** ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH**

**TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN**

**KHOA CÔNG NGHỆ THÔNG TIN**

**BÁO CÁO LAB-02:**

**FREQUENT ITEMSETS AND ASSOCIATION RULES**

**Sinh viên thực hiện:**

**Trần Thanh Tùng – 18120258**

**Trần Hữu Chí Bảo - 18120288**

**T**

**Môn: Khai thác dữ liệu và ứng dụng**

**Thành phố Hồ Chí Minh - 2020**

# MỤC LỤC

[MỤC LỤC 2](#_Toc56275611)

[PHẦN I. DATA 3](#_Toc56275612)

[I. Mô tả tập dữ liệu: 3](#_Toc56275613)

[II. Mô hình phân cấp của dữ liệu: 4](#_Toc56275614)

[2. Set grouping hierarchy: 5](#_Toc56275615)

[PHẦN II. CODE 6](#_Toc56275616)

[I. Tiền xử lý dữ liệu: 6](#_Toc56275617)

[PHẦN III. EXPERIMENTS 11](#_Toc56275618)

[I. Mục đích phân tích dữ liệu: 11](#_Toc56275619)

[II. Tiến hành thử nghiệm: 11](#_Toc56275620)

[2.1 Thử nghiệm 1: xét trường hợp với tất cả các thuộc tính. 11](#_Toc56275621)

[2.2 Thử nghiệm 2: xét trường hợp với những khách hàng churn 15](#_Toc56275622)

[2.3 Thử nghiệm 3: xét trường hợp với những khách hàng sử dụng Int’l Plan hoặc Vmail Plan 19](#_Toc56275623)

[2.4 Thử nghiệm 4: xét thói quen gọi của khách hàng. 23](#_Toc56275624)

[PHẦN IV. TÓM TẮT KẾT QUẢ 27](#_Toc56275625)

[I. Tiêu chí đánh giá kết quả: 27](#_Toc56275626)

[II. 1 số luật tốt rút ra được: 27](#_Toc56275627)

[III. Điểm mạnh và điểm yếu: 29](#_Toc56275628)

[ Điểm mạnh: 29](#_Toc56275629)

[ Điểm yếu: 29](#_Toc56275630)

[PHẦN V. TÀI LIỆU THAM KHẢO 30](#_Toc56275631)

# PHẦN I. DATA

## Mô tả tập dữ liệu:

|  |  |  |
| --- | --- | --- |
| **Cột** | **Loại thuộc tính** | **Mục đích** |
| State | categorical | Ký hiệu viết tắt của các bang |
| Account Length | numeric | Thời gian thuê bao đã hoạt động |
| Area code | numeric | Mã khu vực |
| Phone | Numeric | Số điện thoại của thuê bao |
| Int’l Plan | Yes/No | Gói quốc tế đã được kích hoạt hay chưa? |
| Vmail Plan | Yes/No | Gói thư thoại đã được kích hoạt hay chưa? |
| Day Mins | Numeric | Tổng số phút sử dụng ban ngày |
| Day Calls | Numeric | Tổng số cuộc gọi ban ngày |
| Day Charge | Numeric | Tổng số tiền cước ban ngày |
| Eve Mins | Numeric | Tổng số phút sử dụng buổi tối |
| Eve Calls | Numeric | Tổng số cuộc gọi buổi tối |
| Eve Charge | Numeric | Tổng số tiền cước buổi tối |
| Night Mins | Numeric | Tổng số phút sử dụng buổi đêm |
| Night Calls | Numeric | Tổng số cuộc gọi buổi đêm |
| Night Charge | Numeric | Tổng số tiền cước buổi đêm |
| Intl Mins | Numeric | Tổng số phút sử dụng quốc tế |
| Intl Calls | Numeric | Tổng số cuộc gọi quốc tế |
| Intl Charge | Numeric | Tổng số tiền cước quốc tế |
| CustServ Calls | Numeric | Tổng số cuộc gọi trung tâm chăm sóc khách hàng |
| Churn | Yes/No | Khách hàng có hủy thuê bao không ? |

## Mô hình phân cấp của dữ liệu:

1. **Schema hierarchy:**

Phân cấp theo dịch vụ:

### Set grouping hierarchy:

* Gom nhóm theo các cuộc gọi vào các thời điểm khác nhau trong ngày và các cuộc gọi quốc tế:

{Day, Eve, Night, Intl}

{Day Mins, Day Calls, Day Charge} Day

{Eve Mins, Eve Calls, Eve Charge} Eve

{Night Mins, Night Calls, Night Charge} Night

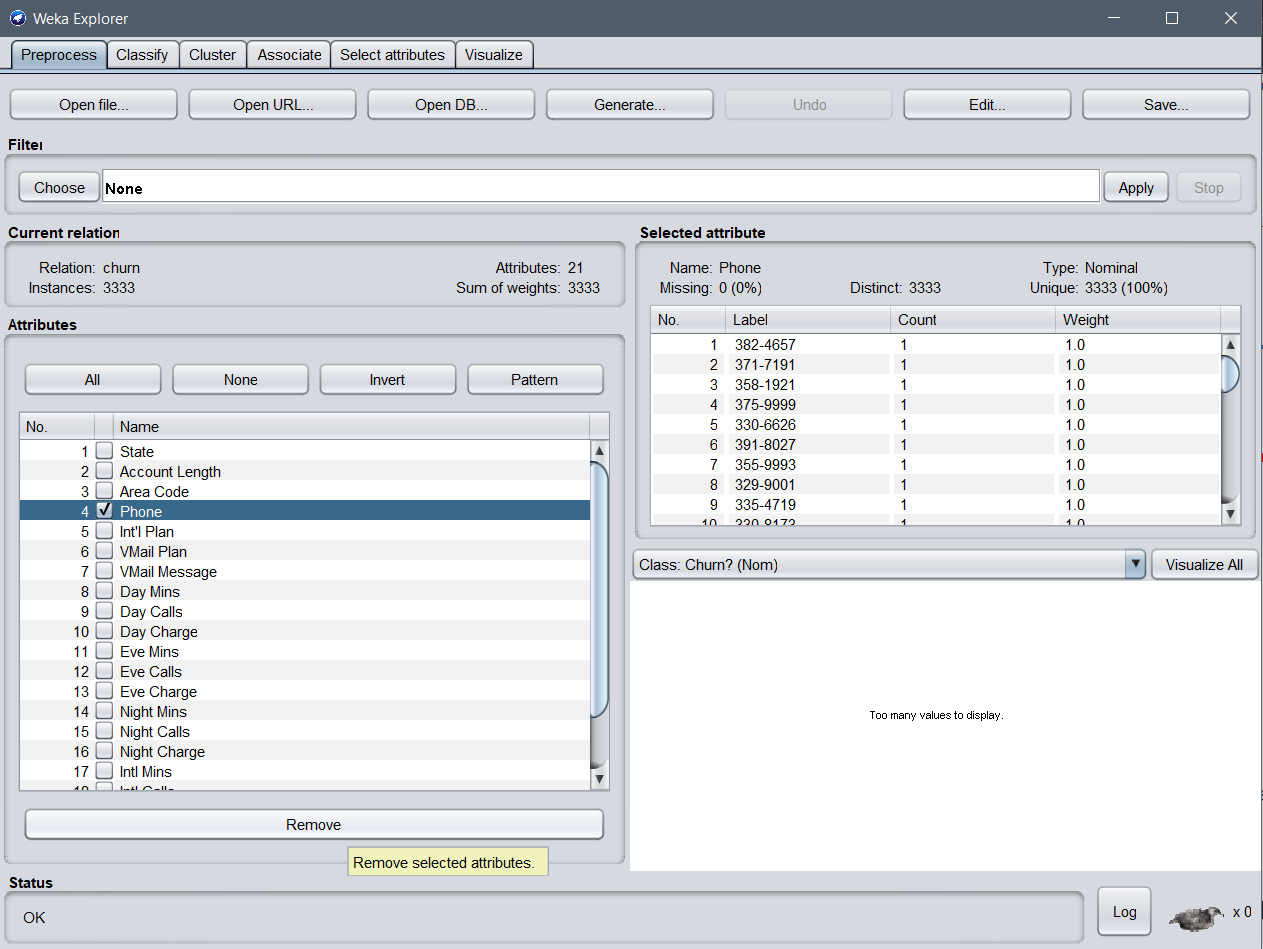
{Intl Mins, Intl Calls, Intl Charge} Intl

# PHẦN II. CODE

## Tiền xử lý dữ liệu:

* **Mục đích:**
* Các thuộc tính charge có thể được suy ra từ các thuộc tính minute thông qua một hàm tuyến tính vì thế nên bỏ qua.
* Bỏ qua thuộc tính phone vì mỗi số điện thoại là 1 mẫu riêng biệt nên không thể rút ra được luật.
* Phân nhóm thuộc tính tính “Account Length”, 'Day Mins', 'Day Calls', 'Eve Mins', 'Eve Calls', 'Night Mins', 'Night Calls', 'Intl Mins', 'Intl Calls' do các thuộc tính này là các thuộc tính.
* **Tác dụng:**
* Kích thước của không gian giải pháp được giảm xuống, do đó các thuật toán khai thác dữ liệu nhất định có thể tìm ra giải pháp tối ưu toàn cầu hiệu quả hơn.
* **Cách thực hiện:**

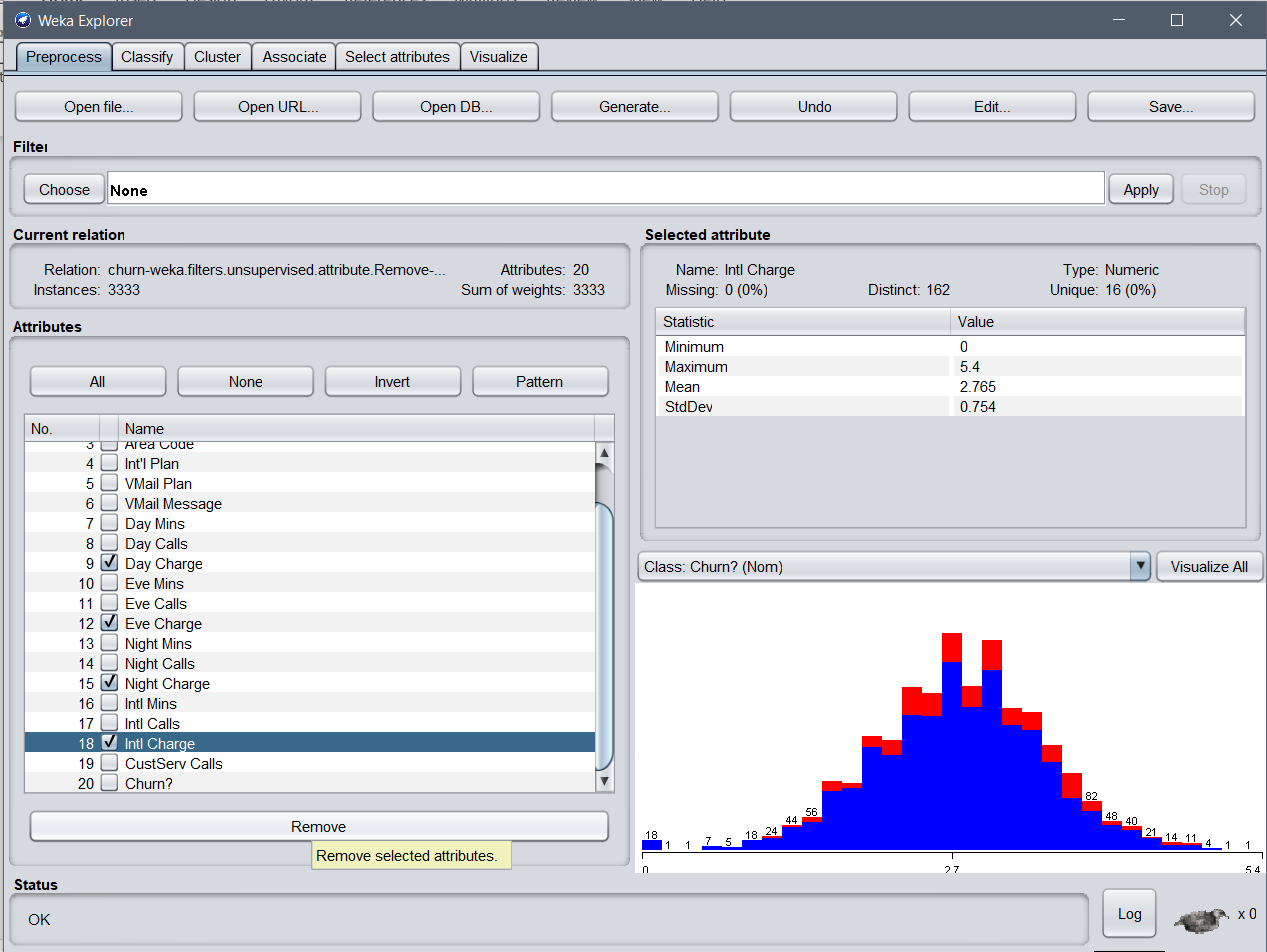
1. Xóa Phone:

* Chọn thuộc tính Phone 🡪 Remove.

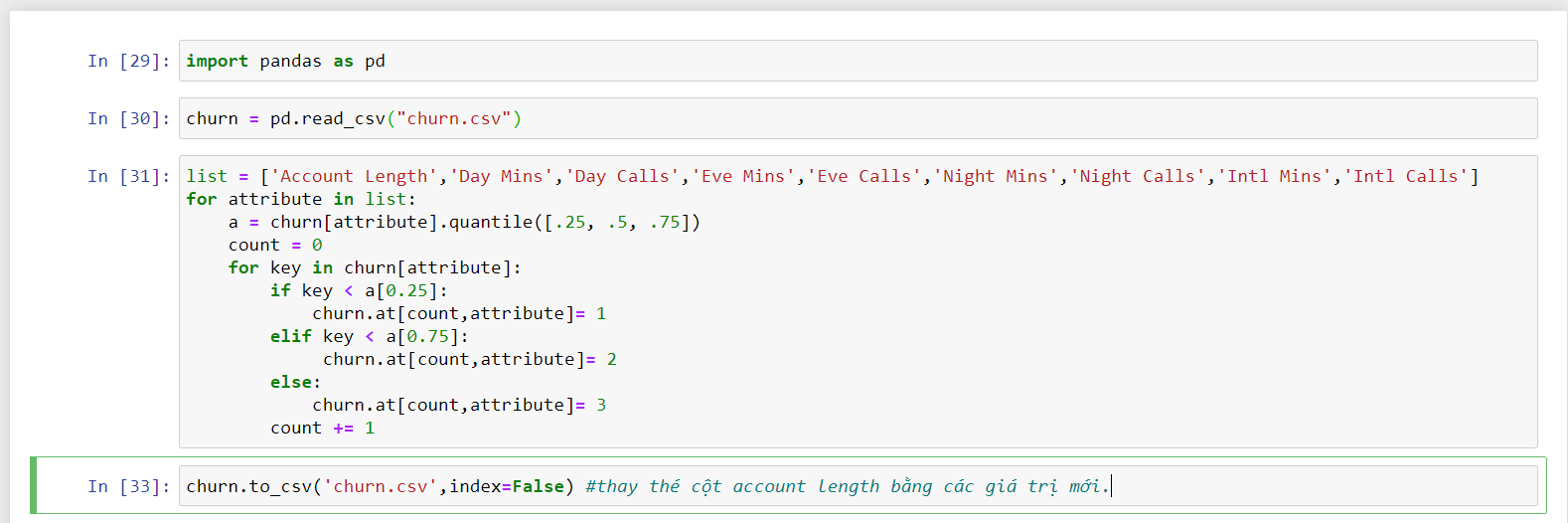
Hình 2.1. Xóa thuộc tính phone trong weka.

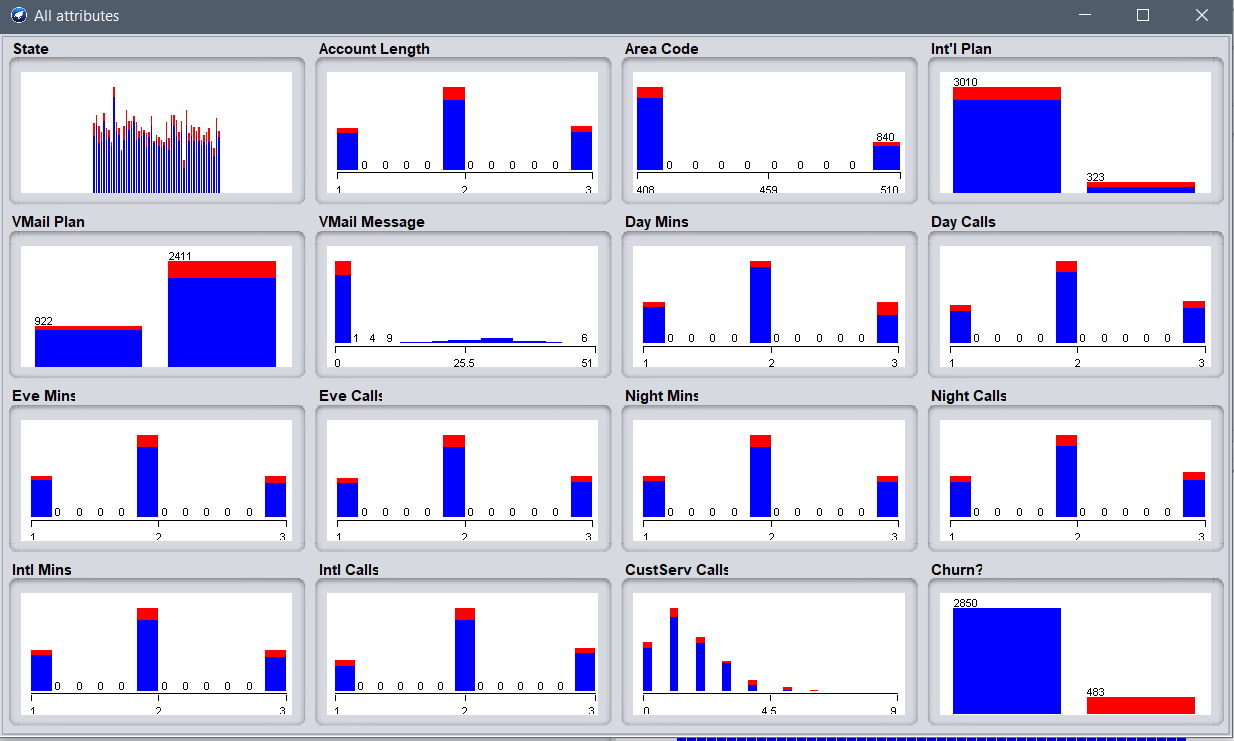
1. Xóa 4 thuộc tính Day Charge, Eve Charge, Night Charge, Intl Charge:

* Chọn 4 thuộc tính Day Charge, Eve Charge, Night Charge, Intl Charge 🡪 Remove

Hình 2.2. Xóa các 4 thuộc tính như trong weka.

1. Phân nhóm thuộc tính “Account Length”, 'Day Mins', 'Day Calls', 'Eve Mins', 'Eve Calls', 'Night Mins', 'Night Calls', 'Intl Mins', 'Intl Calls'.

* Dựa vào 2 phân vị của thuộc tính là lower quartile và upper quartile ta phân nhóm khách hàng thành 3 loại:
* Loại 1: Dưới lower quartile là khách hàng mới sử dụng dịch vụ và gán nhãn là 1.
* Loại 2: Từ lower quartile 🡪 upper quartile là khách hàng có thời gian sử dụng dịch vụ trung bình và gán nhãn là 2.
* Loại 3: Trên upper quartile là khách hàng sử dụng dịch vụ lâu năm và gán nhãn là 3.
* Xử lí dữ liệu bằng python 3 với thư viện pandas được viết trên IDE jupteter notebook.

Hình 2.3. Code thực hiện xóa 4 thuộc tính sử dụng thử viện pandas và lưu kết quả xuống file churn\_process.csv

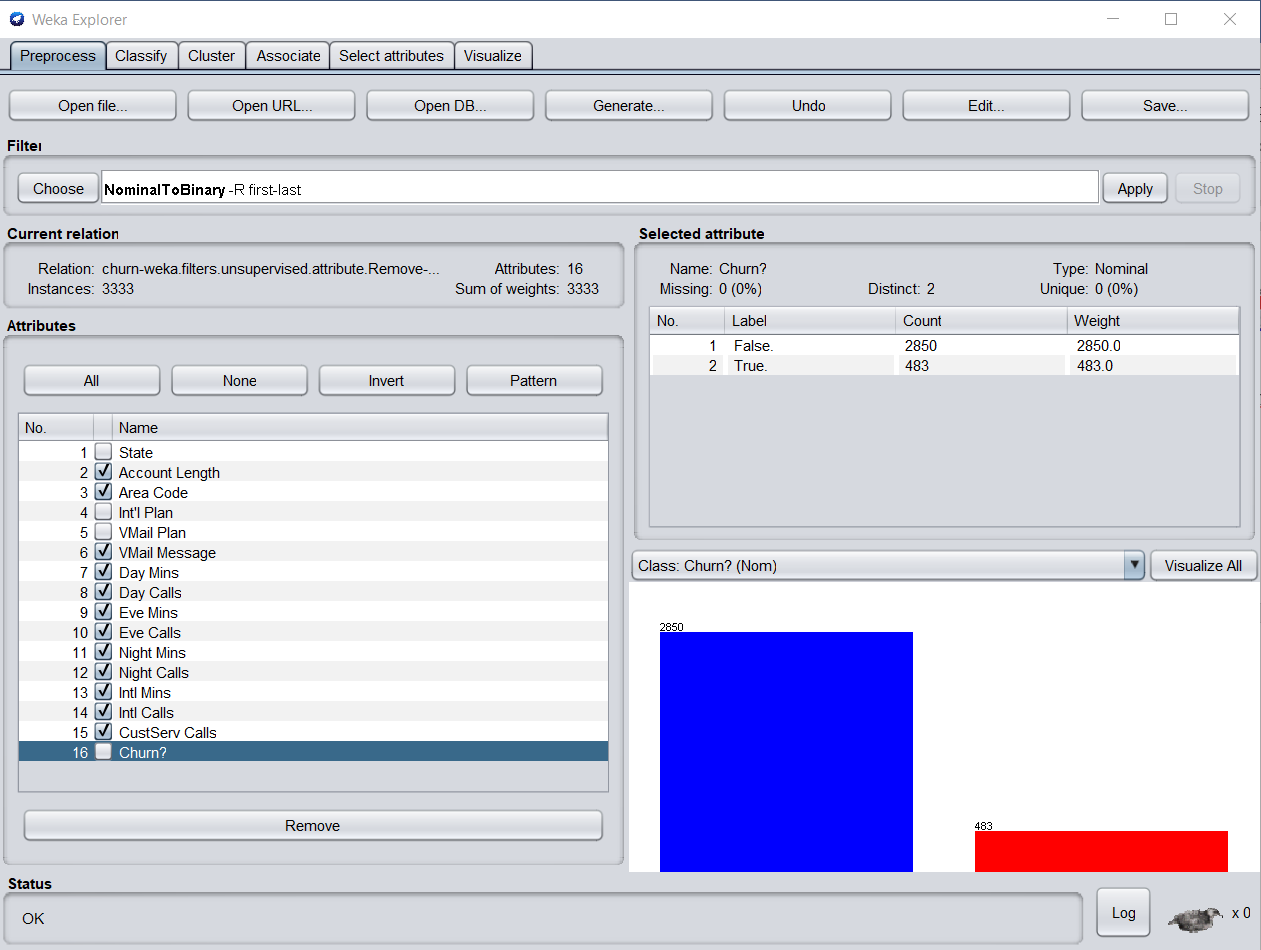
Hình 2.4. Đồ thị các thuộc tính sau khi phân nhóm.

1. Chuyển các thuộc tính numeric 🡪 nominal:

Bước 1: Chọn các thuộc tính cần chuyển( numeric).

Bước 2: Chọn filter 🡪 unsupervised 🡪 attribute 🡪 NumericToNominal.

Buước 3: Chọn apply.



Hình 2.5. Chuyển đổi tất cả các thuộc tính có kiểu dữ liệu numeric 🡪 nominal

# PHẦN III. EXPERIMENTS

## Mục đích phân tích dữ liệu:

Từ bài toán kinh doanh thực tế:

Khách hàng ngày nay thường phải đắn đo suy nghĩ trước khi đưa ra quyết định có đăng ký bất kỳ dịch vụ nào của một hãng viễn thông hay không.

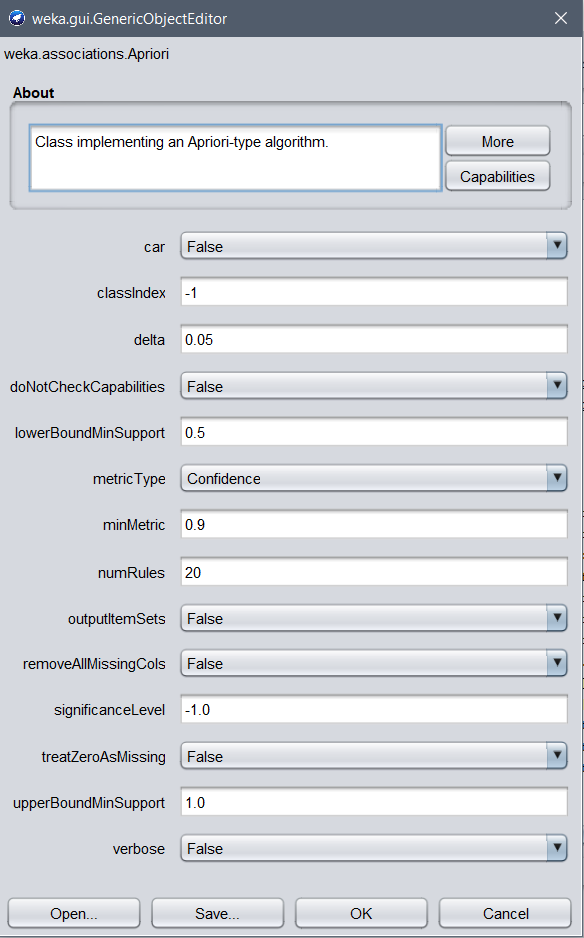
Các dịch vụ đưa ra bởi sale của hãng viễn thông không thực sự khác biệt và có tính hấp dẫn đối với khách hàng.

Sự trung thành của khách hàng trở thành một vấn đề. Bởi vậy, hãng viễn thông cần đưa ra một phương pháp để nhận diện, phân cụm nhóm khách hàng trung thành, nhóm khách hàng có khả năng hủy đăng ký dịch vụ. Từ đó đưa ra các quyết định kinh doanh hợp lý để giữ chân khách hàng có khả năng hủy vào tối ưu hóa lợi nhuận nhóm khách hàng trung thành.

Tối ưu hóa tập khách hàng theo vùng địa lý (theo bang), để đưa ra các chính sách kinh doanh phù hợp cho mỗi bang khách nhau.

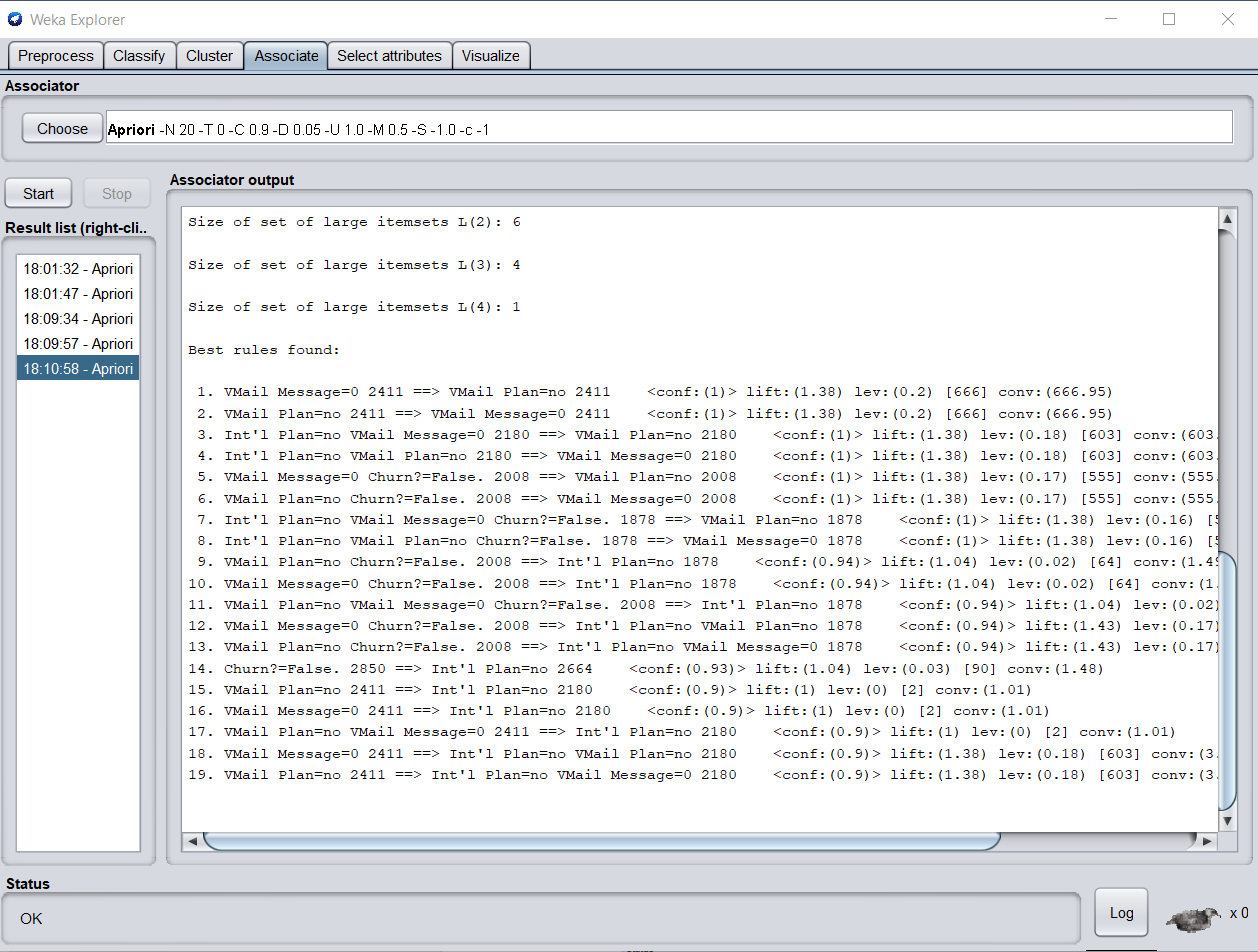
## Tiến hành thử nghiệm:

### **Thử nghiệm 1: xét trường hợp với tất cả các thuộc tính**.

*  **Tham số hệ thống:**

Hình 3.1. Bảng thông số hệ thống.

**Ý nghĩa**:



Hình 3.2. Sử dụng thuật toán Apiori để rút ra các tập luật

* **Kết hợp 2 luật 1,2 :**

*1. VMail Message=0 2411 ==> VMail Plan=no 2411 <conf:(1)> lift:(1.38) lev:(0.2) [666] conv:(666.95)*

*2. VMail Plan=no 2411 ==> VMail Message=0 2411 <conf:(1)> lift:(1.38) lev:(0.2) [666] conv:(666.95)*

* Với độ confidence 100% ta có thể rút ra được kết luận rằng nếu khách hàng không đăng ký Vmail plan sẽ không nhận Vmail Message và ngược lại.
* **Kết hợp 2 luật 3,4**

*3. Int'l Plan=no VMail Message=0 2180 ==> VMail Plan=no 2180 <conf:(1)> lift:(1.38) lev:(0.18) [603] conv:(603.05)*

*4. Int'l Plan=no VMail Plan=no 2180 ==> VMail Message=0 2180 <conf:(1)> lift:(1.38) lev:(0.18) [603] conv:(603.05)*

* Cũng với confidence 100%, từ luật này ta có thể rút ra được kết luận rằng : các khách hàng thường không dùng cả 2 dịch vụ Intl Plan và Vmail Message chứ không dùng một trong hai cái.
* **Kết hợp 2 luật 5,6:**

*5. VMail Message=0 Churn?=False. 2008 ==> VMail Plan=no 2008 <conf:(1)> lift:(1.38) lev:(0.17) [555] conv:(555.47)*

*6. VMail Plan=no Churn?=False. 2008 ==> VMail Message=0 2008 <conf:(1)> lift:(1.38) lev:(0.17) [555] conv:(555.47)*

* Từ các luật này ta kết luận được rằng những khách hàng không dùng dịch vụ thư thoại thì cũng không hủy thuê bao.
* **Kết hợp các luật 7,8,9,10,11,12,13:**

*7. Int'l Plan=no VMail Message=0 Churn?=False. 1878 ==> VMail Plan=no 1878 <conf:(1)> lift:(1.38) lev:(0.16) [519] conv:(519.51)*

*8. Int'l Plan=no VMail Plan=no Churn?=False. 1878 ==> VMail Message=0 1878 <conf:(1)> lift:(1.38) lev:(0.16) [519] conv:(519.51)*

*9. VMail Plan=no Churn?=False. 2008 ==> Int'l Plan=no 1878 <conf:(0.94)> lift:(1.04) lev:(0.02) [64] conv:(1.49)*

*10. VMail Message=0 Churn?=False. 2008 ==> Int'l Plan=no 1878 <conf:(0.94)> lift:(1.04) lev:(0.02) [64] conv:(1.49)*

*11. VMail Plan=no VMail Message=0 Churn?=False. 2008 ==> Int'l Plan=no 1878 <conf:(0.94)> lift:(1.04) lev:(0.02) [64] conv:(1.49)*

*12. VMail Message=0 Churn?=False. 2008 ==> Int'l Plan=no VMail Plan=no 1878 <conf:(0.94)> lift:(1.43) lev:(0.17) [564] conv:(5.3)*

*13. VMail Plan=no Churn?=False. 2008 ==> Int'l Plan=no VMail Message=0 1878 <conf:(0.94)> lift:(1.43) lev:(0.17) [564] conv:(5.3)*

* Ta rút ra được kết luận rằng nếu khách hàng không dùng cả 2 dịch vụ gọi quốc tế và thư thoại thì cũng không hủy thuê bao.
* **Luật 14:**

*14. Churn?=False. 2850 ==> Int'l Plan=no 2664 <conf:(0.93)> lift:(1.04) lev:(0.03) [90] conv:(1.48)*

* Cho ta thấy rằng những khách hàng không hủy thuê bao thì thường không dùng dịch vụ gọi quốc tế.
* **Kết hợp các luật 15,16,17,18,19:**

*15. VMail Plan=no 2411 ==> Int'l Plan=no 2180 <conf:(0.9)> lift:(1) lev:(0) [2] conv:(1.01)*

*16. VMail Message=0 2411 ==> Int'l Plan=no 2180 <conf:(0.9)> lift:(1) lev:(0) [2] conv:(1.01)*

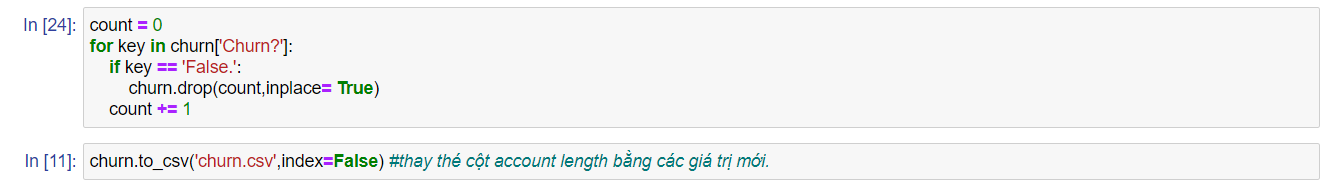
*17. VMail Plan=no VMail Message=0 2411 ==> Int'l Plan=no 2180 <conf:(0.9)> lift:(1) lev:(0) [2] conv:(1.01)*

*18. VMail Message=0 2411 ==> Int'l Plan=no VMail Plan=no 2180 <conf:(0.9)> lift:(1.38) lev:(0.18) [603] conv:(3.6)*

*19. VMail Plan=no 2411 ==> Int'l Plan=no VMail Message=0 2180 <conf:(0.9)> lift:(1.38) lev:(0.18) [603] conv:(3.6)*

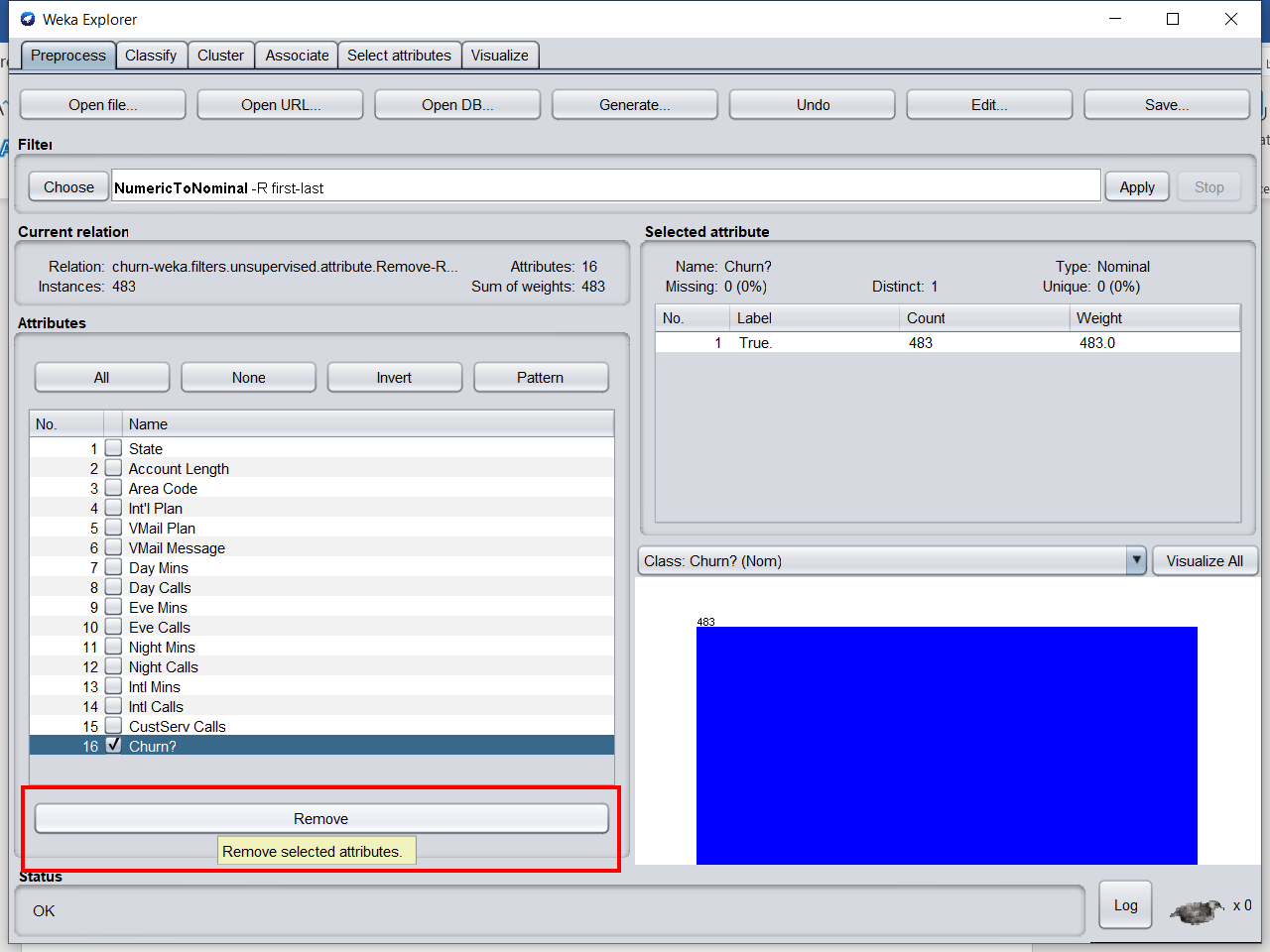
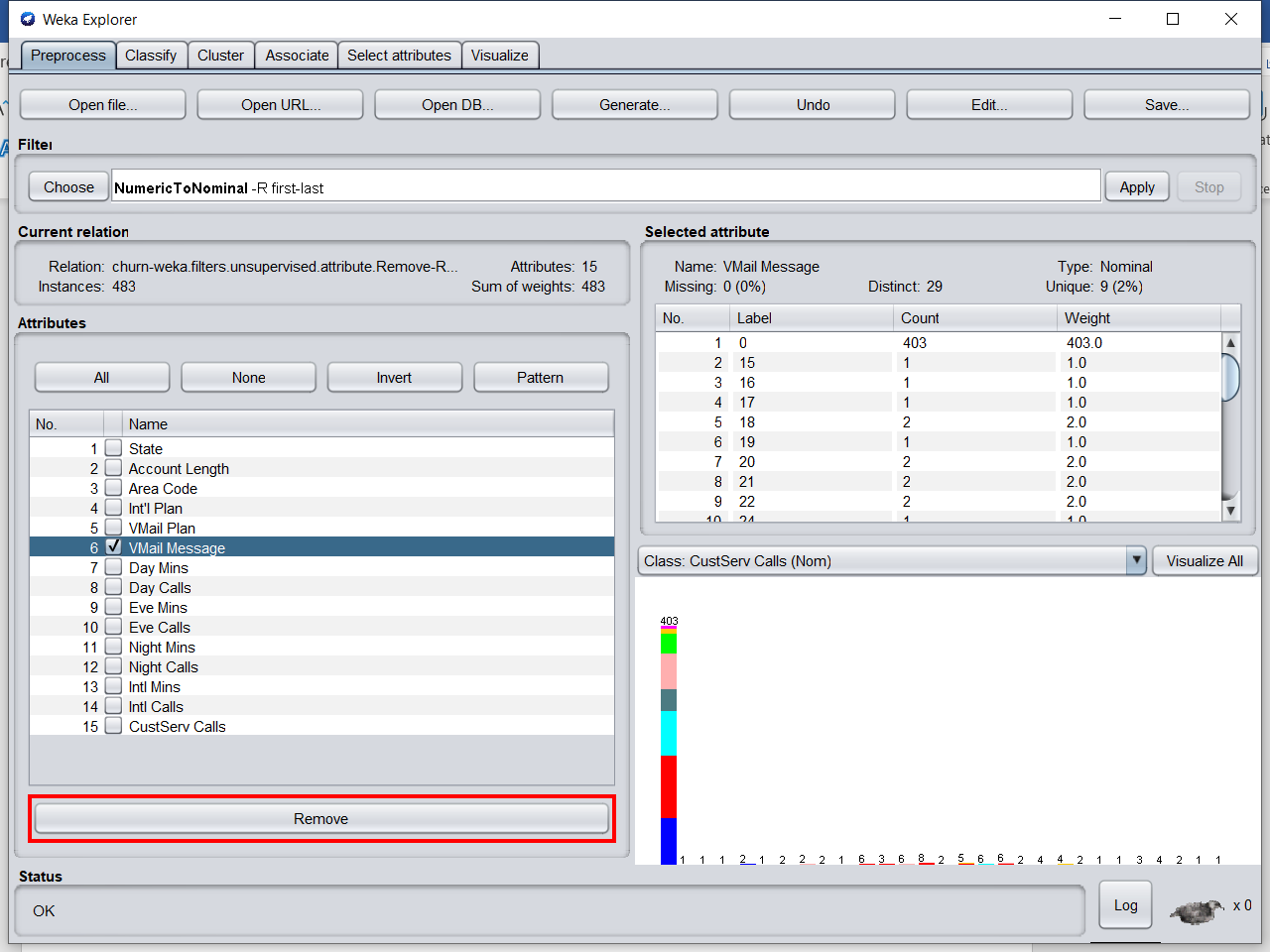
* Rút ra được kết luận rằng nếu không dùng dịch vụ thư thoại thì cũng thường không dùng dịch vụ gọi quốc tế (độ confidence là 0.9)

### Thử nghiệm 2: xét trường hợp với những khách hàng churn

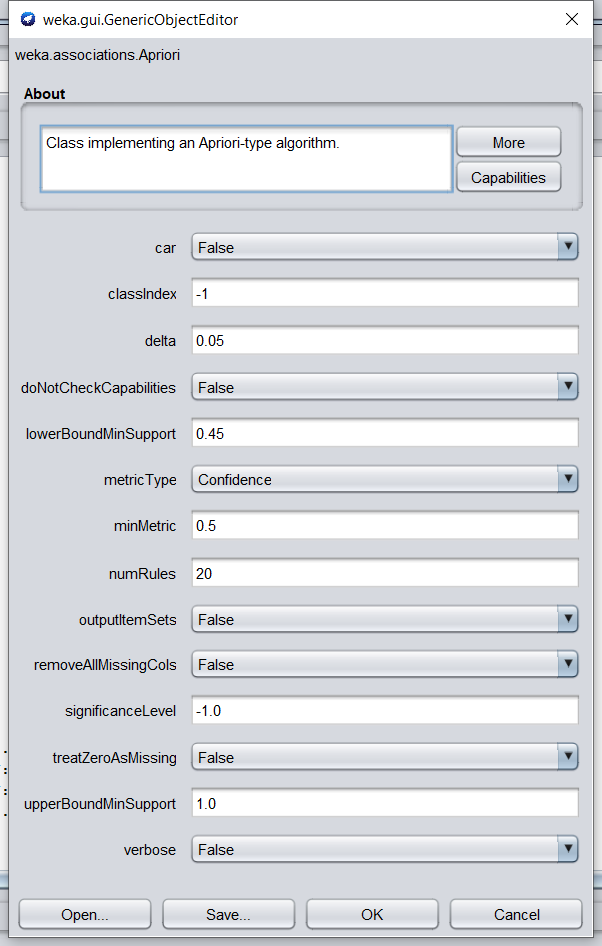
* **Thể hiện dữ liệu:**
* Điều này có nghĩa ta sẽ xét các thuộc tính của tập khách hàng churn, tìm ra điểm chung của những người này.
* **Tiền xử lý:**Dùng python 3 với thư viện pandas để xóa các dòng mà churn = false.

Hình 3.3. Code xóa tất cả các dòng có churn = False trong jupyter notebook.

* Trong Weka Explorer Remove thuộc tính churn (do bây giờ churn chỉ còn 1 giá trị duy nhất là true).

 Hình 3.4. Remove thuộc tính churn trong weka

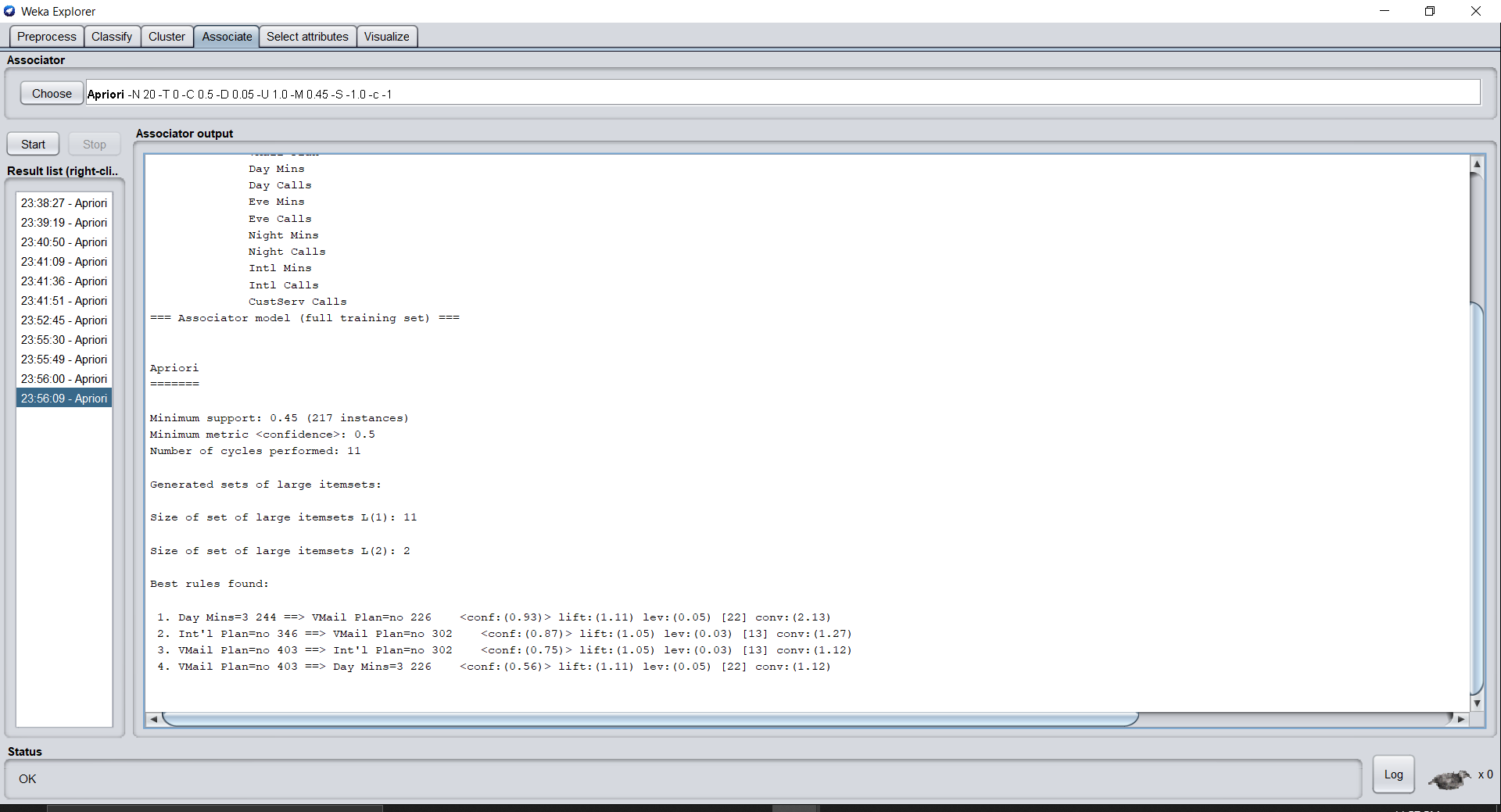
Hình 3.5. Remove thuộc tính Vmail Message trong weka.

* Remove thêm cột Vmail Message vì ở trên ta đã biết rằng nếu những người có đăng ký dịch vụ thư thoại (Vmail Plan) thì chắc chắc sẽ nhận Vmail Message, ta Remove cột này đi để tránh khai thác lại những luật “Những người có Vmail Plan sẽ có Vmail Message và ngược lại”.
* **Tham số hệ thống:**

Hình 3.6. Bảng tham số hệ thống.

**Ý nghĩa**: Ta cho cả min support là 45% vì theo quan điểm cổ điển của xác suất “Khi gieo ngẫu nhiên một đồng xu, nếu số lần gieo là vô hạn, thì tỉ lệ xuất hiện của mặt sấp và mặt ngửa đều là 0.5”. Tương tự cho dữ liệu của ta, nếu những thuộc tính nào đồng xuất hiện với nhau trên 0.5 thì chắc chắn phải có quy luật nào đó, do dữ liệu của chúng ta không phải là vô hạn nên ta chọn bên dưới của 0.5 là 0.45 làm min support.

* Tư thông số trên ta khai thác được các luật:



Hình 3.7. Sử dụng thuật toán Apiori để rút ra các tập luật

* **Kết hợp luật 1 và 4:**

*1. Day Mins=3 244 ==> VMail Plan=no 226 <conf:(0.93)> lift:(1.11) lev:(0.05) [22] conv:(2.13)*

*4. VMail Plan=no 403 ==> Day Mins=3 226 <conf:(0.56)> lift:(1.11) lev:(0.05) [22] conv:(1.12)*

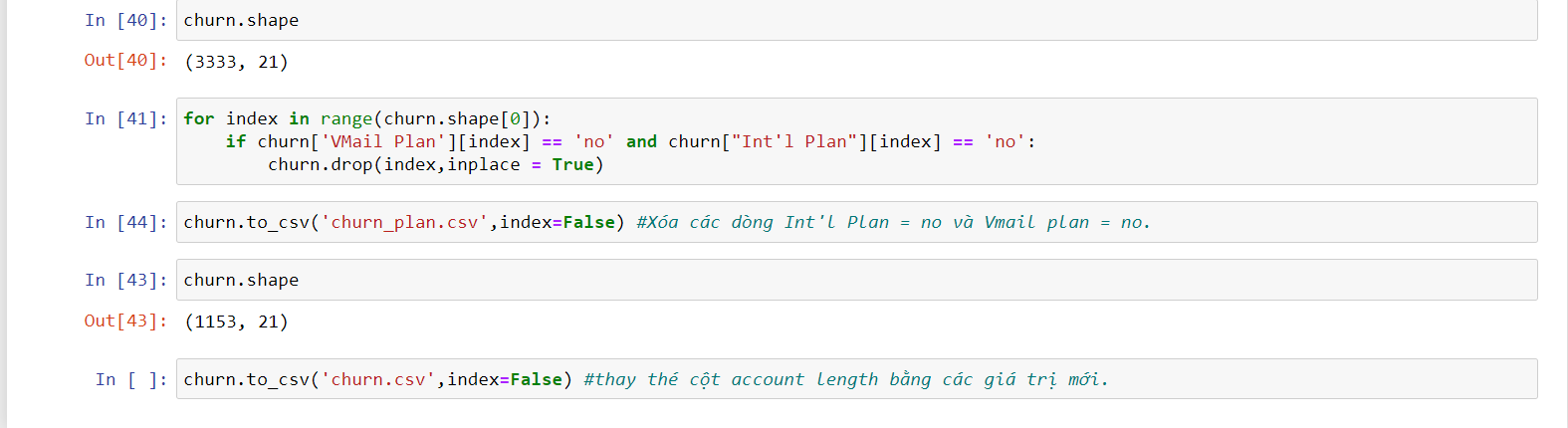
* Cho ta thấy rằng nếu một người hủy thuê bao thường gọi nhiều vào ban ngày và không đăng ký Vmail Plan.
* **Kết hợp 2 luật 2,3:**

*2. Int'l Plan=no 346 ==> VMail Plan=no 302 <conf:(0.87)> lift:(1.05) lev:(0.03) [13] conv:(1.27)*

*3. VMail Plan=no 403 ==> Int'l Plan=no 302 <conf:(0.75)> lift:(1.05) lev:(0.03) [13] conv:(1.12)*

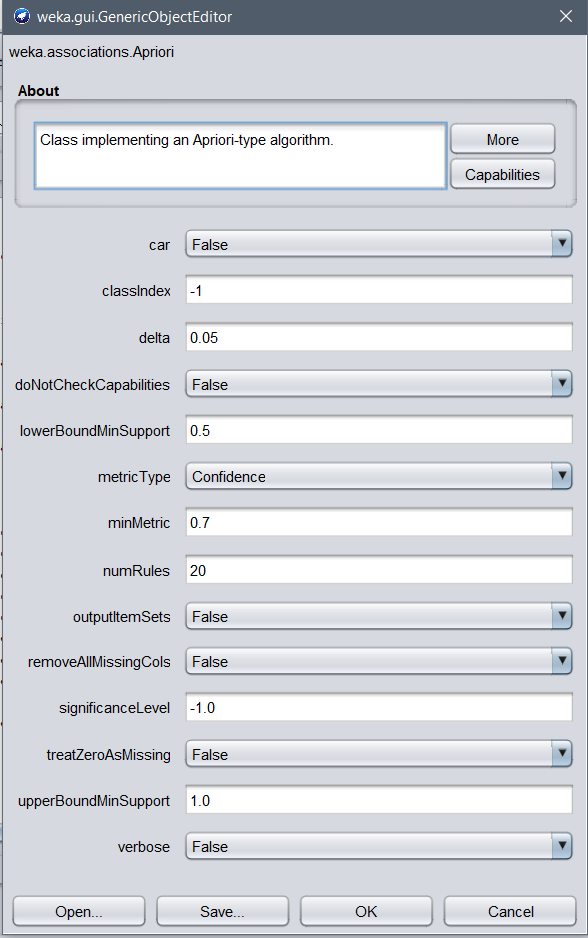
* Ta cũng thấy được rằng người huỷ thuê bao cũng thường không có Vmail Plan và Intl Plan.

### Thử nghiệm 3: xét trường hợp với những khách hàng sử dụng Int’l Plan hoặc Vmail Plan:

* **Thể hiện dữ liệu:**
* Xét tập khách hàng có sử dụng 1 trong 2 dịch vụ Int’l Plan và Vmail Plan hoặc cả hai để tìm luật. Nguyên nhân ta xét vì đây là tập khách hàng tiềm năng(sẵn sàng chi tiền thêm để sử dụng dịch vụ).
* **Tiền xử lý:**
* Dùng python 3 kết hợp với thư viện pandas trên IDE jupyter notebook để chọn ra các khách hàng thỏa mãn yêu cầu.

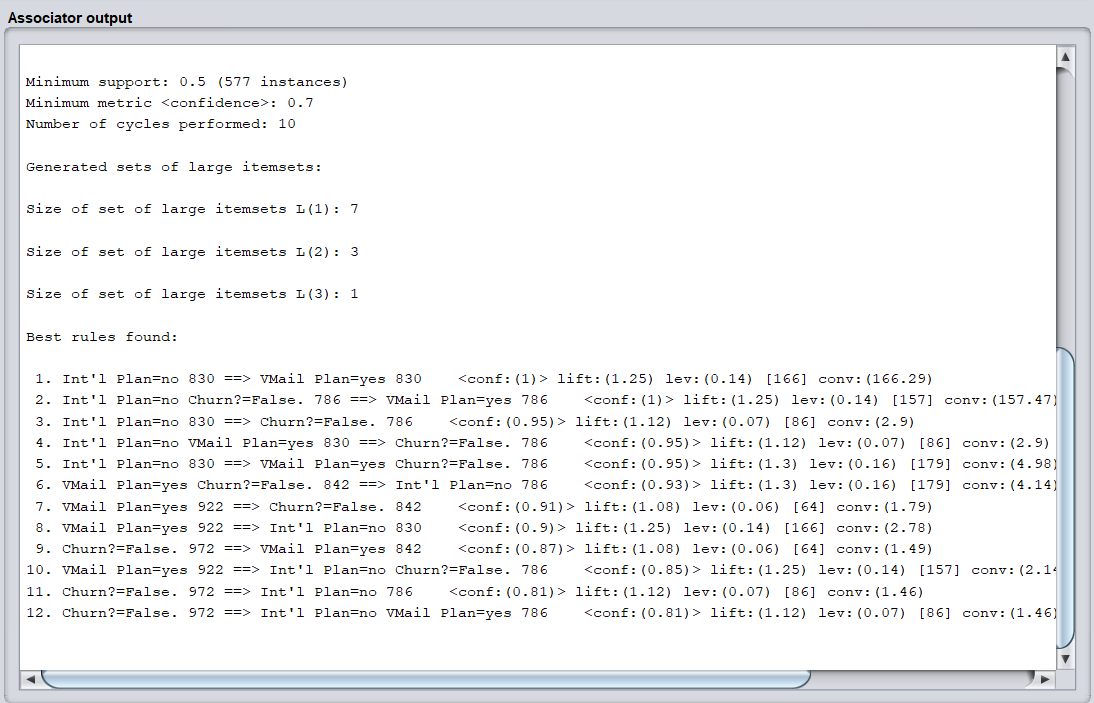
Hình 3.8. Code remove các giá trị “no” của 2 thuộc tính “Vmail Plan” và “Int’l Plan” .

* **Tham số hệ thống:**



Hình 3.9. Bảng thông số hệ thống.

**Ý nghĩa**: Ta cho min support là 50% vì theo quan điểm cổ điển của xác suất “Khi gieo ngẫu nhiên một đồng xu, nếu số lần gieo là vô hạn, thì tỉ lệ xuất hiện của mặt sấp và mặt ngửa đều là 0.5”. Tương tự cho dữ liệu của ta, nếu những thuộc tính nào đồng xuất hiện với nhau trên 0.5 thì chắc chắn phải có quy luật nào đó, min confidence là 0.7 vì số lượng khách hàng có sử dụng 1 trong 2 hoặc cả 2 dịch vụ chiếm tới 1/3 số mẫu ban đầu.



Hình 3.10. Sử dụng thuật toán Apiori để rút ra các tập luật.

* **Kết hợp 2 Luật 1, 9 :**

*1. Int'l Plan=no 830 ==> VMail Plan=yes 830 <conf:(1)> lift:(1.25) lev:(0.14) [166] conv:(166.29)*

*8. VMail Plan=yes 922 ==> Int'l Plan=no 830 <conf:(0.9)> lift:(1.25) lev:(0.14) [166] conv:(2.78)*

Tỷ lệ khách hàng dùng Vmail Plan mà không dùng Intl Plan so với số lượng khách hàng dùng cả 2 dịch vụ.

* **Kết hợp Luật 2:**

*2. Int'l Plan=no Churn?=False. 786 ==> VMail Plan=yes 786 <conf:(1)> lift:(1.25) lev:(0.14) [157] conv:(157.47)*

Nếu 1 khách hàng không hủy gói và không dùng Intl Plan thì thường dùng Vmail Plan

* **Kết hợp Luật 3,10:**

*3. Int'l Plan=no 830 ==> Churn?=False. 786 <conf:(0.95)> lift:(1.12) lev:(0.07) [86] conv:(2.9)*

*11. Churn?=False. 972 ==> Int'l Plan=no 786 <conf:(0.81)> lift:(1.12) lev:(0.07) [86] conv:(1.46)*

* Từ luật này, ta có thể nhận xét rằng nếu một khách hàng có đăng ký Intl Plan thì có tỷ lệ hủy thuê bao cao.
* Cần cải thiện gói Intl Plan để hạn chế việc khách hàng hủy thuê bao.
* **Kết hợp 3 luật 4,5,6:**

*4. Int'l Plan=no VMail Plan=yes 830 ==> Churn?=False. 786 <conf:(0.95)> lift:(1.12) lev:(0.07) [86] conv:(2.9)*

*5. Int'l Plan=no 830 ==> VMail Plan=yes Churn?=False. 786 <conf:(0.95)> lift:(1.3) lev:(0.16) [179] conv:(4.98)*

*6. VMail Plan=yes Churn?=False. 842 ==> Int'l Plan=no 786 <conf:(0.93)> lift:(1.3) lev:(0.16) [179] conv:(4.14)*

* Cho ta biết rằng nếu một khách hàng có sử dụng Vmail Plan và không sử dụng Intl Plan thì tỷ lệ hủy thuê bao thấp.
* **Kết hợp 2 Luật 7,9:**

7. VMail Plan=yes 922 ==> Churn?=False. 842 <conf:(0.91)> lift:(1.08) lev:(0.06) [64] conv:(1.79)

*9. Churn?=False. 972 ==> VMail Plan=yes 842 <conf:(0.87)> lift:(1.08) lev:(0.06) [64] conv:(1.49*)

* Người dùng Vmail Plan thường không hủy thuê bao 🡪 Gói Vmail Plan đem lại trải nghiệm tốt cho khách hàng.
* **Kết hợp 2 luật 10, 12:**

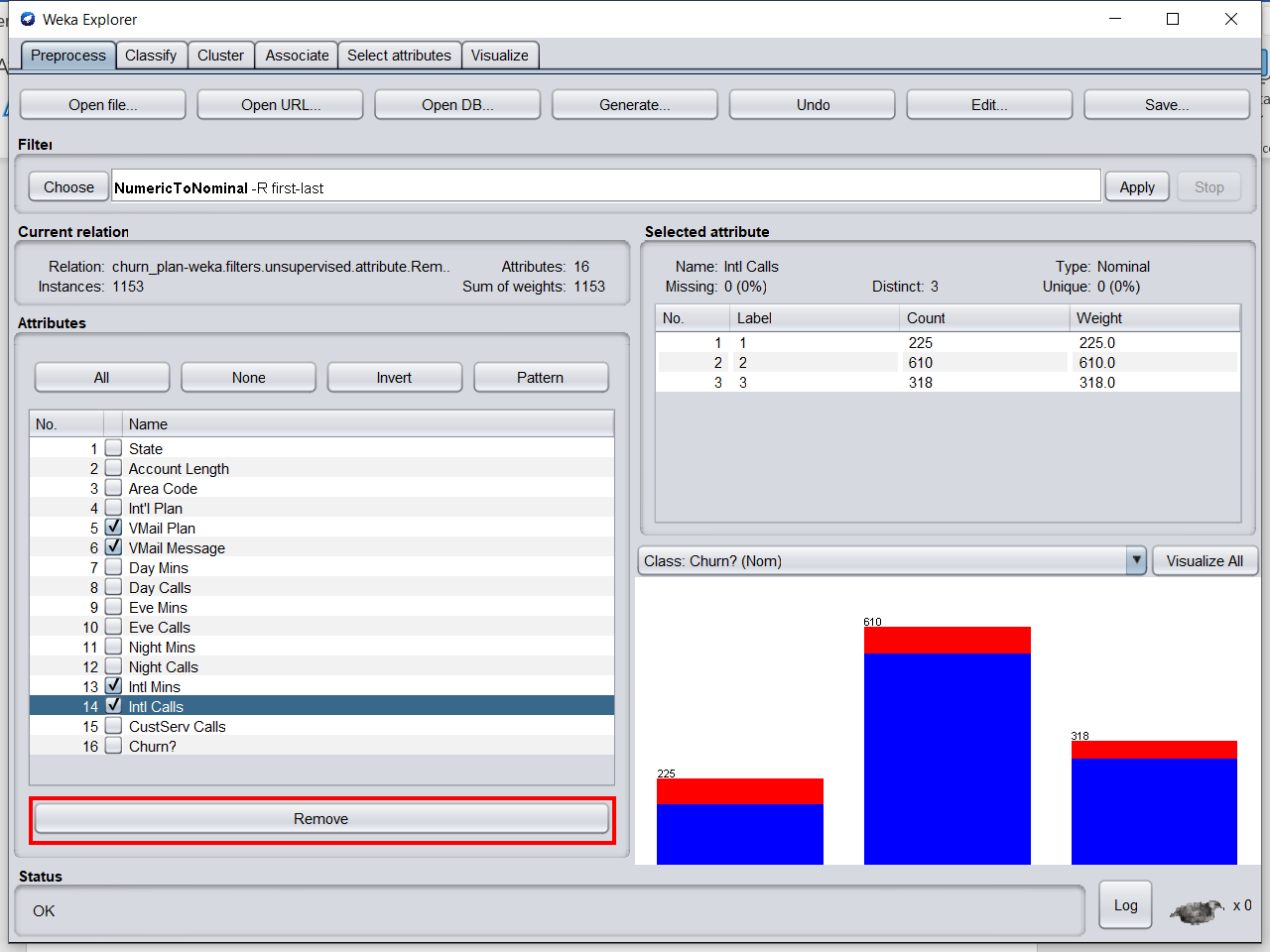
*10. VMail Plan=yes 922 ==> Int'l Plan=no Churn?=False. 786 <conf:(0.85)> lift:(1.25) lev:(0.14) [157] conv:(2.14)*

*12. Churn?=False. 972 ==> Int'l Plan=no VMail Plan=yes 786 <conf:(0.81)> lift:(1.12) lev:(0.07) [86] conv:(1.46)*

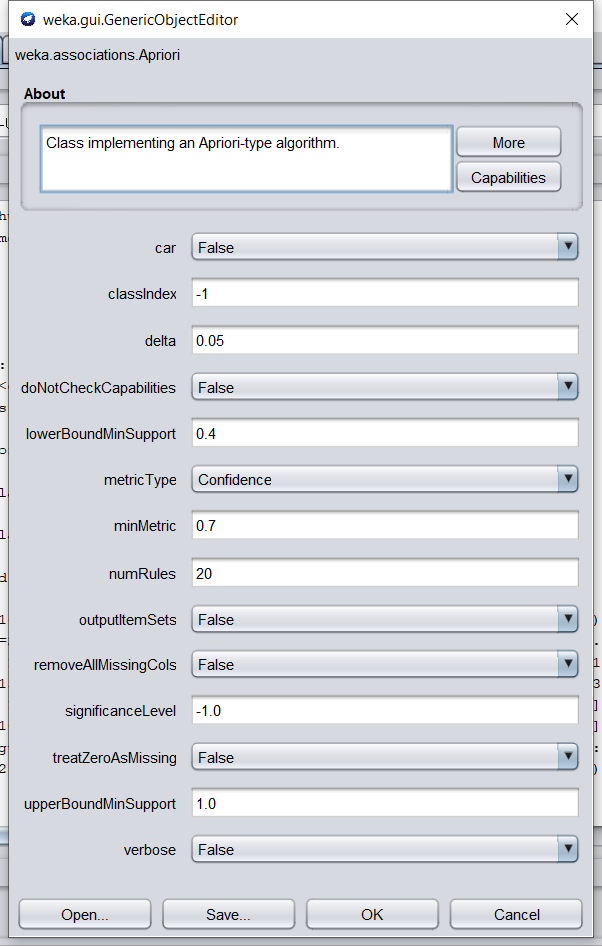
* 2 luật này càng củng cố thêm lập luận rằng : Người dùng Vmail thường không dùng Intl Plan và thường không hủy gói.

### Thử nghiệm 4: xét thói quen gọi của khách hàng.

* **Thể hiện dữ liệu:**
* Ta tập trung xét thói quen nghe gọi của khách hàng để rút ra các luật. Tức trong thử nghiệm này ta không xét các thuộc tính Vmail Message, Vmail Plan, Intl Plan, Int’l Call, Int’l Mins.
* **Tiền xử lý dữ liệu:**

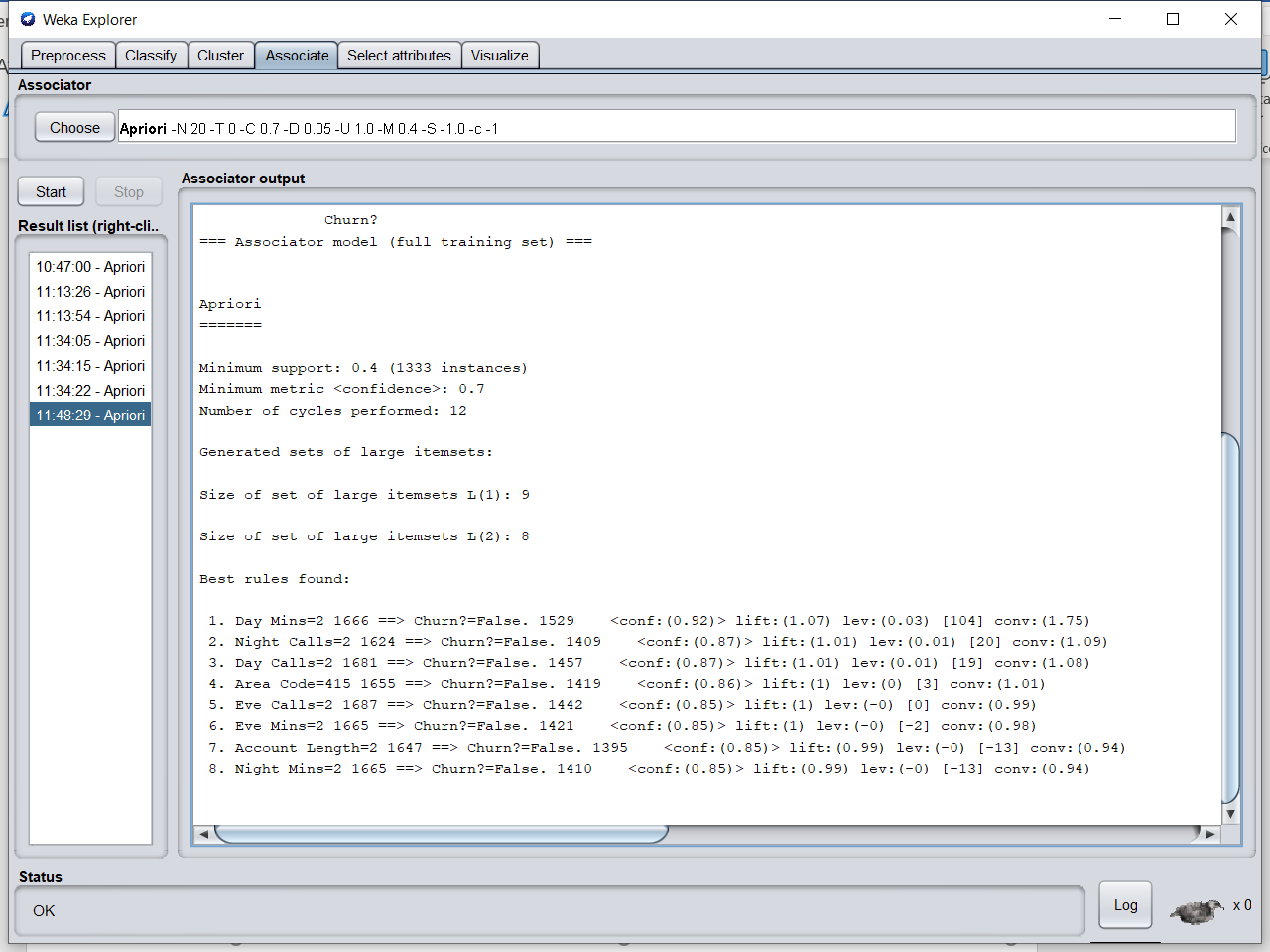
Remove các cột Vmail Message, Vmail Plan, Intl Plan, Int’l Call, Int’l Mins bằng Weka:

Bảng 3.11. Xóa các thuộc tính trong weka.

* **Thông số hệ thống:**

Hình 3.12. Bảng thông số hệ thống.

**Ý nghĩa:** Ta chọn min Support là 0.4 vì số lượng các khoảng của các thuộc tính là 3, lấy cận trên của 1/3 để rút được những luật có ý nghĩa chứ không phải biến ngẫu nhiên, min Confidence là 0.7 các luật đáng tin cậy.

* Rút ra được các luật:

Bảng 3.13. Sử dụng thuật toán Apiori để rút ra các tập luật

* **Luật 1:**

*1. Day Mins=2 1666 ==> Churn?=False. 1529 <conf:(0.92)> lift:(1.07) lev:(0.03) [104] conv:(1.75)*

* + - Khách hàng gọi vào ban ngày với thời gian trung bình thường không hủy thuê bao. Có thể coi đây là dạng khách hàng cơ bản bền vững.
* **Luật 2,3:**

*2. Night Calls=2 1624 ==> Churn?=False. 1409 <conf:(0.87)> lift:(1.01) lev:(0.01) [20] conv:(1.09)*

*3. Day Calls=2 1681 ==> Churn?=False. 1457 <conf:(0.87)> lift:(1.01) lev:(0.01) [19] conv:(1.08)*

* Khách hàng có số cuộc gọi trung bình vào buổi ngày và buổi đêm ít hủy thuê bao.
* **Luật 4:**

*4. Area Code=415 1655 ==> Churn?=False. 1419 <conf:(0.86)> lift:(1) lev:(0) [3] conv:(1.01)*

* Nhóm khách hàng có mã vùng 415 ít hủy thuê bao.
* **Kết hợp 2 luật 5,6:**

*5. Eve Calls=2 1687 ==> Churn?=False. 1442 <conf:(0.85)> lift:(1) lev:(-0) [0] conv:(0.99)*

*6. Eve Mins=2 1665 ==> Churn?=False. 1421 <conf:(0.85)> lift:(1) lev:(-0) [-2] conv:(0.98)*

* Những khách hàng có số cuộc buổi tối trung bình hoặc số phút gọi buổi tối trung bình ít hủy thuê bao.
* **Luật 7:**

*7. Account Length=2 1647 ==> Churn?=False. 1395 <conf:(0.85)> lift:(0.99) lev:(-0) [-13] conv:(0.94)*

* Những khách hàng có đã sử dụng thuê bao với thời gian trung bình thường ít hủy gói.
* **Luật 8:**

*8. Night Mins=2 1665 ==> Churn?=False. 1410 <conf:(0.85)> lift:(0.99) lev:(-0) [-13] conv:(0.94)*

* Khách hàng có thời gian gọi buổi đêm trung bình thường không hủy gói.

# PHẦN IV. TÓM TẮT KẾT QUẢ

## Tiêu chí đánh giá kết quả:

Dựa vào độ min support và min confidence để xét tính đúng đắn của luật.

Ngoài ra còn phải loại bỏ các luật mà có các thuộc tính có thể tự suy diễn ra nhau, như là:

*1. VMail Message=0 2411 ==> VMail Plan=no 2411 <conf:(1)> lift:(1.38) lev:(0.2) [666] conv:(666.95)*

*2. VMail Plan=no 2411 ==> VMail Message=0 2411 <conf:(1)> lift:(1.38) lev:(0.2) [666] conv:(666.95)*

Vì nếu không đăng ký Vmail Plan thì Vmail Message sẽ bằng 0 hoặc ngược lại.

## 1 số luật tốt rút ra được:

Đối với tập dữ liệu tổng thể:

*Churn?=False. 2850 ==> Int'l Plan=no 2664 <conf:(0.93)> lift:(1.04) lev:(0.03) [90] conv:(1.48)*

* Khách hàng không hủy thuê bao thì thường không dùng dịch vụ quốc tế.

🡪 Cần cải thiện chất lượng dịch vụ quốc tế để tránh việc khách hàng hủy thuê bao.

*Int'l Plan=no VMail Message=0 Churn?=False. 1878 ==> VMail Plan=no 1878 <conf:(1)> lift:(1.38) lev:(0.16) [519] conv:(519.51)*

-khách hàng không dùng cả 2 dịch vụ quốc tế và thư thoại thường không hủy thuê bao

🡪 có thể phân lớp những khách hàng này là khách hàng đơn giản, dễ đáp ứng.

Đối với những khách hàng hủy thuê bao:

*Day Mins=3 244 ==> VMail Plan=no 226 <conf:(0.93)> lift:(1.11) lev:(0.05) [22] conv:(2.13)*

*VMail Plan=no 403 ==> Day Mins=3 226 <conf:(0.56)> lift:(1.11) lev:(0.05) [22] conv:(1.12)*

-Một khách hàng hủy thuê bao thường gọi nhiều vào ban ngày và không đăng ký Vmail Plan

*Int'l Plan=no 346 ==> VMail Plan=no 302 <conf:(0.87)> lift:(1.05) lev:(0.03) [13] conv:(1.27)*

*VMail Plan=no 403 ==> Int'l Plan=no 302 <conf:(0.75)> lift:(1.05) lev:(0.03) [13] conv:(1.12)*

* Khách hàng huỷ thuê bao cũng thường không có Vmail Plan và Intl Plan.

🡪 Giới thiệu Vmail Plan đối với những khách hàng chưa dùng để tăng trải nghiệm nhằm níu giữ khách hàng tốt hơn.

* **Đối với khách hàng sử dụng 1 trong 2 dịch vụ Vmail Plan, Int’l Plan hoặc cả 2:**

VMail Plan=yes 922 ==> Churn?=False. 842 <conf:(0.91)> lift:(1.08) lev:(0.06) [64] conv:(1.79)

*Churn?=False. 972 ==> VMail Plan=yes 842 <conf:(0.87)> lift:(1.08) lev:(0.06) [64] conv:(1.49*)

- Người dùng Vmail Plan thường không hủy thuê bao

*VMail Plan=yes 922 ==> Int'l Plan=no Churn?=False. 786 <conf:(0.85)> lift:(1.25) lev:(0.14) [157] conv:(2.14)*

*Churn?=False. 972 ==> Int'l Plan=no VMail Plan=yes 786 <conf:(0.81)> lift:(1.12) lev:(0.07) [86] conv:(1.46)*

-Người dùng Vmail Plan và không dùng Intl Plan thường không hủy thuê bao

* **Đối với thói quen nghe gọi của khách hàng:**

*Day Mins=2 1666 ==> Churn?=False. 1529 <conf:(0.92)> lift:(1.07) lev:(0.03) [104] conv:(1.75)*

-Khách hàng gọi vào ban ngày với thời gian trung bình thường không hủy thuê bao

*Night Calls=2 1624 ==> Churn?=False. 1409 <conf:(0.87)> lift:(1.01) lev:(0.01) [20] conv:(1.09)*

*Day Calls=2 1681 ==> Churn?=False. 1457 <conf:(0.87)> lift:(1.01) lev:(0.01) [19] conv:(1.08)*

-Khách hàng có số cuộc gọi trung bình vào ban ngày hoặc ban đêm ít hủy thuê bao

*Eve Calls=2 1687 ==> Churn?=False. 1442 <conf:(0.85)> lift:(1) lev:(-0) [0] conv:(0.99)*

*Eve Mins=2 1665 ==> Churn?=False. 1421 <conf:(0.85)> lift:(1) lev:(-0) [-2] conv:(0.98)*

- Khách hàng có số cuộc buổi tối trung bình hoặc số phút gọi buổi tối trung bình ít hủy thuê bao

*Account Length=2 1647 ==> Churn?=False. 1395 <conf:(0.85)> lift:(0.99) lev:(-0) [-13] conv:(0.94)*

-khách hàng có đã sử dụng thuê bao với thời gian trung bình thường ít hủy gói

*Night Mins=2 1665 ==> Churn?=False. 1410 <conf:(0.85)> lift:(0.99) lev:(-0) [-13] conv:(0.94)*

* Khách hàng có thời gian gọi buổi đêm trung bình thường không hủy gói.

🡪 Có thể coi đây là tập khách hàng bền vững, có nhu cầu bình thường 🡪 Hãng đáp ứng tốt cho nhu cầu gọi bình thường.

*Area Code=415 1655 ==> Churn?=False. 1419 <conf:(0.86)> lift:(1) lev:(0) [3] conv:(1.01)*

* Nhóm khách hàng có mã vùng 415 ít hủy thuê bao.

🡪 Có thể đây là bằng chứng cho việc chi nhánh ở khu vực có mã vùng 415 hỗ trợ khách hàng tốt khi tỉ lệ hủy gói ở đây thấp.

## Điểm mạnh và điểm yếu:

### Điểm mạnh:

* Bài tập có sử dụng nhiều kiến thức miền (domain knowledge) thích hợp cho các sinh viên có domain knowledge mạnh. Việc biết được mô hình kinh doanh của một hãng viễn thông giúp cho sinh viên có thể phân nhóm được các dịch vụ, từ đó khám phá dữ liệu tốt hơn.
* Bài tập tập trung vào khả năng tự tìm hiểu, tra cứu tài liệu cũng là một điểm mạnh khác của nhóm.

### Điểm yếu:

* Do việc chưa làm quen được với ngôn ngữ lập trình java cũng như bộ thư viện của Weka khiến cho nhóm có hạn chế một số phần trong công tác xử lý dữ liệu.

# PHẦN V. TÀI LIỆU THAM KHẢO

* Tài liệu giáo trình Khai thác dữ liệu và ứng dụng.
* <https://docs.weka.io/>
* <https://pandas.pydata.org/docs/>