TRƯỜNG ĐẠI HỌC BÁCH KHOA TP. HCM

KHOA KHOA HỌC VÀ KỸ THUẬT MÁY TÍNH

---------- oOo ----------

Báo cáo môn Học Máy Và Ứng Dụng

Tìm Hiểu Cách Sử Dụng Cây Quyết Định (J 4.8) Của Phần Mềm WEKA Để Giải Quyết Bài Toán Chuẩn Đoán Bệnh Ung Thư Vú

GVHD: PGS.TS Đặng Tuấn Anh

Học Viên: Nguyễn Xuân Vĩnh Hưng (1970589)

TP. Hồ Chí Minh, năm 2021

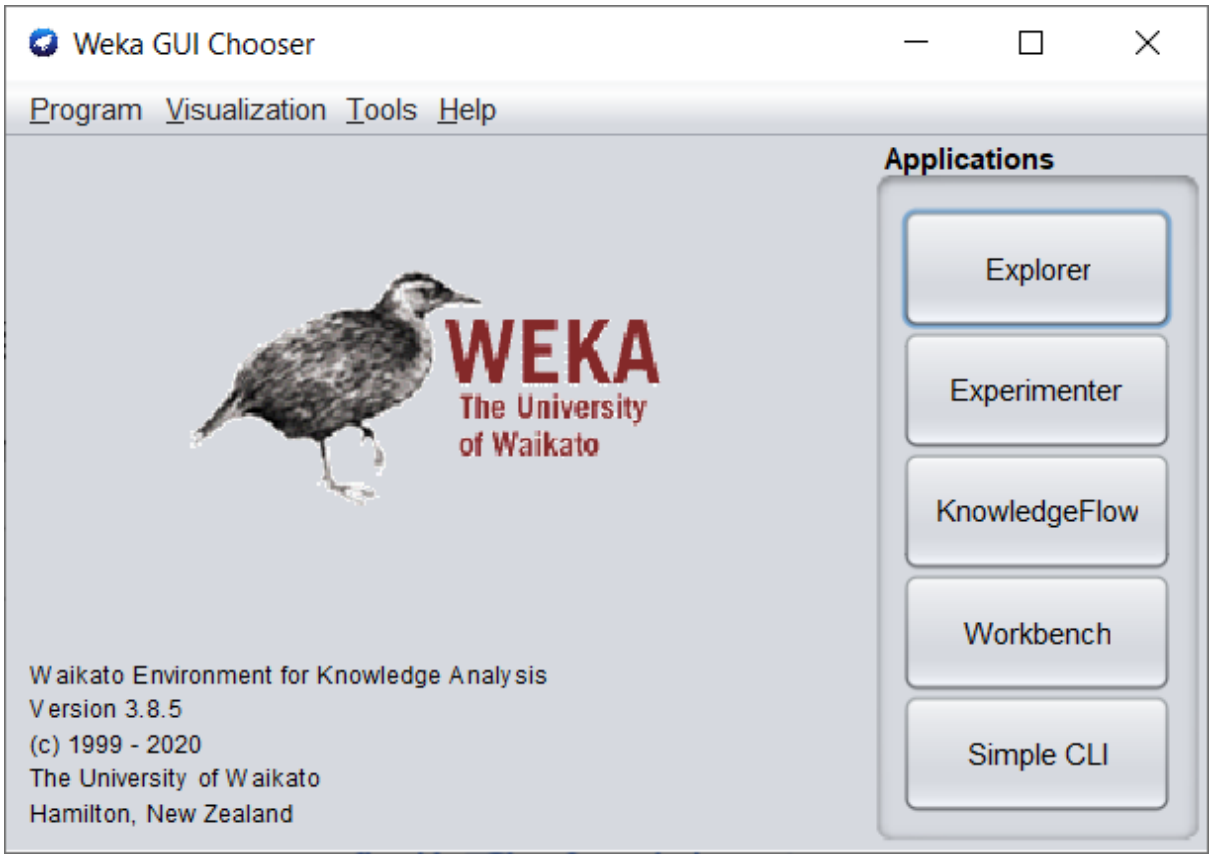
MỤC LỤC

I. Mở đầu

1.1 Giới thiệu Weka

Weka là viết tắt của Waikato Environment for Knowledge Analysis, được phát triển bởi Đại Học Waikato, New Zealand. Nó được phát hành dưới dạng mã nguồn mở và được phát hành theo giấy phép **GNU General Public License**. Weka có nhiều thuật toán Máy Học đã được xây dựng sẵn từ linear regression cho tới neutral networks. Nó cho phép triển khai các thuật toán phức tạp trên những tập dữ liệu chỉ với vài cú click chuột. Ngoài ra, Weka còn có thể tích hợp với Python và R, những ngôn ngữ phổ biến trong Machine Learning.

1.2 Khám phá các chức năng Weka



Hình 1: Giao diện của phần mền Weka

Ưu điểm:

- Có tính tổng quát, bao gồm nhiều công cụ học máy và khai phá dữ liệu thông dụng.
- Là một công cụ học máy được ưa chuộng trong môi trường học thuật.

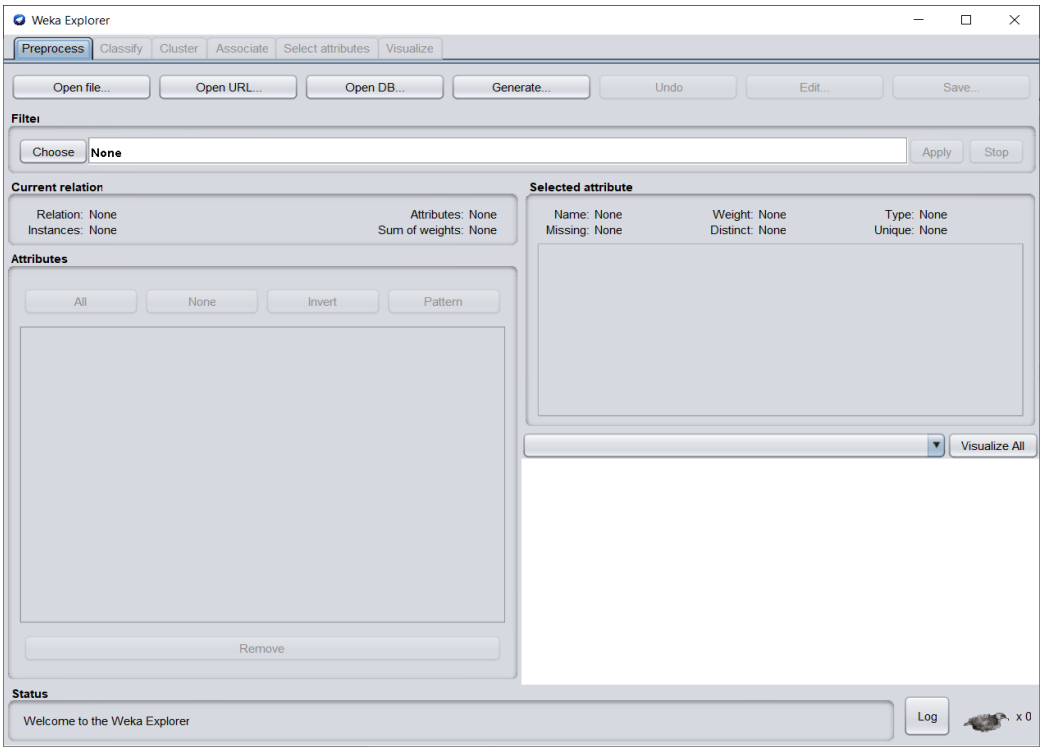
Khuyết điểm:

- Còn hạn chế để làm việc với dữ liệu lớn, khai phá văn bản và học bán giám sát.
- Còn yếu khi xử lý dữ liệu chuỗi thời gian.

Trong màn hình chính của Weka có 5 phần:

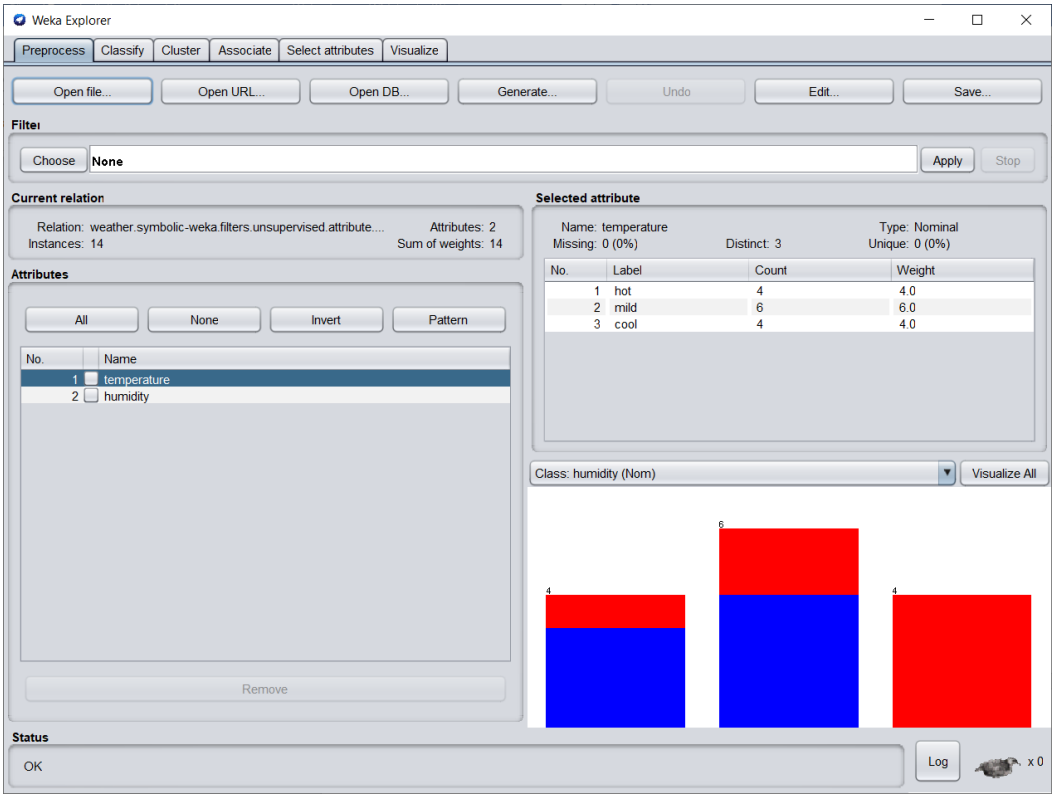
1.2.1 Chức năng của Weka Explorer

- Được sử dụng cho những bộ dữ liệu vừa và nhỏ.
- Được chia làm 6 tab, mỗi tab có một chức năng riêng.



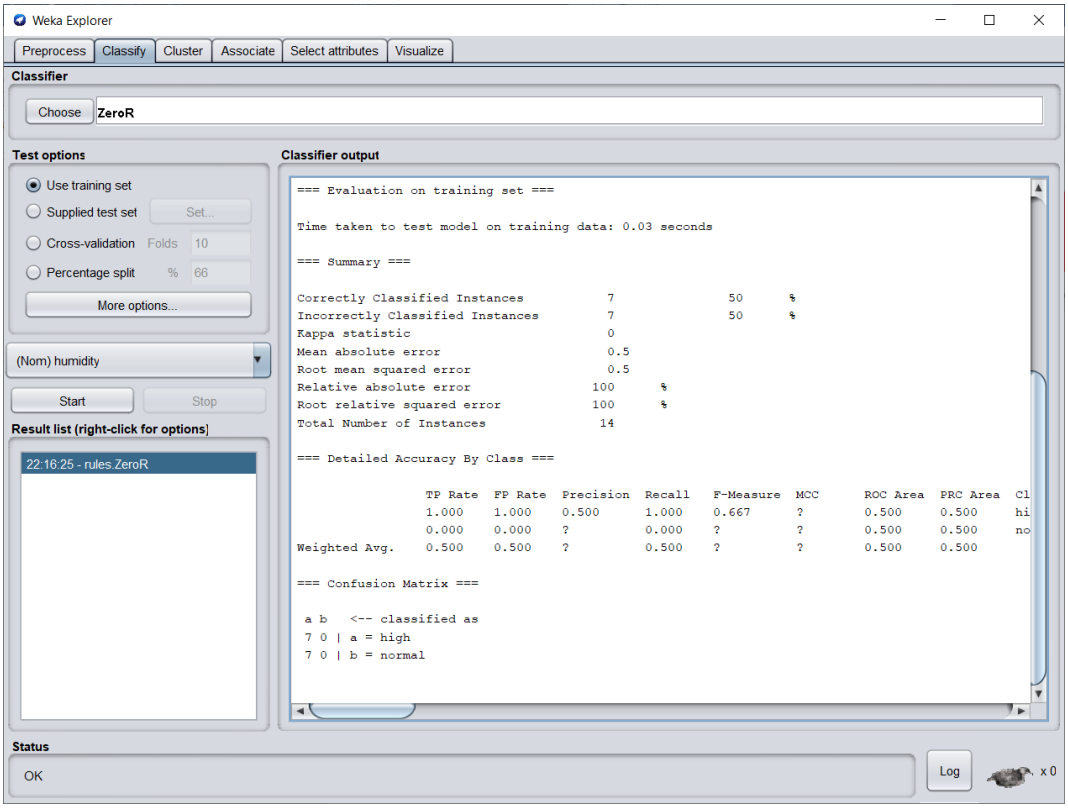
Hình 2: Giao diện Weka Explorer

- Preprocess: Cho phép mở và chỉnh sửa, lọc, lưu lại tập dữ liệu, và chứa cái chức năng tiền xử lý dữ liệu.



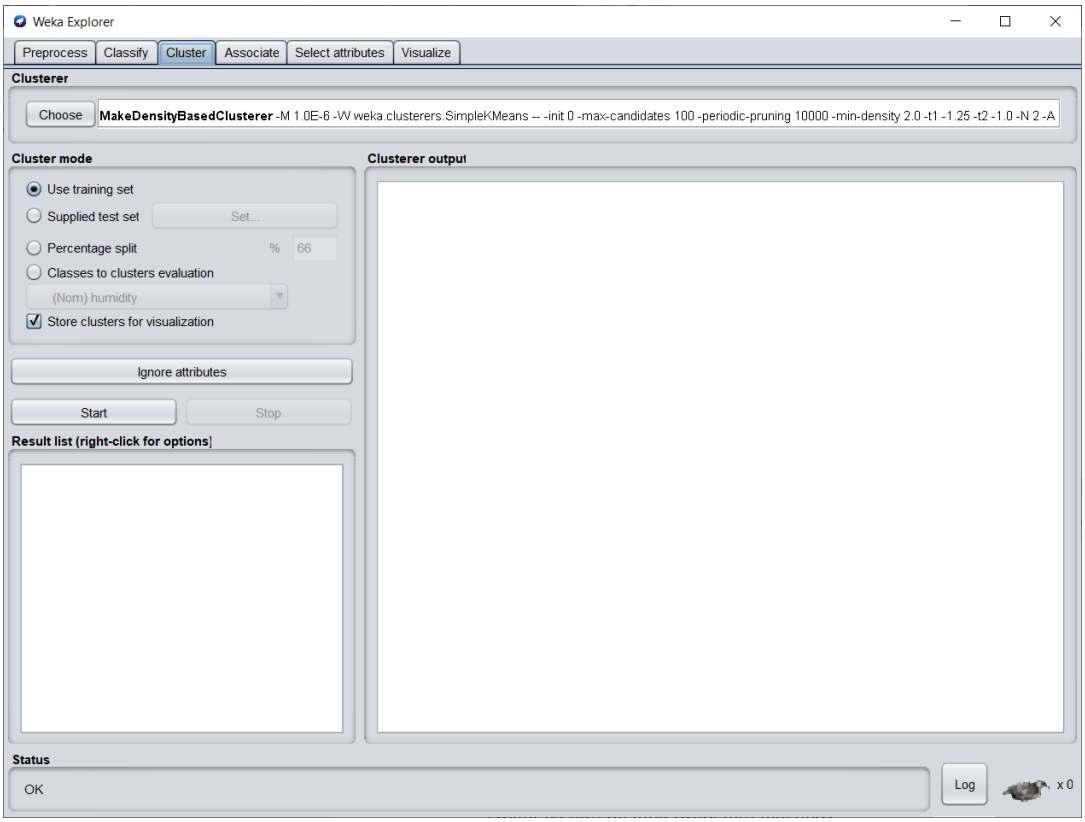
Hình 3: Giao diện tab Preprocess

- Classify: Cho phép người dùng sử dụng và đánh giá hiệu năng các thuật toán phân lớp như ZeroR, Bayes, Random Forest...



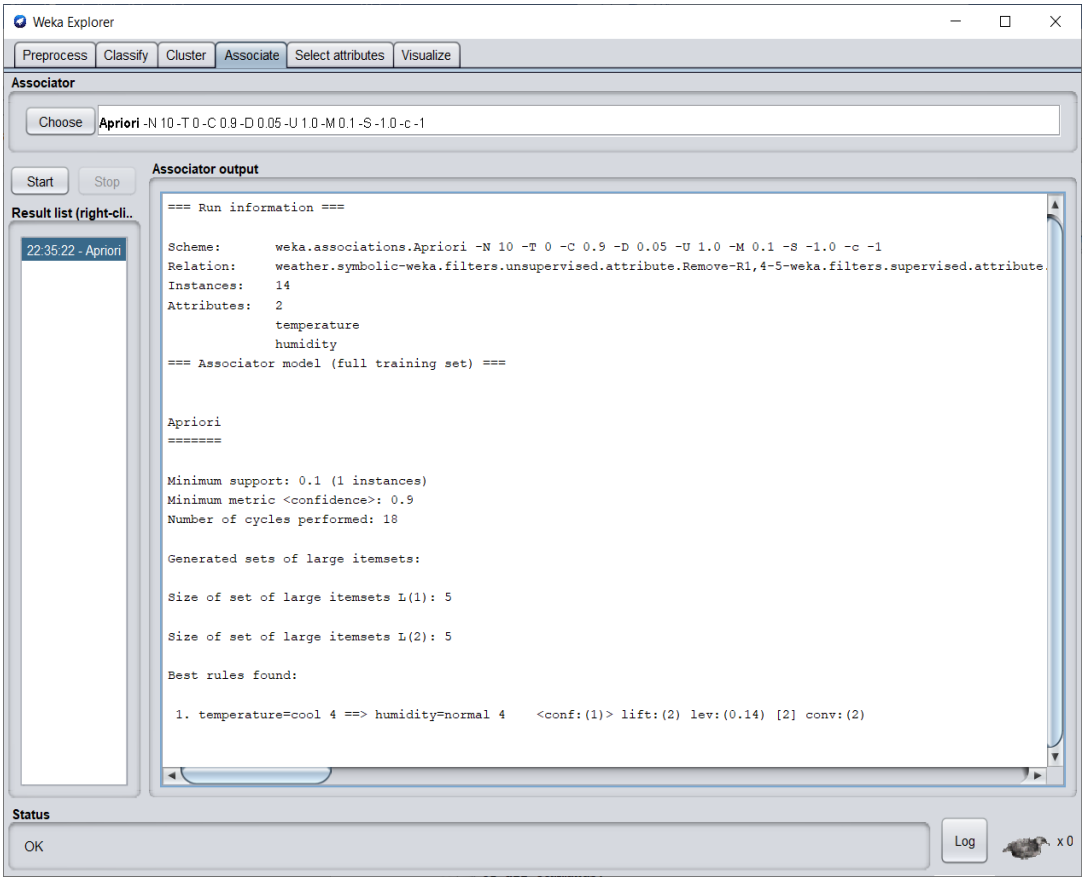
Hình 4: Giao Diện của tab Classify

- Cluster: Cho phép người dùng sử dụng và đánh giá hiệu năng những thuật toán gom cụm như K-Mean, Expectation Maximization…



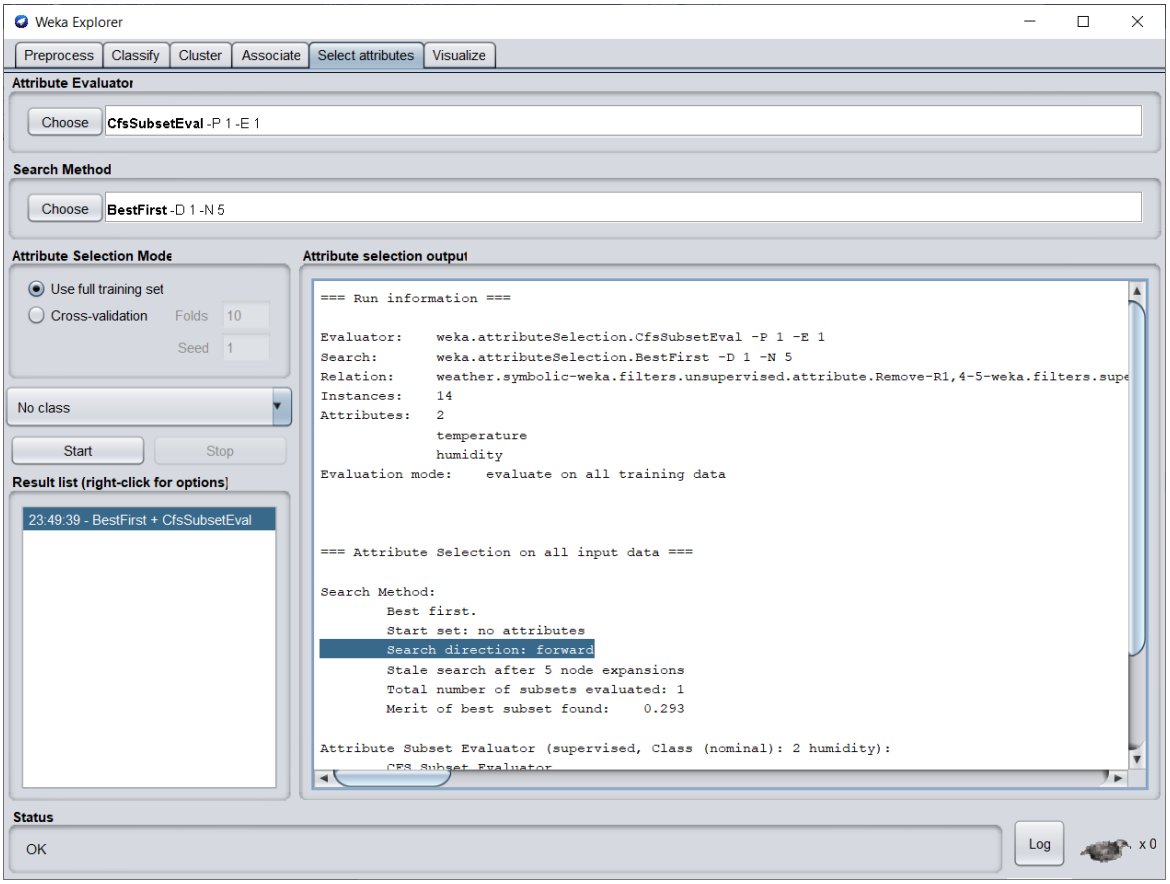
Hình 5: Giao diện tab cluster

- Associate: Chức năng này cho phép tự động tìm ra mối liên kết trong tập dữ liệu. Kỹ thuật này thường được sử dụng cho bài toán Data Minning mà yêu cầu dữ liệu phải được phân loại.



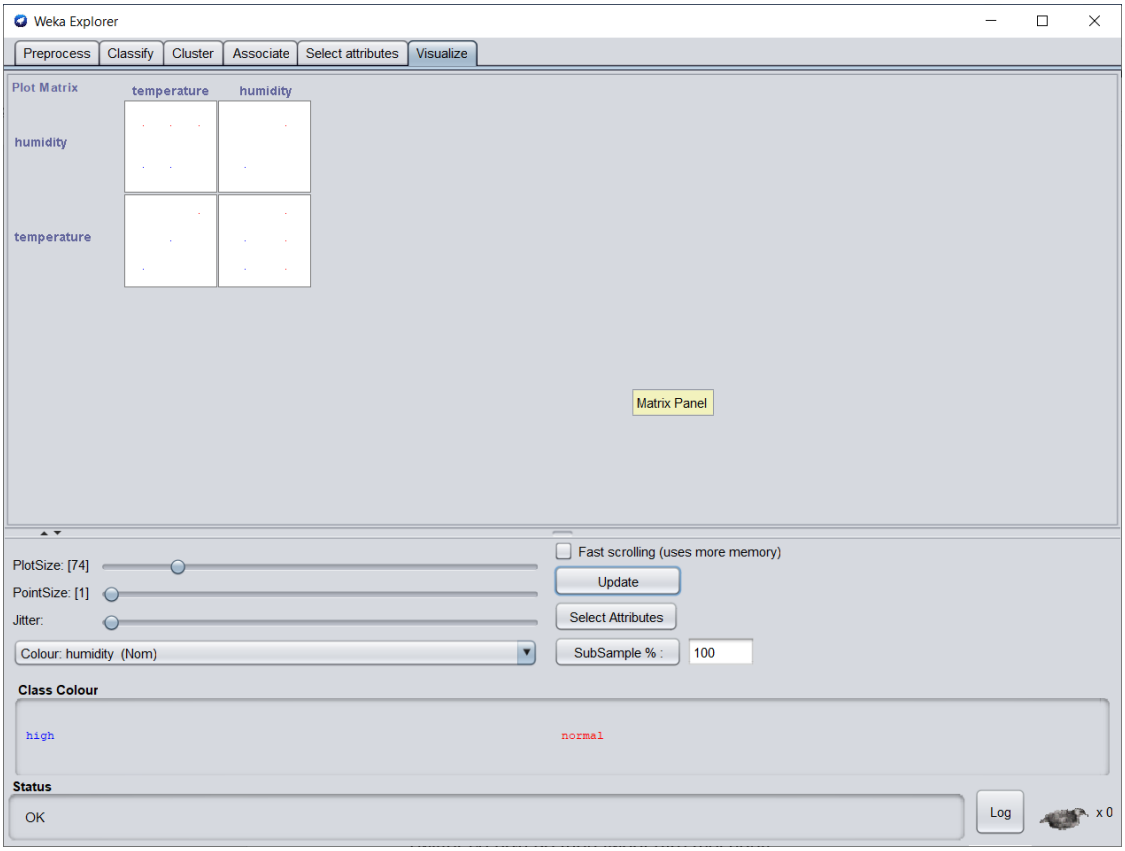
Hình 6: Giao diện của tab Associate

- Select attributes: lựa chọn các thuộc tính thích hợp nhất trong tập dữ liệu



Hình 7: Giao diện tab select attributes

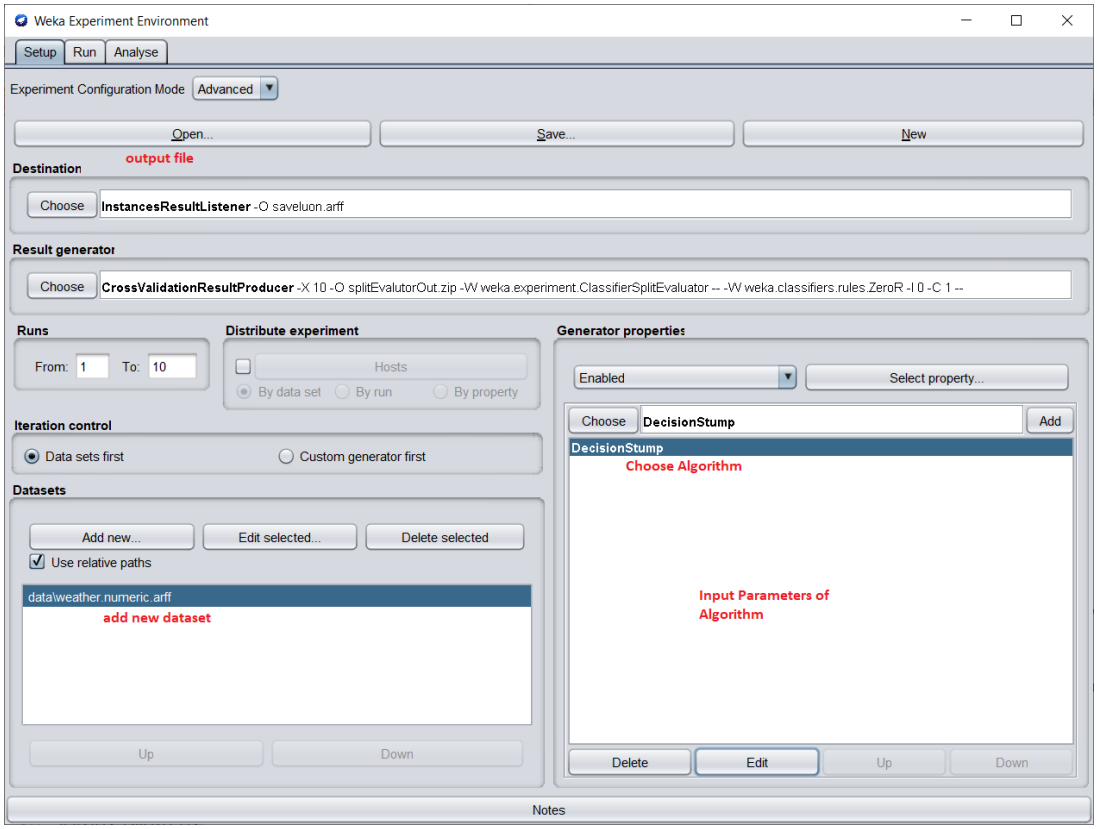
Visualize: Thể hiện dữ liệu dưới dạng biểu đồ. Nó hiển thị ma trận giữa các thuộc tính. Nó hữu dụng để chúng ta có thể so sánh được quan hệ giữa các thuộc tính đó với nhau.



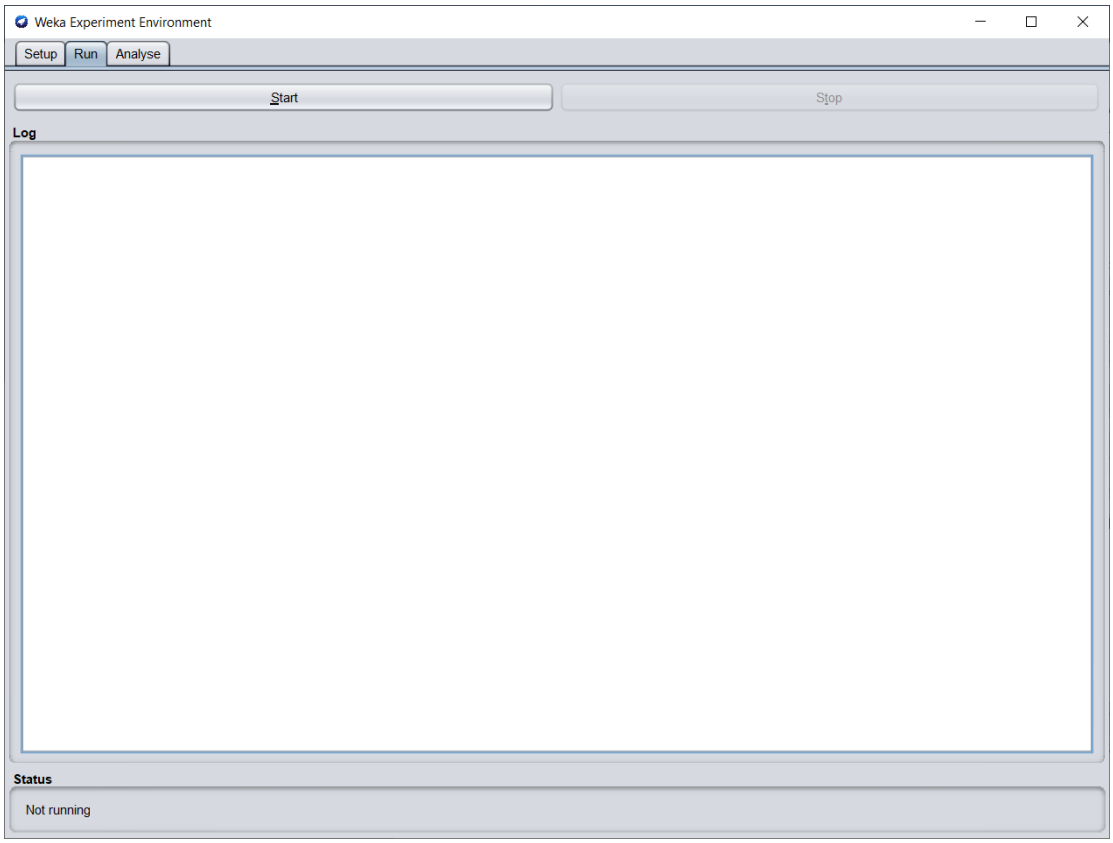
Hình 8: Giao diện tab Visualize

1.2.2 Chức năng của Experimenter

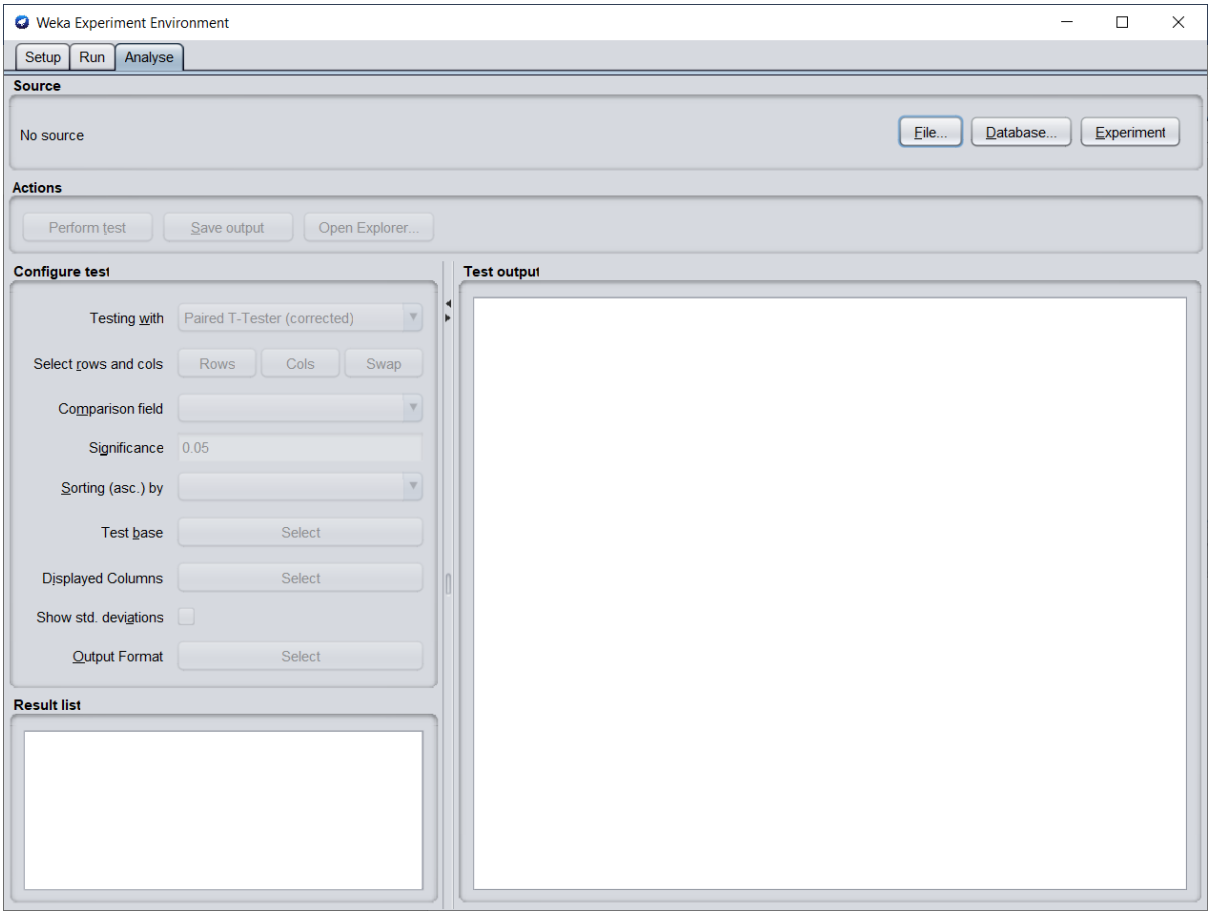
- Experimenter cho phép cài đặt với bộ dữ liệu lớn và phân tích số liệu thống kê sau khi chạy xong.
- Nó có thể cho ta tự động quá trình chạy.
- Số liệu thống kê sau khi phân tích có thể lưu lại dưới dạng định dạng ARFF.
- Cho phép xử lý trên nhiều máy tính qua Java RMI.



Hình 8: Giao diện tab Setup



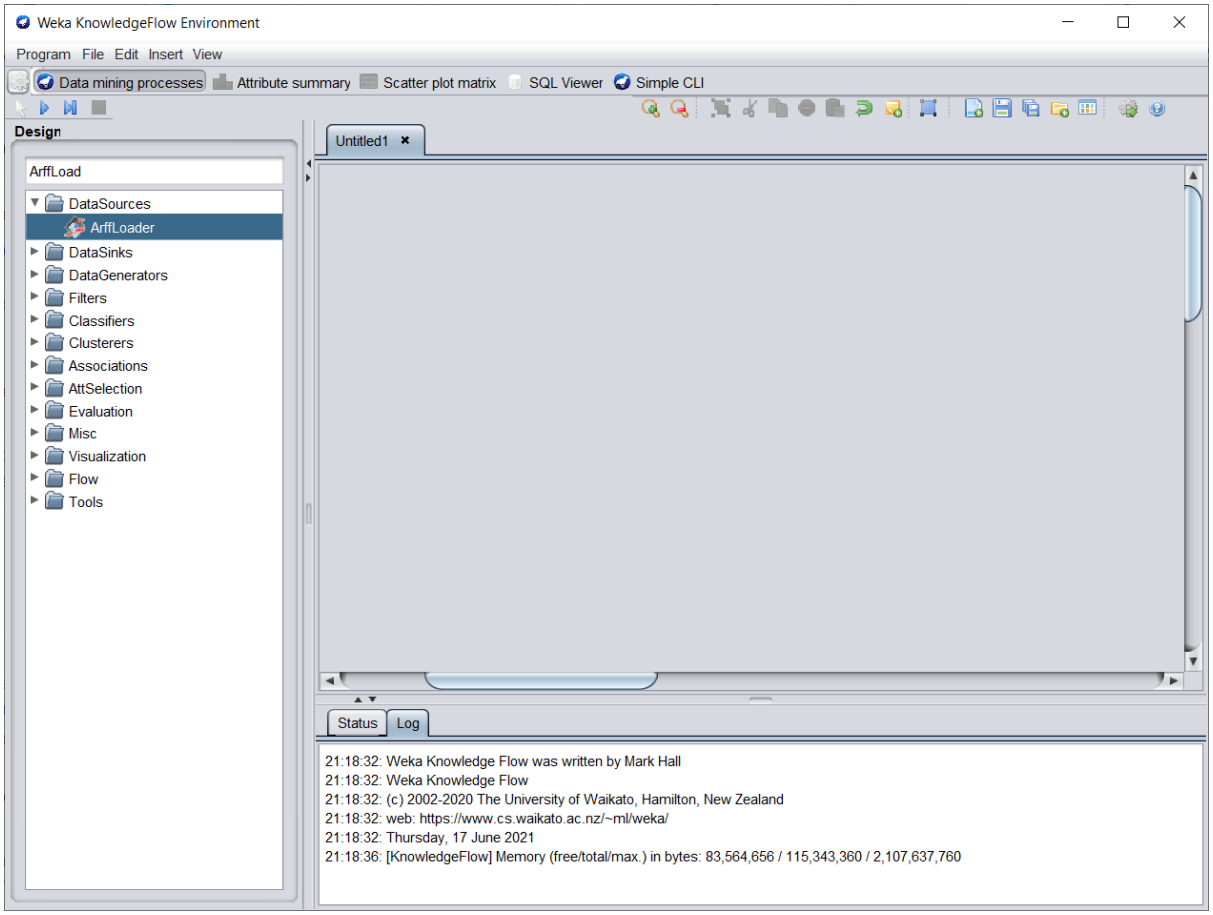
Hình 10: Giao diện tab Setup



Hình 11: Giao diện tab Analyse

1.2.3 Chức năng của Knowledge Flow

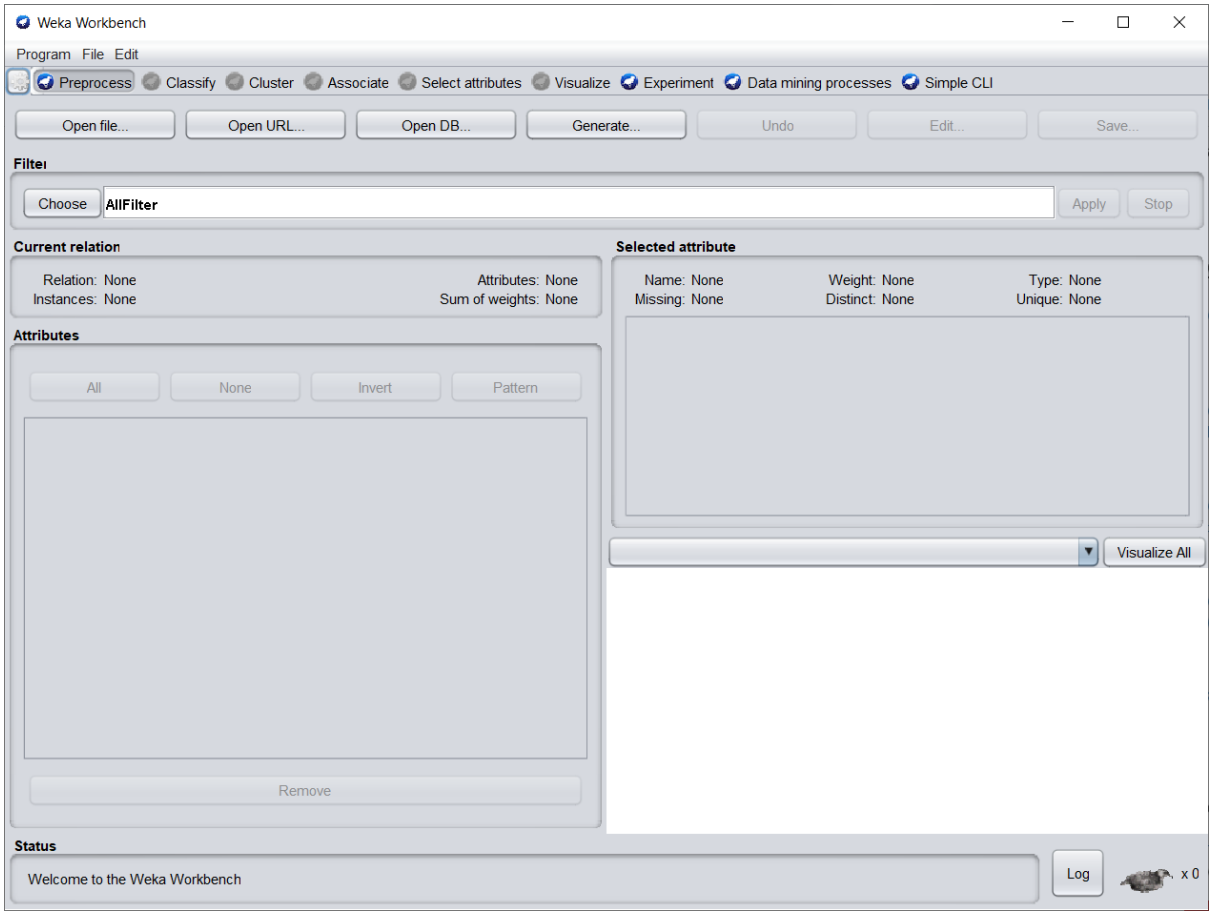
- Cho phép người dùng chọn những Weka components từ toolbar, và đặt nó trong 1 cái canvas và liên kết chúng lại với nhau để xử lý và phân tích dữ liệu theo luồng.
- Xử lý dữ liệu hàng loạt và tăng dần.
- Giúp cho việc trực quan hóa luồn dữ liệu.



Hình 12: Giao diện KnowledgeFlow

1.2.4 Chức năng của Workbench

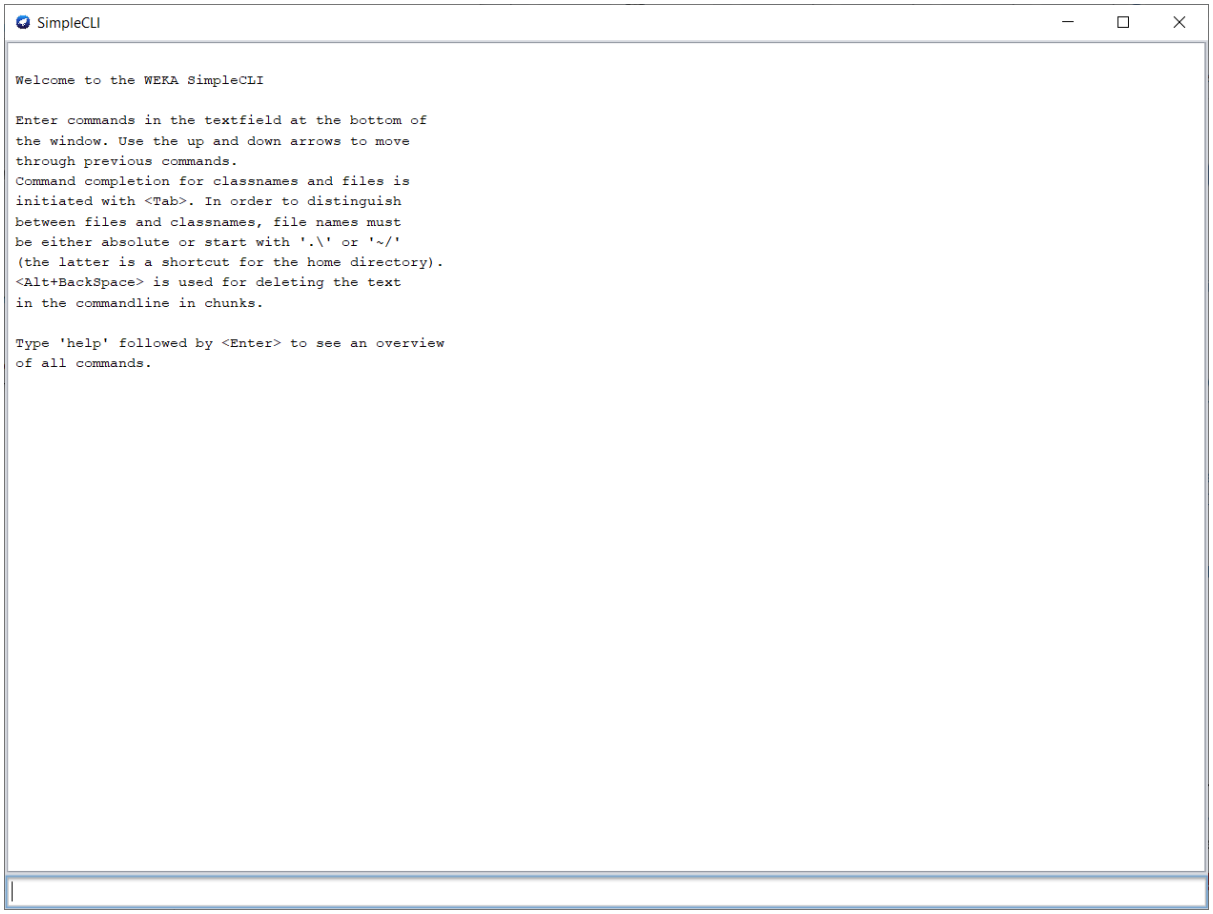
Workbench là môi trường gom tất cái giao diện màn hình vào một giao diện duy nhất. Nó hữu dụng khi bạn cần nhiều giao diện như Explorer and the Experiment.



Hình 13: Giao diện Workbench

1.2.5 Chức năng của SimpleCLI

Weka có thể được sử dụng từ giao diện gõ lệnh. Công cụ này rất mạnh cho viết script để có thể sử dụng toàn bộ api với tham số. cho phép bạn xây dựng model, chạy thử nghiệm và dự đoán mà không cần tới giao diện.



Hình 14: Giao diện SimpleCLI

1.3 Bài Toán

1.3.1 Giới thiệu về bài toán

- Chúng ta sẽ giải quyết bài toán phân lớp (Classification) trên tập dữ liệu Breast Cancer Wisconsin (Diagnostic).

1.3.2 Giới thiệu về tập dữ liệu

- Tập dữ liệu về ung thư vú được đóng góp bởi các tác giả

    1. Tiến sĩ. William H. Wolberg, General Surgery Dept., University of Wisconsin, Clinical Sciences Center, Madison, WI 53792 wolberg@eagle.surgery.wisc.edu.

    2. W. Nick Street, Computer Sciences Dept., University of Wisconsin, 1210 West Dayton St., Madison, WI 53706 street@cs.wisc.edu 608-262-6619.

3. Olvi L. Mangasarian, Computer Sciences Dept., University of Wisconsin, 1210 West Dayton St., Madison, WI 53706 olvi@cs.wisc.edu.

- Thuộc tính của bộ dữ liệu
  - Thuộc tính định danh ID
  - Kết quả chuẩn đoán:
    1. N: âm tính
    2. R: dương tính
  - Có mười thuộc tính được tính toán cho mỗi nhân tế bào:
    1. radius: là bán kính.
    2. texture: độ lệch chuẩn của giá trị màu thang xám.
    3. perimeter: chu vi.
    4. area: diện tích.
    5. smoothness: (local variation in radius lengths).
    6. compactness: được tính bằng công thức perimeter ^ 2 / (area - 1.0).
    7. concavity: Mức độ nghiệm trong của các phần lõm.
    8. concave points: số phần lõm của đường viền.
    9. symmetry: giá trị đối xứng.
    10. fractal dimension ("coastline approximation" - 1).
- Mỗi thuộc tính trên sẽ đi đôi với 3 thuộc tính nữa: mean, standard error, worst. Tổng cộng ta có 33 thuộc tính trong tập dữ liệu.
- Tập dữ liệu có 198 mẫu.

II. Hiện Thực Bài Toán

2.1 Hướng giải quyết bài toán

2.1.1 Giới thiệu về cây quyết định

- Cây quyết định là môt loại cây mà mỗi nút nội bộ tương ứng với với 1 quyết định và nút lá tương ứng với kết quả hay nhãn của lớp.
- Mỗi nút nội bộ sẽ kiểm tra một hay nhiều thuộc tính, để dẫn đến một hay nhiều nhánh. Mỗi nhánh tương ứng với mỗi giá trị của quyết định. Các nhánh hoàn toàn khác biệt.

- Cây quyết định là công cụ tuyệt vời để lựa chọn giữa cách hành động. Chúng cung cấp một cấu trúc hiệu quả cao mà bạn có thể đưa ra các tùy chọn và điều tra kết quả có thể có của việc lựa chọn các tùy chọn này.
- Bộ phân lớp của cây quyết định là học quy nạp.

Bộ dữ liệu “Ung thư vú” trong bài tiểu luận là dữ liệu phi tuyến tính và kết quả trả về là “Dương tính” và “Âm Tính”. Chúng ta sẽ sử dụng cây quyết định (Decision Trees) để phân lớp.

2.2 Hiện Thực Bài Toán

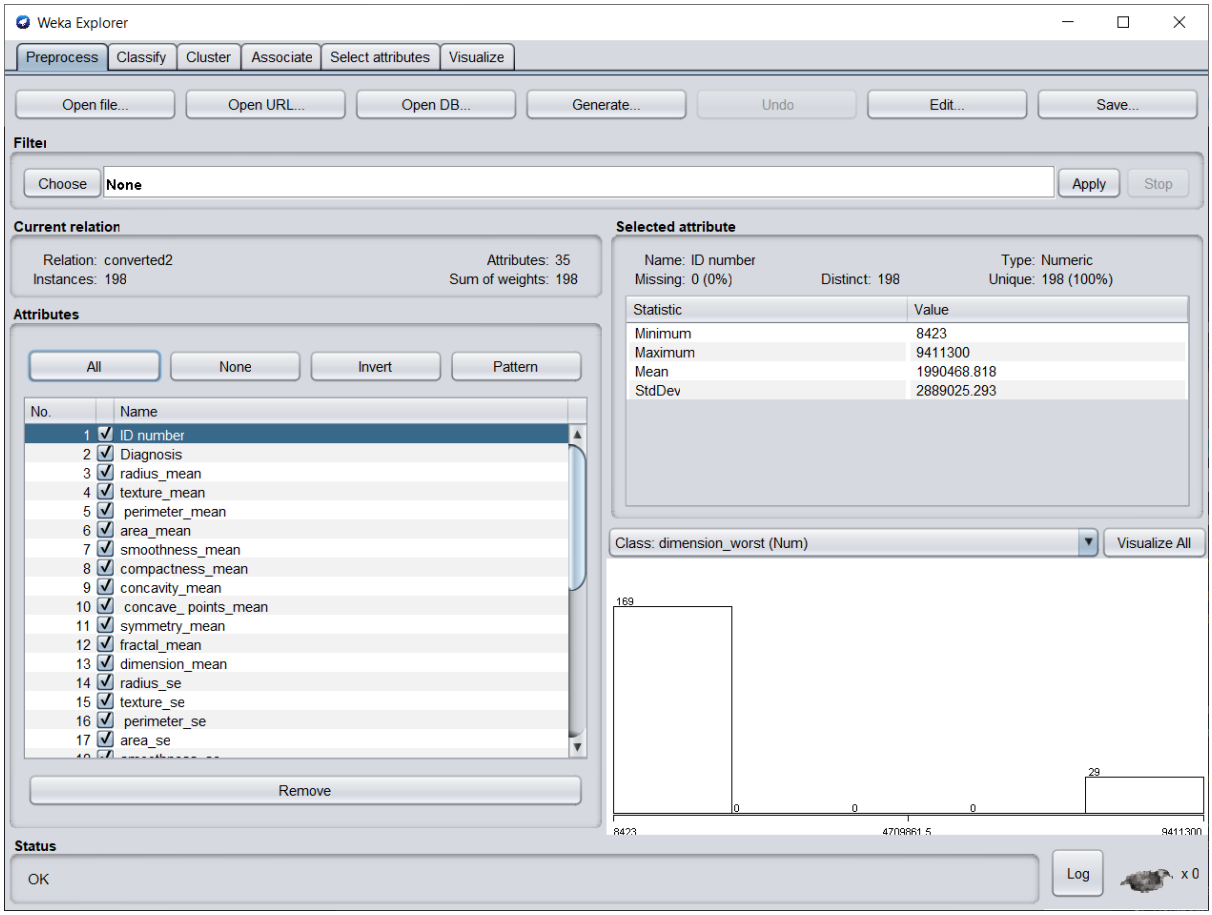
2.2.1 Tiền xử lý dữ liệu

Giữ liệu được cung câp đang ở định dạng C4.5. Chúng ta sẽ chuyển đổi tập dữ liệu file CSV để Weka dễ dàng xử lý.

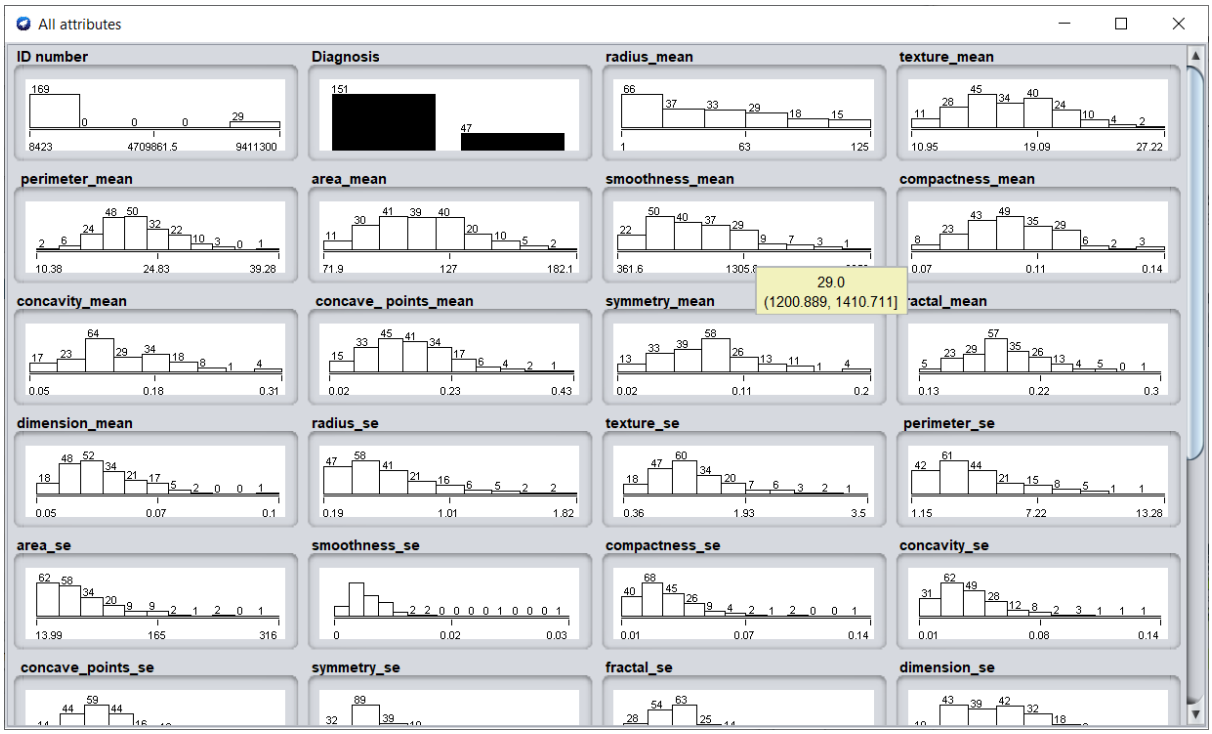
- Tạo 1 file csv.
- Mở file “wdbc.names”, copy toàn bộ nội dụng vào file csv và xuống hàng.
- Mở file “wdbc.data” và copy toàn bộ nồi dung vào file csv. Save lại.

2.2.2 Hiện thực chi tiết

- Mở file csv ở bước tiền xử lý dữ liệu. Chúng ta có thể thấy mốt số thống kế trong tab Preprocess. Ví dụ khi ta chọn vào một thuộc tính, ta có thể thấy bảng thống kê bên tay phải như giá trị lớn nhất, nhỏ nhât, trung bình và độ lệch chuẩn.
- Chọn thuộc tính ID number, và bấm nút Remove.



Hình 15: Sau khi mở file, mọi thuộc tính trong tập dữ liệu sẽ được hiển thị trong ô bên trái ở dưới.

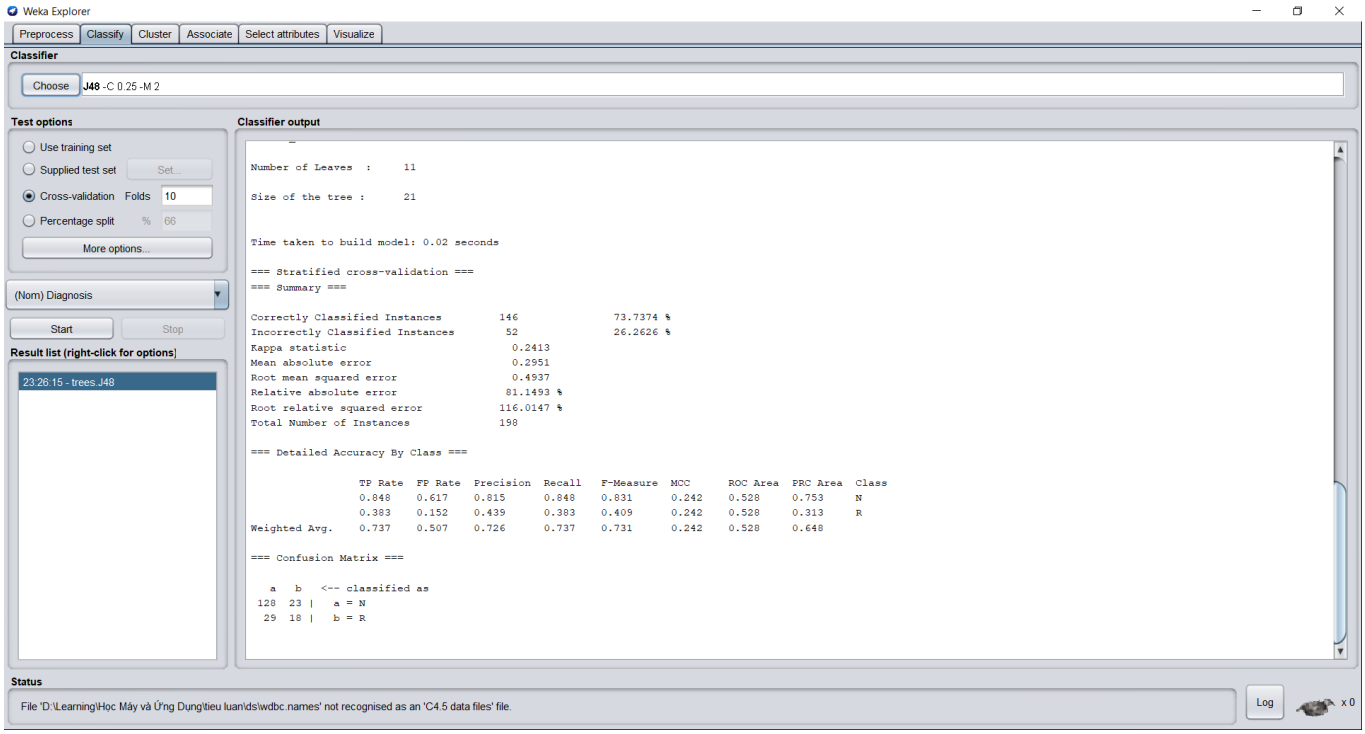


Hình 16: Đồ thị dữ liệu.

- Chuyển qua tab Classify, chọn thuật toán **J48**

> **J48** *là một thuật toán để sinh ra cây quyết định. Nó là một sự thay đổi của thuật toán của C4.5. Nó là thế hệ sau của thuật toán ID3.*

- Ta chọn Cross Validation Folds: 10 nghĩa là chia tập dự liệu thành 10 phần, và mỗi lần chạy sẽ lấy một phần làm dữ liệu test, phần còn lại làm dữ liệu huấn luyện.
- Chọn Norm: Diagnosis



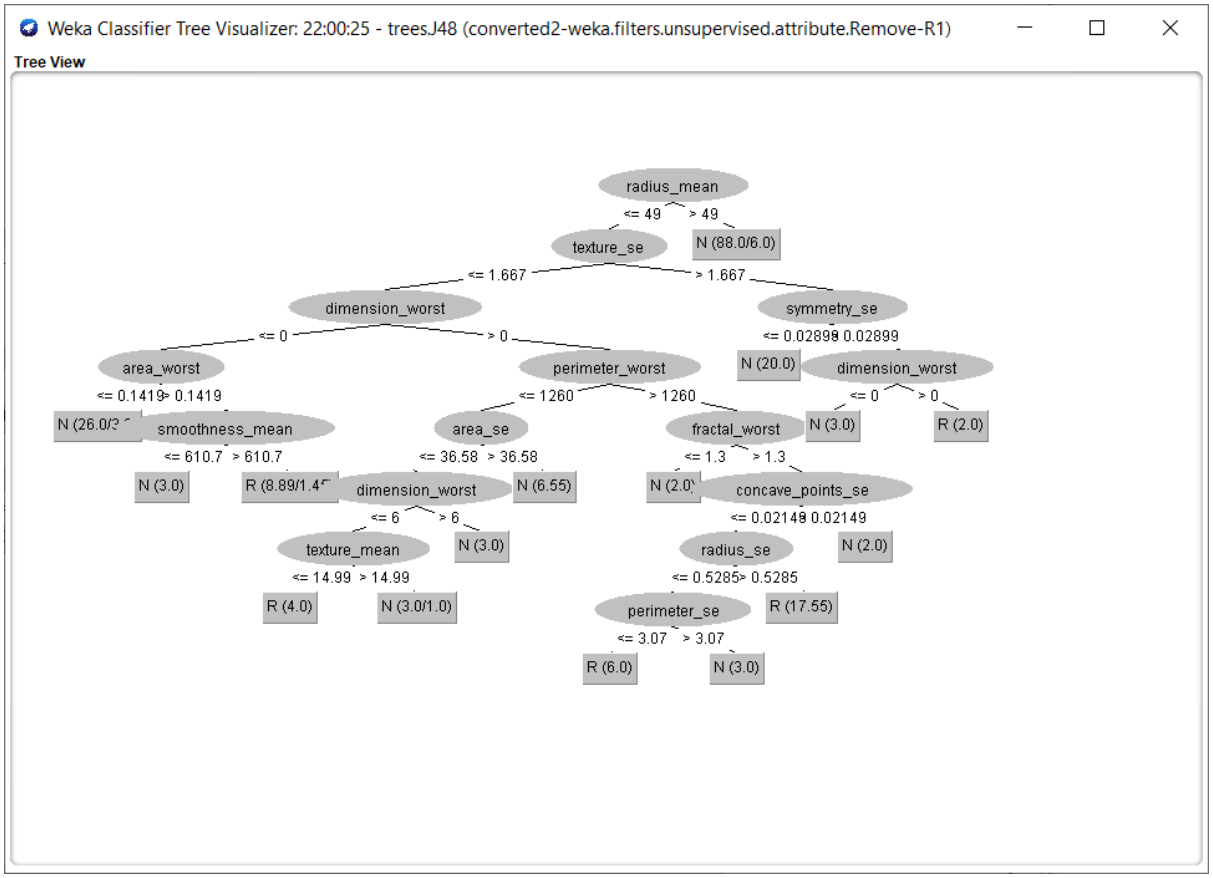
Hình 17: Kết sau khi chạy thuật toán xong

- Ma trận đúng sai:

  a b <-- classified as

  129 22 | a = N

  26 21 | b = R
- Số Mẫu được phân lớp chính xác: 75.7576%
- Số Mẫu được phân lớp sai: 24.2424%



Hình 18: Hình cây quyết định

- Kết quả như ta thấy trên hình. Thuộc tính chính là radius_mean, nếu radius_mean > 49 thì chuẩn đoán là âm tính với bệnh. Nếu radius_mean <= 49, thì phân tích texture_se, nếu texture_se > 1.667, thì xét tiếp symmetry_se <=0.0289, thì kết quả là âm tính…

III. Kết Quả Đề Tài

Qua đề tài này, em đã nghiên cứu những tính năng phần mềm mã nguồn mở Weka. Ngoài ra em cũng tìm hiểu bộ phân lớp cây quyết định, và thuật toán J48 để và áp dụng vào một bài toán thực tế là “Chuẩn Đoán Bệnh Ung Thư Vú”. Trong tương lai, em có thể dung nghiên cứu, áp dụng và so sánh nhiều hơn các thuật toán cây quyết định khác hoặc các thuật toán máy học khác có thể so sánh hiệu năng, độ chính xác để có thể làm tốt hơn giải pháp.

Tài liệu tham khảo

1. PGS Tiến Sĩ Dương Tuấn Anh, Chương 5, Cây Quyết Định, Khoa Khoa Học Máy Tính, Đại Học Bách Khoa TP HCM, 2021.

2. Aniruddha Bhandari, Build a Decision Tree in Minutes using Weka (No Coding Required!), analyticsvidhya.com, 2020.
3. DataminingTools Inc, WEKA: The Experimenter, 2010.
4. Jason Brownlee, A Tour of the Weka Machine Learning Workbench, machinelearningmastery.com, 2016.
5. Dr. William H. Wolberg, W. Nick Street, Olvi L. Mangasarian, Breast Cancer Wisconsin (Diagnostic) Data Set, University of California Irvine, 1995.