ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH
ĐẠI HỌC KHOA HỌC TỰ NHIÊN
KHOA CÔNG NGHỆ THÔNG TIN

DATA MINING

LAB 2: ASSOCIATION RULES

Lớp: KTDL & UD 18_21
Nhóm thực hiện:
18120078 – Ngô Phù Hữu Đại Sơn
18120253 – Mai Ngọc Tú

MỤC LỤC

I. THÔNG TIN KHÁI QUÁT

I. Thông tin nhóm

MSSV	Họ tên	Vai trò
18120078	Ngô Phù Hữu Đại Sơn	Nhóm trưởng
18120253	Mai Ngọc Tú	Thành viên

II. Bảng phân công công việc

MSSV	Công việc phụ trách
18120253	Làm sạch dữ liệu
18120078	Định nghĩa phân cấp và cài đặt mã nguồn
18120078	Khai thác luật kết hợp bằng thuật toán Apriori
18120253	Các luật có ích từ kết quả
18120078	Rút trích tri thức từ các luật

III. Github:

https://github.com/IrisStream/Data-Mining

B. NỘI DUNG

I. Mục tiêu của đồ án:

Khai thác luật kết hợp trên tập dữ liệu nhiều thuộc tính. Từ đó, muốn thu được các tập luật nhằm phục vụ cho việc ra quyết định, hiểu rõ hơn về dữ liệu, hay gia tăng lợi nhuận của công ty. Khai thác các luật cho đến khi bạn đạt được một tập luật thỏa mãn mục tiêu đó.

II. Triển khai

1. Dữ liệu

1.1. Mô tả tập dữ liệu:

Thuộc tính phân lớp là *Churn* biểu thị rằng 1 khách hàng có từ bỏ dịch vụ của 1 công ty để chuyển sang 1 công ty khác hay không.

20 thuộc tính còn lại của dữ liệu đầu vào được mô tả như sau:

STT	TÊN THUỘC TÍNH	KIỂU DỮ LIỆU	Ý NGHĨA
1	State	Phân loại	Mã 50 tiểu bang của Columbia
2	Account length	Phân loại	Tài khoản được được kích hoạt bao lâu.
3	Area code	Phân loại	Mã vùng
4	Phone number	Phân loại	Dùng như ID của khách hàng
5	International Plan	Nhị phân	Có tham gia chương trình quốc tế không
6	VoiceMail Plan	Nhị phân	Có tham gia chương trình thư thoại không
7	VMail Messages	Số	Số lượng thư thoại
8	Day Mins	Số	Số phút đã gọi/ngày
9	Day Calls	Số	Số cuộc gọi/ngày
10	Day Charge	Số	Số tiền phải nạp/ngày
11	Eve Mins	Số	Số phút đã gọi/tối
12	Eve Calls	Số	Số cuộc gọi/tối
13	Eve Charge	Số	Số tiên phải nạp/tối
14	Night Mins	Số	Số phút gọi/đêm
15	Night Calls	Số	Số cuộc gọi/đêm
16	Night Charge	Số	Số tiền phải nạp/đêm
17	Intl Mins	Số	Số phút gọi quốc tế
18	Intl Calls	Số	Số cuộc gọi quốc tế
19	Intl Charge	Số	Số tiền phải nạp để gọi quốc tế
20	Serv Cust Calls	Số	Số cuộc gọi chăm sóc dịch vụ khách hàng

1.2. Tiền xử lý

- *Xóa các thuộc tính phụ thuộc vào thuộc tính khác (Tương quan):*

Tồn tại các thuộc tính có tương quan với nhau trong tập dữ liệu khai thác có thể dẫn đến các thuộc tính này bị nhấn mạnh quá mực. Hoặc tệ hơn có thể gây ra lỗi trong quá trình khám phá tri thức.

Ta thấy được thuộc tính *Day Charge* có tương quan với thuộc tính *Day Mins* theo 1 hàm tuyến tính. Dễ nhận thấy là nếu số phút gọi càng nhiều thì số tiền phải trả cũng tăng theo. Từ cơ sở trên, ta có thể loại bỏ thuộc tính *Day Charge* khỏi tập dữ liệu khai thác.

Tương tự với các thuộc tính *Eve Charge, Night Charge* và *Intl Charge*.

Việc xóa các thuộc tính trên 1 phần giúp giảm kích thước không gian dữ liệu (Từ 20 thuộc tính xuống còn 16 thuộc tính. Giúp tăng tốc các thuật toán khai thác.

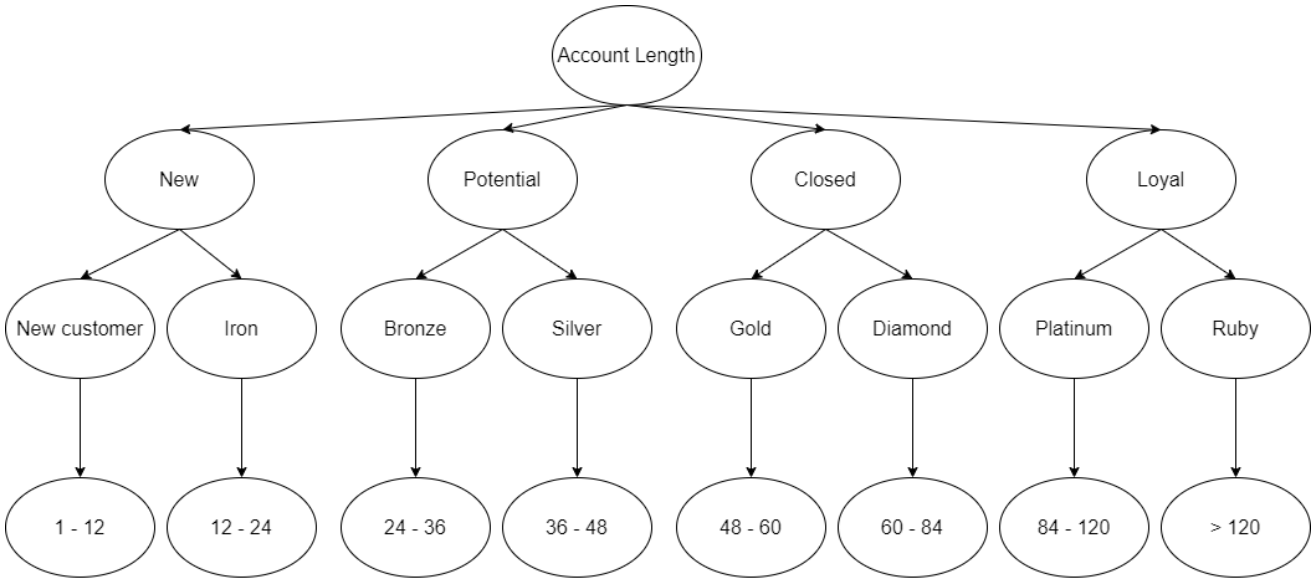
- *Xóa các thuộc tính dị thường:*

Xét thuộc tính *Area code*. Mặc dù thuộc tính chứa số nhưng lại có thể xem như kiểu dữ liệu phân loại, Vì chúng có thể phân loại khách hàng theo vị trí địa lý. Ta có thể thấy rằng *Area code* chỉ gồm 3 giá trị khác nhau là 408, 415 và 510. Cả ba đều thuộc California. Nhưng ta có thể thấy rằng các giá trị này nằm rải rác theo các bang khác nhau của thuộc tính *State* chứ không chỉ thuộc California. Đây có thể là miền dữ liệu bị lỗi. Từ cơ sở này, rất có thể thuộc tính *State* hoặc *Area code* đang bị lỗi. Vì vậy ta có thể xóa 2 thuộc tính này khỏi tập dữ liệu khai thác.

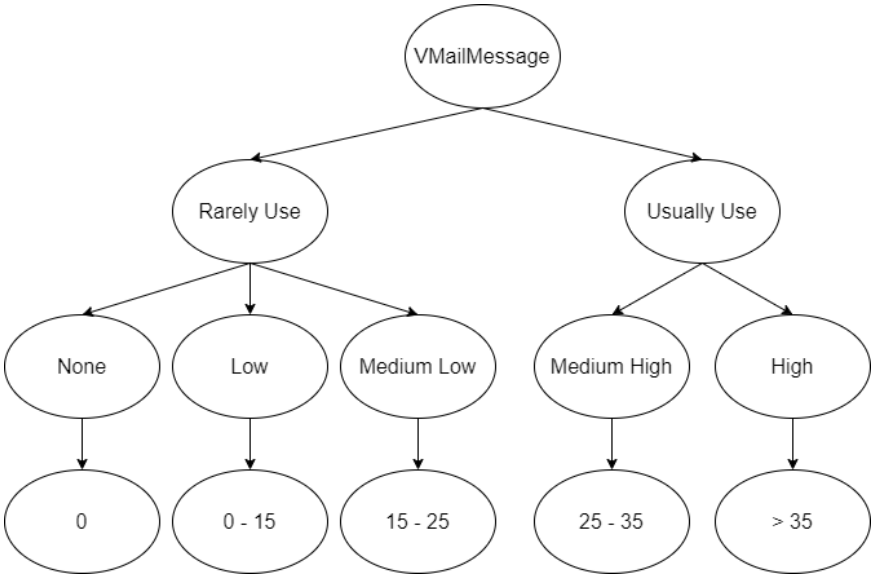
1.3. Các khái niệm phân cấp:

Đối với các thuộc toán khai thác tập phổ biến và luật kết hợp. Ta phải đưa tập dữ liệu về dạng các Transation. Để làm điều đó ta phải chuyển đổi được các thuộc tính kiểu số thành kiểu phân loại. Ta sẽ sử dụng khái niệm phân cấp cho các dữ liệu số.

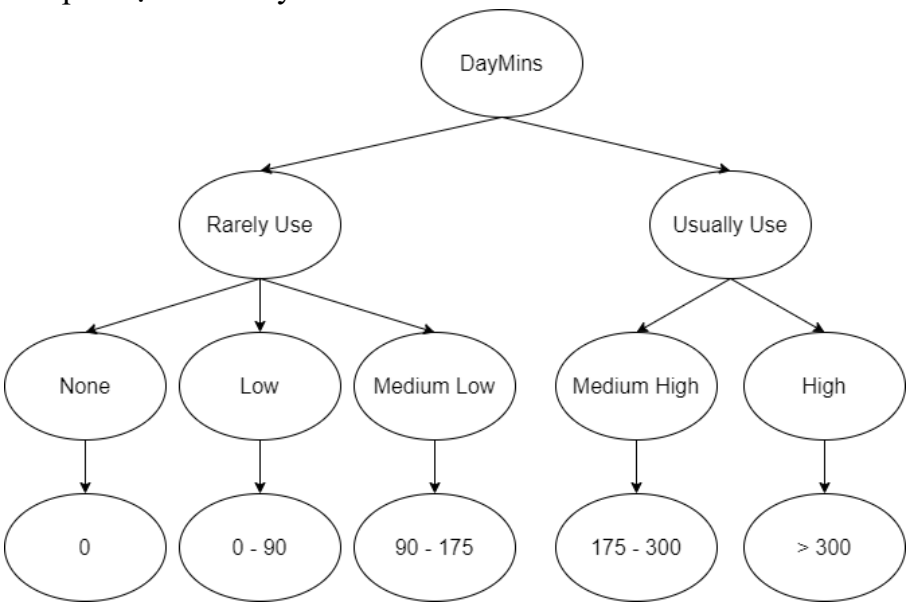
- Cây phân cấp thuộc tính Account length:



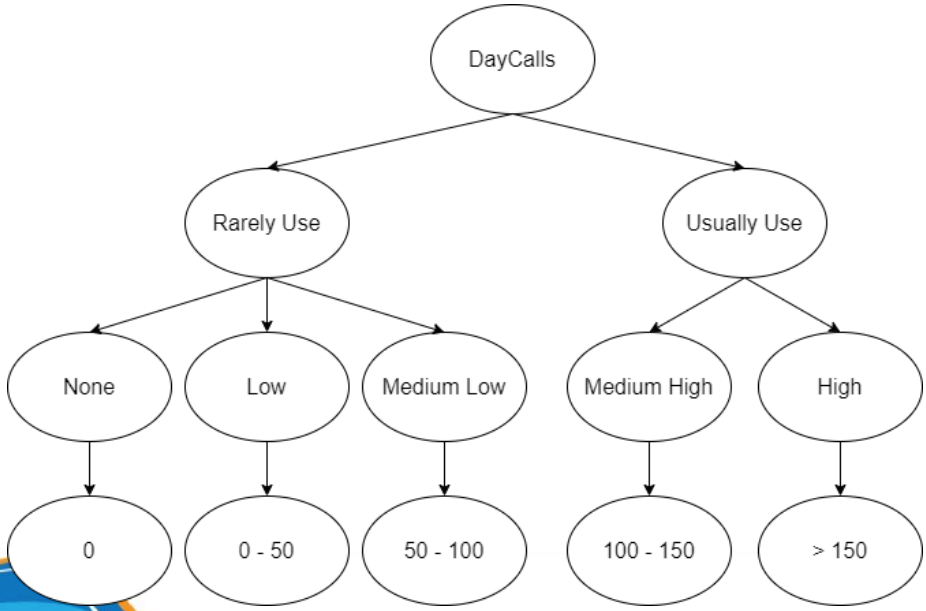
- Cây phân cấp thuộc tính VMailMessage



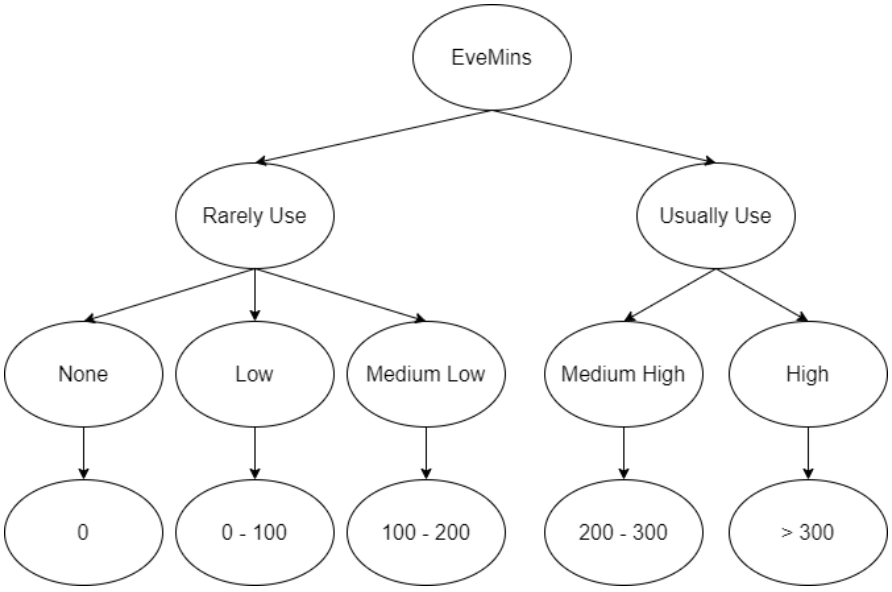
- Cây phân cấp thuộc tính DayMins



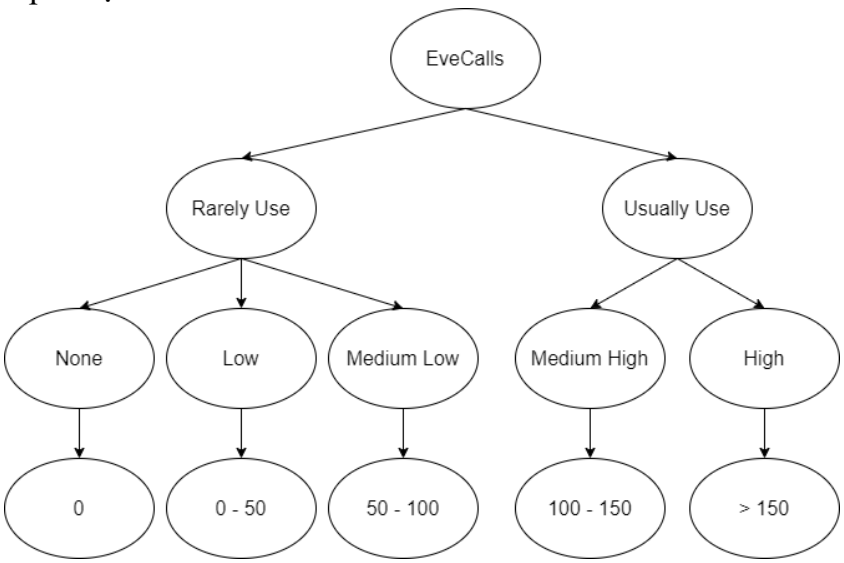
- Cây phân cấp thuộc tính DayCalls



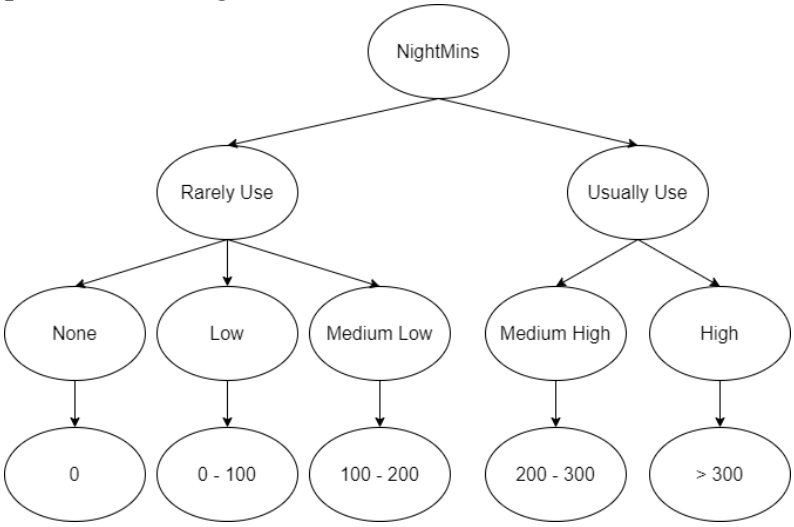
- Cây phân cấp thuộc tính EveMins



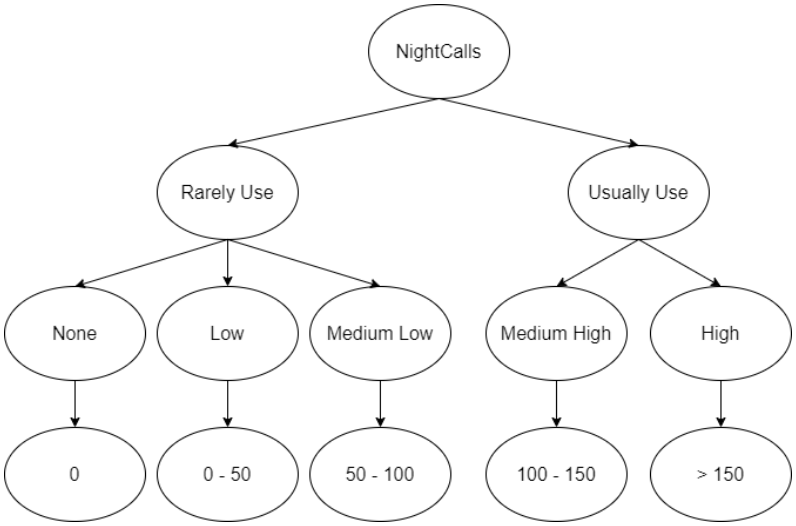
- Cây phân cấp thuộc tính EveCalls



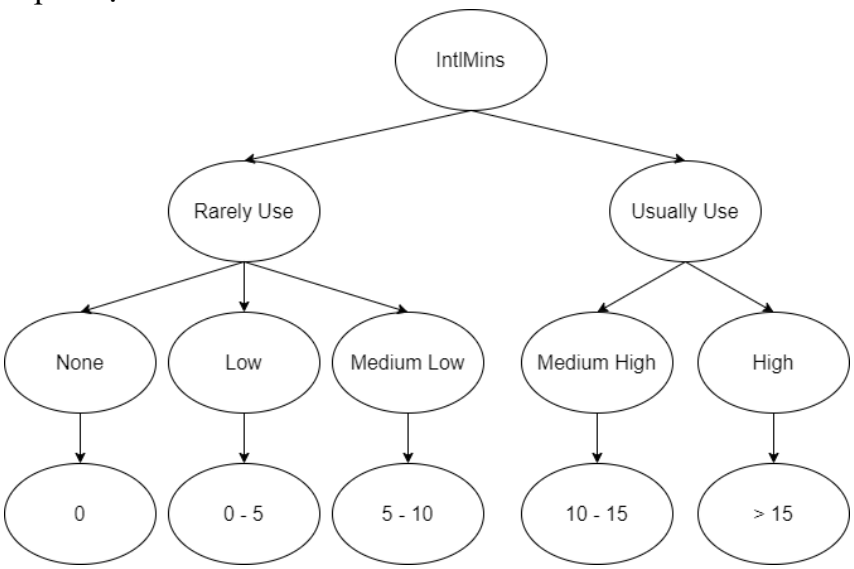
- Cây phân cấp thuộc tính NightMins



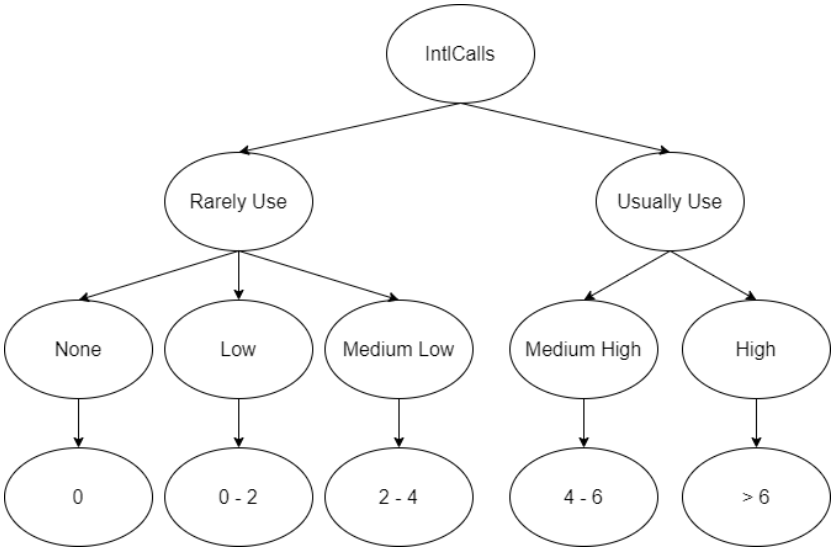
- Cây phân cấp thuộc tính NightCalls



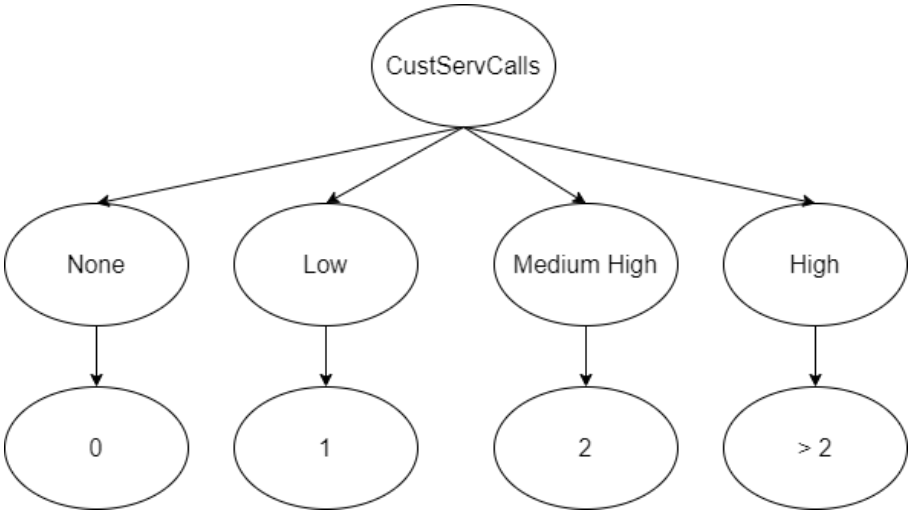
- Cây phân cấp thuộc tính IntlMins



- Cây phân cấp thuộc tính IntlCalls



- Cây phân cấp thuộc tính CustServCalls



2. Mã nguồn

2.1. Đầu vào

- Cú pháp: *python hiearchies.py -i churn.csv -hi input1.txt input2.txt… -o output.csv*
- *Ý nghĩa:*
  - *Churn.csv là tập dữ liệu đầu vào*
  - *Input1.txt input2.txt …: là các file định nghĩa phân cấp cho các thuộc tính*
  - Output.csv là tập dữ liệu sau khi phân cấp
- Cấu trúc file định nghĩa phân cấp:
  - Dòng đầu tiên là tên thuộc tính muốn phân cấp
  - Với mỗi 3 dòng tiếp theo sẽ chức thông tin:
    - Nhãn của khoảng giá trị muốn gán.
    - Chặn dưới của khoảng giá trị.
    - Chặn trên của khoảng giá trị.
  - Kết thúc file là dấu "." và 1 dòng trống.

2.2. Xử lý

Thuật toán duyệt qua các ô giá trị và gán lại nhãn cho các giá trị cần phân cấp theo định nghĩa trong các file định nghĩa phân cấp.

2.3. Đầu ra

output.csv là tập dữ liệu sau khi đươc phân phân cấp.

3. Thực nghiệm

Thuật toán sử dụng: *Apriori*.
Các tham số truyền vào:

- classIndex = -1 : Chỉ mục của thuộc tính phân lớp.
- lowerBoundMinSupport(-M) : chặn dưới của MinSupport
- upperBoundMinSupport(-U) : chặn trên của MinSupport. Bắt đầu giảm min support từ giá trị này
- delta (-D): hệ số giảm support khi lặp, giảm support đến khi đạt min support hay đã phát sinh đủ luật
- metricType (-T): độ đo tính quan trọng/ lý thú của luật. Ta chỉ quan tâm độ đo Confidence.
- minMetric (-C): độ tin cậy nhỏ nhất. Chỉ xét những luật có điểm lớn hơn giá trị này.
- NumRules (-N): số luật cần phải tìm.
- outputItemSets (-I): xuất ra nội dung các tập hạng mục

Các bộ test sử dụng:

STT	TEST
1	**Apriori** -I -N 10 -T 0 -C 0.6 -D 0.05 -U 0.4 -M 0.1 -S -1.0 -c -1
2	**Apriori** -I -N 10 -T 0 -C 0.7 -D 0.05 -U 0.4 -M 0.1 -S -1.0 -c -1
3	**Apriori** -I -N 10 -T 0 -C 0.8 -D 0.05 -U 0.4 -M 0.1 -S -1.0 -c -1
4	**Apriori** -I -N 10 -T 0 -C 0.9 -D 0.05 -U 0.4 -M 0.1 -S -1.0 -c -1
5	**Apriori** -I -N 10 -T 0 -C 0.6 -D 0.05 -U 0.7 -M 0.4 -S -1.0 -c -1
6	**Apriori** -I -N 10 -T 0 -C 0.7 -D 0.05 -U 0.7 -M 0.4 -S -1.0 -c -1
7	**Apriori** -I -N 10 -T 0 -C 0.8 -D 0.05 -U 0.7 -M 0.4 -S -1.0 -c -1
8	**Apriori** -I -N 10 -T 0 -C 0.9 -D 0.05 -U 0.7 -M 0.4 -S -1.0 -c -1
9	**Apriori** -I -N 10 -T 0 -C 0.6 -D 0.05 -U 0.9 -M 0.7 -S -1.0 -c -1
10	**Apriori** -I -N 10 -T 0 -C 0.7 -D 0.05 -U 0.9 -M 0.7 -S -1.0 -c -1
11	**Apriori** -I -N 10 -T 0 -C 0.8 -D 0.05 -U 0.9 -M 0.7 -S -1.0 -c -1
12	**Apriori** -I -N 10 -T 0 -C 0.9 -D 0.05 -U 0.9 -M 0.7 -S -1.0 -c -1

Ta sẽ tiến hành tổ chức dữ liệu phân cấp trước khi thực nghiệm từ cụ thể đến tổng quát:

3.1. *Phân cấp thấp nhất cho các thuộc tính DayMins, EveMins, NightMins và IntlMins. Chuyển các thuộc tính DayCalls, EveCalls, NightCalls, IntlCalls và CustServCalls thành thuộc tính phân loại.*

Tập luật hữu ích:

LHS	*Sup*	*RHS*	*Sup*	*Confidence*
VMailPlan=no	*2411*	*Churn?=False.*	*2008*	*<conf:(0.83)>*
IntlPlan=no	*3010*	*Churn?=False.*	*2664*	*<conf:(0.89)>*

3.2. *Phân cấp thấp nhất cho tất cả các thuộc tính kiểu số.*

Tập luật hữu ích:

LHS	Sup	RHS	Sup	Confidence
VMailPlan=no	*2411*	*Churn?=False.*	*2008*	*<conf:(0.83)>*
IntlPlan=no	*3010*	*Churn?=False.*	*2664*	*<conf:(0.89)>*

3.3. *Nâng cấp cho tất cả các thuộc tính kiểu số.*

Tập luật hữu lích:

LHS	Sup	RHS	Sup	Confidence
VMailMessage=Rarely Use	*2695*	*Churn?=False.*	*2279*	*<conf:(0.85)>*
IntlPlan=no VMailMessage=Rarely Use	*2440*	*Churn?=False.*	*2133*	*<conf:(0.87)>*
IntlPlan=no	*3010*	*Churn?=False.*	*2664*	*<conf:(0.89)>*

Nhận xét:

Ta thấy các luật mà ta rút ra chỉ toàn là các luật liên quan đến các khách hàng không rời bỏ dịch vụ của công ty. Với các luật này thì ta có thể quan tâm, cải thiện hoặc loại bỏ các dịch vụ liên quan. Ví dụ như Dịch vụ gọi quốc tế và Dịch vụ hộp thư thoại.

Mặc khác, ta cần quan tâm đến những khách hàng rời bỏ công ty. Ta cần biết lý do tại sao họ lại chuyển sang các dịch vụ của công ty khác. Các bộ test ở trên không khai thác được các luật liên quan đến loại người dùng này là do số lượng người rời bỏ công ty quá ích so với số người ở lại. Từ đó mà thuật Apriori chỉ thêm các bộ có churn?=False vào tập phổ biến mà không thêm các bộ có churn?=True.

Để khắc phục tình trạng này. Ta sẽ xóa hết các bộ có churn?=False khỏi tập dữ liệu khai thác. Từ đó ta chỉ quan tâm đến các khác hàng rời bỏ công ty. Thực hiện lại các test với các bộ dữ liệu ở muc 3.1, 3.2 và 3.3 với tập các bộ có churn?=True. Ta có được tập các luật hữu ích

3.1

LHS	Sup	RHS	Sup	Confidence
IntlPlan=no, VMailPlan=no	*302*	*Churn?=True.*	*302*	*<conf:(1)>*
IntlPlan=no, VMailMessage=0	*302*	*Churn?=True.*	*302*	*<conf:(1)>*
IntlPlan=no, VMailPlan=no, VMailMessage=0	*302*	*Churn?=True.*	*302*	*<conf:(1)>*
IntlPlan=no, VMailMessage=0	*302*	*Churn?=True.*	*302*	*<conf:(1)>*
IntlPlan=no, VMailPlan=no	*302*	*Churn?=True.*	*302*	*<conf:(1)>*
DayMins=Medium High	*266*	*Churn?=True.*	*266*	*<conf:(1)>*
VMailPlan=no	*403*	*Churn?=True.*	*403*	*<conf:(1)>*
VMailMessage=0	*403*	*Churn?=True.*	*403*	*<conf:(1)>*
VMailPlan=no, VMailMessage=0	*403*	*Churn?=True.*	*403*	*<conf:(1)>*
VMailMessage=0	*403*	*Churn?=True.*	*403*	*<conf:(1)>*
VMailPlan=no	*403*	*Churn?=True.*	*403*	*<conf:(1)>*
IntlPlan=no	*346*	*Churn?=True.*	*346*	*<conf:(1)>*

3.2

LHS	Sup	RHS	Sup	Confidence
VMailPlan=no, NightCalls=Medium High	190	Churn?=True.	190	<conf:(1)>
VMailMessage=None, NightCalls=Medium High	190	Churn?=True.	190	<conf:(1)>
VMailPlan=no, VMailMessage=None, NightCalls=Medium High	190	Churn?=True.	190	<conf:(1)>
VMailMessage=None, NightCalls=Medium High	190	Churn?=True.	190	<conf:(1)>
VMailPlan=no, NightCalls=Medium High	190	Churn?=True.	190	<conf:(1)>
IntlPlan=no, VMailMessage=None	302	Churn?=True.	302	<conf:(1)>
IntlPlan=no, VMailPlan=no VMailMessage=None	302	Churn?=True.	302	<conf:(1)>
IntlPlan=no, VMailMessage=None	302	Churn?=True.	302	<conf:(1)>
IntlPlan=no, VMailPlan=no	302	Churn?=True.	302	<conf:(1)>
VMailMessage=None	403	Churn?=True.	403	<conf:(1)>
VMailPlan=no, VMailMessage=None	403	Churn?=True.	403	<conf:(1)>
VMailMessage=None	403	Churn?=True.	403	<conf:(1)>
VMailPlan=no	403	Churn?=True.	403	<conf:(1)>

3.3

LHS	Sup	RHS	Sup	Confidence
EveMins=Rarely Use	193	Churn?=True.	193	<conf:(1)>
AccountLength=Loyal, DayMins=Usually Use	192	Churn?=True.	192	<conf:(1)>
AccountLength=Loyal, IntlMins=Usually Use	191	Churn?=True.	191	<conf:(1)>
IntlPlan=no, DayCalls=Usually Use	191	Churn?=True.	191	<conf:(1)>
VMailMessage=Rarely Use, DayCalls=Rarely Use	189	Churn?=True.	189	<conf:(1)>
DayMins=Usually Use, IntlCalls=Rarely Use	188	Churn?=True.	188	<conf:(1)>
AccountLength=Loyal, VMailPlan=no, IntlCalls=Rarely Use	188	Churn?=True.	188	<conf:(1)>
AccountLength=Loyal	328	Churn?=True.	328	<conf:(1)>
IntlCalls=Rarely Use	316	Churn?=True.	316	<conf:(1)>
IntlPlan=no, VMailMessage=Rarely Use	307	Churn?=True.	307	<conf:(1)>
IntlPlan=no, VMailPlan=no, VMailMessage=Rarely Use	302	Churn?=True.	302	<conf:(1)>
IntlPlan=no, VMailPlan=no	302	Churn?=True.	302	<conf:(1)>
DayMins=Usually Use	298	Churn?=True.	298	<conf:(1)>
EveMins=Usually Use	290	Churn?=True.	290	<conf:(1)>
VMailMessage=Rarely Use	416	Churn?=True.	416	<conf:(1)>
VMailPlan=no, VMailMessage=Rarely Use	403	Churn?=True.	403	<conf:(1)>
VMailPlan=no	403	Churn?=True.	403	<conf:(1)>
VMailMessage=Rarely Use	416	Churn?=True.	403	<conf:(0.97)>

- *Ghi chú*: kết quả của các thực nghiệm được lưu vào thư mục */test/*

4. Kết quả

Tìm được bô luật "tốt", đánh giá dựa trên confidence của các luật (>95% với các luật Churn=True và >80% đối với các luật Churn=False).

Bộ luật tốt nhất:

LHS	***Sup***	***RHS***	***Sup***	***Confidence***
IntlPlan=no, VMailPlan=no, VMailMessage=0	*302*	*Churn?=True.*	*302*	*<conf:(1)>*
DayMins=Medium High	*266*	*Churn?=True.*	*266*	*<conf:(1)>*
VMailPlan=no, VMailMessage=0	*403*	*Churn?=True.*	*403*	*<conf:(1)>*
IntlPlan=no	*346*	*Churn?=True.*	*346*	*<conf:(1)>*
AccountLength=Loyal	*328*	*Churn?=True.*	*328*	*<conf:(1)>*
IntlCalls=Rarely Use	*316*	*Churn?=True.*	*316*	*<conf:(1)>*
IntlPlan=no, VMailMessage=Rarely Use	*307*	*Churn?=True.*	*307*	*<conf:(1)>*
IntlPlan=no, VMailPlan=no, VMailMessage=Rarely Use	*302*	*Churn?=True.*	*302*	*<conf:(1)>*
DayMins=Usually Use	*298*	*Churn?=True.*	*298*	*<conf:(1)>*
EveMins=Usually Use	*290*	*Churn?=True.*	*290*	*<conf:(1)>*
VMailMessage=Rarely Use	*416*	*Churn?=True.*	*416*	*<conf:(1)>*

Nhận xét:

- Các khách hàng ít hoặc không sử dụng dịch vụ tin nhắn thoại có tỉ lệ rời bỏ công ty rất cao.
- Tiếp đến là các khách hàng trung thành. Đã dùng các dịch vụ của công ti(>7 năm) có xu hướng chuyển hướng sang công ty khác.
- Đa số các khách hàng rời bỏ dịch vụ của công ty là các khách hàng không tham gia dịch vụ gọi quốc tế.
- Các khách hàng rời bỏ dịch vụ có thói quen gọi điện thoại vào buổi sáng và buổi tối cao.

III. Đánh giá đồ án

1. Mức độ hoàn thành của các thành viên

MSSV	Mức độ hoàn thành công việc	Đóng góp
18120078	100%	50%
18120253	100%	50%

2. Mức độ hoàn thành đồ án:

- Làm sạch được dữ liệu, xóa các thuộc tính tương quan và các thuộc tính dị thường
- Rời rạc hóa được dữ liệu dựa trên khai niệm phân cấp (hiearchies).
- Sử dụng thuật toán Apriori khai thác được tập luật kết hợp.
- Các mức phân cấp chỉ mang tính chủ quan, không phản ánh được 1 cách chính xác việc phân loại khách hàng thực tế của 1 doanh nghiệp. Cần nhiều thông tin hơn về nghiệp vụ của công ty cụ thể.

VII. Nguồn tham khảo

- Slide bài giảng
- https://www.slideshare.net/mauliktogadiya/data-mining-65738602
- https://weka.sourceforge.io/doc.dev/weka/associations/Apriori.html
- http://ce.sharif.edu/courses/85-86/1/ce925/assignments/files/assignDir4/Churn.pdf