**Data**

**1.Tìm hiểu Data**

Số lượng instances: 3333

Số lượng attribute: 21

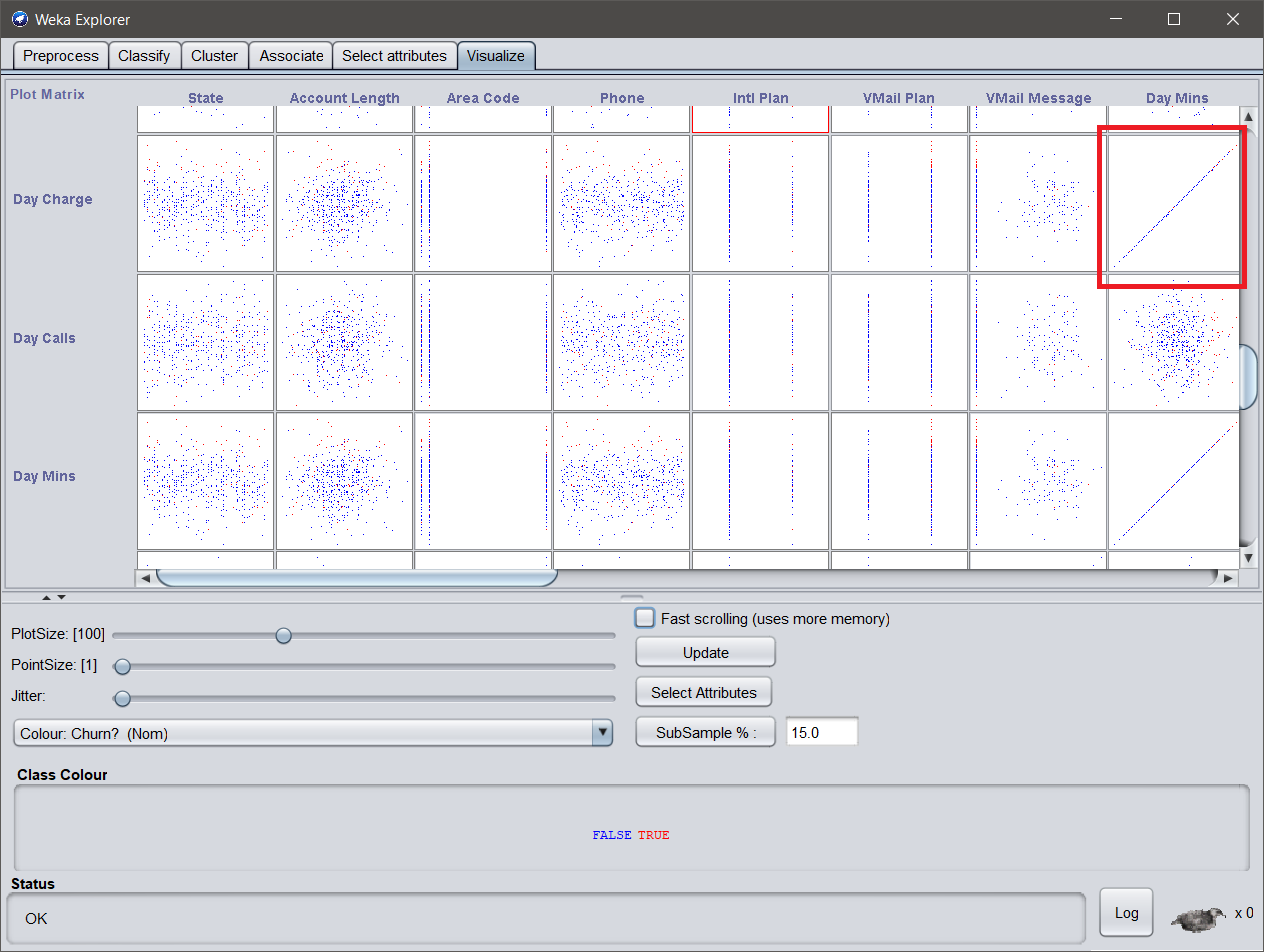
Class attribute: Churn

Ý nghĩa các attribute:

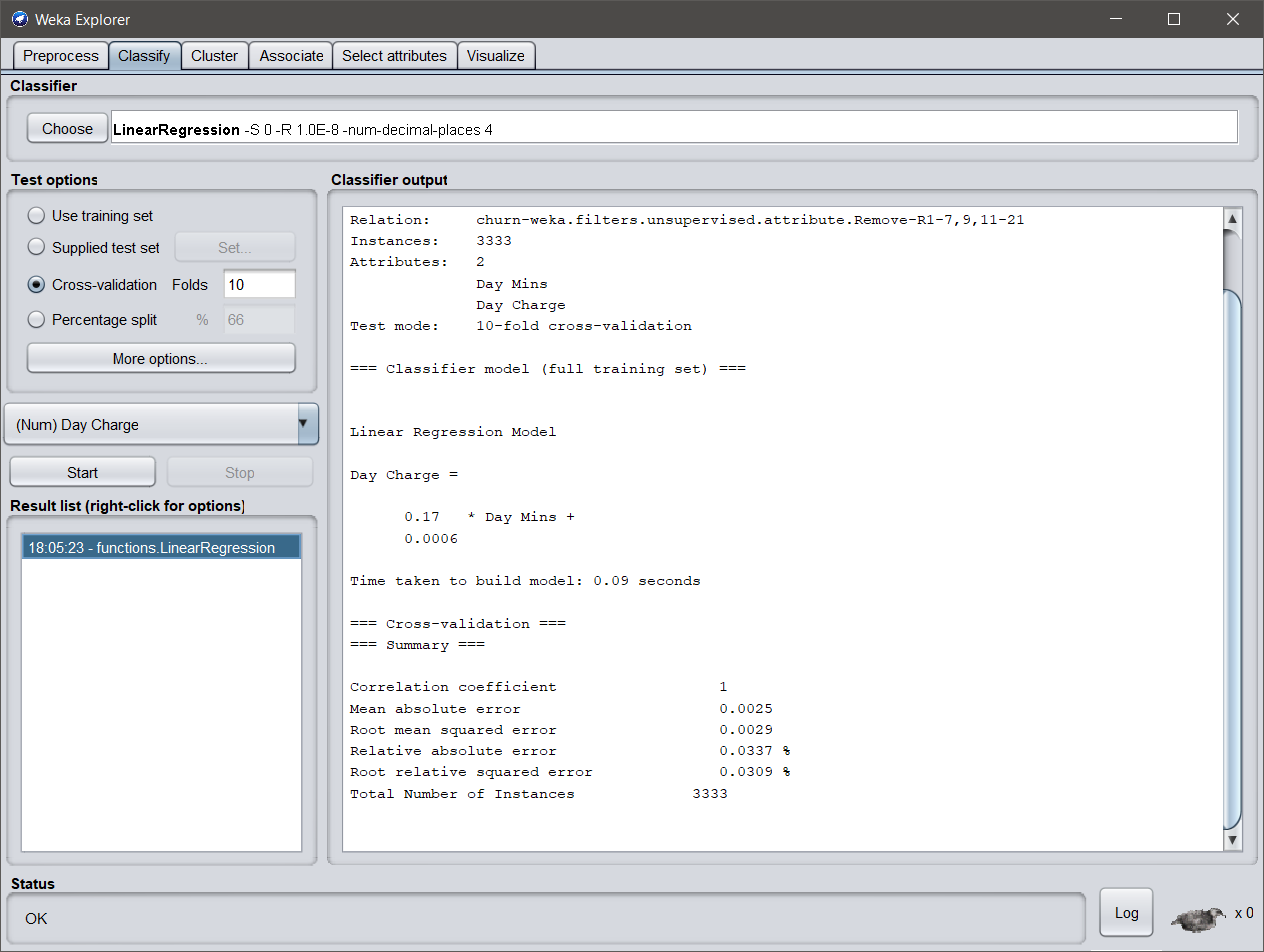
1. **State:** tên các bang, quận ở Columbia (nom)
2. **Account length:** thời gian kích hoạt của tài khoản (num)
3. **Area code:** mã vùng (nom)
4. **Phone number:** số điện thoại, có thể thay thế cho ID khách hàng (num)
5. **International Plan:** dịch vụ cuộc gọi quốc tế (Y/N)
6. **VoiceMail Plan:** dịch vụ thư thoại (Y/N)
7. **Number of voice mail messages:** số tin nhắn thoại (num)
8. **Total day minutes:** tổng thời gian sử dụng vào ban ngày
9. **Total day calls:** tổng số cuộc gọi vào ban ngày
10. **Total day charge**: tổng cước vào ban ngày, dựa vào 2 biến trên
11. **Total evening minutes**: tổng thời gian sử dụng vào buổi tối
12. **Total evening calls**: tổng số cuộc gọi vào buổi tối
13. **Total evening charge**: tổng cước vào buổi tối, dựa vào 2 biến trên
14. **Total night minutes**: tổng thời gian sử dụng vào ban đêm
15. **Total night calls**: tổng số cuộc gọi vào ban đêm
16. **Total night charge**: tổng cước vào ban đêm, dựa vào 2 biến trên
17. **Total international minutes**: tổng thời gian gọi quốc tế
18. **Total international calls**: tổng số cuộc gọi quốc tế
19. **Total international charge**: tổng cước quốc tế, dựa vào 2 biến trên
20. **Number of calls to customer service**: số cuộc gọi chăm sóc khách hàng
21. **Churn:** Có hủy dịch vụ hay không

**2.Tiền xử lý**

**Loại bỏ các thuộc tính tương quan nhau:**

****

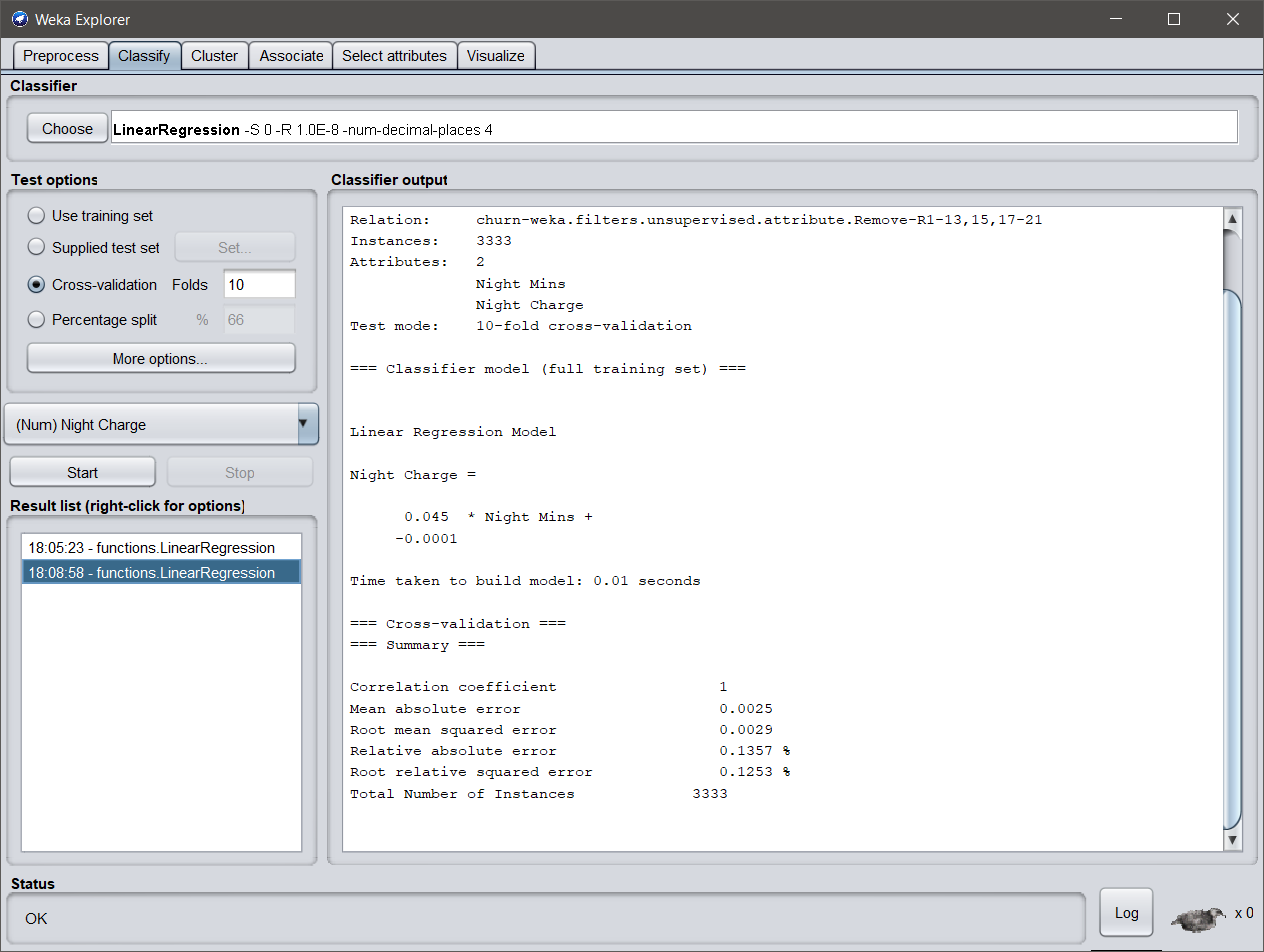
Tab Visualize cho ta thấy có vẻ như Data Mins va Data Charge có liên quan đến nhau vì vậy, ta sử dụng Classification>Choose>Function>Linear Regression để kiểm tra

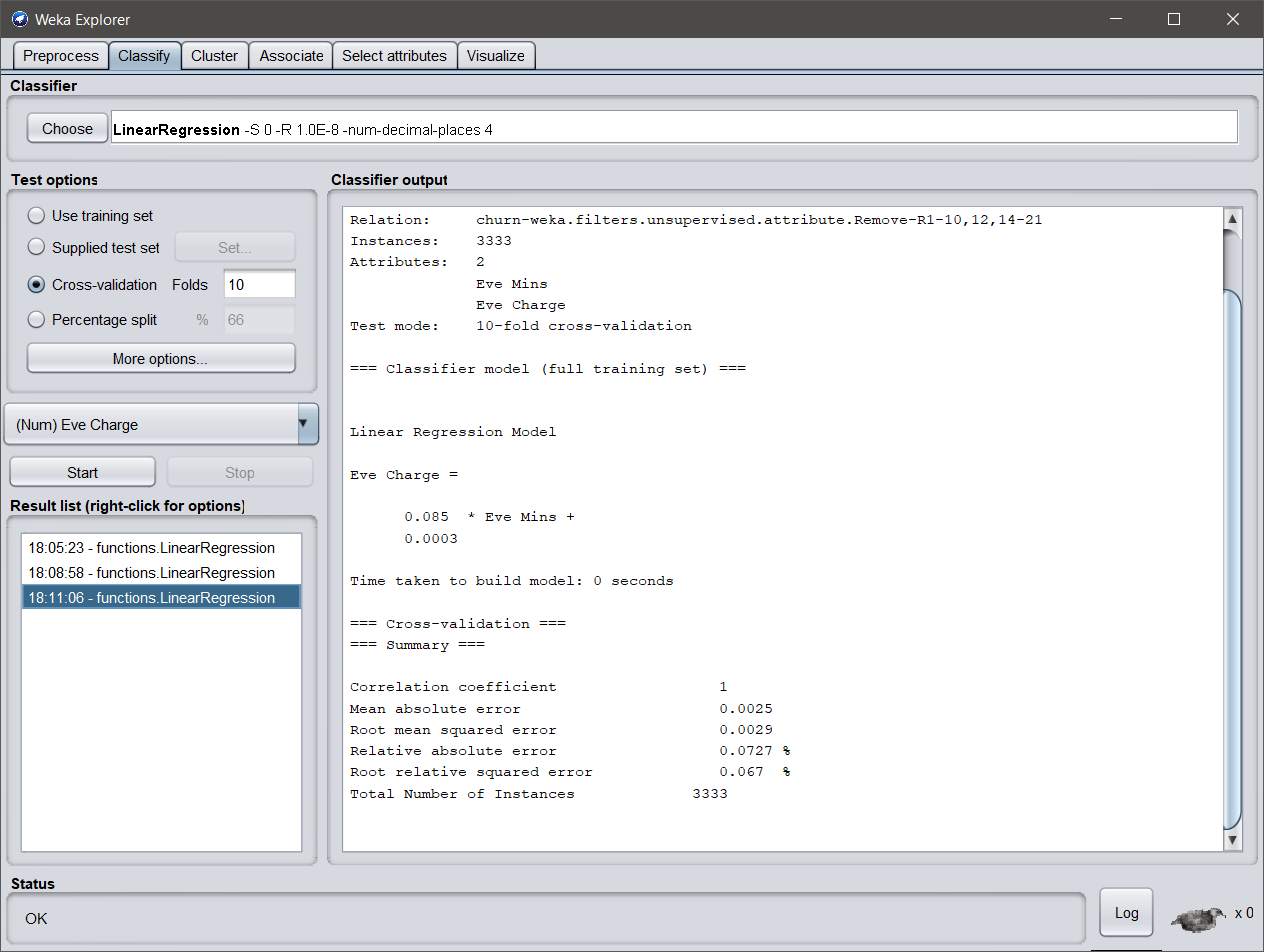


Như vậy, Day Charge phụ thuộc vào Day Mins theo hàm

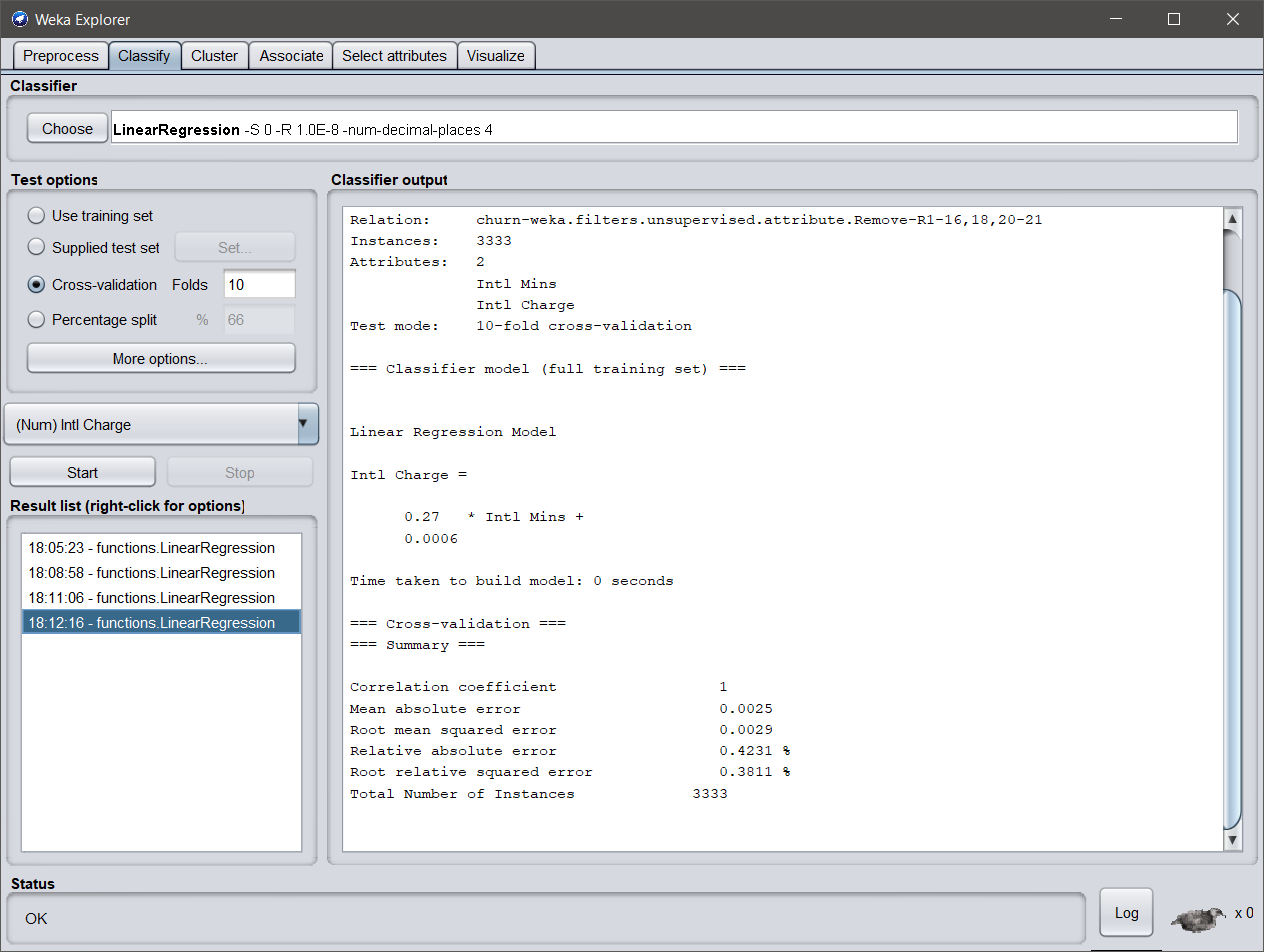
Day Charge = 0.17 \* Day Mins +0.0006

Tương tự

 Night Charge = 0.045 \* Night Mins - 0.0001



Eve Charge = 0.085 \* Eve Mins + 0.0003



Intl Charge = 0.27 \* Intl Mins + 0.0006

Việc giữ lại các thuộc tính tương quan sẽ:

Khiến một thuộc tính nào đó bị nhấn mạnh quá mức

Khiến mô hình không ổn đỉnh và cho ra những kết quá không tin cậy

Như vậy, ta có thể xóa các 4 thuộc tính Day, Night Charge, Eve Charge, Intl Charge.

**Loại bỏ các thuộc tính dị thường:**

Thuộc tính State có vẻ như là viết tắt của 51 bang của Mỹ, nhưng thuộc tính Area Code thì chỉ có 3 giá trị: 408, 415, 510 đều là Area Code thuộc California.

Có 2 cách giải quyết:

Xóa

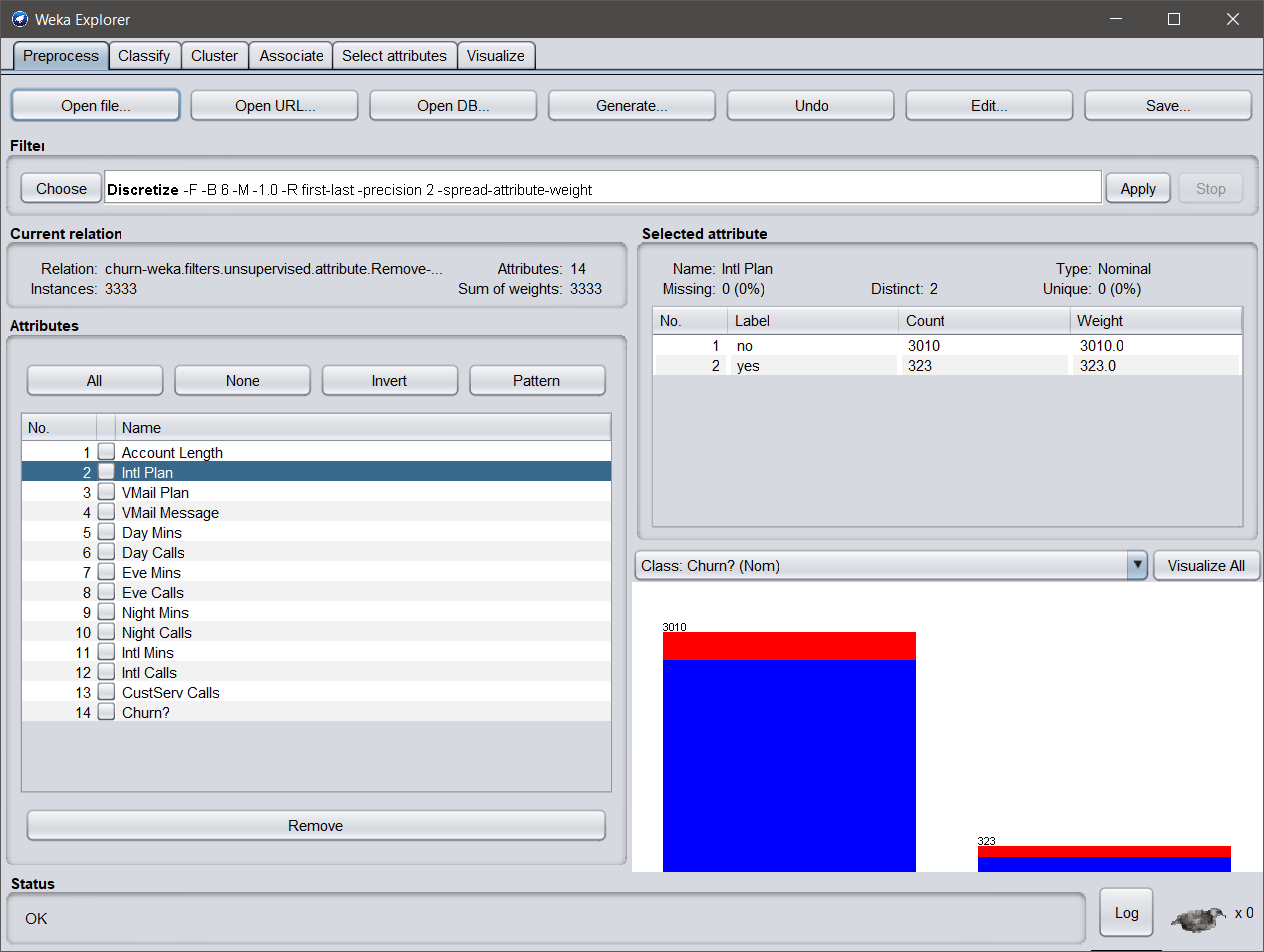
Tham khảo ý kiến của các chuyên gia hoặc người có kiến thức thuộc lĩnh vực này để hiểu được ý nghĩa của thuộc tính

**Loại bỏ thuộc tính Index**

Phone (số điện thoại): có thể được xem như thuộc tính Index, loại bỏ

**Khám phá dữ liệu kiểu nominal**

Các dữ liệu kiểu nominal:

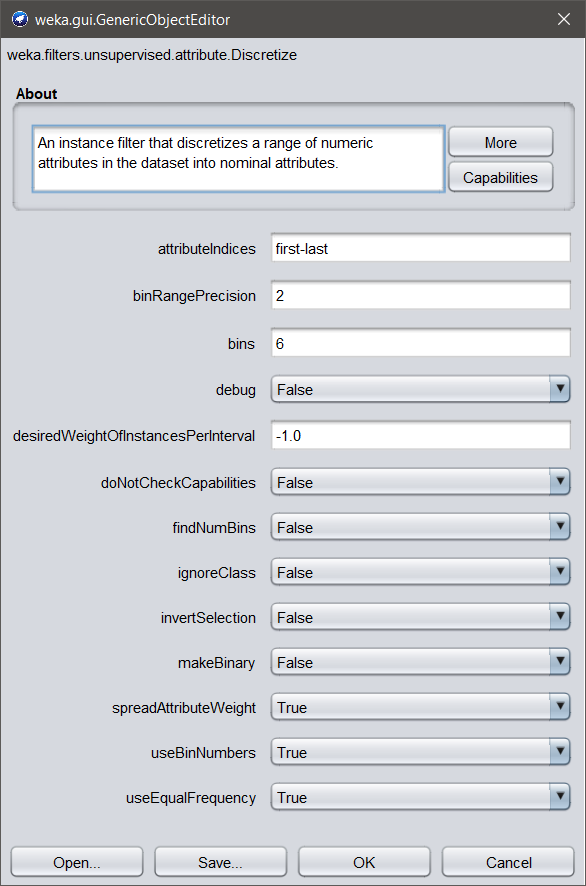


Khách hàng sử dụng Intl Plan tuy ít hơn (323 vs 3010) nhưng có tỉ lệ hủy dịch vụ cao nhiều lần so với khách hàng không sử dụng

**Khám phá dữ liệu kiểu numeric**

Đưa về nominal -> Rời rạc hóa

Sử dụng Filter> unsupervised>Attributes>Discretize

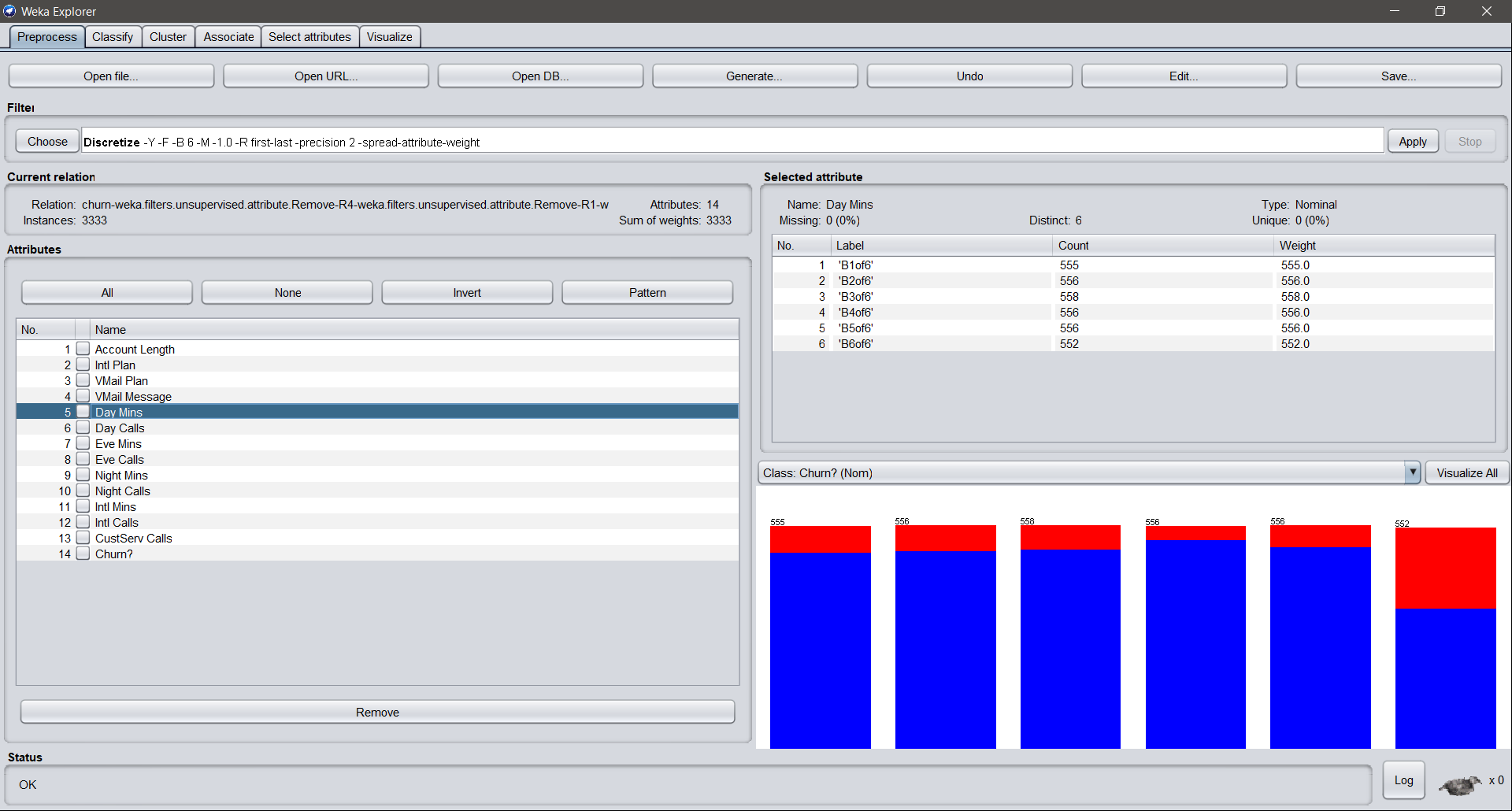


Ta chú ý

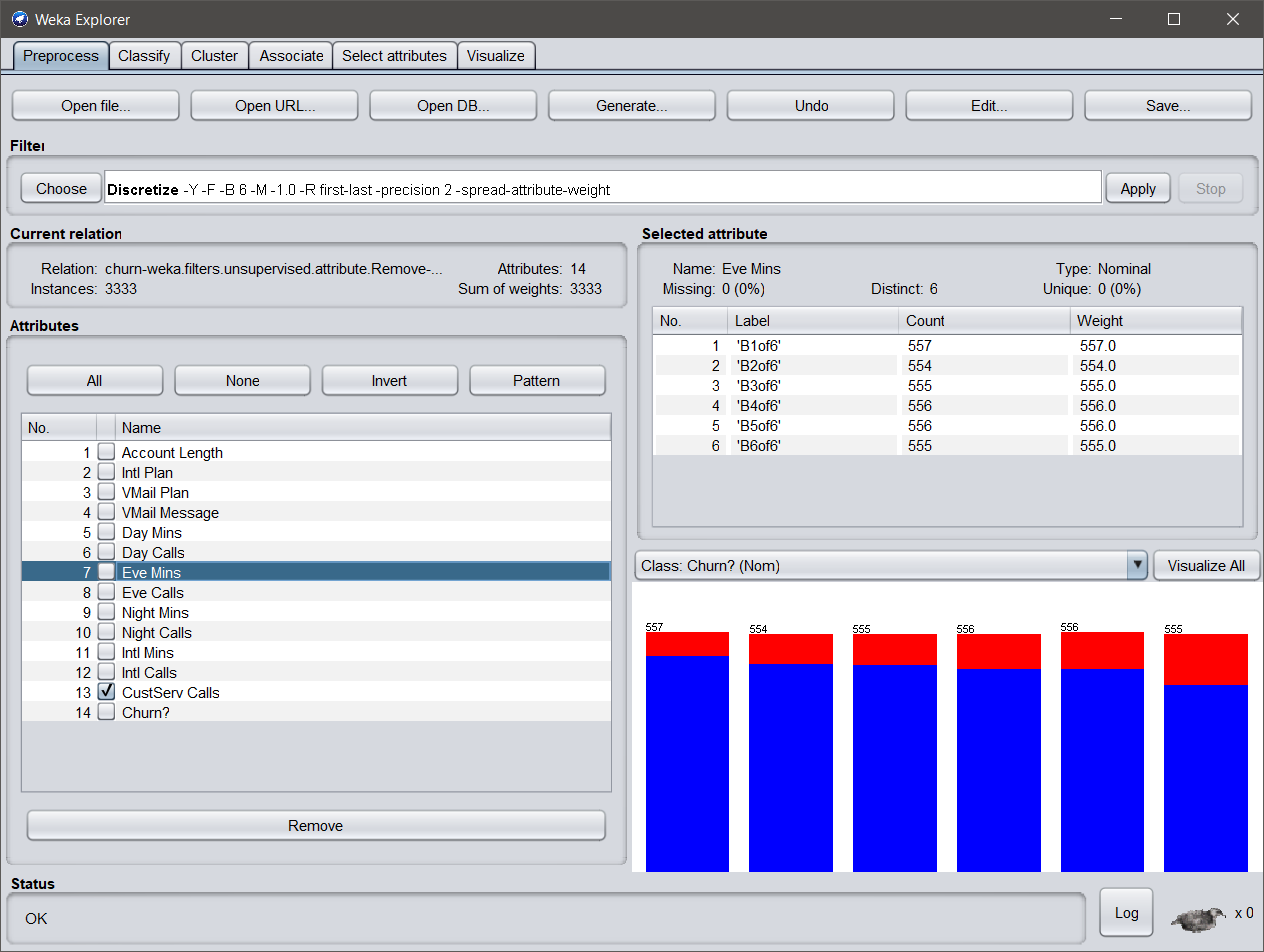
Bins: dữ liệu được rời rạc thành mấy phần

useEqualFrequency: thay vì chia miền giá trị thành các phần bằng nhau (False) thì chia miền giá trị thành các phần có số mẫu bằng nhau(True)

Sau áp Filter các cột Numeric được chuyển về thành Nominal, tham khảo 1 số cột, ta phát hiện



Khách hàng có Day Mins càng cao có tỉ lệ hủy dịch vụ đáng kể

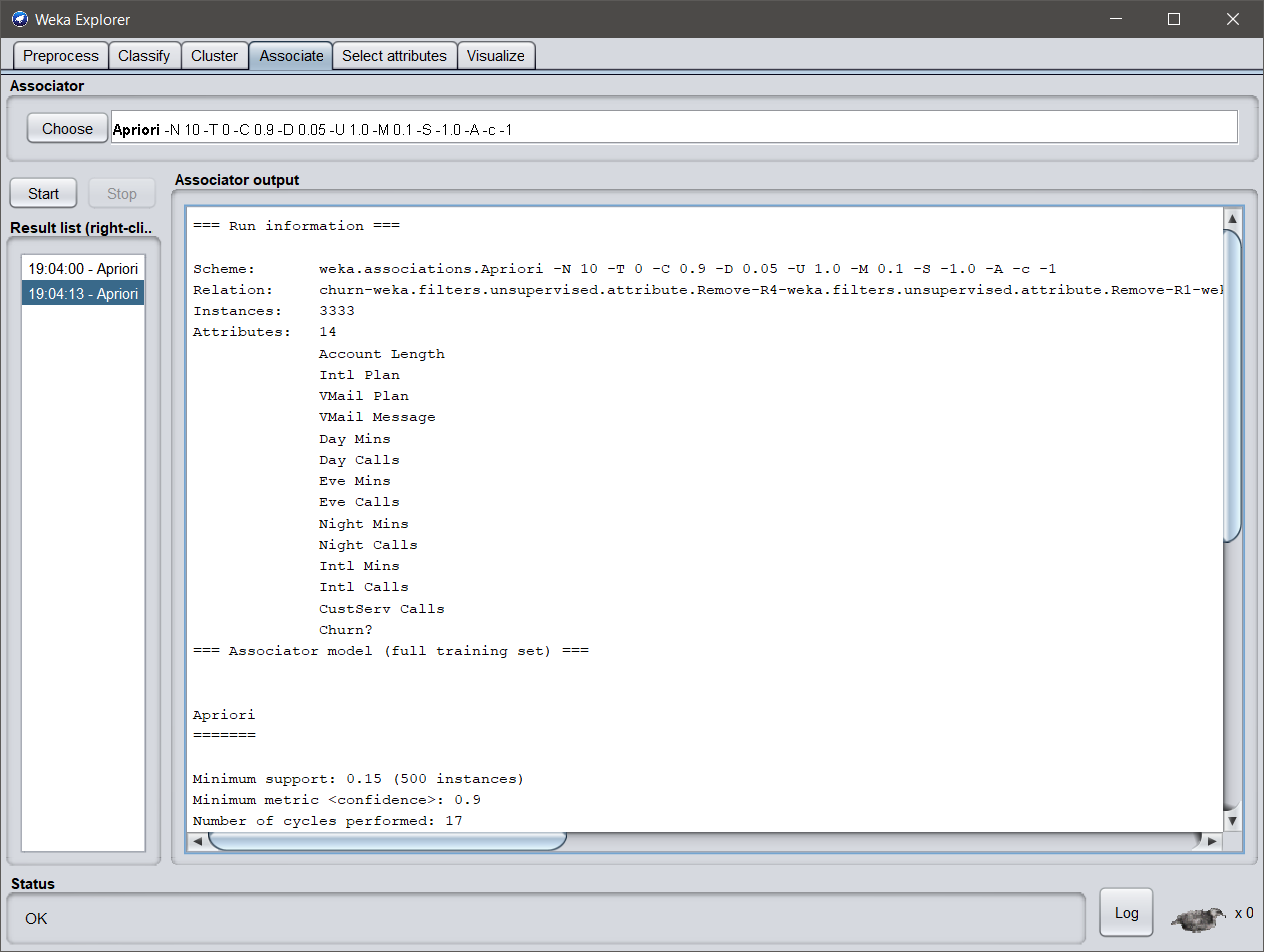


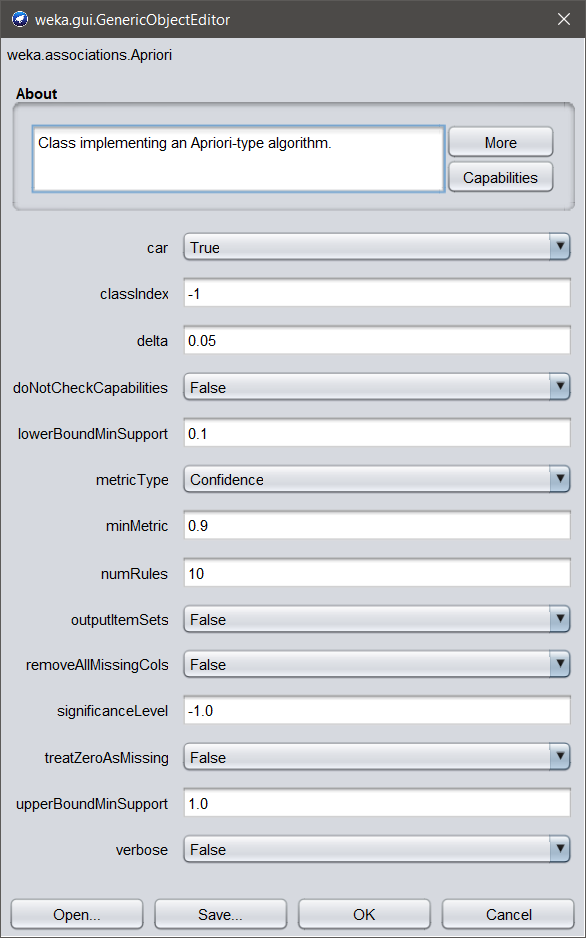
Eve Mins tăng thì tỉ lệ huy dịch vụ cũng tăng nhẹ theo

**Khai thác luật kết hợp:**

Sử dụng tab Associate

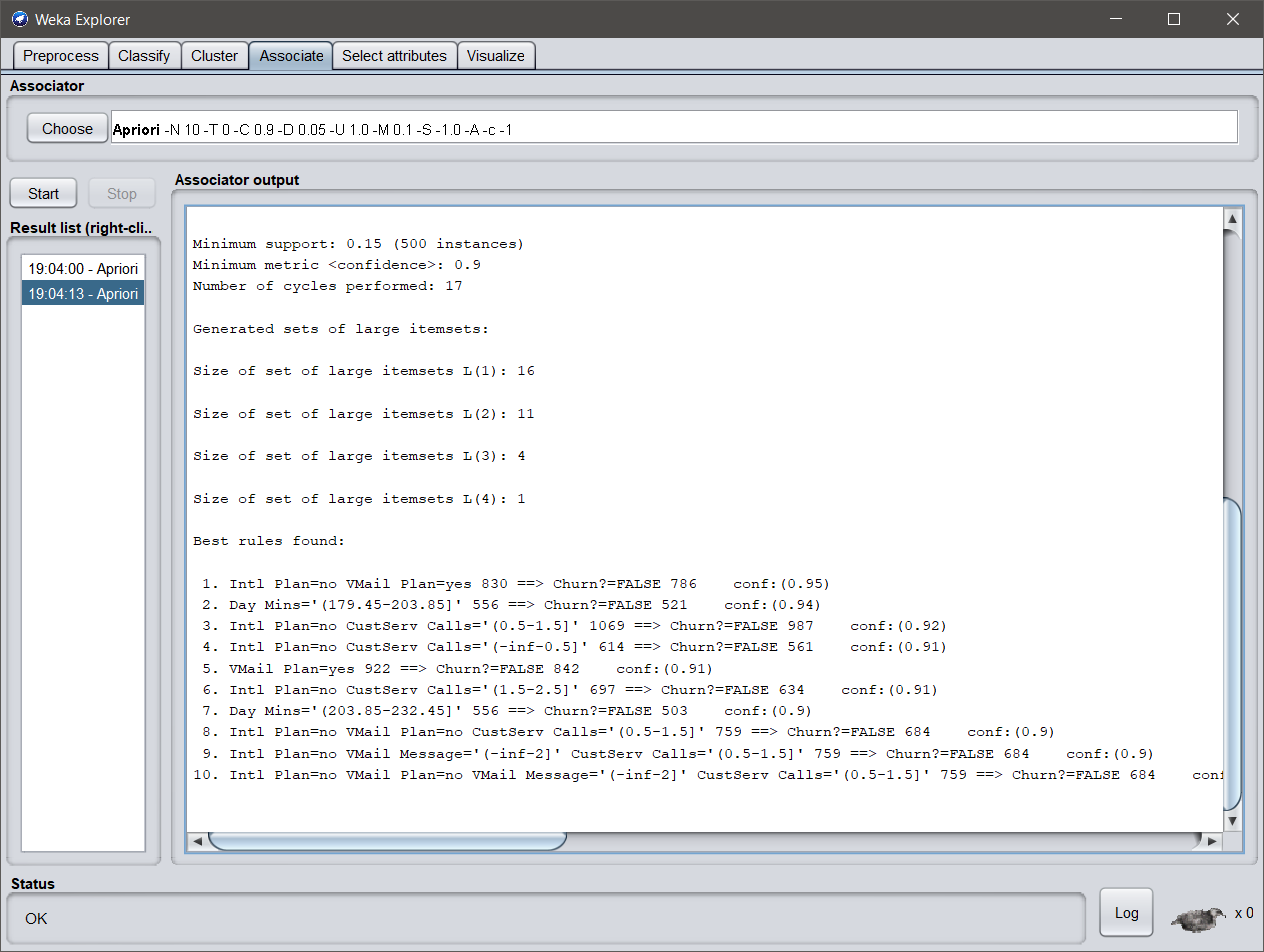
Chọn thuật toán Apriori





Các thông số của thuật toán

|  |  |
| --- | --- |
| **minMetric** | Minimum metric score. Consider only rules with scores higher than this value.  Mức tối thiểu mà luật được chấp nhận |
| verbose | If enabled the algorithm will be run in verbose mode. |
| numRules | Number of rules to find. |
| **lowerBoundMinSupport** | Lower bound for minimum support.(Độ support sẽ giảm dần về mức này) |
| classIndex | Index of the class attribute. If set to -1, the last attribute is taken as class attribute. |
| outputItemSets | If enabled the itemsets are output as well |
| **car** | If enabled class association rules are mined instead of (general) association rules.  Nếu chọn thì luật kết hợp sẽ bao gồm thuộc tính class (churn) |
| doNotCheckCapabilities | If set, associator capabilities are not checked before associator is built (Use with caution to reduce runtime) |
| removeAllMissingCols | Remove columns with all missing values |
| significanceLevel | Significance level. Significance test (confidence metric only). |
| treatZeroAsMissing | If enabled, zero (that is, the first value of a nominal) is treated in the same way as a missing value |
| delta | Iteratively decrease support by this factor. Reduces support until min support is reached or required number of rules has been generated. |
| **metricType** | Set the type of metric by which to rank rules. **Confidence is the proportion of the examples covered by the premise that are also covered by the consequence (Class association rules can only be mined using confidence).** Lift is confidence divided by the proportion of all examples that are covered by the consequence. This is a measure of the importance of the association that is independent of support. Leverage is the proportion of additional examples covered by both the premise and consequence above those expected if the premise and consequence were independent of each other. The total number of examples that this represents is presented in brackets following the leverage. Conviction is another measure of departure from independence. Conviction is given by P(premise)P(!consequence) / P(premise, !consequence). |
| **upperBoundMinSupport** | Upper bound for minimum support. Start iteratively decreasing minimum support from this value. Độ support sẽ giảm dần từ mức này) |



Bên dưới rút cho ta được 10 luật với min metric 90%

Có 16 bộ L(1) (bộ phổ biến có độ dài là 1)

Có 11 bộ L(2) (bộ phổ biến có độ dài là 2)

…

Có 1 bộ L(4) (bộ phổ biến có độ dài là 4)