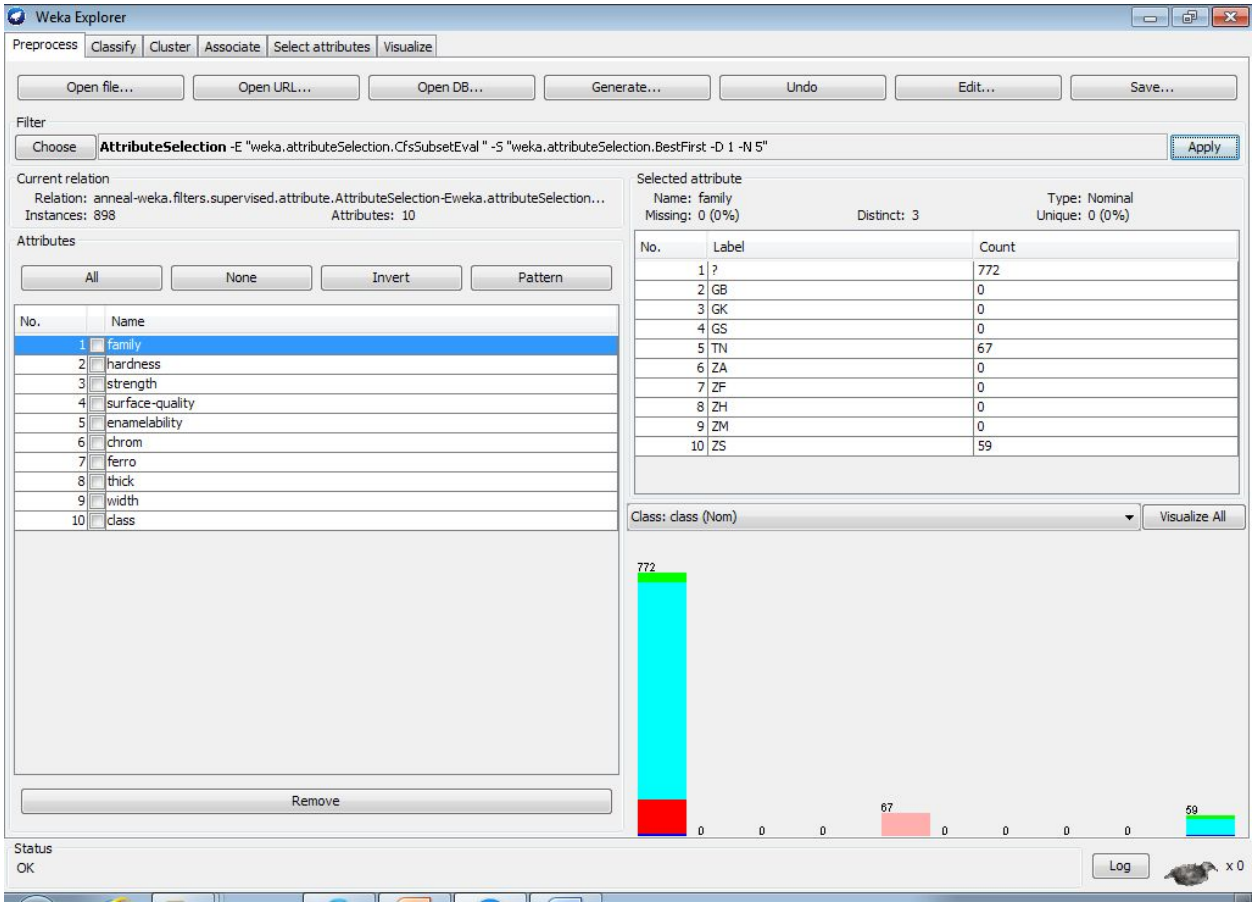
Bước 2. Sau khi đã chọn ở tab “preprocess” sẽ hiện ra các thông tin như sau:



Hình 11. Hiển thông tin về dữ liệu, các thuộc tính của dữ liệu và thống kê về các thuộc tính

- Tên dữ liệu: anneal
- Số lượng đối tượng: 898
- Số lượng thuộc tính: 39
- Lựa chọn các thuộc tính sẽ hiện thị thông về các thuộc tính. Ví dụ như ở hình 11, lựa chọn thuộc tính “family”, các thông tin về thuộc tính này bao gồm: không có dữ liệu lỗi, có ba giá trị trong 10 giá trị của thuộc tính tồn tại trong dữ liệu. Tuy nhiên, giá trị “?” của thuộc tính này thực sự là missing value, nhưng do kiểu dữ liệu là Nominal nên giá trị được xem như giá trị đúng. Ngoài ra, tần suất xuất hiện của giá trị này là rất nhiều (772/898). Từ đó, chúng ta có thể loại bỏ thuộc tính này ra khỏi dữ liệu nhằm mục đích xây dựng mô hình học tốt hơn.

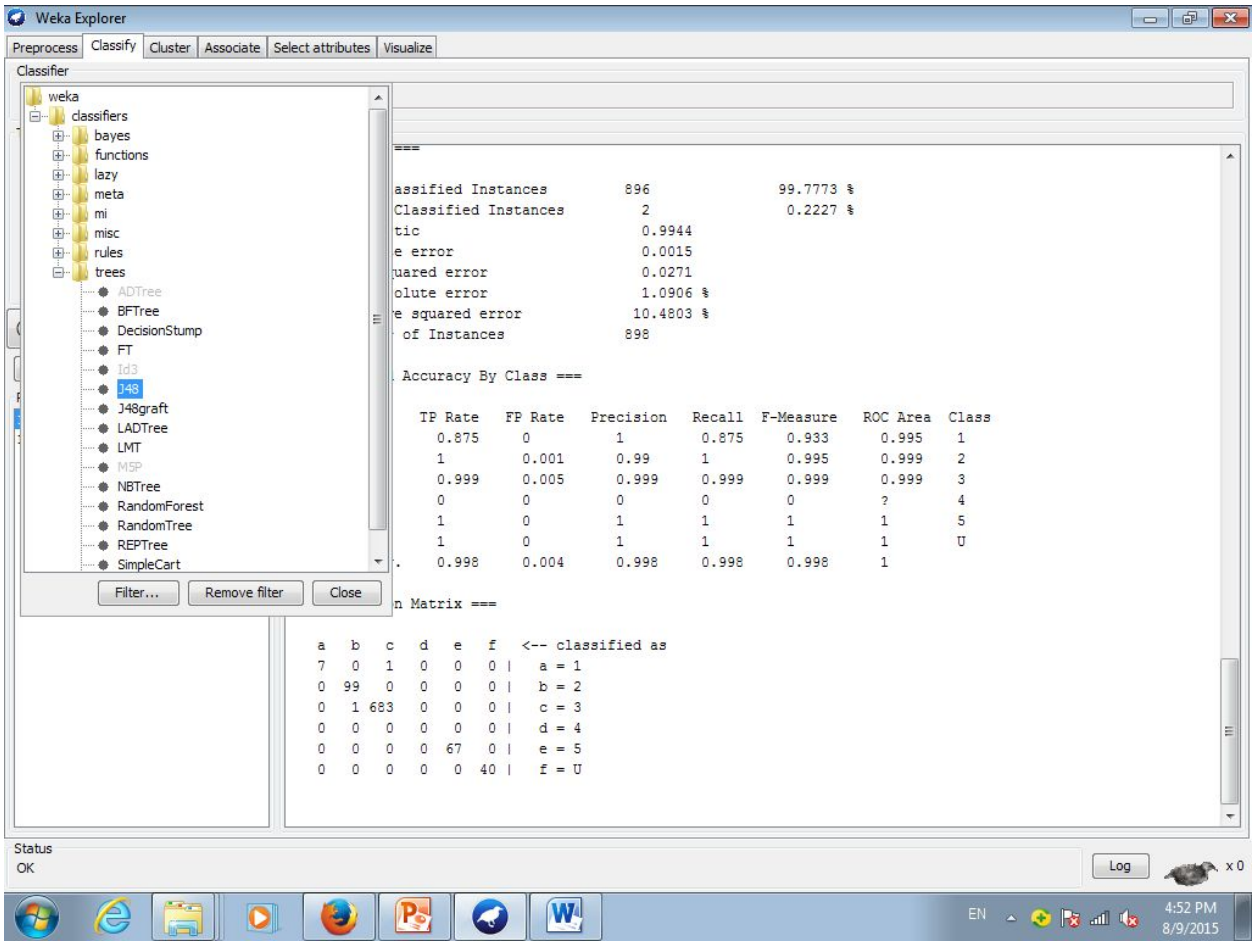
Bước 3. Lọc dữ liệu: vì là dữ liệu có giám sát (dữ liệu có nhãn) nên chọn các bộ lọc có sử dụng nhãn trong quá trình lọc. Ví dụ, chọn bộ lọc “AttributeSelection”. Kết quả thu được như sau:



Hình 12. Thông tin về dữ liệu sau khi lọc

Sau khi lọc, dữ liệu chỉ còn lại 10 thuộc tính có mối tương quan lớn tới nhãn class.

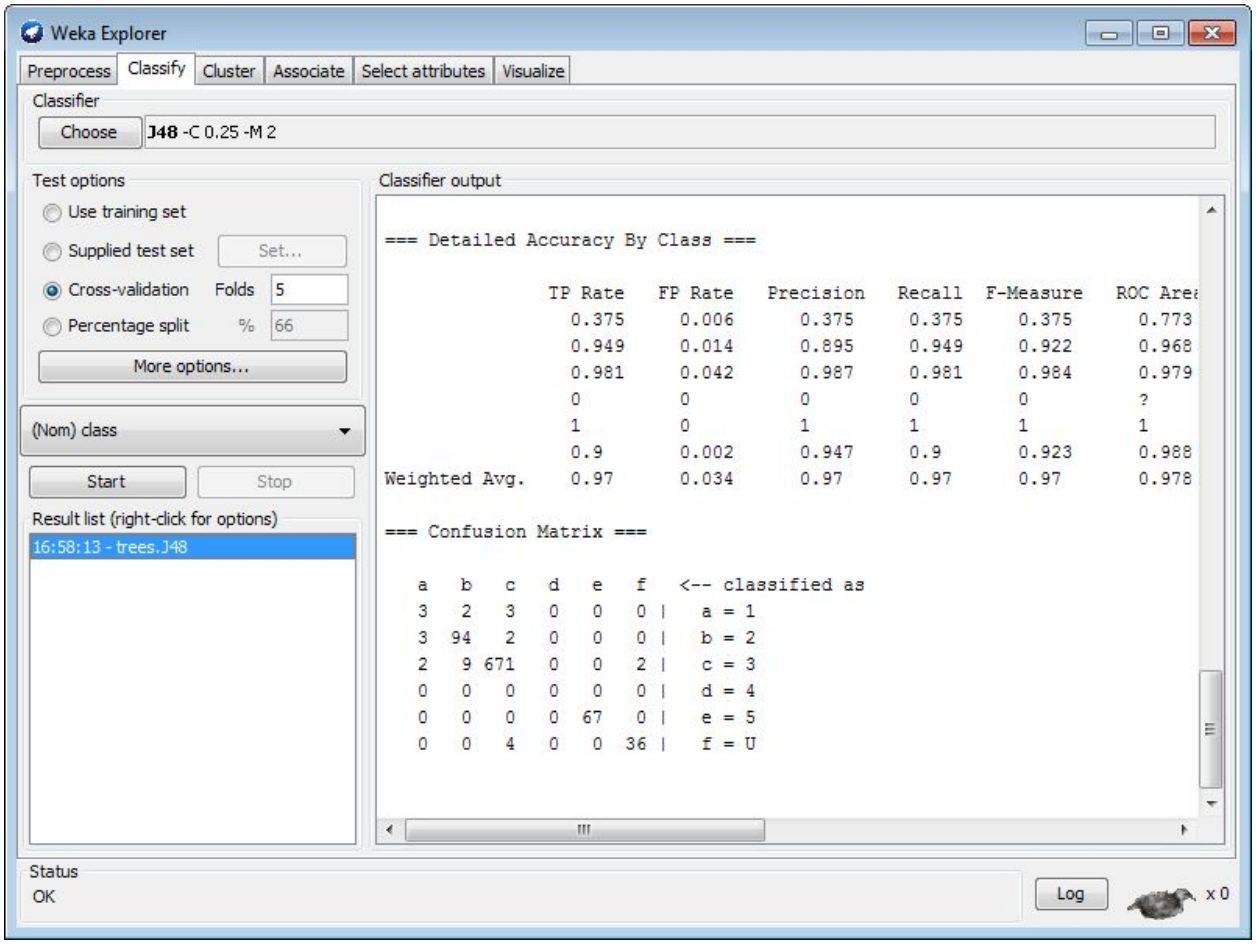
Bước 4. Từ tab “classify”, trong hộp thoại “classifier”, chọn Choose sẽ hiển thị một cửa số chứa các bộ phân lớp như sau:



Hình 13. Các bộ phân lớp

Chọn thuật toán J48 để xây dựng cây quyết định. Từ đây có thể thiết lập các tham số: Confidence Factor, và minNumObj. Trong ví dụ này được để mặc định cho Confidence Factor, và minNumObj lần lượt là 0.25 và 2.

Bước 5. Lựa chọn cách học và kiểm tra mô hình: Chọn cross validation với 5 folds. Sau khi chạy để học mô hình bằng cách chọn start, kết quả được hiển thị như sau:



Hình 14. Kết quả sau khi học mô hình và kiểm định chéo với 5 folds.

Bước 6. Phân tích và đánh giá kết quả:

- Kết quả hiển thị thông tin cơ bản về dữ liệu, phương pháp sử dụng để học mô hình là J48. Mô hình được học là một cây có 9 mức, số lượng nút lá là 28, kích thước cây là 44.

- Các thông tin về độ chính xác của mô hình:

Correctly Classified Instances	871	96.9933 %
Incorrectly Classified Instances	27	3.0067 %
Kappa statistic	0.9254	
Mean absolute error	0.0143	
Root mean squared error	0.098	
Relative absolute error	10.5941 %	
Root relative squared error	37.9404 %	
Total Number of Instances	898	

Chúng ta thấy rằng về trung bình thì độ chính xác của mô hình sau 5 lần test là 96.9933%. Có thể thấy rằng phương pháp J48 đã xây dựng được mô hình tốt cho tập dữ liệu này. Ngoài ra, các sai số trung bình, bình phương sai số, sai số tương đối, bình phương sai số tương đối giữa giá trị nhãn được đoán nhận và giá trị nhãn thực là tương đối nhỏ. Điều này chứng tỏ độ chính xác cao của phương pháp khi xây dựng mô hình trong 5 lần kiểm test.

- Phân tích cụ thể hơn về mô hình:

TP Rate FP Rate Precision Recall F-Measure ROC Area Class

0.375	0.006	0.375	0.375	0.375	0.773	1
0.949	0.014	0.895	0.949	0.922	0.968	2
0.981	0.042	0.987	0.981	0.984	0.979	3
0	0	0	0	0	?	4
1	0	1	1	1	1	5
0.9	0.002	0.947	0.9	0.923	0.988	U
Weighted Avg.	0.97	0.034	0.97	0.97	0.97	0.978

Dựa vào bảng trên chúng ta thấy rằng, độ chính xác để mô hình quyết định phân vào class 1 là rất thấp (0.375%). Ngoài ra, do không có dữ liệu được gán nhãn là class 4 nên các độ đo độ chính xác của mô hình khi phân vào class 4 là bằng 0. Dựa vào bảng thống kê này, chúng ta có thể đưa ra một số kết luận như sau: số lượng dữ liệu được gán cho class 1 là 8 đối tượng, không đủ mẫu để học mô hình.

- Confusion matrix

```
  a   b   c   d   e   f   <-- classified as
  3   2   3   0   0   0 |   a = 1
  3  94   2   0   0   0 |   b = 2
  2   9 671   0   0   2 |   c = 3
  0   0   0   0   0   0 |   d = 4
  0   0   0   0  67   0 |   e = 5
  0   0   4   0   0  36 |   f = U
```

Chúng ta thấy rằng, 8 đối tượng được gãn nhãn là 1 được mô hình dự đoán như sau:

- 03 đối tượng được dự đoán là nhãn 1
- 02 đối tượng được dự đoán là nhãn 2
- 3 đối tượng được dự đoán là nhãn 3

Phân tích tương tự cho các đối tượng thuộc các nhãn khác

Bài tập thực hành 2: cho tập dữ liệu auto93.names.arff, sử dụng thuật toán linear regression để xây dựng mô hình dự đoán. Phân tích và đánh giá kết quả của mô hình.