Ho Chi Minh University of Science

Information Technology (HQ Program)

**KHAI THÁC DỮ LIỆU VÀ ỨNG DỤNG**

**PROJECT 2 – Classification and Clustering**

**Lớp:** 19KHMT **– Nhóm:** 17

**December 26, 2021**

**PROGRESS:** 100%

**GIẢNG VIÊN**

Lê Hoài Bắc

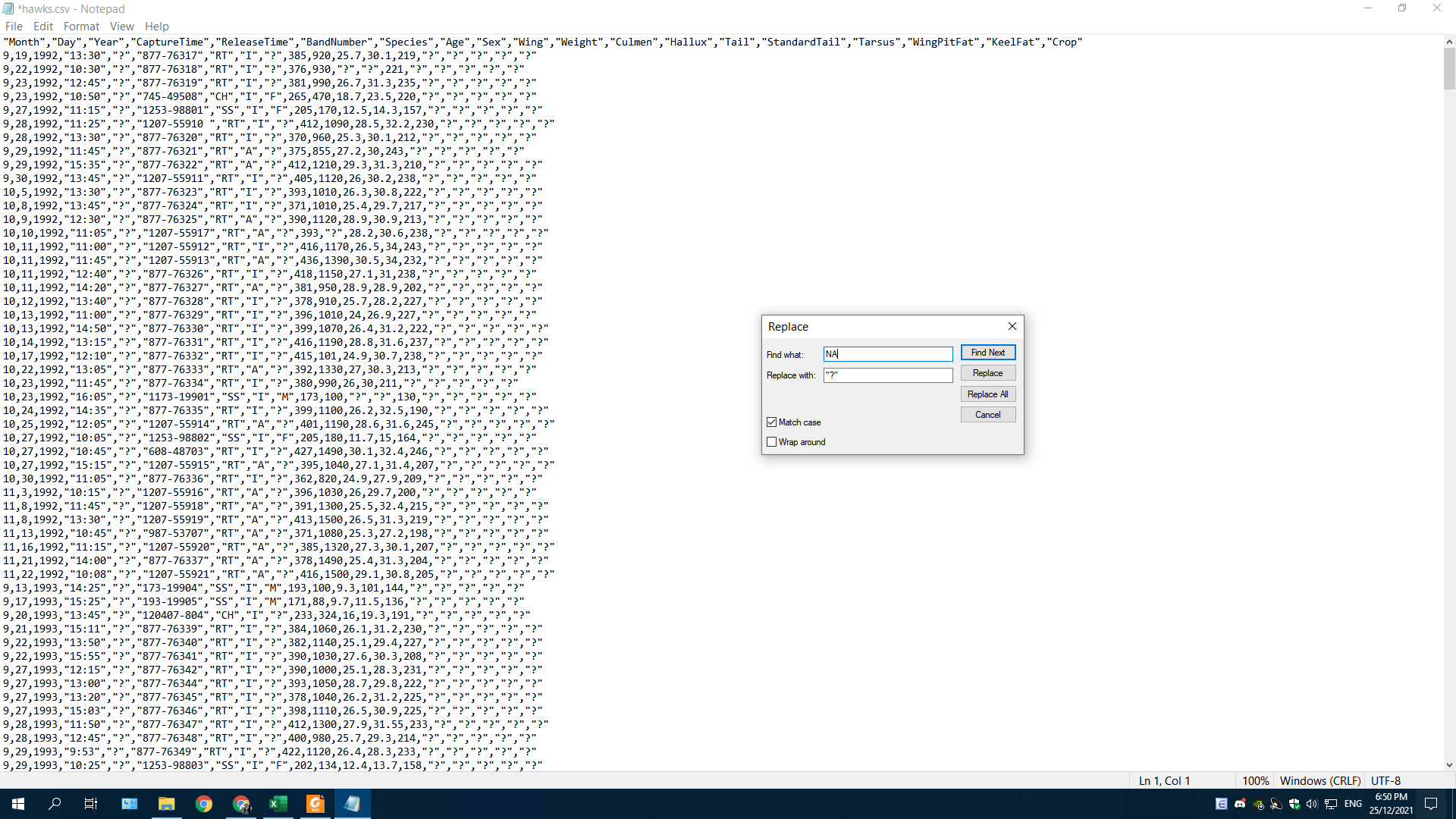
Nguyễn Khánh Toàn

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Project 2** | **Classification** | | | **Clustering** | **Report** | **Percentage** |
| Explorer | Experimenter | Observation |
| **Ngô Văn Anh Kiệt**  19127191 | x |  | x | x | x | 100% |
| **Triệu Nguyên Phát**  19127505 |  | x | x | x | x | 100% |

# Preprocessing hawks dataset with Weka

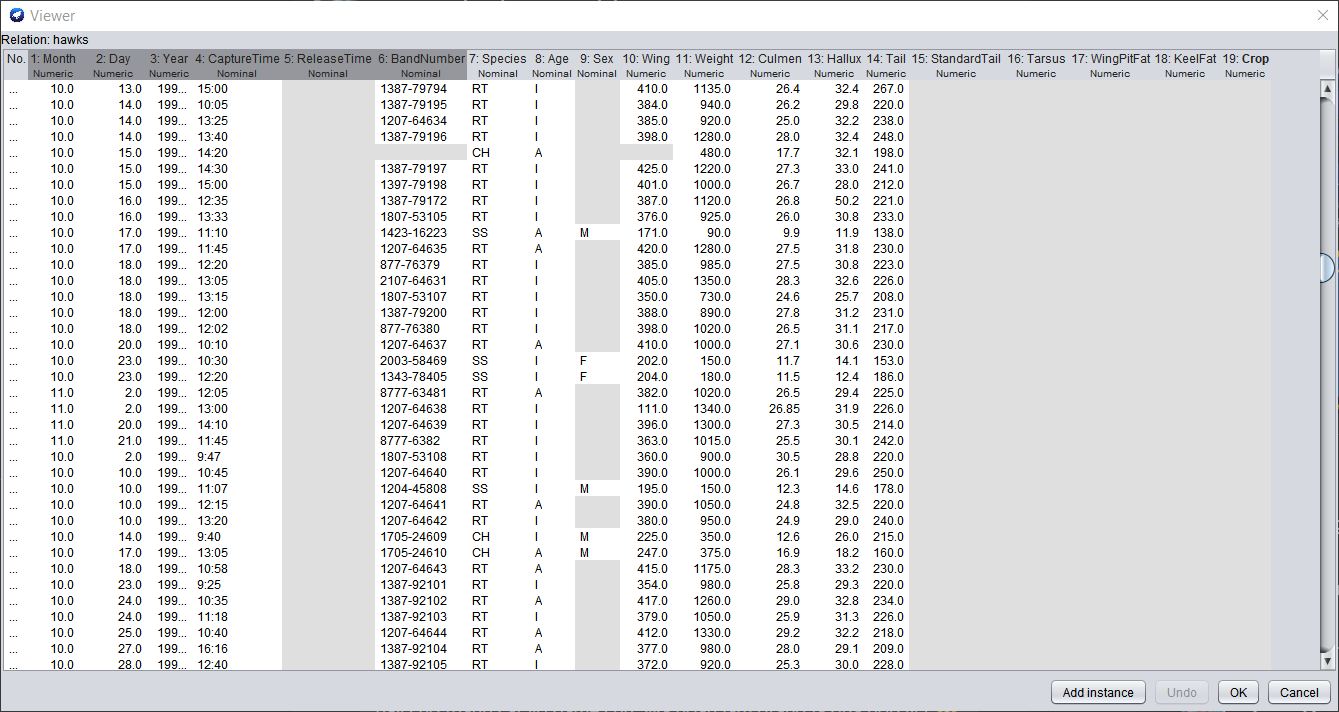
* **Bước 1**: Bởi vì các dữ liệu bị thiếu trong file “hawks.csv” được ghi bằng chuỗi “NA”, “” hoặc “ “ nên ta cần phải chuyển về cùng một định dạng đánh dấu dữ liệu bị thiếu mà Weka có thể đọc được là “?”.

Do file data này tương đối nhẹ, ta chỉ cần bật file csv bằng Notepad (Windows 10) hoặc các trình soạn thảo văn bản tương tự để dùng chức năng Replace All thay đổi tất cả chuỗi không hợp định dạng thành “?”.



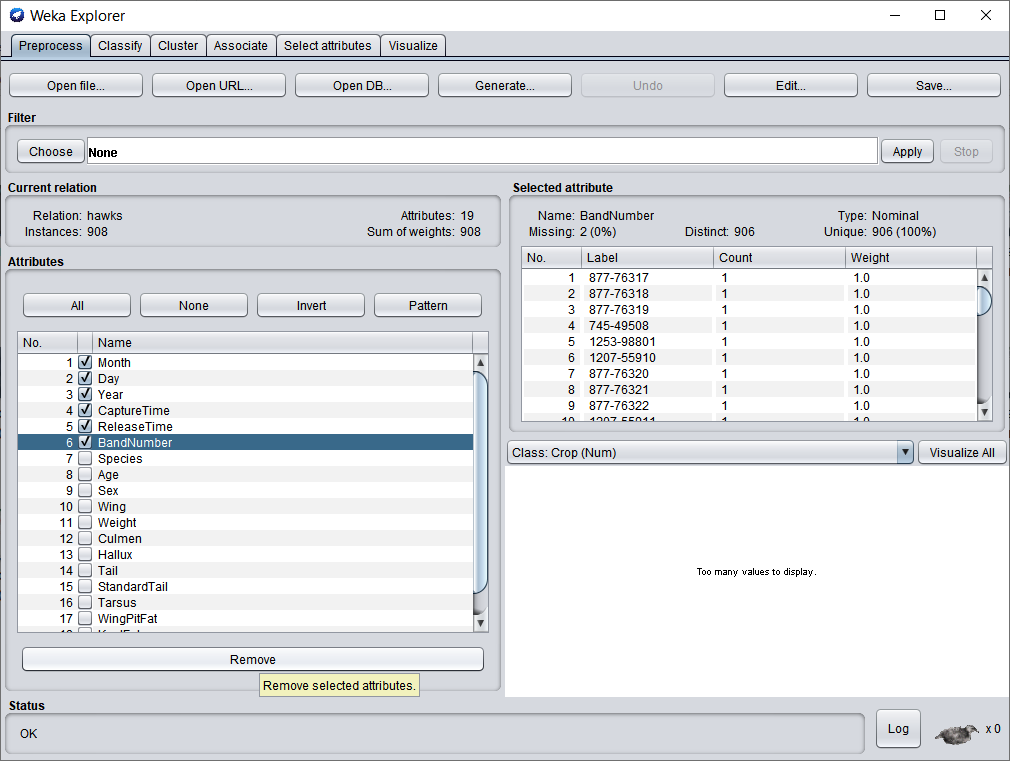
Hình 1. Thay đổi các giá trị bị thiếu từ NA thành "?"

* **Bước 2**: Ta mở file csv bằng Weka Explorer, vào giao diện Edit của tab Preprocessing để kiểm tra lại các cột có đúng kiểu dữ liệu được mô tả trong trang web nguồn của tập dữ liệu hawks.

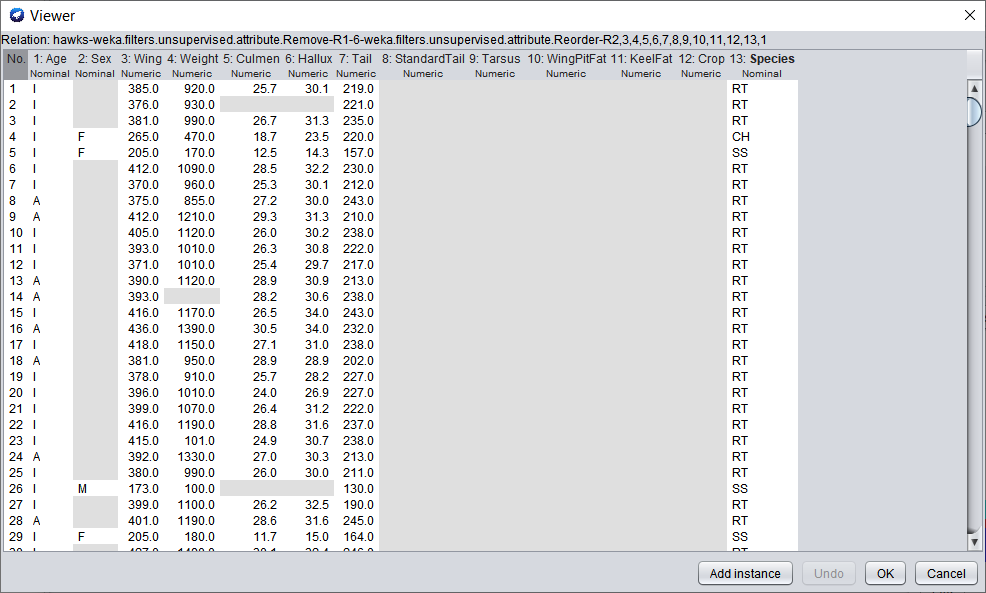


Hình 2. Các thuộc tính đúng kiểu dữ liệu được mô tả và thấy rõ các cột thiếu dữ liệu

* **Bước 3**: Nhận thấy trong các thuộc tính của tập dữ liệu hawks, có một vài thuộc tính không mang nhiều ý nghĩa cho việc phân loại chủng diều hâu, ví dụ như: ngày, tháng, năm ghi dữ liệu, thời điểm bắt, thả chim, mã số của chim. Ta sẽ loại bỏ các thuộc tính đó ra khỏi tập dữ liệu cần dùng, đồng thời đặt thuộc tính “Species” là thuộc tính lớp cần phân loại.

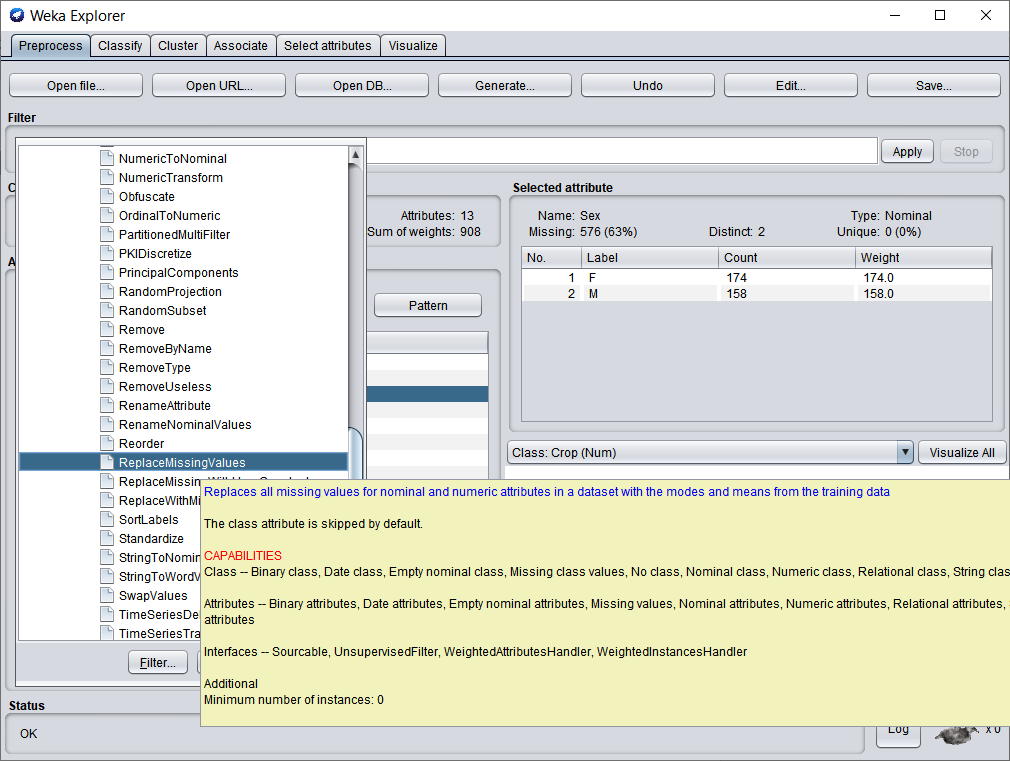


Hình 3. Loại bỏ một số thuộc tính khỏi tập dữ liệu gốc



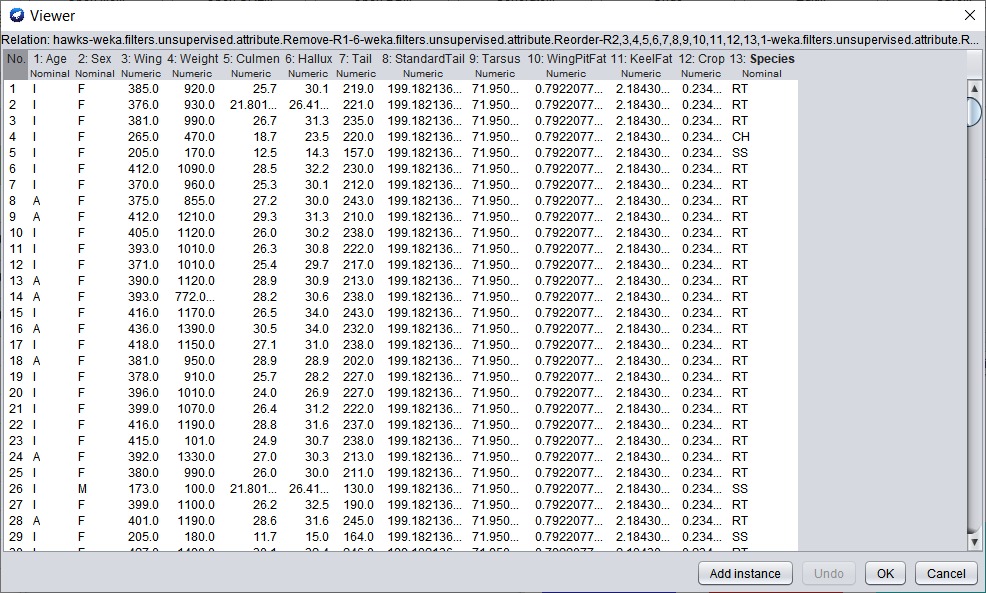
Hình 4. Đặt Species làm class

* **Bước 4**: Bây giờ, ta bắt đầu tiền xử lý các dữ liệu bị thiếu. Sử dụng bộ lọc weka.filters.unsupervised.attribute.ReplaceMissingValues để thay thế các dữ liệu bị thiếu bằng giá trị mean của thuộc tính (nếu là kiểu số) hoặc mode (nếu là kiểu định danh).



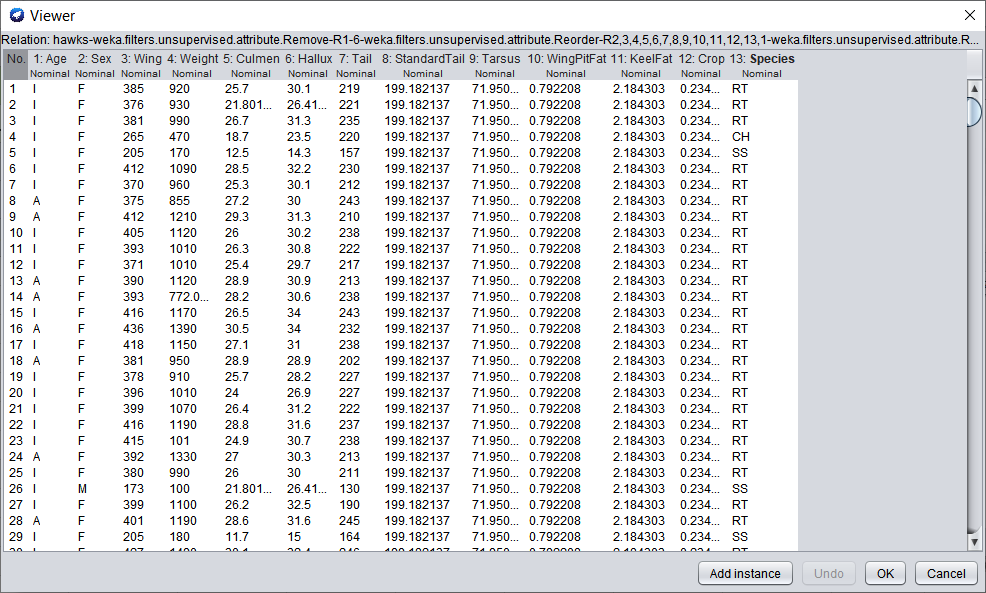
Hình 5. Sử dụng filter ReplaceMissingValues để điền dữ liệu thiếu

Như vậy, ta có bảng dữ liệu sau. Lưu vào file “preprocessed\_hawks.arff”



Hình 6. Tập dữ liệu preprocessed\_hawks.arff

* **Bước 5**: Do trong thí nghiệm A không sử dụng tập dữ liệu kiểu rời rạc, trong khi thuật toán ID3 thì lại không hoạt động được với dữ liệu liên tục, ta dùng bộ lọc weka.filters.unsupervised.attribute.NumericToNominal để tạm biến đổi các thuộc tính số thành thuộc tính định danh để có thể chạy ID3 cho phần này. Lưu tập dữ liệu này thành file “preprocessed\_numeric2nominal\_hawks.arff”. Vì tập này chỉ được dùng tạm để chạy ID3 trong thí nghiệm A, kết quả của nó sẽ được ghi cùng dòng với tập “preprocessed\_hawks.arff”.



Hình 7. Thuộc tính numeric biến đổi thành nominal để chạy ID3 trong thí nghiệm A

# Observation

## Which classification method typically has the best result?

PP phân lớp thường cho kết quả cao nhất là Use training set.

## Which method does not work well and why?

PP phân lớp Percentage split cho kết qủa không tốt, vì khi tách tập dữ liệu theo tỷ lệ  
(thông số) không phù hợp sẽ làm mất đi tính liên quan của dữ liệu, nhiều trường hợp xấu  
có thể làm mất/sai lệch thông tin của một lớp phân loại.  
Hơn nữa, với dữ liệu hawk này chỉ có 907 mẫu, cộng thêm số lượng dữ liệu bị thiếu quá  
nhiều nên pp này cho kết quả không đồng đều.

## Why should we use the discretized version of the data set instead of the original one?

Vì rời rạc hóa các thuộc tính số để các thuộc tính được mô tả đúng như ý nghĩa của nó.

## Do the discretization process and method affect the classification results? If yes then how?

Qua quá trình thực nghiệm bằng cách rời rạc hóa các thuộc tình không phải là lớp (B và C)  
thì nhận thấy việc rời rạc và cách rời rạc hầu như không ảnh hưởng đến kết quả phân lớp.

## Which evaluation strategy tends to overestimate the accuracy and why?

Hai thuật toán NaiveBayesSimple và J48 cho kết quả rất tốt, nhưng suy cho cùng thì J48 tốt  
hơn cả.  
Bởi lẽ, các thuật toán như J48, C4.5 có hiệu quả hơn đối với các dữ liệu Qualititive value  
(ordinal, Binary, nominal).

## Which evaluation strategy tends to underestimate the accuracy and why?

Khi đánh giá bằng thuật toán Id3 ta nhận thấy kết quả chênh lệch ‘rất lớn’ giữa các loại  
thwucj nghiệm.  
Thuật toán ID3 và CART cho hiệu quả phân lớp rất cao đối với các trường dữ liệu số  
(quantitative value).