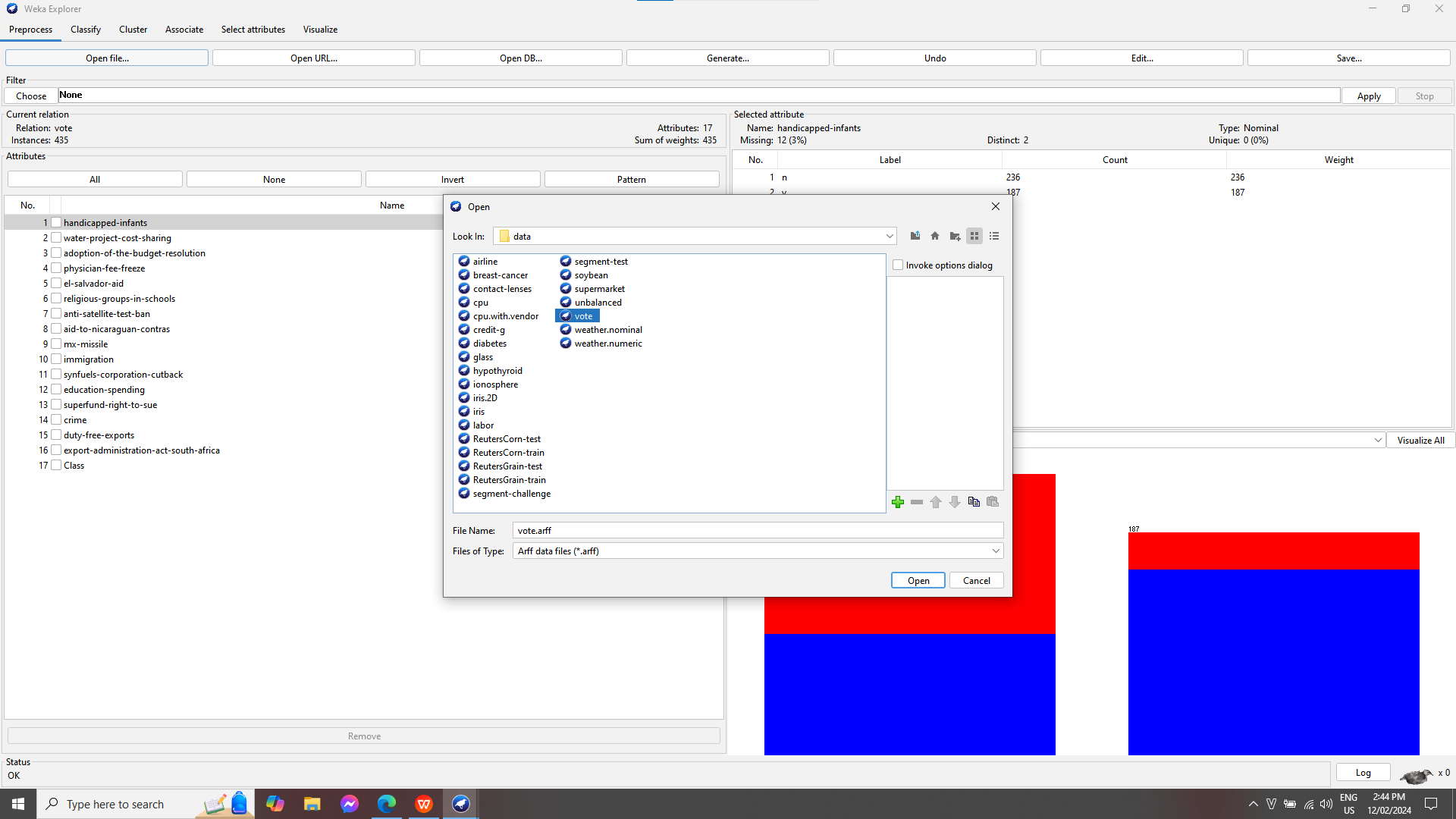
***Bài TH2: Khai phá luật kết hợp trên WEKA***

***Hà Anh Tú - 2151062887***

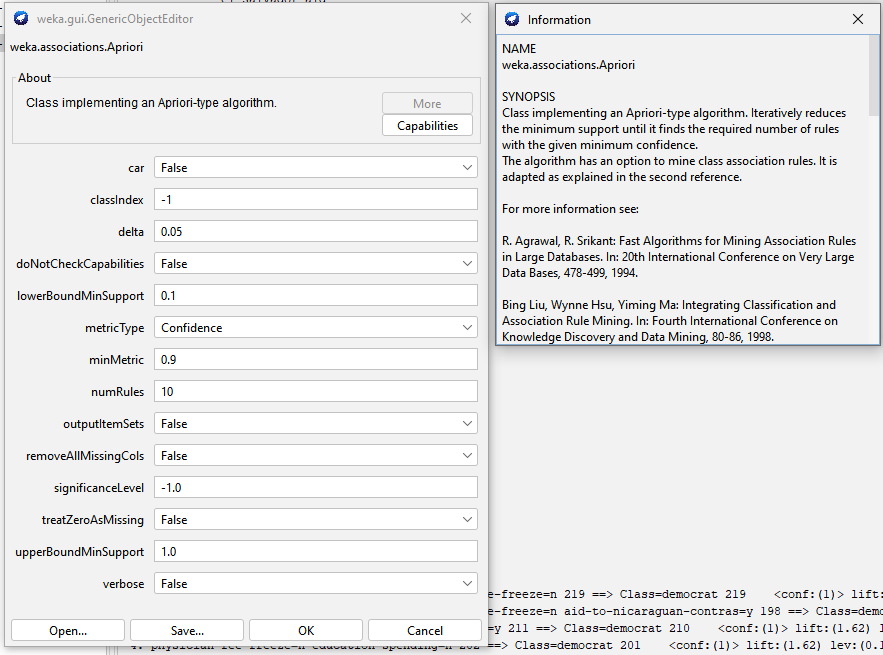
**- Mở data vote.arff:**



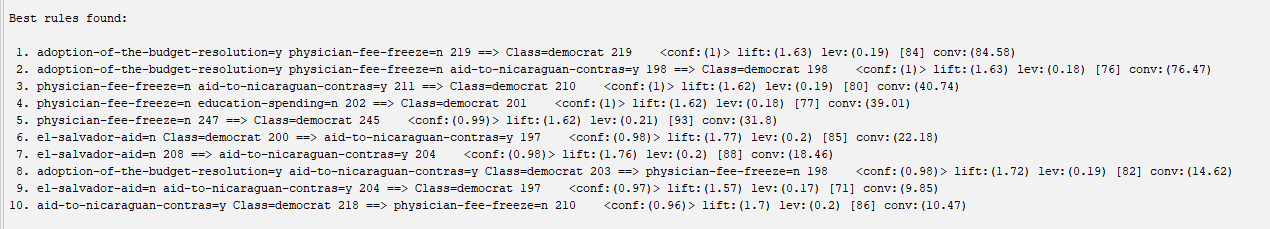
**- Chạy Apriori (start):**



**-Thiết lập thông số:**



**Câu hỏi số 1:**



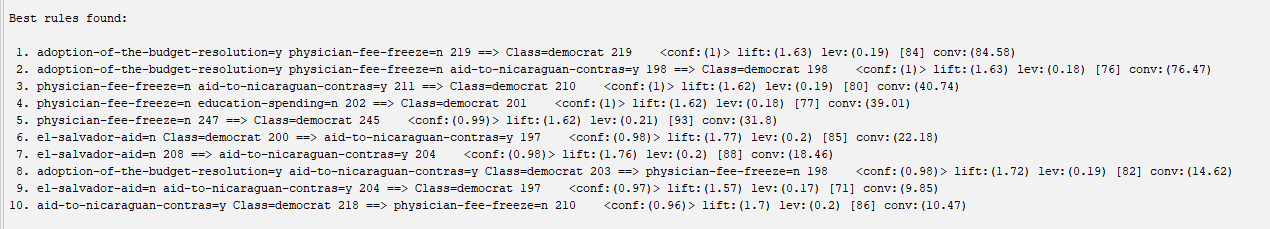
**Quan sát kết quả:**

* Tất cả các luật được sinh ra đều có RHS là Class=democrat.
* Không có luật nào chứa Class=republican.

**Lý do:**

* Tần suất xuất hiện thấp: Số lượng các vote thuộc Class=republican có thể quá ít ( <min\_sup=45% hoặc 196 instances). Nên không đủ điều kiện sinh ra luật.
* Tham số cấu hình: Tham số confidence tối thiểu 0,9. Các luật Class=republican có thể không đạt đủ yêu cầu.

**Câu hỏi số 2:**



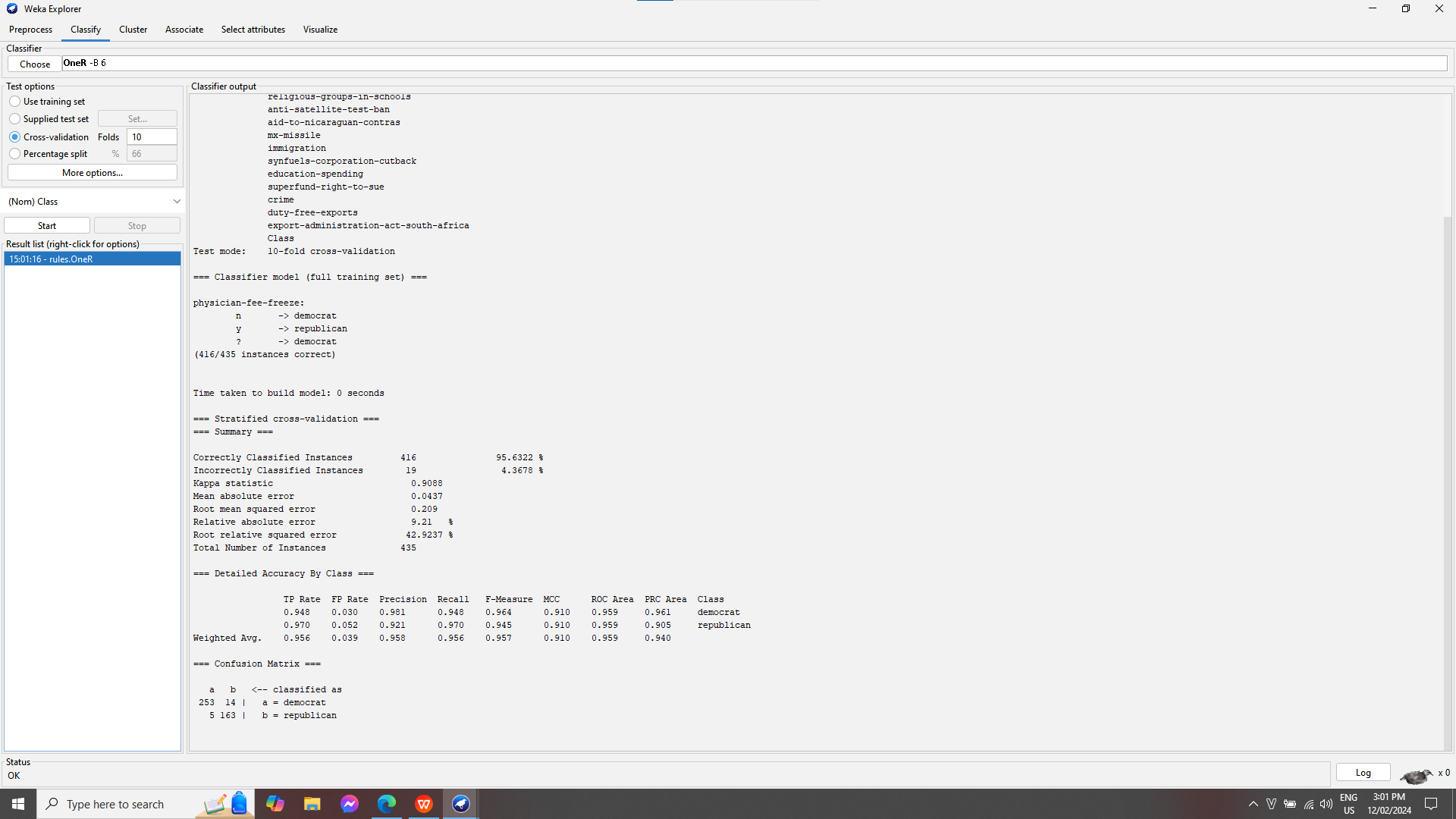
**Quan sát kết quả:**

* Luật đầu tiên:1. adoption-of-the-budget-resolution=y physician-fee-freeze=n 219 ==> Class=democrat 219 có **lift** = 1.63, là giá trị cao nhất.
* Từ đó thấy adoption-of-the-budget-resolution và physician-fee-freeze là hai thuộc tính có tác động mạnh nhất đến Class=democrat.

**Giải thích:**

* Lift = 1.63 nghĩa là mối quan hệ giữa hai thuộc tính này và Class=democrat mạnh hơn ngẫu nhiên 1.63 lần.
* Thuộc tính này xuất hiện trong nhiều luật có giá trị confidence = 1, cho thấy nó luôn dẫn đến lớp democrat.

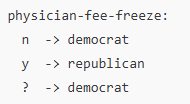
**Câu hỏi số 3:**



**Mô hình OneR đã chọn thuộc tính:**

* physician-fee-freeze làm thuộc tính quyết định cho phân loại.

**Quy tắc OneR sinh ra:**



**Từ đó có nghĩa:**

* Nếu physician-fee-freeze=n, dự đoán là democrat.
* Nếu physician-fee-freeze=y, dự đoán là republican.
* Nếu giá trị bị thiếu (?), dự đoán là democrat.

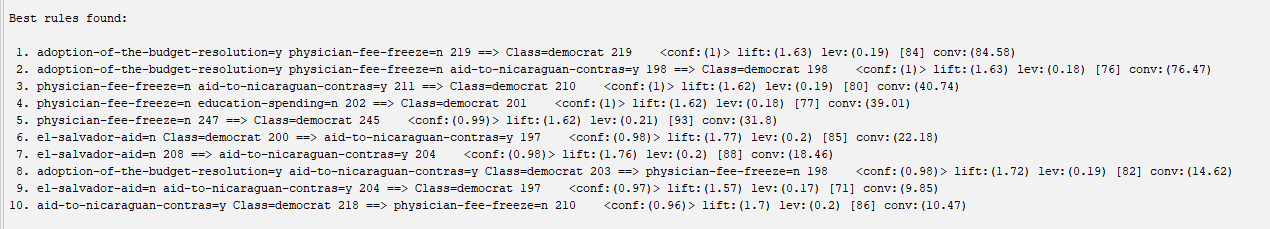
**Accuracy:**

* Correctly Classified Instances: 416/435
* Độ chính xác = (416 / 435) × 100% = 95.63%

**Giá trị thống kê:**

* Kappa statistic: 0.9088 (thể hiện độ mạnh của phân loại).
* Precision và Recall của cả hai lớp (democrat và republican) đều cao, với giá trị trung bình lần lượt là 0.958 và 0.956.

**Câu hỏi số 4:**



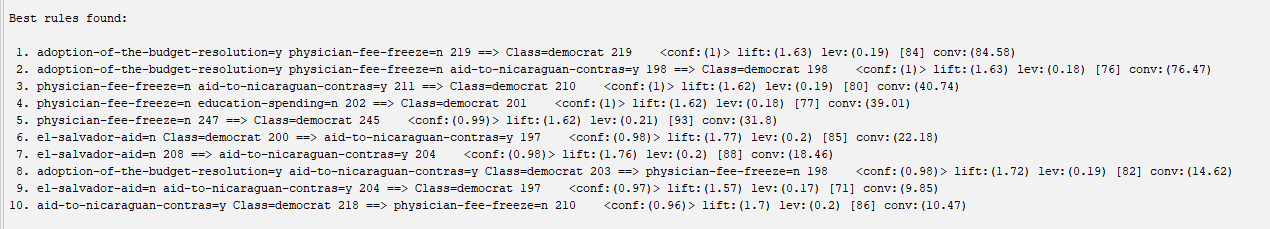
**Quan sát kết quả:**

* Luật 6: el-salvador-aid=n ==> aid-to-nicaraguan-contras=y có lift = 1.77, là cao nhất trong các luật liên quan.
* Điều này cho thấy el-salvador-aid là thuộc tính tác động mạnh nhất đến việc bầu chọn cho aid-to-nicaraguan-contras.

**Lý do:**

* Lift = 1.77 nghĩa là khi el-salvador-aid=n, khả năng xảy ra aid-to-nicaraguan-contras=y cao hơn ngẫu nhiên 1.77 lần.

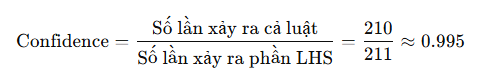
**Câu hỏi số 5:**



**Luật số 3: physician-fee-freeze=n, aid-to-nicaraguan-contras=y ==> Class=democrat**

* Số lần luật xảy ra (Cả LHS và RHS): 210.
* Số lần phần bên trái (LHS) xảy ra: 211.

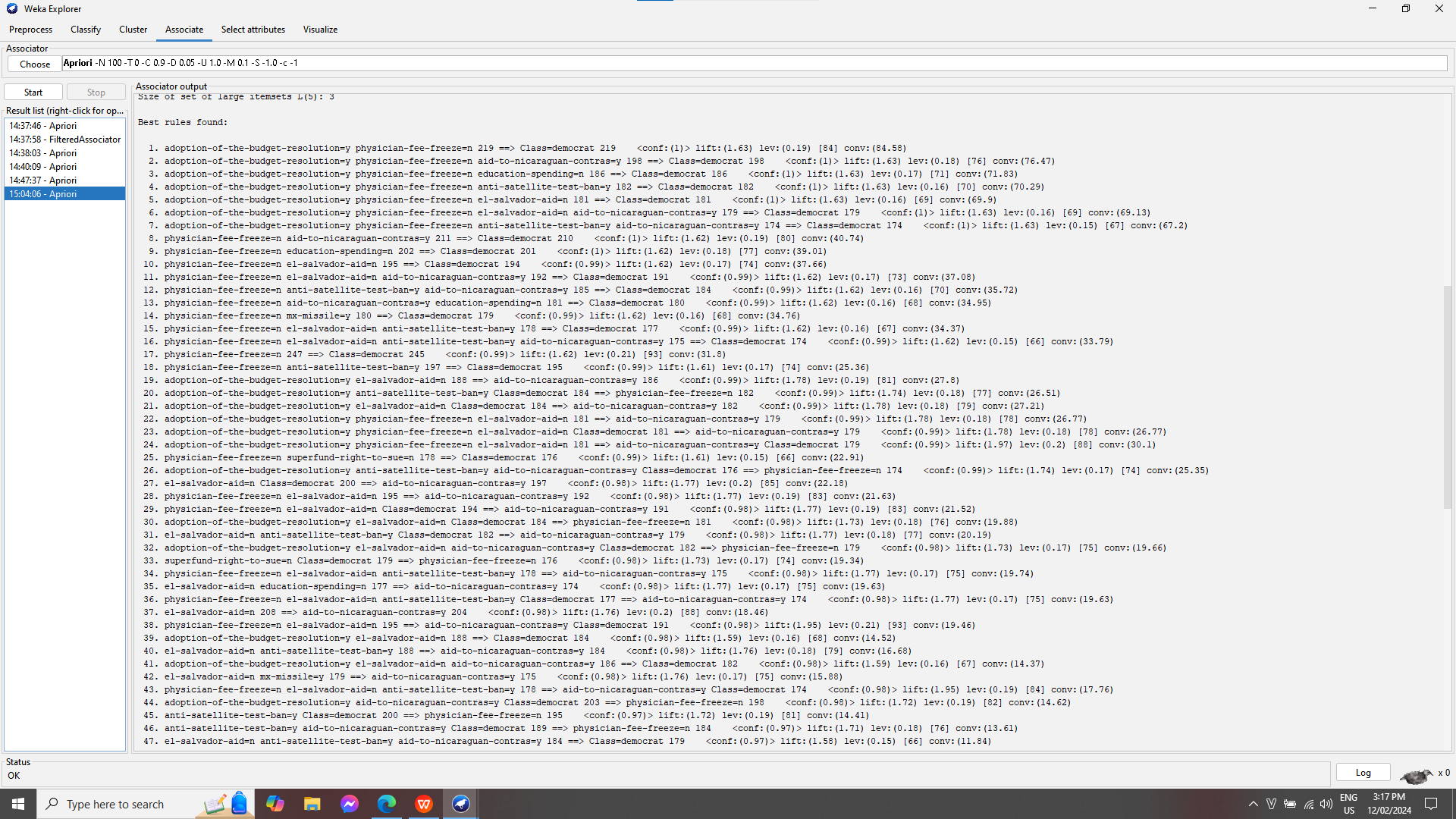
**Tính Confidence:**

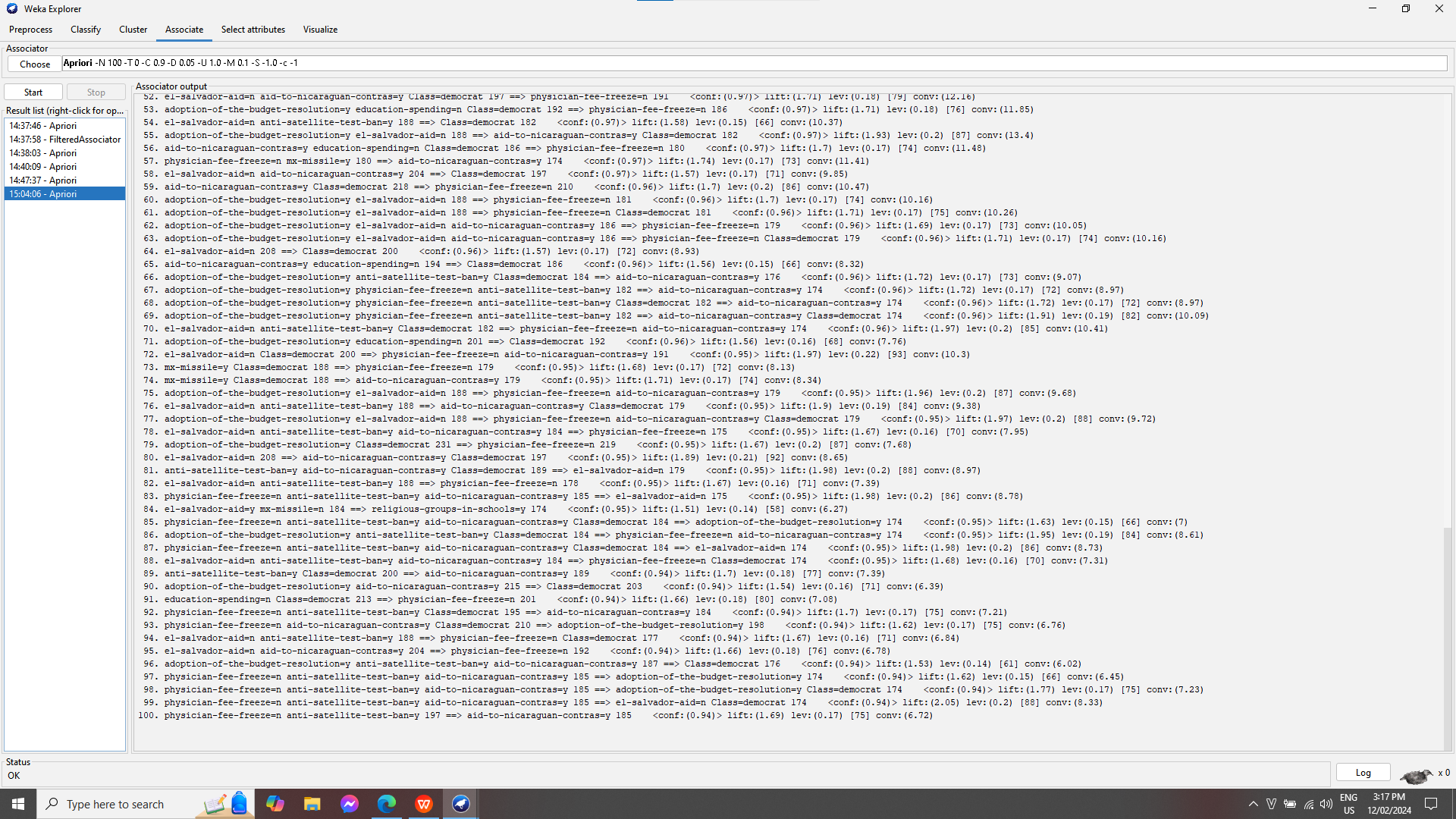


**Kết luận:**

* Trùng với kết quả start được là 0,995.

**Câu hỏi số 6:**





**Các thông số ban đầu:**

* Minimum support (hỗ trợ tối thiểu): 0.4 nghĩa là các mẫu phải xuất hiện ít nhất ở 40% trong tổng số 435 mẫu (174 mẫu) để được coi là "phổ biến".
* Minimum confidence (độ tin cậy tối thiểu): 0.9, đảm bảo rằng các luật được khai thác có độ chính xác ít nhất 90%.
* Số vòng lặp (Number of cycles): 12, đây là số lần thuật toán lặp lại để khám phá tất cả các tập phổ biến (itemsets).

**Gọi luật này là A lần 1:**

* 3. physician-fee-freeze=n aid-to-nicaraguan-contras=y 211 ==> Class=democrat 210 <conf:(1)> lift:(1.62) lev:(0.19) [80] conv:(40.74)

**Sau khi thay đổi numRules từ 10 lên 100:**

* 8. physician-fee-freeze=n aid-to-nicaraguan-contras=y 211 ==> Class=democrat 210 <conf:(1)> lift:(1.62) lev:(0.19) [80] conv:(40.74)

**Kết luận:**

* Luật tụt từ vị trí thứ 3 xuống vị trí thứ 8, do lift của các luật xuất hiện cao hơn.