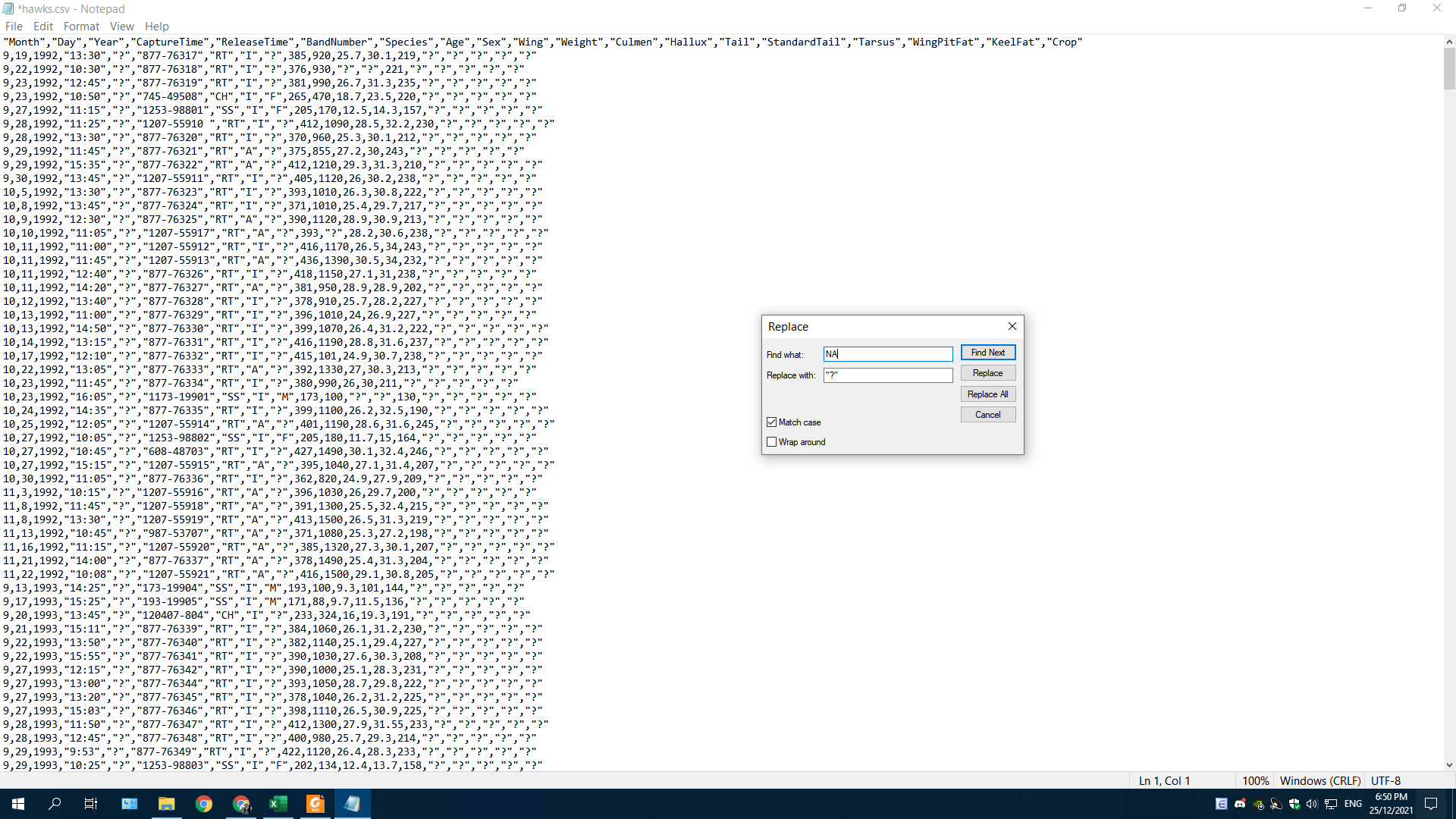
**HAWKS DATASET PREPROCESSING WITH WEKA**

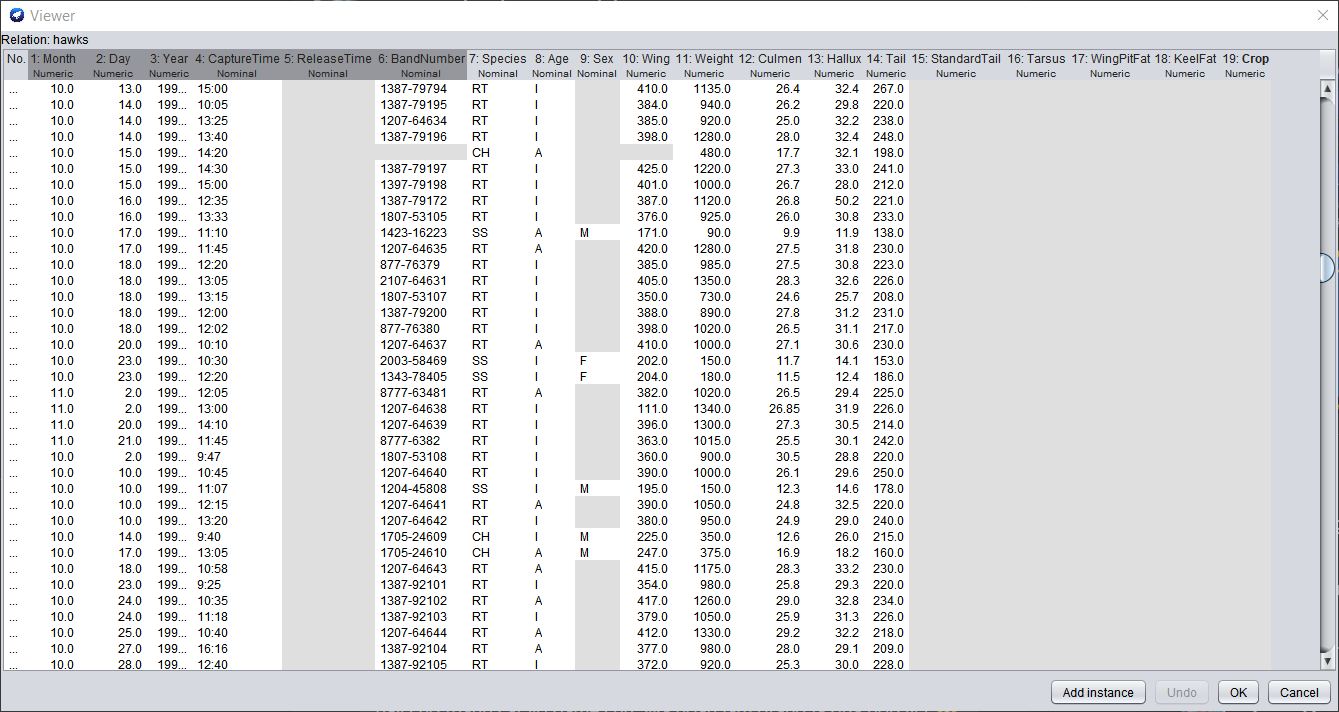
* Bước 1: Bởi vì các dữ liệu bị thiếu trong file “hawks.csv” được ghi bằng chuỗi “NA”, “” hoặc “ “ nên ta cần phải chuyển về cùng một định dạng đánh dấu dữ liệu bị thiếu mà Weka có thể đọc được là “?”.

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 “?”.



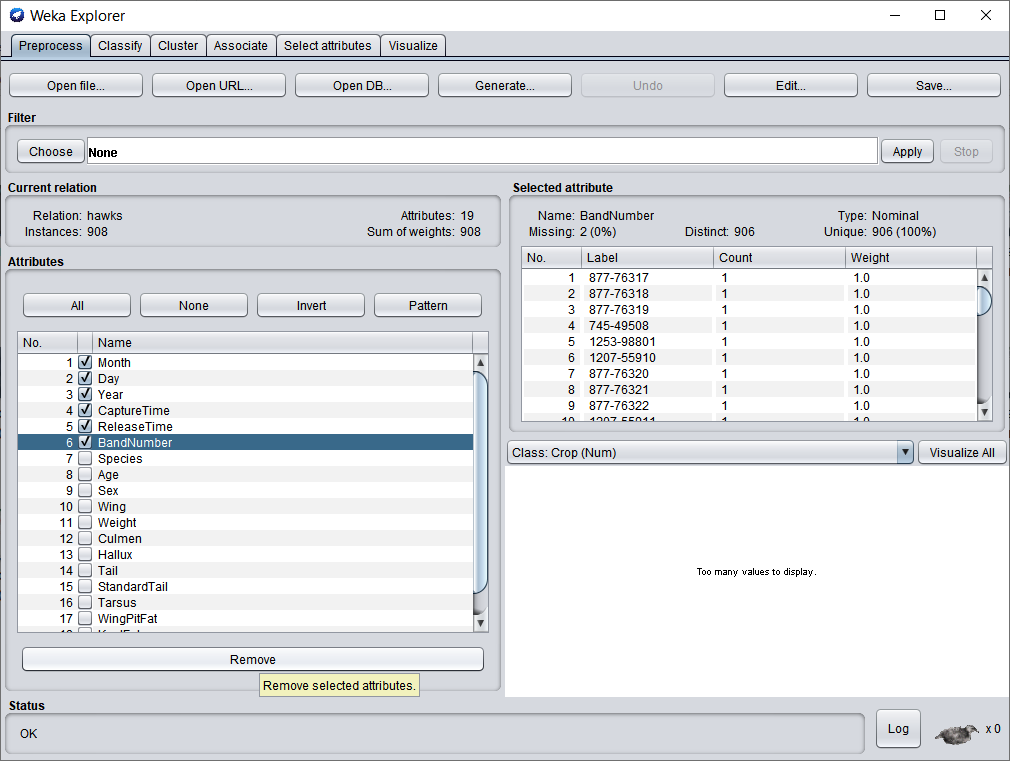
Hình 1. Thay đổi các giá trị bị thiếu từ NA thành "?"

* 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.

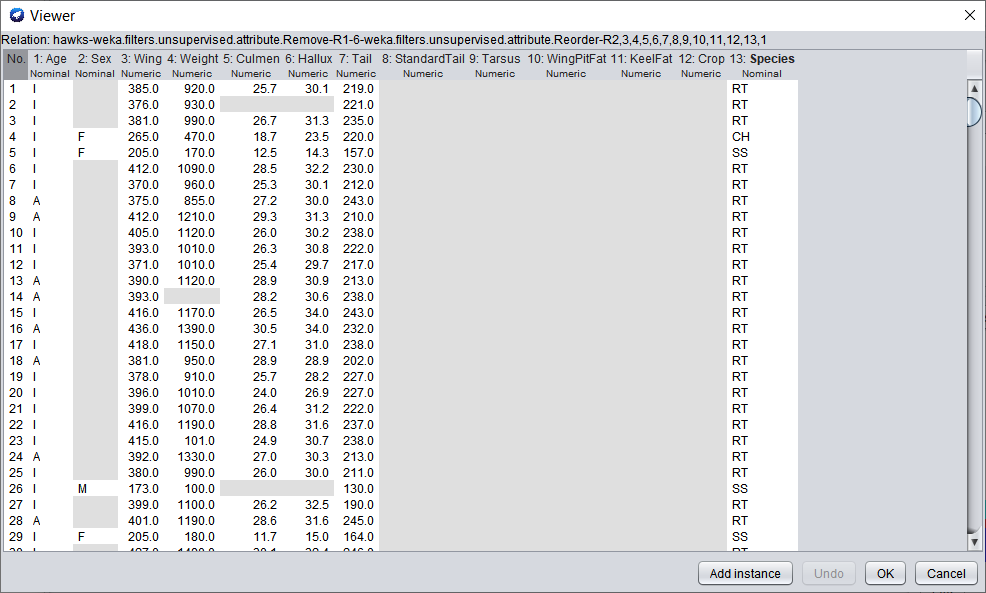


Hình 2. Các thuộc tính đúng kiểu dữ liệu được mô tả và thấy rõ các cột thiếu dữ liệu

* 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.

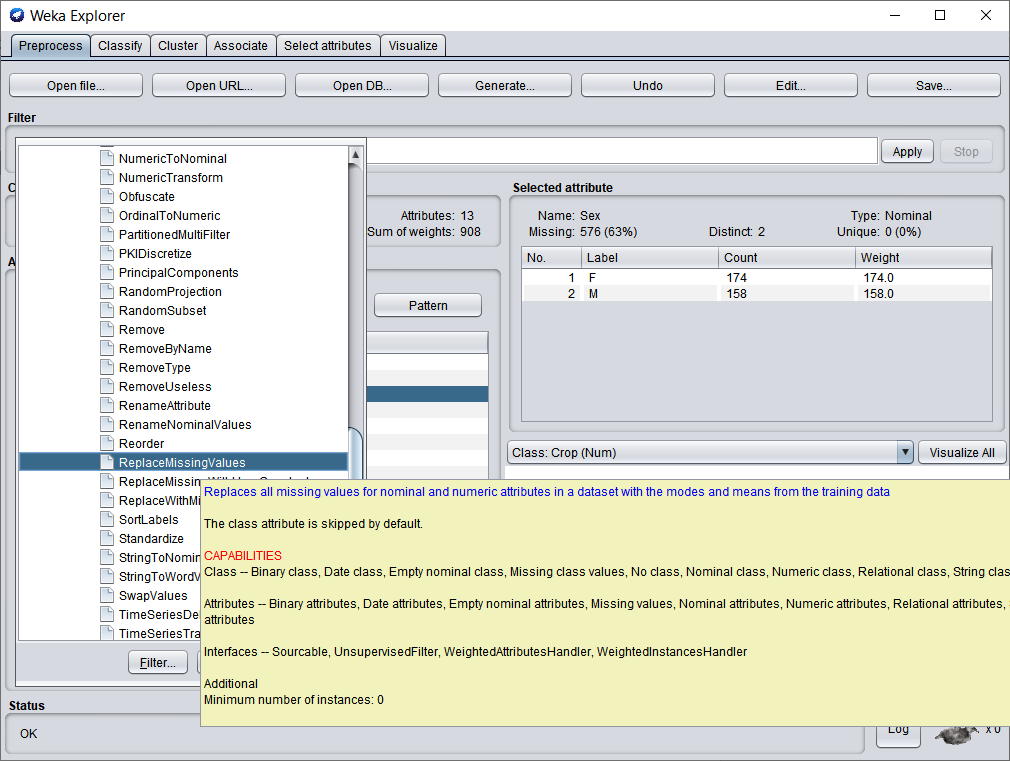


Hình 3. Loại bỏ một số thuộc tính khỏi tập dữ liệu gốc



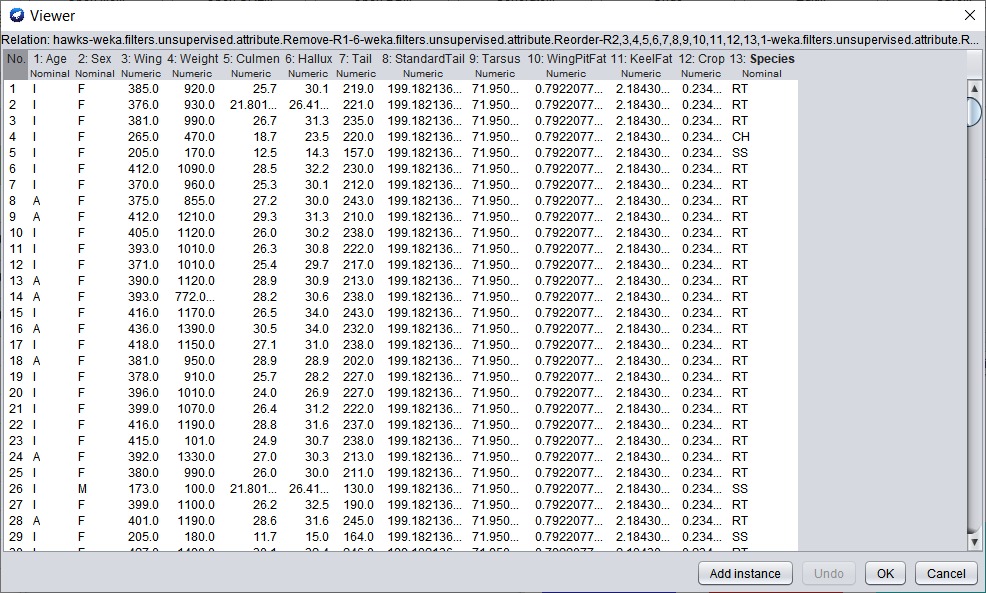
Hình 4. Đặt Species làm class

* 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).



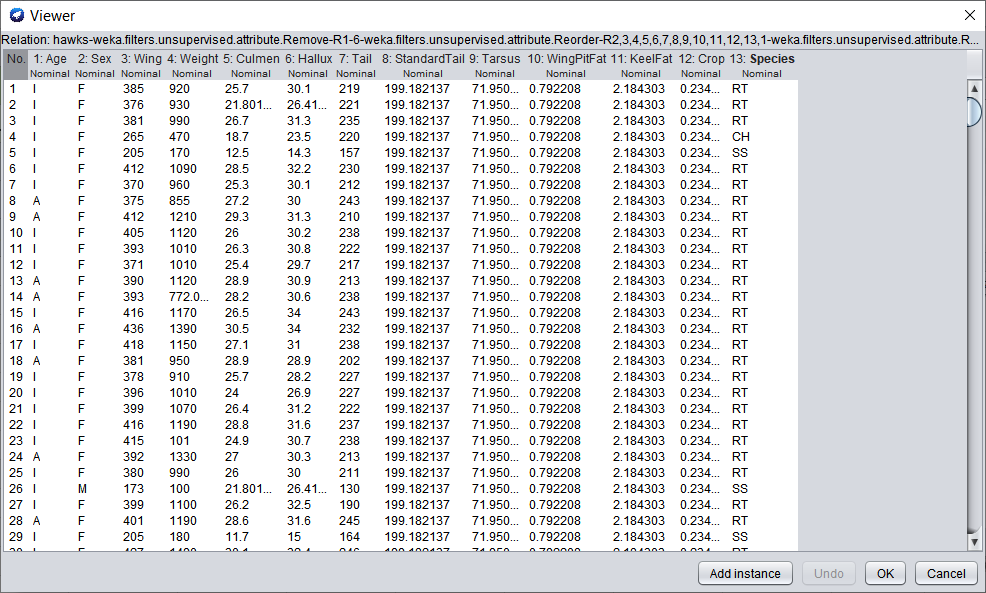
Hình 5. Sử dụng filter ReplaceMissingValues để điền dữ liệu thiếu

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