TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN **TÀI LIỆU HƯỚNG DẪN THỰC HÀNH**

KHOA HỆ THỐNG THÔNG TIN Môn: Khai thác dữ liệu

Bài thực hành 3

PHÂN LỚP DỮ LIỆU

I. Mục tiêu

1. Hiểu và vận dụng giải thuật cây quyết định để phân lớp dữ liệu.
2. Hiểu và vận dụng giải thuật Naïve Bayes để phân lớp dữ liệu.

II. Thời gian

1. Thực hành: 10 tiết
2. Bài tập làm thêm: 12 tiết

III. Hướng dẫn chung

Cho bảng dữ liệu về đánh giá rủi ro hồ sơ vay tín dụng tại một ngân hàng như sau

Thu nhập	Nghề nghiệp	Tình trạng hôn nhân	Sở hữu nhà	*Nguy cơ rủi ro*
Thấp	Vận tải	Độc thân	Rồi	*Cao*
Thấp	Vận tải	Độc thân	Chưa	*Cao*
Cao	Vận tải	Độc thân	Rồi	*Thấp*
Trung bình	Truyền thông	Độc thân	Rồi	*Thấp*
Trung bình	Kinh doanh	Đã kết hôn	Rồi	*Thấp*
Trung bình	Kinh doanh	Đã kết hôn	Chưa	*Cao*
Cao	Kinh doanh	Đã kết hôn	Chưa	*Thấp*
Thấp	Truyền thông	Độc thân	Rồi	*Cao*
Thấp	Kinh doanh	Đã kết hôn	Rồi	*Thấp*
Trung bình	Truyền thông	Đã kết hôn	Rồi	*Thấp*
Thấp	Truyền thông	Đã kết hôn	Chưa	*Thấp*
Cao	Truyền thông	Độc thân	Chưa	*Thấp*
Cao	Vận tải	Đã kết hôn	Rồi	*Thấp*
Trung bình	Truyền thông	Độc thân	Chưa	*Cao*
Cao	Kinh doanh	Độc thân	Rồi	*Cao*

Hình 4.1: Bảng đánh giá rủi ro tính dụng

Trong đó cột dữ liệu *Nguy cơ rủi ro* là thuộc tính quyết định

1. Tính giá trị độ lợi thông tin (information gain) của các thuộc tính và vẽ cây quyết định, rút luật từ cây quyết định theo thuật toán **ID3** cho tập dữ liệu trên
2. Tính giá trị chỉ số Gini (*gini index*) của các thuộc tính và vẽ cây quyết định theo thuật toán CART cho dữ liệu trên.
3. Sử dụng một trong hai cây quyết định ở trên để tiên đoán giá trị *Nguy cơ* của những hồ sơ sau

Thu nhập	Nghề nghiệp	Tình trạng hôn nhân	Sở hữu nhà
Trung bình	Vận tải	Độc thân	Rồi
Cao	Truyền thông	Độc thân	Rồi
Thấp	Vận tải	Đã kết hôn	Rồi

4. Sử dụng thuật toán phân lớp Bayes với kỹ thuật làm trơn **Laplace** để tính phân lớp các đối tượng ở câu 3

Hướng dẫn:

1. **Tính độ lợi thông tin** và vẽ cây quyết định bằng thuật toán ID3

- **Bước 1**: Tính Information của thuộc tính quyết định

$$I(p,n) = -\frac{p}{p+n}\log_2\frac{p}{p+n} - \frac{n}{p+n}\log_2\frac{n}{p+n}$$

  - Gọi ***P*** , ***N*** là hai lớp và **S** là một tập dữ liệu có ***p*** phần tử thuộc lớp ***P*** và ***n*** phần tử thuộc lớp ***N***

- **Bước 2**: Tính Entropy cho từng thuộc tính

$$E(A) = \sum_{i=1}^{v}\frac{p_i + n_i}{p+n} I(p_i, n_i)$$

  - Gọi {S1, S2 , …, Sv} là một phân hoạch của S, khi sử dụng thuộc tính A
  - Với mỗi Si chứa pi mẫu thuộc lớp P và ni mẫu thuộc lớp N.
  - Entropy, hay thông tin mong muốn cần thiết để phân lớp các đối tượng trong tất cả các cây con Si

- **Bước 3**: Tính độ lợi thông tin của từng thuộc tính theo công thức

$$Gain(A) = I(p,n) - E(A)$$

- **Thực hiện bước 1**: Tính information cho thuộc tính quyết định

  Ban đầu tập S bao gồm toàn bộ 15 dòng dữ liệu đã cho, trong đó có 9 dòng Rủi ro thấp, 6 dòng Rủi ro cao. Vậy độ bất định của tập S lúc này là

$$I(9,6) = -\frac{9}{15}\log_2\frac{9}{15} - \frac{6}{15}\log_2\frac{6}{15} = 0.97$$

- **Thực hiện bước 2 và 3**: Tính độ lợi thông tin cho từng thuộc tính
  - Thuộc tính "**Thu nhập**" có 3 giá trị phân biệt là: Thấp, Trung bình, Cao.
  - Giá trị **Thấp** có 5 dòng trong đó **3 dòng** được gán nhãn nguy cơ rủi ro "**Cao**", 2 dòng được gán nhãn nguy cơ rủi ro "**Thấp**".
  - Giá trị Trung bình có 5 dòng trong đó 2 dòng được gán nhãn nguy cơ rủi ro "**Cao**", 3 dòngđược gán nhãn Rủi ro "**Thấp**"
  - Giá trị Cao có 5 dòng trong đó 1 dòng được gán nhãn nguy cơ rủi ro "**Cao**", 4 dòng được gán nhãn nguy cơ rủi ro "**Thấp**".

Thu nhập	Cao (rủi ro)	Thấp (rủi ro)	I(Cao,Thấp)
Thấp	3	2	I(2,3)= 0.971
Trung bình	2	3	I(2,3)=0.971
Cao	1	4	I(1,4)= 0.722

    - $I(3,2) = -\frac{3}{5}\log_2\frac{3}{5} - \frac{2}{5}\log_2\frac{2}{5} = 0.971$
    - $I(2,3) = -\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5} = 0.971$
    - $I(1,4) = -\frac{1}{5}\log_2\frac{1}{5} - \frac{4}{5}\log_2\frac{4}{5} = 0.722$
    - $E(\text{Thu nhập}) = \frac{5}{15}\ I(2,3) + \frac{5}{15}I(2,3) + \frac{5}{15}I(1,4) = 0.887$
    - $\text{Gain(Thu nhập)} = I(9,6) - E(\text{thu nhập}) = 0.97 - 0.887 = 0.083$
  - Xét thuộc tính "***nghề nghiệp***" ta có 3 giá trị phân biệt là: Vận tải, Truyền thông, kinh doanh
    - Giá trị Vận tải có 4 dòng trong đó 2 dòng được gán nhãn Rủi ro cao, 2 dòng đượcgán nhãn Rủi ro thấp.
    - Giá trị Truyền thông có 6 dòng trong đó 2 dòng được gán nhãn Rủi ro cao, 4 dòngđược gán nhãn Rủi ro thấp.
    - Giá trị Kinh doanh có 5 dòng trong đó 2 dòng được gán nhãn Rủi ro cao, 3 dòngđược gán nhãn Rủi ro thấp.

Nghề nghiệp	Cao (rủi ro)	Thấp (rủi ro)	I(Cao,Thấp)
Vận tải	2	2	I(2,2)= 1
Truyền thông	2	4	I(2,4)=0.918
Kinh doanh	2	3	I(2,3)=0.971

- $I(2,4) = -\frac{2}{6}\log 2\frac{2}{6} - \frac{4}{6}\log 2\frac{4}{6} = 0.918$
- $I(2,3) = -\frac{2}{5}\log 2\frac{2}{5} - \frac{3}{5}\log 2\frac{3}{5} = 0.971$
- $E(\text{Nghề nghiệp}) = \frac{4}{15} I(2,2) + \frac{6}{15} I(2,4) + \frac{5}{15} I(2,3) = 0.957$
- $\text{Gain(Nghề nghiệp)} = I(9,6) - E(\text{thu nhập}) = 0.97 - 0.957 = 0.013$

- Tiếp tục xét thuộc tính "***Tình trạng hôn nhân***" ta có:
  - Thuộc tính Tình trạng hôn nhân có 2 giá trị phân biệt là: Độc thân, Đã kết hôn.
  - Giá trị Độc thân có 8 dòng trong đó 5 dòng được gán nhãn Rủi ro cao, 3 dòng được gán nhãn Rủi ro thấp.
  - Giá trị Đã kết hôn có 7 dòng trong đó 1 dòng được gán nhãn Rủi ro cao, 6 dòngđược gán nhãn Rủi ro thấp

Tình trạng hôn nhân	Cao (rủi ro)	Thấp (rủi ro)	I(Cao,Thấp)
Độc thân	5	3	I(5,3)= 0,954
Đã kết hôn	1	6	I(1,6)= 0,592

- $I(5,3) = -\frac{5}{8}\log 2\frac{5}{8} - \frac{3}{8}\log 2\frac{3}{8} = 0.954$
- $I(1,6) = -\frac{1}{7}\log 2\frac{1}{7} - \frac{6}{7}\log 2\frac{6}{7} = 0.592$
- $E(\text{Tình trạng hôn nhân}) = \frac{8}{15} I(5,3) + \frac{7}{15} I(1,6) = 0.785$
- $\text{Gain(Tình trạng hôn nhân)} = I(9,6) - E(\text{Tình trạng hôn nhân}) = 0.97 - 0.785 = 0.185$

- Xét thuộc tính Sở hữu nhà
  - Thuộc tính Sở hữu nhà có 2 giá trị phân biệt là: Chưa, Rồi.
  - Giá trị Chưa có 6 dòng trong đó 3 dòng được gán nhãn Rủi ro cao, 3 dòng được gán nhãn Rủi ro thấp.

- Giá trị Rồi có 9 dòng trong đó 3 dòng được gán nhãn Rủi ro cao, 6 dòng được gánnhãn Rủi ro thấp.

Sở hữu nhà	Cao (rủi ro)	Thấp (rủi ro)	I(Cao,Thấp)
Chưa	3	3	I(3,3)= 1
Rồi	3	6	I(3,6)= 0,918

  - $I(3,3) = -\frac{3}{6}\log 2\frac{3}{6} - \frac{3}{6}\log 2\frac{3}{6} = 1.0$
  - $I(3,6) = -\frac{3}{9}\log 2\frac{3}{9} - \frac{6}{9}\log 2\frac{6}{9} = 0.918$
  - $E(\text{Sở hữu nhà}) = \frac{6}{15} I(3,3) + \frac{9}{15} I(3,6) = 0.95$
  - Gain(Sở hữu nhà) = I(9,6) - E(Sở hữu nhà) = 0.97 - 0.95 = 0,02

- Độ lợi thông tin các thuộc tính tính được là
  - Gain(***Thu nhập***) = I(9,6) - E(thu nhập) = 0.97 - 0.888 = 0.083
  - Gain(***Nghề nghiệp***) = I(9,6) - E(thu nhập) = 0.97 - 0.888 = 0.013
  - Gain(**Tình trạng hôn nhân**) = I(9,6) - E(thu nhập) = 0.97 - 0.785 = **0.185**
  - Gain(***Sở hữu nhà***) = I(9,6) - E(thu nhập) = 0.97 - 0.785 = 0,02
  - Thuộc tính "***Tình trạng hôn nhân***" có độ lợi thông tin **lớn nhất** nên ta chọn thuộc tính này làm thuộc tính phân lớp

Thu nhập	Nghề nghiệp	Tình trạng hôn nhân	Sở hữu nhà	*Nguy cơ rủi ro*
Thấp	Vận tải	Độc thân	Rồi	*Cao*
Thấp	Vận tải	Độc thân	Chưa	*Cao*
Cao	Vận tải	Độc thân	Rồi	*Thấp*
Trung bình	Truyền thông	Độc thân	Rồi	*Thấp*
Thấp	Truyền thông	Độc thân	Rồi	*Cao*
Cao	Truyền thông	Độc thân	Chưa	*Thấp*
Trung bình	Truyền thông	Độc thân	Chưa	*Cao*
Cao	Kinh doanh	Độc thân	Rồi	*Cao*

Hình 4.2: Bảng con 1

Thu nhập	Nghề nghiệp	Tình trạng hôn nhân	Sở hữu nhà	*Nguy cơ rủi ro*
Trung bình	Kinh doanh	Đã kết hôn	Rồi	*Thấp*
Trung bình	Kinh doanh	Đã kết hôn	Chưa	*Cao*
Cao	Kinh doanh	Đã kết hôn	Chưa	*Thấp*
Thấp	Kinh doanh	Đã kết hôn	Rồi	*Thấp*
Trung bình	Truyền thông	Đã kết hôn	Rồi	*Thấp*
Thấp	Truyền thông	Đã kết hôn	Chưa	*Thấp*
Cao	Vận tải	Đã kết hôn	Rồi	*Thấp*

Hình 4.3: Bảng con 2

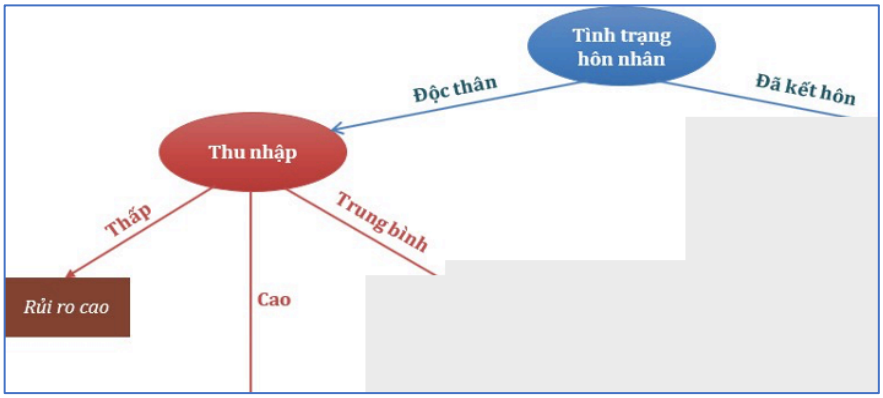
- Xét bảng con 1 ở hình 4.2 sau khi loại bỏ thuộc tính "***Tình trạng hôn nhân***"

Thu nhập	Nghề nghiệp	Sở hữu nhà	*Nguy cơ rủi ro*
Thấp	Vận tải	Rồi	*Cao*
Thấp	Vận tải	Chưa	*Cao*
Cao	Vận tải	Rồi	*Thấp*
Trung bình	Truyền thông	Rồi	*Thấp*
Thấp	Truyền thông	Rồi	*Cao*
Cao	Truyền thông	Chưa	*Thấp*
Trung bình	Truyền thông	Chưa	*Cao*
Cao	Kinh doanh	Rồi	*Cao*

- Ta quay lại Bước 1 , bước 2, bước 3 cho bảng dữ liệu này
    - $I(5,3) = -\frac{5}{8}\log_2\frac{5}{8} - \frac{3}{8}\log_2\frac{3}{8} = 0.954$
- Xét thuộc tính **Thu nhập**, tính toán tương tự ta có:
    - $I(\text{Thấp}) = I(3,0) = 0;$
    - $I(\text{Trung bình}) = I(1,1) = 1;$
    - $I(\text{Cao}) = I(1,2) \approx 0,92;$
    - $E(\text{Thu nhập}) = \frac{3}{8} I(3,0) + \frac{2}{8} I(1,1) + \frac{3}{8} I(1,2) = 0.594$
    - $G(\text{Thu nhập}) \approx 0.954 - 0.594 = 0,36$
- Với thuộc tính "***Nghề nghiệp"*** ta có:
    - $I(\text{Vận tải}) = I(2,1) \approx 0,92;$
    - $I(\text{Truyền thông}) = I(2,2) = 1;$
    - $I(\text{Kinh doanh}) = I(1,0) = 0;$
    - $E(\text{Nghề nghiệp}) = \frac{3}{8} I(2,1) + \frac{4}{8} I(2,2) + \frac{1}{8} I(1,0) = 0.845$
    - $G(\text{Nghề nghiệp}) = I(5,3) - E(\text{nghề nghiệp}) = 0.954 - 0.845 \approx 0.11$

- Xét thuộc tính "***Sở hữu nhà***" ta có
    - $I(\text{Chưa}) = I(2,1) \approx 0{,}92;$
    - $I(\text{rồi}) = I(3,2) \approx 0{,}971;$
    - $E(\text{Sở hữu nhà}) = \frac{3}{8} I(2,1) + \frac{5}{8} I(3,2) = 0.952$
    - $G(\text{Sở hữu nhà}) = I(5,3) - E(\text{Sở hữu nhà}) = 0.954 - 0.952 \approx 0{,}002$

- Vậy ta chọn thuộc tính "***Thu nhập***" có độ lợi thông tin lớn nhất, ta có cây quyết định cho nhánh "**Thu nhập**" như sau



- Vậy bảng con 1 còn lại các đối tượng chưa phân hoạch nằm ở bảng con ở hình 4.4 và bảng con ở hình 4.5 bên dưới

Thu nhập	Nghề nghiệp	Sở hữu nhà	*Nguy cơ rủi ro*
Cao	Vận tải	Rồi	*Thấp*
Cao	Truyền thông	Chưa	*Thấp*
Cao	Kinh doanh	Rồi	*Cao*

Hình 4.4: Bảng con 1.1

Thu nhập	Nghề nghiệp	Sở hữu nhà	*Nguy cơ rủi ro*
Trung bình	Truyền thông	Rồi	*Thấp*
Trung bình	Truyền thông	Chưa	*Cao*

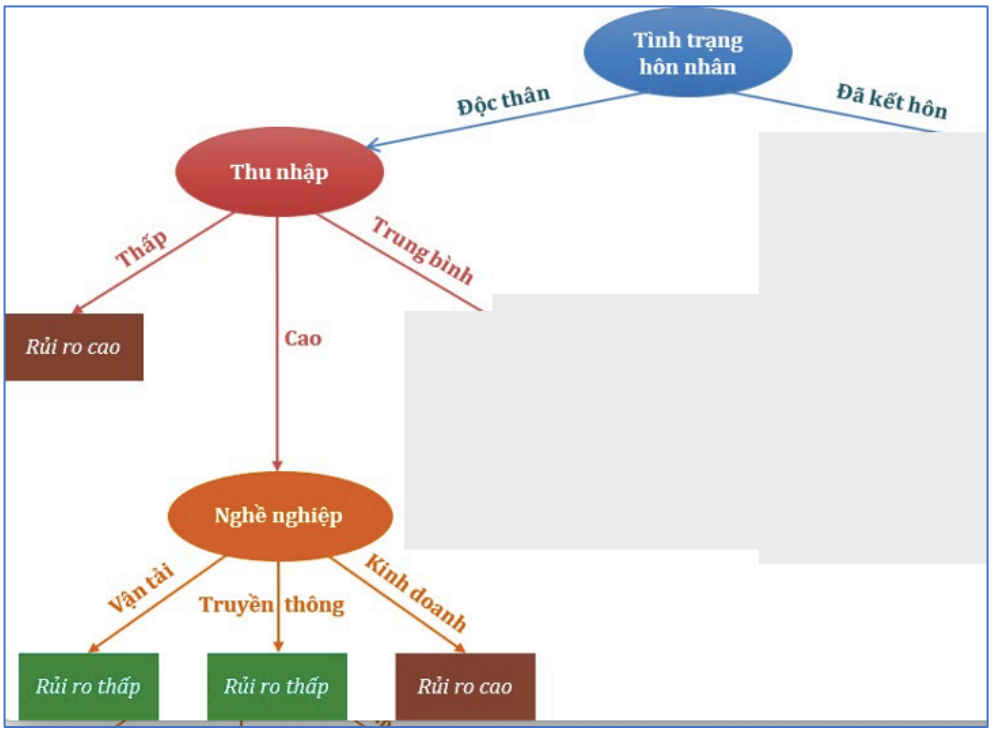
Hình 4.5: Bảng con 1.2

- **Xét *bảng con 1.1*** khi bỏ thuộc tính "***Thu nhập***"ta tiếp tục lại các Bước 1, bước 2, bước 3 để đi tìm độ lợi thông tin

Nghề nghiệp	Sở hữu nhà	*Nguy cơ rủi ro*
Vận tải	Rồi	*Thấp*
Truyền thông	Chưa	*Thấp*
Kinh doanh	Rồi	*Cao*

- Tính Information cho "***Bảng con 1.1***"
  - $I(1,2) = 0.92$
  - Xét thuộc tính "**Nghề nghiệp**" ta đi tính độ lợi thông tin
  - $I(\text{Vận tải}) = I(0,1) = 0;$
  - $I(\text{Truyền thông}) = I(0,1) = 0$
  - $I(\text{Kinh doanh}) = I(1,0) = 0$
  - $E(\text{nghề nghiệp}) = \frac{1}{3} I(0,1) + \frac{1}{3} I(0,1) + \frac{1}{3} I(1,0) = 0$
  - $G(\text{nghề nghiệp}) = I(1,2) - E(\text{nghề nghiệp}) = 0.92 - 0 \approx 0.92$
- Xét thuộc tính "**Sở hữu nhà**" ta đi tính độ lợi thông tin
  - $I(\text{rồi}) = I(1,1) = 1;$
  - $I(\text{Chưa}) = I(0,1) = 0$
  - $E(\text{Sở hữu nhà}) = \frac{2}{3} I(1,1) + \frac{1}{3} I(0,1) = 0.67$
  - $G(\text{Sở hữu nhà}) = I(1,2) - E(\text{Sở hữu nhà}) = 0.92 - 0.67 \approx 0.25$

- Vậy ta chọn thuộc tính "**Nghề nghiệp**" làm thuộc tính phân hoạch

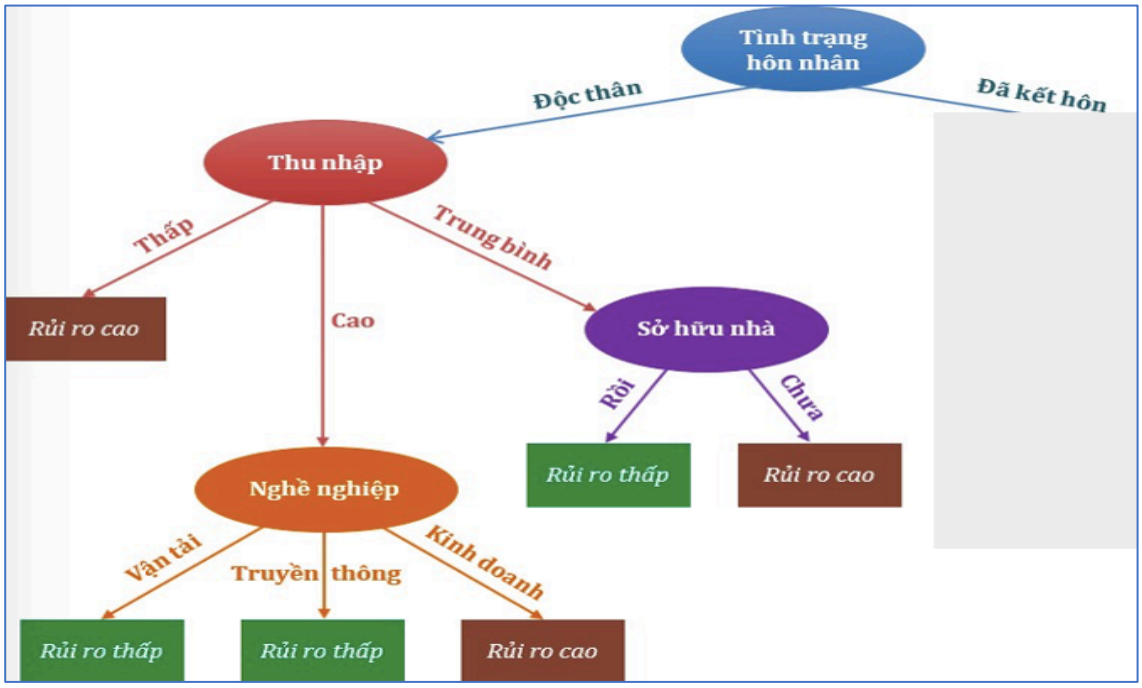


- **Xét *bảng con 1.2*** khi bỏ thuộc tính "***Thu nhập***"ta tiếp tục lại các Bước 1, bước 2, bước 3 để đi tìm độ lợi thông tin

Nghề nghiệp	Sở hữu nhà	*Nguy cơ rủi ro*
Truyền thông	Rồi	*Thấp*
Truyền thông	Chưa	*Cao*

- Tính information cho thuộc tính kết luận
  - I(1,1) = 1
- Xét thuộc tính "***Nghề nghiệp***"
  - I(Truyền thông) = I(1,1) = 1
  - E(Nghề nghiệp)= $\frac{2}{2}$ I(1,1) =1
  - G(Nghề nghiệp) = I(1,1)-E(Nghề nghiệp)=1-1=0

- Xét thuộc tính "***Sở hữu nhà***"
  - I(rồi) = I(0,1) = 0
  - I(chưa)= I(1,0)=0
  - E(Sở hữu nhà)= $\frac{1}{2}$ I(0,1) + $\frac{1}{2}$ I(1,0)=0
  - G(Sở hữu nhà) = I(1,1)-E(Sở hữu nhà)=1-0=1
- Vậy ta chọn thuộc tính " ***Sở hữu nhà*** " để phân hoạch cho các đối tượng ở ***bảng con 1.2***

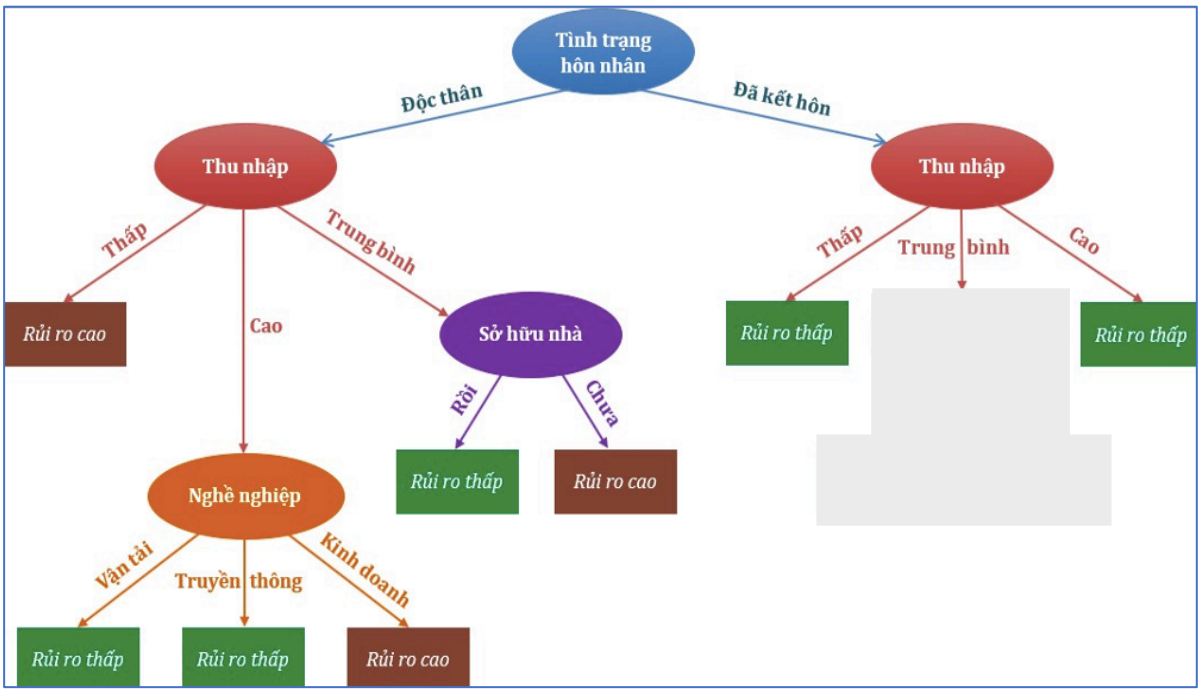


- Xét bảng con 2 ở hình 4.3 khi ta loại bỏ thuộc tính "***Tình trạng hôn nhân***"

Thu nhập	Nghề nghiệp	Sở hữu nhà	*Nguy cơ rủi ro*
Trung bình	Kinh doanh	Rồi	*Thấp*
Trung bình	Kinh doanh	Chưa	*Cao*
Cao	Kinh doanh	Chưa	*Thấp*
Thấp	Kinh doanh	Rồi	*Thấp*
Trung bình	Truyền thông	Rồi	*Thấp*
Thấp	Truyền thông	Chưa	*Thấp*
Cao	Vận tải	Rồi	*Thấp*

- Chúng ta tiếp tục lặp lại Bước 1 , Bước 2, Bước 3 để phân hoạch cho bảng con này bằng cách tính ***độ lợi thông tin*** cho các thuộc tính
- Giá trị Information cho thuộc tính quyết định
  - I(1,6) = 0.592
- Xét thuộc tính "***thu nhập***"
  - I(Trung bình) = I(1,2) = 0.918
  - I(Cao)= I(0,2)=0
  - I(Thấp) =I(0,2)=0
  - E(Thu nhập)= $\frac{3}{7}$ I(1,2) + $\frac{2}{7}$ I(0,2)+ $\frac{2}{7}$ I(0,2) ≈0.393
  - G(Thu nhập) = I(1,6)-E(Thu nhập)=0.592-0.393≈0.199
- Xét thuộc tính "***Nghề nghiệp***"
  - I(Kinh doanh) = I(1,3) = 0.81
  - I(Truyền thông)= I(0,2)=0
  - I(Vận tải) =I(0,1)=0
  - E(Nghề nghiệp)= $\frac{4}{7}$ I(1,3) + $\frac{2}{7}$ I(0,2)+ $\frac{1}{7}$ I(0,1) ≈0.463
  - G(Nghề nghiệp) = I(1,6)-E(Nghề nghiệp)=0.592-0.463≈0.129

- Xét thuộc tính "***Sở hữu nhà***"
  - I(rồi) = I(0,4) = 0
  - I(Chưa)= I(1,2)= 0.918
  - E(Sở hữu nhà)= $\frac{4}{7}$ I(0,4) + $\frac{3}{7}$ I(1,2)≈0.393
  - G(Sở hữu nhà) = I(1,7)-E(Sở hữu nhà)=0.592-0.393≈0.199
- Thuộc tính "***Thu nhập***" và thuộc tính "***Sở hữu nhà***" có độ lợi thông tin lớn nhất và bằng nhau nên ta chọn thuộc tính "***Thu nhập***" làm thuộc tính phân hoạch. Ta còn lại các dòng chưa được phân hoạch ở bảng 2.1 (hình 2.1)



Thu nhập	Nghề nghiệp	Sở hữu nhà	*Nguy cơ rủi ro*
Trung bình	Kinh doanh	Rồi	*Thấp*
Trung bình	Kinh doanh	Chưa	*Cao*
Trung bình	Truyền thông	Rồi	*Thấp*

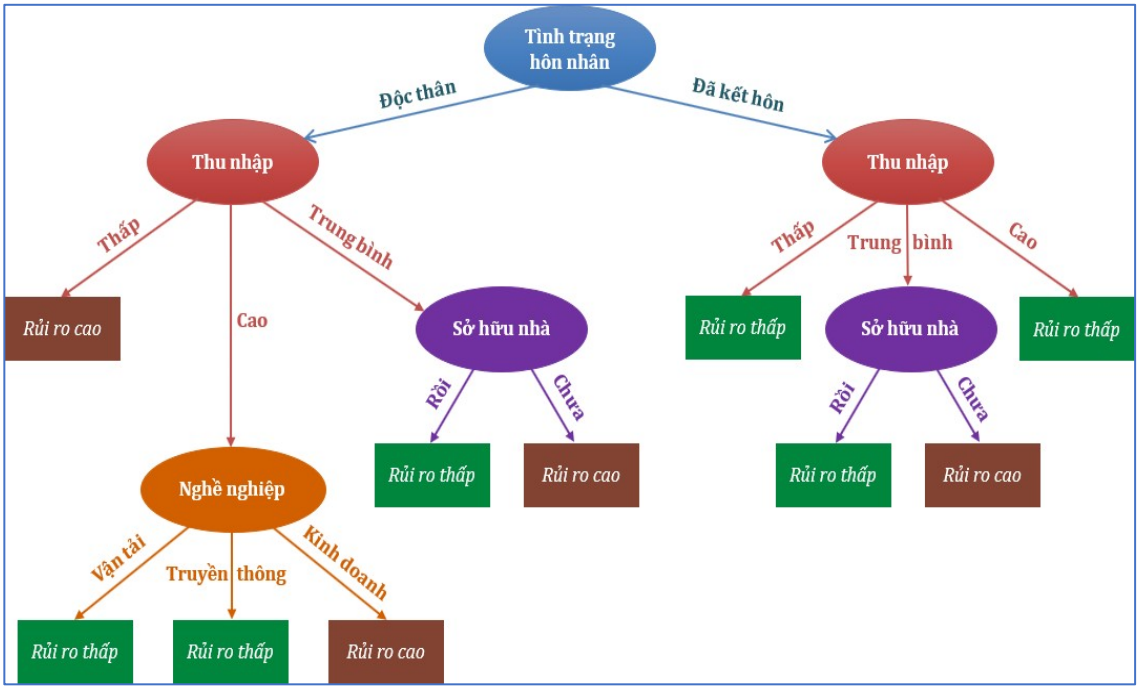
Hình 4.6: Bảng con 2.1

- Chúng ta tiếp tục lặp lại Bước 1 , Bước 2, Bước 3 để phân hoạch cho bảng con này bằng cách tính ***độ lợi thông tin*** cho các thuộc tính ở bảng con 2.1 khi loại bỏ thuộc tính "***Thu nhập***"

Nghề nghiệp	Sở hữu nhà	*Nguy cơ rủi ro*
Kinh doanh	Rồi	*Thấp*
Kinh doanh	Chưa	*Cao*
Truyền thông	Rồi	*Thấp*

- Giá trị Information cho thuộc tính quyết định
  - I(1,2) = 0.918
- Xét thuộc tính "***Nghề nghiệp***"
  - I(Kinh doanh) = I(1,1) = 1
  - I(Truyền thông)= I(0,1)=0
  - E(Nghề nghiệp)= $\frac{2}{3}$ I(1,1) + $\frac{1}{3}$ I(0,1) ≈ 0.667
  - G(Nghề nghiệp) = I(1,2)-E(Nghề nghiệp)=0.918-0.667 ≈0.251
- Xét thuộc tính "***Sở hữu nhà***"
  - I(rồi) = I(0,2) = 0
  - I(Chưa)= I(1,0)=0
  - E(Sở hữu nhà)= $\frac{2}{3}$ I(0,2) + $\frac{1}{3}$ I(1,0) = 0
  - G(Sở hữu nhà) = I(1,2)-E(Sở hữu nhà)=0.918-0 ≈0.918

- Vậy chọn thuộc tính "***Sở hữu nhà***" làm thuộc tính phân lớp cho bảng con 2.1 vì có độ lợi thông tin lớn nhất



- **Các luật rút từ cây quyết định**
- **R1**: IF "***Tình trạng hôn nhân***" đọc thân và "***thu nhập***" thấp THEN Rủi ro cao
- **R2**:IF "***Tình trạng hôn nhân***" đọc thân và "***thu nhập***" cao và "***nghề nghiệp***" vận tải THEN Rủi ro **thấp**
- **R3**: IF "***Tình trạng hôn nhân***" đọc thân và "***thu nhập***" cao và "***nghề nghiệp***" truyền thông THEN Rủi ro **thấp**
- **R4**: IF "***Tình trạng hôn nhân***" đọc thân và "***thu nhập***" cao và "***nghề nghiệp***" kinh doanh THEN Rủi ro **cao**
- **R5**: IF "***Tình trạng hôn nhân***" đọc thân và "***thu nhập***" trung bình và "***Sở hữu nhà***" rồi THEN Rủi ro **thấp**
- **R6**: IF "***Tình trạng hôn nhân***" đọc thân và "***thu nhập***" trung bình và "***Sở hữu nhà***" chưa THEN Rủi ro **cao**
- **R7**: IF "***Tình trạng hôn nhân***" đã kết hôn và "***thu nhập"*** thấp THEN Rủi ro

thấp

- **R8**: IF "***Tình trạng hôn nhân***" đã kết hôn và "***thu nhập***" trung bình và ***"Sở hữu nhà***" rồi THEN Rủi ro **thấp**
- **R9**: IF "***Tình trạng hôn nhân***" đã kết hôn và "***thu nhập***" trung bình và "***Sở hữu nhà***" chưa THEN Rủi ro **cao**
- **R10**: IF "***Tình trạng hôn nhân***" đã kết hôn và "***thu nhập***" cao THEN Rủi ro **thấp**

2. Chỉ số **Gini** dùng để đánh giá thuộc tính phân nhánh được tính theo công thức sau. Chỉ số Gini của tập huấn luyện S:

$$Gini(S) = 1 - \sum_j p(j|S)^2$$

  - Với $p(j|S)$ là tần suất của lớp j trong S.
  - Trong ví dụ "***Hồ sơ vay tính dụng***"ở trên: 15 mẫu
  - Phân lớp yes: 9
  - Phân lớp no: 6
  - Gini(S)=1 - $(9/15)^2$ - $(6/15)^2$ = 0.48
- Khi phân chia nút A thành k nhánh, chất lượng của phép chia được tính bằng công thức:

$$Gini_A(S) = \sum_{i=1}^{k} \frac{n_i}{n} Gini(i)$$

- Trong đó:
  - n_i là số mẫu trong nút i
  - n là số mẫu trong nút A
- Theo những thống kê từ câu 1, ta tính chỉ số Gini của lần lượt từng thuộc tính để tìm ra thuộc tính phân nhánh có **lợi nhất**. Chọn thuộc tính có giá trị GINI

nhỏ nhất để phân hoạch các đối tượng

- Xét thuộc tính "***Thu nhập***", ta có

$$Gini(S_{Thấp}) = 1 - \left(\frac{3}{5}\right)^2 - \left(\frac{2}{5}\right)^2 = 0{,}48$$

$$Gini(S_{Trung\ bình}) = 1 - \left(\frac{2}{5}\right)^2 - \left(\frac{3}{5}\right)^2 = 0{,}48$$

$$Gini(S_{Cao}) = 1 - \left(\frac{1}{5}\right)^2 - \left(\frac{4}{5}\right)^2 = 0{,}32$$

$$Gini_{Thu\ nhập}(S) = \frac{5}{15} \times 0{,}48 + \frac{5}{15} \times 0{,}48 + \frac{5}{15} \times 0{,}32 = 0{,}427$$

- Xét thuộc tính "***Nghề nghiệp***" ta có:

$$Gini(S_{Vận\ tải}) = 1 - \left(\frac{2}{4}\right)^2 - \left(\frac{2}{4}\right)^2 = 0{,}5$$

$$Gini(S_{Truyền\ thông}) = 1 - \left(\frac{2}{6}\right)^2 - \left(\frac{4}{6}\right)^2 \approx 0{,}444$$

$$Gini(S_{Kinh\ doanh}) = 1 - \left(\frac{2}{5}\right)^2 - \left(\frac{3}{5}\right)^2 = 0{,}48$$

$$Gini_{Nghề\ nghiệp}(S) \approx \frac{4}{15} \times 0{,}5 + \frac{6}{15} \times 0{,}444 + \frac{5}{15} \times 0{,}48 \approx 0{,}471$$

- Xét thuộc tính "***Tình trạng hôn nhân***", ta có:

$$Gini(S_{\text{Độc thân}}) = 1 - \left(\frac{5}{8}\right)^2 - \left(\frac{3}{8}\right)^2 \approx 0{,}469$$

$$Gini(S_{\text{Đã kết hôn}}) = 1 - \left(\frac{1}{7}\right)^2 - \left(\frac{6}{7}\right)^2 \approx 0{,}245$$

$$Gini_{\text{Tình trạng hôn nhân}}(S) \approx \frac{8}{15} \times 0{,}469 + \frac{7}{15} \times 0{,}245 \approx 0{,}403$$

- Xét thuộc tính "Sở hữu nhà", ta có:

$$Gini(S_{Chưa}) = 1 - \left(\frac{3}{6}\right)^2 - \left(\frac{3}{6}\right)^2 = 0{,}5$$

$$Gini(S_{Rồi}) = 1 - \left(\frac{3}{9}\right)^2 - \left(\frac{6}{9}\right)^2 \approx 0{,}444$$

$$Gini_{\text{Tình trạng hôn nhân}}(S) \approx \frac{6}{15} \times 0{,}5 + \frac{9}{15} \times 0{,}444 \approx 0{,}467$$

- Chọn thuộc tính có chỉ số Gini **thấp nhất** là "***Tình trạng hôn nhân***" để phân hoạch các đối tượng

Thu nhập	Nghề nghiệp	Tình trạng hôn nhân	Sở hữu nhà	*Nguy cơ rủi ro*
Thấp	Vận tải	Độc thân	Rồi	*Cao*
Thấp	Vận tải	Độc thân	Chưa	*Cao*
Cao	Vận tải	Độc thân	Rồi	*Thấp*
Trung bình	Truyền thông	Độc thân	Rồi	*Thấp*
Thấp	Truyền thông	Độc thân	Rồi	*Cao*
Cao	Truyền thông	Độc thân	Chưa	*Thấp*
Trung bình	Truyền thông	Độc thân	Chưa	*Cao*
Cao	Kinh doanh	Độc thân	Rồi	*Cao*

Hình 4.7: Bảng con 3

Thu nhập	Nghề nghiệp	Tình trạng hôn nhân	Sở hữu nhà	*Nguy cơ rủi ro*
Trung bình	Kinh doanh	Đã kết hôn	Rồi	*Thấp*
Trung bình	Kinh doanh	Đã kết hôn	Chưa	*Cao*
Cao	Kinh doanh	Đã kết hôn	Chưa	*Thấp*
Thấp	Kinh doanh	Đã kết hôn	Rồi	*Thấp*
Trung bình	Truyền thông	Đã kết hôn	Rồi	*Thấp*
Thấp	Truyền thông	Đã kết hôn	Chưa	*Thấp*
Cao	Vận tải	Đã kết hôn	Rồi	*Thấp*

Hình 4.8: Bảng con 4

- Xét bảng con 3 khi ta loại bỏ thuộc tính "***Tình trạng hôn nhân***"

Thu nhập	Nghề nghiệp	Sở hữu nhà	*Nguy cơ rủi ro*
Thấp	Vận tải	Rồi	*Cao*
Thấp	Vận tải	Chưa	*Cao*
Cao	Vận tải	Rồi	*Thấp*
Trung bình	Truyền thông	Rồi	*Thấp*
Thấp	Truyền thông	Rồi	*Cao*
Cao	Truyền thông	Chưa	*Thấp*
Trung bình	Truyền thông	Chưa	*Cao*
Cao	Kinh doanh	Rồi	*Cao*

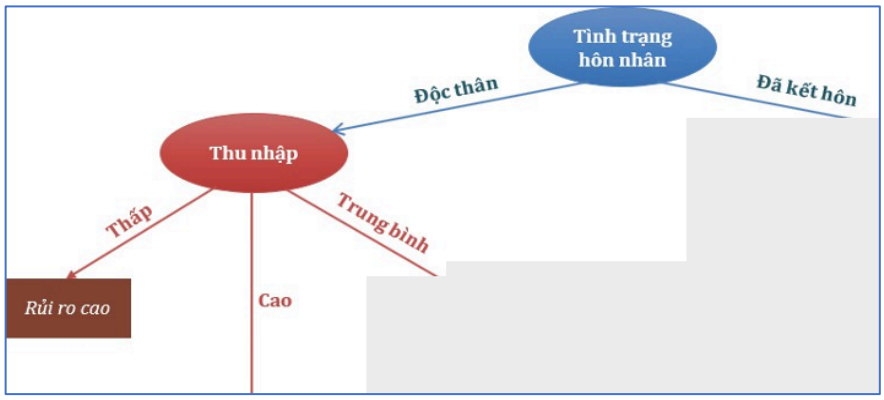
- Tính Gini cho tập các đối tượng "**S**" trong "***Bảng con 3***". Có 5 mẫu "**Cao**" và 3 mẫu "**Thấp**"
    - Gini(S) = 1 - $(5/8)^2$ - $(3/8)^2$ = 0.469
- Tính chỉ mục Gini cho từng đối tượng ở bảng con 3

$$Gini_{Thu\ nhập}(S_{Độc\ thân}) = \frac{3}{8} \times 0 + \frac{2}{8} \times 0{,}5 + \frac{3}{8} \times 0{,}444 \approx 0{,}292$$

$$Gini_{Nghề\ nghiệp}(S_{Độc\ thân}) = \frac{3}{8} \times 0{,}444 + \frac{4}{8} \times 0{,}5 + \frac{1}{8} \times 0 \approx 0{,}417$$

$$Gini_{Sở\ hữu\ nhà}(S_{Độc\ thân}) = \frac{3}{8} \times 0{,}444 + \frac{5}{8} \times 0{,}48 \approx 0{,}467$$

- Vậy ta chọn thuộc tính "**Thu nhập**" để phân hoạch cho các đối tượng ở bảng con 3



- Vậy bảng con 3 còn lại các đối tượng chưa phân hoạch nằm ở bảng con trong hình 4.9 và bảng con ở hình 4.10 bên dưới

Thu nhập	Nghề nghiệp	Sở hữu nhà	*Nguy cơ rủi ro*
Cao	Vận tải	Rồi	*Thấp*
Cao	Truyền thông	Chưa	*Thấp*
Cao	Kinh doanh	Rồi	*Cao*

Hình 4.9: Bảng con 3.1

Thu nhập	Nghề nghiệp	Sở hữu nhà	*Nguy cơ rủi ro*
Trung bình	Truyền thông	Rồi	*Thấp*
Trung bình	Truyền thông	Chưa	*Cao*

Hình 4.10: Bảng con 3.2

- Xét bảng con 3.1 khi loại bỏ thuộc tính "***Thu nhập***"

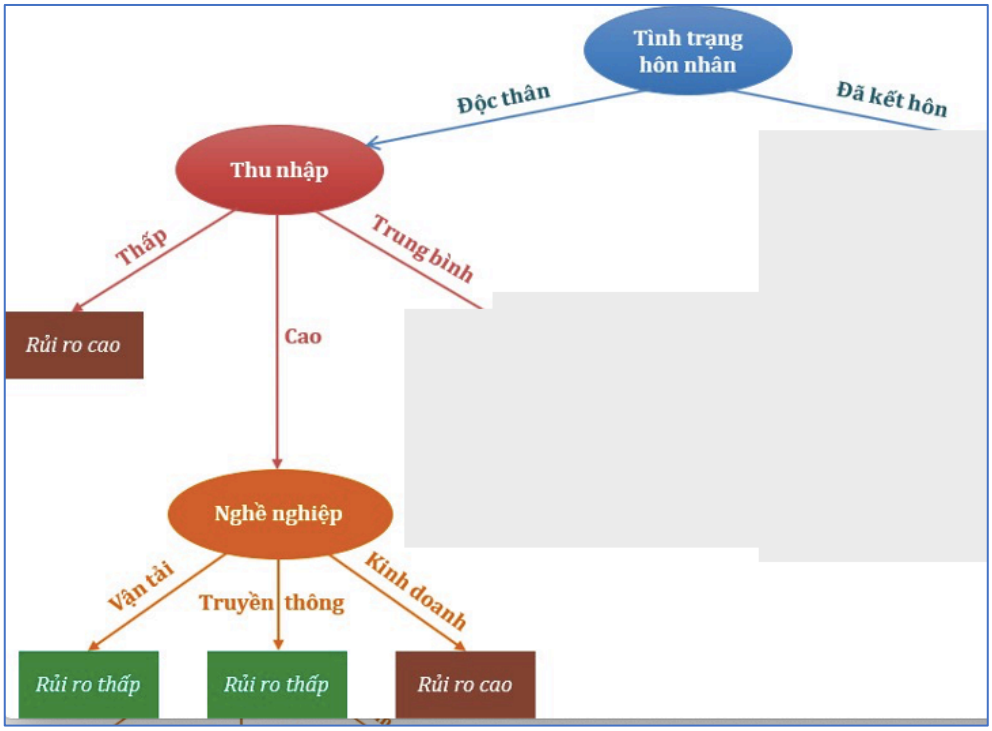
Nghề nghiệp	Sở hữu nhà	*Nguy cơ rủi ro*
Vận tải	Rồi	*Thấp*
Truyền thông	Chưa	*Thấp*
Kinh doanh	Rồi	*Cao*

- Ta tiếp tục tính chỉ mục Gini cho các thuộc tính "Nghề nghiệp" và sở hữu nhà cho bảng con này

$$Gini_{Nghề\ nghiệp}(S_{Cao}) = \frac{1}{3} \times 0 + \frac{1}{3} \times 0 + \frac{1}{3} \times 0 = 0$$

$$Gini_{Sở\ hữu\ nhà}(S_{Cao}) = \frac{1}{3} \times 0 + \frac{2}{3} \times 0{,}5 \approx 0{,}333$$

- Vậy ta cho thuộc tính "***Nghề nghiệp***" để phân hoạch cho các mẫu ở nhánh này



- Xét bảng con 3.2 khi ta loại bỏ thuộc tính "***thu nhập***"

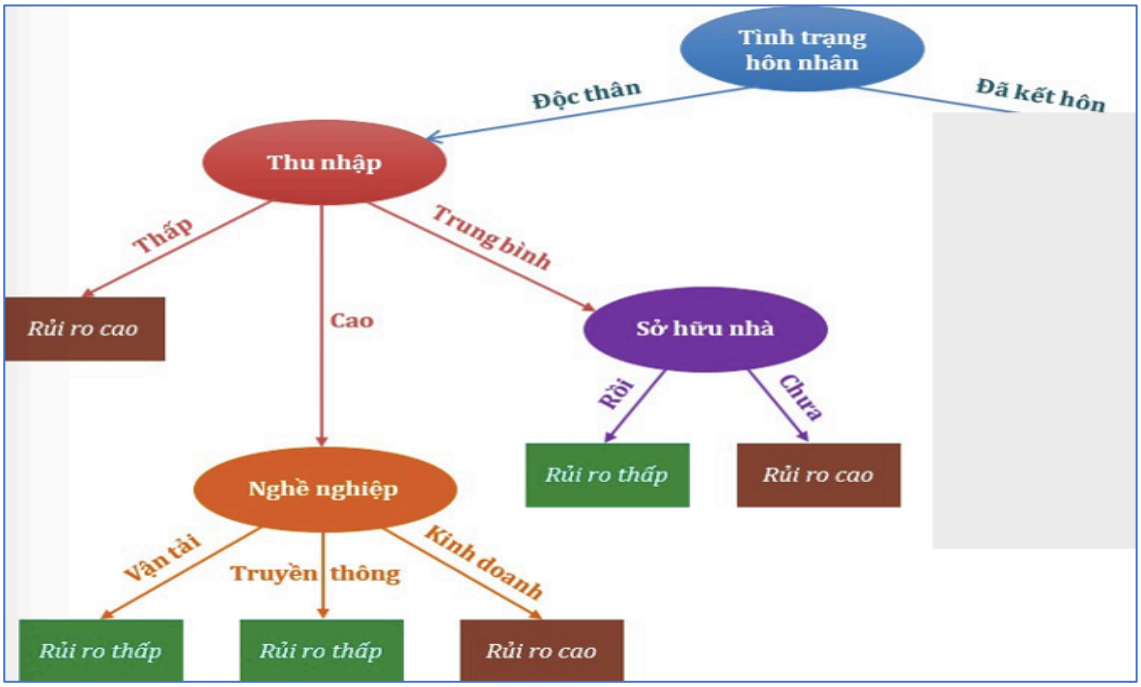
Nghề nghiệp	Sở hữu nhà	*Nguy cơ rủi ro*
Truyền thông	Rồi	*Thấp*
Truyền thông	Chưa	*Cao*

- Xét thuộc tính "***Nghề nghiệp***"

$$Gini_{Nghề\ nghiệp}(S_{Trung\ bình}) = \frac{2}{2} \times 0{,}5 = 0{,}5$$

$$Gini_{Sở\ hữu\ nhà}(S_{Trung\ bình}) = \frac{1}{2} \times 0 + \frac{1}{2} \times 0 = 0$$

- Vậy ta chọn thuộc tính "Sở hữu nhà" để phân hoạch các mẫu ở bảng con 3.2



- Xét bảng con 4 khi ta loại bỏ thuộc tính "***Tình trạng hôn nhân***".

Thu nhập	Nghề nghiệp	Sở hữu nhà	*Nguy cơ rủi ro*
Trung bình	Kinh doanh	Rồi	*Thấp*
Trung bình	Kinh doanh	Chưa	*Cao*
Cao	Kinh doanh	Chưa	*Thấp*
Thấp	Kinh doanh	Rồi	*Thấp*
Trung bình	Truyền thông	Rồi	*Thấp*
Thấp	Truyền thông	Chưa	*Thấp*
Cao	Vận tải	Rồi	*Thấp*

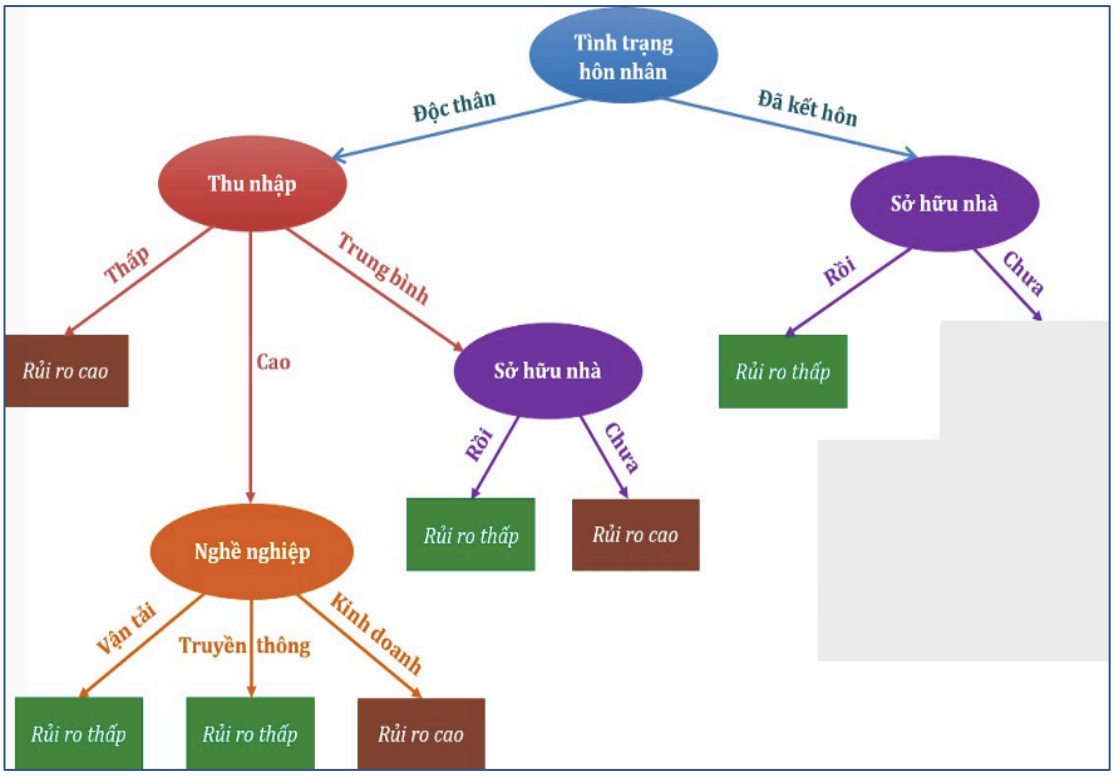
- Ta đi tính chỉ mục Gini cho các thuộc tính ở bảng con này

$$Gini_{Thu\ nhập}(S_{Đã\ kết\ hôn}) = \frac{2}{7} \times 0 + \frac{3}{7} \times 0{,}444 + \frac{3}{7} \times 0{,}5 \approx 0{,}19$$

$$Gini_{Nghề\ nghiệp}(S_{Đã\ kết\ hôn}) = \frac{1}{7} \times 0 + \frac{2}{7} \times 0 + \frac{4}{7} \times 0 \approx 0{,}214$$

$$Gini_{Sở\ hữu\ nhà}(S_{Đã\ kết\ hôn}) = \frac{3}{7} \times 0{,}444 + \frac{4}{7} \times 0 \approx 0{,}19$$

- Chỉ số Gini của hai thuộc tính "*Thu nhập*" và "***Sở hữu nhà***" thấp ngang nhau, ta lựa chọn thuộc tính "***Sở hữu nhà***" để phân hoạch
- Nhánh với giá trị Rồi nối đến nút lá Rủi ro thấp, xét nhánh Chưa ta có



- Các mẫu chưa được phân hoạch là

Thu nhập	Nghề nghiệp	Sở hữu nhà	*Nguy cơ rủi ro*
Trung bình	Kinh doanh	Chưa	*Cao*
Cao	Kinh doanh	Chưa	*Thấp*
Thấp	Truyền thông	Chưa	*Thấp*

Hình 4.11: Bảng con 4.1

- Xét bảng con khi loại thuộc tính "***Sở hữu nhà***"

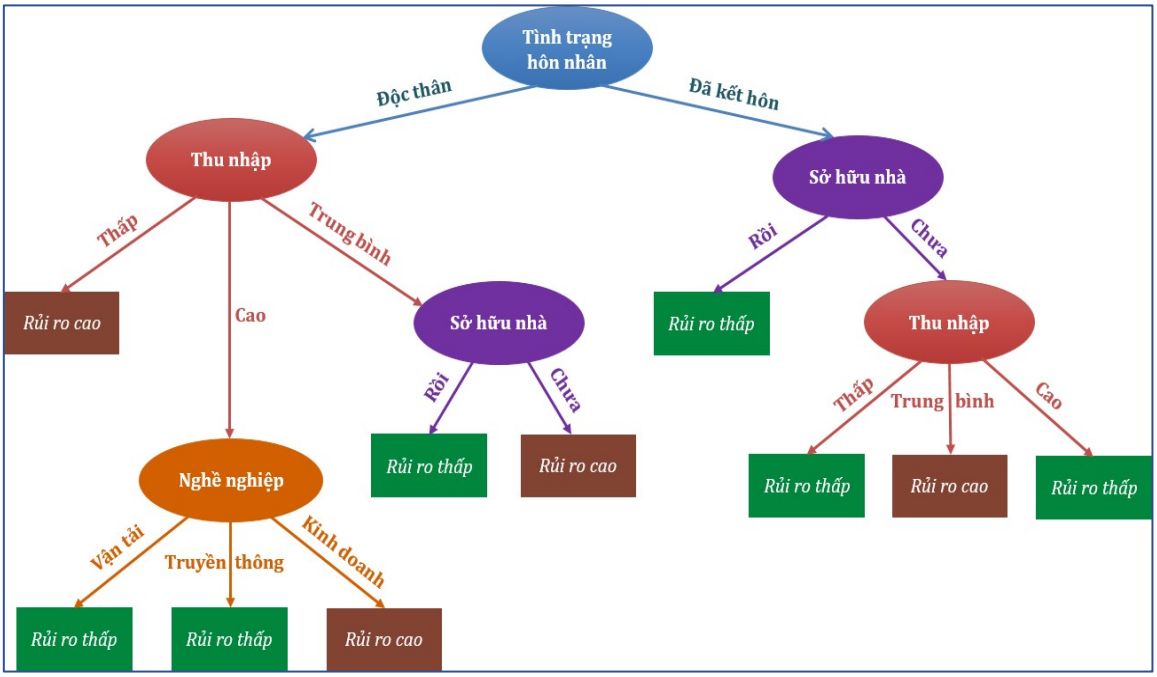
Thu nhập	Nghề nghiệp	*Nguy cơ rủi ro*
Trung bình	Kinh doanh	*Cao*
Cao	Kinh doanh	*Thấp*
Thấp	Truyền thông	*Thấp*

- Tính chỉ mục Gini cho bảng này

$$Gini_{Nghề\ nghiệp}(S_{Chưa}) = \frac{1}{3} \times 0 + \frac{2}{3} \times 0{,}5 \approx 0{,}333$$

$$Gini_{Thu\ nhập}(S_{Chưa}) = \frac{1}{2} \times 0 + \frac{1}{2} \times 0 = 0$$

- Vậy ta chọn thuộc tính "***Thu nhập***" là thuộc tính phân hoạch cho các mẫu ở ***bảng con 4.1***



- Rút luật từ cây quyết định
- **R1**: IF "***Tình trạng hôn nhân***" đọc thân và "***thu nhập***" thấp THEN Rủi ro cao
- **R2**:IF "***Tình trạng hôn nhân***" đọc thân và "***thu nhập***" cao và "***nghề nghiệp***" vận tải THEN Rủi ro **thấp**
- **R3**: IF "***Tình trạng hôn nhân***" đọc thân và "***thu nhập***" cao và "***nghề nghiệp***" truyền thông THEN Rủi ro **thấp**
- **R4**: IF "***Tình trạng hôn nhân***" đọc thân và "***thu nhập***" cao và "***nghề nghiệp***" kinh doanh THEN Rủi ro **cao**
- **R5**: IF "***Tình trạng hôn nhân***" đọc thân và "***thu nhập***" trung bình và "***Sở hữu nhà***" rồi THEN Rủi ro **thấp**
- **R6**: IF "***Tình trạng hôn nhân***" đọc thân và "***thu nhập***" trung bình và "***Sở hữu nhà***" chưa THEN Rủi ro **cao**
- **R7**: IF "***Tình trạng hôn nhân***" đã kết hôn và "***Sở hữu nhà"*** rồi THEN Rủi ro **thấp**
- **R8**: IF "***Tình trạng hôn nhân***" đã kết hôn và "***Sở hữu nhà"*** chưa và thu nhập

thấp THEN Rủi ro **thấp**

- **R9**: IF "***Tình trạng hôn nhân***" đã kết hôn và "***Sở hữu nhà"*** chưa và "***Thu nhập***" trung bình THEN Rủi ro **Cao**
- **R10**: IF "***Tình trạng hôn nhân***" đã kết hôn và "***Sở hữu nhà"*** chưa và "***Thu nhập***" cao THEN Rủi ro **Thấp**

3. Sử dụng cây quyết định bằng kỹ thuật tính chỉ mục **Gini** tiên đoán giá trị ***Nguy cơ rủi ro*** của những hồ sơ sau

- **R3**: IF "***Tình trạng hôn nhân***" đọc thân và "***thu nhập***" cao và "***nghề nghiệp***" truyền thông THEN Rủi ro **thấp**
- **R5**: IF "***Tình trạng hôn nhân***" đọc thân và "***thu nhập***" trung bình và "***Sở hữu nhà***" rồi THEN Rủi ro **thấp**
- **R7**: IF "***Tình trạng hôn nhân***" đã kết hôn và "***Sở hữu nhà"*** rồi THEN Rủi ro **thấp**

Thu nhập	Nghề nghiệp	Tình trạng hôn nhân	Sở hữu nhà	Nguy cơ rủi ro
Trung bình	Vận tải	Độc thân	Rồi	*Thấp (R5)*
Cao	Truyền thông	Độc thân	Rồi	*Thấp (R3)*
Thấp	Vận tải	Đã kết hôn	Rồi	*Thấp (R7)*

4. Sử dụng thuật toán phân lớp Bayes với kỹ thuật làm trơn **Laplace** để phân lớp các đối tượng ở câu 3

- Xét mẫu X1 = {**Thu nhập** = Trung bình, **Nghề nghiệp** = Vận tải, **Tình trạng hôn nhân** = Độc thân, **Sở hữu nhà** = Rồi}
- Ta đi tính xác suất của X thuộc lớp "***Nguy cơ rủi ro***" Cao hay Thấp
- **P(Nguy cơ rủi ro=Cao|X1)=P(Nguy cơ rủi ro=Cao) *P(X|Nguy cơ rủi ro=Cao) = P(Nguy cơ rủi ro=Cao)***P(**Thu nhập** = Trung bình| **Nguy cơ rủi ro=Cao**)*P(**Nghề nghiệp** = Vận tải | **Nguy cơ rủi ro=Cao**)*P(**Tình trạng hôn nhân** = Độc thân |**Nguy cơ rủi ro=Cao**)* P(**Sở hữu nhà** = Rồi |

Nguy cơ rủi ro=Cao)

- Thuật toán Bayes với kỹ thuật làm trơn **Laplace**
  - Để tránh trường hợp P(Xk|Ci)=0, áp dụng công thức Laplace
  - **P(Ci)=(|$C_{i,D}$|+1)/(|D|+m)**
  - **P(Xk|Ci)=(# $C_{i,D}${xk}+1)/(|$C_{i,D}$|+r)**
  - **Với**
    - m: số phân lớp
    - r: số giá trị rời rạc của thuộc tính
- Ta đi tính xác suất X1 thuộc lớp "**Nguy cơ rủi ro**"=Cao
  - P(Nguy cơ rủi ro=Cao) = $\frac{6+1}{15+2} = \frac{7}{17}$
  - P(Thu nhập=trung bình|Nguy cơ rủi ro=Cao) = $\frac{2+1}{6+3} = \frac{3}{9}$
  - P(Nghề nghiệp=Vận tải|Nguy cơ rủi ro=Cao) = $\frac{2+1}{6+3} = \frac{3}{9}$
  - P(Tình trạng hôn nhân = độc thân| Nguy cơ rủi ro=Cao) = $\frac{5+1}{6+2} = \frac{6}{8}$
  - P(Sở hữu nhà =rồi| Nguy cơ rủi ro=Cao) = $\frac{3+1}{6+2} = \frac{4}{8}$
  - P(Nguy cơ rủi ro=Cao) |X1) = $\frac{7}{17} \times \frac{3}{9} \times \frac{3}{9} \times \frac{6}{8} \times \frac{4}{8} = 0.017$
- Ta đi tính xác suất X1 thuộc lớp "**Nguy cơ rủi ro**"=Thấp
  - P(Nguy cơ rủi ro=Thấp) = $\frac{9+1}{15+2} = \frac{10}{17}$
  - P(Thu nhập=trung bình|Nguy cơ rủi ro= Thấp) = $\frac{3+1}{9+3} = \frac{4}{12}$
  - P(Nghề nghiệp=Vận tải|Nguy cơ rủi ro= Thấp) = $\frac{2+1}{9+3} = \frac{3}{12}$
  - P(Tình trạng hôn nhân = độc thân| Nguy cơ rủi ro= Thấp) = $\frac{3+1}{9+2} = \frac{4}{11}$
  - P(Sở hữu nhà =rồi| Nguy cơ rủi ro= Thấp) = $\frac{6+1}{9+2} = \frac{7}{11}$
  - P(Nguy cơ rủi ro=Thấp) |X1) = $\frac{10}{17} \times \frac{4}{12} \times \frac{3}{12} \times \frac{4}{11} \times \frac{7}{11} = 0.011$
  - Vậy X1 thuộc lớp "Nguy cơ rủi ro" Cao ví có xác suất lớn hơn

- Xét mẫu X2 = {Thu nhập = Cao, Nghề nghiệp = Truyền thông, Tình trạng hôn nhân = Độc thân, Sở hữu nhà = Rồi}
- Ta đi tính xác suất X2 thuộc lớp "**Nguy cơ rủi ro**"=Cao
  - P(Nguy cơ rủi ro=Cao) = $\frac{6+1}{15+2} = \frac{7}{17}$
  - P(Thu nhập = Cao |Nguy cơ rủi ro=Cao) = $\frac{1+1}{6+3} = \frac{2}{9}$
  - P(Nghề nghiệp = Truyền thông |Nguy cơ rủi ro=Cao) = $\frac{2+1}{6+3} = \frac{3}{9}$
  - P(Tình trạng hôn nhân = Độc thân | Nguy cơ rủi ro=Cao) = $\frac{5+1}{6+2} = \frac{6}{8}$
  - P(Sở hữu nhà = Rồi | Nguy cơ rủi ro=Cao) = $\frac{3+1}{6+2} = \frac{4}{8}$
  - P(Nguy cơ rủi ro=Cao) |X2) = $\frac{7}{17} \times \frac{2}{9} \times \frac{3}{9} \times \frac{6}{8} \times \frac{4}{8} = 0.011$
- Ta đi tính xác suất X1 thuộc lớp "**Nguy cơ rủi ro**" = Thấp
  - P(Nguy cơ rủi ro=Thấp) = $\frac{9+1}{15+2} = \frac{10}{17}$
  - P(Thu nhập = Cao |Nguy cơ rủi ro= Thấp) = $\frac{4+1}{9+3} = \frac{5}{12}$
  - P(Nghề nghiệp = Truyền thông |Nguy cơ rủi ro= Thấp) = $\frac{4+1}{9+3} = \frac{5}{12}$
  - P(Tình trạng hôn nhân = Độc thân | Nguy cơ rủi ro= Thấp) = $\frac{6+1}{9+2} = \frac{7}{11}$
  - P(Sở hữu nhà =rồi| Nguy cơ rủi ro= Thấp) = $\frac{6+1}{9+2} = \frac{7}{11}$
  - P(Nguy cơ rủi ro=Thấp) |X2) = $\frac{10}{17} \times \frac{5}{12} \times \frac{5}{12} \times \frac{7}{11} \times \frac{7}{11} = 0.024$
  - Vậy X2 thuộc lớp có "***Nguy cơ rủi ro***" thấp
- Xét mẫu X3 {Thu nhập = Thấp, Nghề nghiệp = Vận tải, Tình trạng hôn nhân = Đã kết hôn, Sở hữu nhà = Rồi}
- Ta đi tính xác suất X1 thuộc lớp "**Nguy cơ rủi ro**"=Cao
  - P(Nguy cơ rủi ro=Cao) = $\frac{6+1}{15+2} = \frac{7}{17}$
  - P(Thu nhập = Thấp |Nguy cơ rủi ro=Cao) = $\frac{3+1}{6+3} = \frac{4}{9}$
  - P(Nghề nghiệp = Vận tải |Nguy cơ rủi ro=Cao) = $\frac{2+1}{6+3} = \frac{3}{9}$

- P(Tình trạng hôn nhân = Đã kết hôn | Nguy cơ rủi ro=Cao) = $\frac{1+1}{6+2} = \frac{2}{8}$
- P(Sở hữu nhà = Rồi | Nguy cơ rủi ro=Cao) = $\frac{3+1}{6+2} = \frac{4}{8}$
- P(Nguy cơ rủi ro=Cao) |X3) = $\frac{7}{17} \times \frac{4}{9} \times \frac{3}{9} \times \frac{2}{8} \times \frac{4}{8} = 0.007$

- Ta đi tính xác suất X3 thuộc lớp "**Nguy cơ rủi ro**"=Thấp
  - P(Nguy cơ rủi ro=Thấp) = $\frac{9+1}{15+2} = \frac{10}{17}$
  - P(Thu nhập = Thấp |Nguy cơ rủi ro= Thấp) = $\frac{2+1}{9+3} = \frac{3}{12}$
  - P(Nghề nghiệp = Vận tải |Nguy cơ rủi ro= Thấp) = $\frac{2+1}{9+3} = \frac{3}{12}$
  - P(Tình trạng hôn nhân=Đã kết hôn|Nguy cơ rủi ro= Thấp) = $\frac{6+1}{9+2} = \frac{7}{11}$
  - P(Sở hữu nhà =rồi| Nguy cơ rủi ro= Thấp) = $\frac{6+1}{9+2} = \frac{7}{11}$
  - P(Nguy cơ rủi ro=Thấp) |X3) = $\frac{10}{17} \times \frac{3}{12} \times \frac{3}{12} \times \frac{7}{11} \times \frac{7}{11} = 0.014$
  - Vậy X3 thuộc lớp "***Nguy cơ rủi ro***" **thấp**

IV. Thực hành

1. Một doanh nghiệp sản xuất đồ chơi cho trẻ em muốn dự đoán doanh số của các sản phẩm sắp đưa ra thị trường, họ thu thập những dữ liệu dưới đây

Loại	Số màu	Kích thước	Chất liệu	*Doanh số bán*
Điều khiển	3	Nhỏ	Nhựa PP	*Cao*
Xếp hình	5	Vừa	Cao su	*Thấp*
Xếp hình	7	To	Nhựa PP	*Thấp*
Điều khiển	5	Nhỏ	Cao su	*Thấp*
Búp bê	3	Vừa	Nhựa PP	*Thấp*
Điều khiển	5	Vừa	Nhựa PP	*Cao*
Búp bê	5	To	Nhựa PP	*Cao*
Điều khiển	7	Vừa	Cao su	*Thấp*
Xếp hình	7	To	Cao su	*Cao*
Xếp hình	3	To	Nhựa PP	*Thấp*
Búp bê	3	Nhỏ	Cao su	*Thấp*
Xếp hình	3	Nhỏ	Nhựa PP	*Cao*
Điều khiển	5	To	Cao su	*Thấp*
Búp bê	5	Vừa	Nhựa PP	*Cao*
Búp bê	7	To	Nhựa PP	*Cao*

Sinh viên giúp doanh nghiệp bằng cách thực hiện những yêu cầu sau

a) Tính giá trị độ lợi thông tin (information gain) của các thuộc tính và vẽ cây quyết định theo thuật toán ID3 cho dữ liệu trên
b) Tính giá trị chỉ số Gini (gini index) của các thuộc tính và vẽ cây quyết định theo thuật toán CART cho dữ liệu trên
c) Sử dụng một trong hai cây quyết định ở trên để tiên đoán giá trị Doanh số bán của những sản phẩm sau:

Loại	Số màu	Kích thước	Chất liệu
Búp bê	3	To	Cao su
Xếp hình	5	To	Nhựa PP
Điều khiển	3	Vừa	Cao su

d) Sử dụng thuật toán Naïve Bayes và làm trơn Laplace để dự đoán giá trị Doanh số bán của những sản phẩm trong Yêu cầu ở câu c

2. Phân tích cảm xúc (sentiment analysis) là một lĩnh vực nghiên cứu rất quan trọng và thú vị trong khai thác dữ liệu văn bản (text mining). Sinh viên có thể làm quen với vấn đề này thông qua bài tập sau. Người ta phân tích các trạng thái trên mạng xã hội và thống kê được số lần xuất hiện của các từ khóa (term) được trình bày trong bảng dữ liệu bên dưới, Cảm xúc là thuộc tính phân lớp

giảm	người	chuyển	yêu	vừa	đi	*Cảm xúc*
0..5	11..20	>20	11..20	>20	0..5	*tốt*
11..20	6..10	6..10	0..5	11..20	11..20	*tốt*
6..10	0..5	6..10	11..20	0..5	6..10	*xấu*
>20	0..5	11..20	6..10	0..5	>20	*bình thường*
0..5	>20	11..20	0..5	6..10	0..5	*xấu*
0..5	6..10	0..5	0..5	11..20	11..20	*xấu*
0..5	6..10	11..20	0..5	6..10	0..5	*tốt*
11..20	>20	0..5	11..20	0..5	11..20	*bình thường*
0..5	0..5	6..10	6..10	6..10	>20	*tốt*
11..20	0..5	11..20	11..20	0..5	11..20	*tốt*
>20	6..10	0..5	0..5	0..5	6..10	*xấu*
0..5	0..5	11..20	0..5	11..20	>20	*bình thường*
6..10	11..20	6..10	>20	0..5	6..10	*bình thường*
11..20	6..10	>20	11..20	0..5	0..5	*xấu*

a) Tính giá trị chỉ số Gini của các thuộc tính và vẽ cây quyết định theo thuật toán CART cho dữ liệu trên

b) Sử dụng cây quyết định và thuật toán Naïve Bayes để dự đoán cảm xúc của những trạng thái sau:

giảm	người	chuyển	yêu	vừa	đi
0..5	6..10	0..5	11..20	6..10	0..5
0..5	0..5	6..10	0..5	11..20	>20
6..10	0..5	11..20	>20	6..10	6..10
6..10	11..20	6..10	6..10	>20	0..5

V. Bài tập thêm

1. Chọn một ngôn ngữ lập trình, sinh viên hãy cài đặt:

a) Thuật toán ID3
b) Thuật toán CART
c) Thuật toán Naïve Bayes

2. Giả sử tồn tại một bảng có số dòng là vô tận do dữ liệu liên tục được thêm vào. Để đọc hết toàn bộ dữ liệu sẽ mất rất nhiều thời gian nên yêu cầu đặt ra là chỉ được đọc tất cả một lần duy nhất.

a) Hãy thiết kế một mô hình để áp dụng có hiệu quả thuật toán Naïve Bayes trên dữ liệu này.
b) Người ta muốn theo dõi, so sánh sự thay đổi của mô hình phân lớp theo thời gian (ví dụ: mô hình phân lớp của tuần trước so với hiện tại...). Sinh viên hãy gợi ý phương pháp thực hiện điều này.

VI. Tài liệu tham khảo

1. Slide bài giảng lý thuyết môn Khai thác dữ liệu.
2. Han, J., Kamber, M. & Pei, J. (2012). Data mining concepts and techniques, third edition Morgan Kaufmann Publishers