BÁO CÁO BÀI THỰC HÀNH SỐ 4

Giảng viên hướng dẫn: ThS. Vũ Minh Sang

Sinh viên thực hiện: Nguyễn Minh Duy

MSSV: 21522005

Lớp: IS252.O21.HTCL

Bài làm

IV. Thực hành

Câu 2: (Cơ bản)

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) Xác định tất cả những mâu thuẫn có thể có trong dữ liệu

Tập dữ liệu trên **không xảy ra mâu thuẫn** vì không tồn tại các dòng dữ liệu có giá trị thuộc tính giống nhau nhưng lại thuộc phân lớp khác nhau.

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

+ Thuộc tính "giảm" :

Giá trị phân biệt	Số dòng gán nhãn "tốt"	Số dòng gán nhãn "bình thường"	Số dòng gán nhãn "xấu"	Tổng số dòng
0..5	3	1	2	6
6..10	0	1	1	2
11..20	2	1	1	4
>20	0	1	1	2

$$Gini_{0..5}(S) = 1 - \left(\frac{3}{6}\right)^2 - \left(\frac{1}{6}\right)^2 - \left(\frac{2}{6}\right)^2 = \frac{11}{18} \approx 0{,}61$$

$$Gini_{6..10}(S) = 1 - \left(\frac{0}{2}\right)^2 - \left(\frac{1}{2}\right)^2 - \left(\frac{1}{2}\right)^2 = \frac{1}{2} = 0{,}5$$

$$Gini_{11..20}(S) = 1 - \left(\frac{2}{4}\right)^2 - \left(\frac{1}{4}\right)^2 - \left(\frac{1}{4}\right)^2 = \frac{5}{8} = 0{,}625$$

$$Gini_{>20}(S) = 1 - \left(\frac{0}{2}\right)^2 - \left(\frac{1}{2}\right)^2 - \left(\frac{1}{2}\right)^2 = \frac{1}{2} = 0{,}5$$

$$Gini_{giảm}(S) = \frac{6}{14} \times \frac{11}{18} + \frac{2}{14} \times \frac{1}{2} + \frac{4}{14} \times \frac{5}{8} + \frac{2}{14} \times \frac{1}{2} \approx 0{,}583$$

+ Thuộc tính "người" :

Giá trị phân biệt	Số dòng gán nhãn "tốt"	Số dòng gán nhãn "bình thường"	Số dòng gán nhãn "xấu"	Tổng số dòng
0..5	2	2	1	5
6..10	2	0	3	5
11..20	1	1	0	2
>20	0	1	1	2

$$Gini_{0..5}(S) = 1 - \left(\frac{2}{5}\right)^2 - \left(\frac{2}{5}\right)^2 - \left(\frac{1}{5}\right)^2 = \frac{16}{25} = 0{,}64$$

$$Gini_{6..10}(S) = 1 - \left(\frac{2}{5}\right)^2 - \left(\frac{0}{5}\right)^2 - \left(\frac{3}{5}\right)^2 = \frac{12}{25} = 0{,}48$$

$$Gini_{11..20}(S) = 1 - \left(\frac{1}{2}\right)^2 - \left(\frac{1}{2}\right)^2 - \left(\frac{0}{2}\right)^2 = \frac{1}{2} = 0{,}5$$

$$Gini_{>20}(S) = 1 - \left(\frac{0}{2}\right)^2 - \left(\frac{1}{2}\right)^2 - \left(\frac{1}{2}\right)^2 = \frac{1}{2} = 0{,}5$$

$$Gini_{người}(S) = \frac{5}{14} \times \frac{16}{25} + \frac{5}{14} \times \frac{12}{25} + \frac{2}{14} \times \frac{1}{2} + \frac{2}{14} \times \frac{1}{2} \approx 0{,}543$$

+ Thuộc tính "chuyển" :

Giá trị phân biệt	Số dòng gán nhãn "tốt"	Số dòng gán nhãn "bình thường"	Số dòng gán nhãn "xấu"	Tổng số dòng
0..5	0	1	2	3
6..10	2	1	1	4
11..20	2	2	1	5
>20	1	0	1	2

$$Gini_{0..5}(S) = 1 - \left(\frac{0}{3}\right)^2 - \left(\frac{1}{3}\right)^2 - \left(\frac{2}{3}\right)^2 = \frac{4}{9} \approx 0{,}44$$

$$Gini_{6..10}(S) = 1 - \left(\frac{2}{4}\right)^2 - \left(\frac{1}{4}\right)^2 - \left(\frac{1}{4}\right)^2 = \frac{5}{8} = 0{,}625$$

$$Gini_{11..20}(S) = 1 - \left(\frac{2}{5}\right)^2 - \left(\frac{2}{5}\right)^2 - \left(\frac{1}{5}\right)^2 = \frac{16}{25} = 0{,}64$$

$$Gini_{>20}(S) = 1 - \left(\frac{1}{2}\right)^2 - \left(\frac{0}{2}\right)^2 - \left(\frac{1}{2}\right)^2 = \frac{1}{2} = 0{,}5$$

$$Gini_{chuyển}(S) = \frac{3}{14} \times \frac{4}{9} + \frac{4}{14} \times \frac{5}{8} + \frac{5}{14} \times \frac{16}{25} + \frac{2}{14} \times \frac{1}{2} \approx 0{,}574$$

+ Thuộc tính "yêu" :

Giá trị phân biệt	Số dòng gán nhãn "tốt"	Số dòng gán nhãn "bình thường"	Số dòng gán nhãn "xấu"	Tổng số dòng
0..5	2	1	3	6
6..10	1	1	0	2
11..20	2	1	2	5
>20	0	1	0	1

$$Gini_{0..5}(S) = 1 - \left(\frac{2}{6}\right)^2 - \left(\frac{1}{6}\right)^2 - \left(\frac{3}{6}\right)^2 = \frac{11}{18} \approx 0{,}61$$

$$Gini_{6..10}(S) = 1 - \left(\frac{1}{2}\right)^2 - \left(\frac{1}{2}\right)^2 - \left(\frac{0}{2}\right)^2 = \frac{1}{2} = 0{,}5$$

$$Gini_{11..20}(S) = 1 - \left(\frac{2}{5}\right)^2 - \left(\frac{1}{5}\right)^2 - \left(\frac{2}{5}\right)^2 = \frac{16}{25} = 0{,}64$$

$$Gini_{>20}(S) = 1 - \left(\frac{1}{1}\right)^2 = 0$$

$$Gini_{yêu}(S) = \frac{6}{14} \times \frac{11}{18} + \frac{2}{14} \times \frac{1}{2} + \frac{5}{14} \times \frac{16}{25} + \frac{1}{14} \times 0 \approx 0{,}562$$

+ Thuộc tính "vừa" :

Giá trị phân biệt	Số dòng gán nhãn "tốt"	Số dòng gán nhãn "bình thường"	Số dòng gán nhãn "xấu"	Tổng số dòng
0..5	1	3	3	7
6..10	2	0	1	3
11..20	1	1	1	3
>20	1	0	0	1

$$Gini_{0..5}(S) = 1 - \left(\frac{1}{7}\right)^2 - \left(\frac{3}{7}\right)^2 - \left(\frac{3}{7}\right)^2 = \frac{30}{49} \approx 0{,}61$$

$$Gini_{6..10}(S) = 1 - \left(\frac{2}{3}\right)^2 - \left(\frac{0}{3}\right)^2 - \left(\frac{1}{3}\right)^2 = \frac{4}{9} \approx 0{,}44$$

$$Gini_{11..20}(S) = 1 - \left(\frac{1}{3}\right)^2 - \left(\frac{1}{3}\right)^2 - \left(\frac{1}{3}\right)^2 = \frac{2}{3}$$

$$Gini_{>20}(S) = 1 - \left(\frac{1}{1}\right)^2 - \left(\frac{0}{1}\right)^2 - \left(\frac{0}{1}\right)^2 = 0$$

$$Gini_{vừa}(S) = \frac{7}{14} \times \frac{30}{49} + \frac{3}{14} \times \frac{4}{9} + \frac{3}{14} \times \frac{2}{3} \approx 0{,}544$$

+ Thuộc tính "đi" :

Giá trị phân biệt	Số dòng gán nhãn "tốt"	Số dòng gán nhãn "bình thường"	Số dòng gán nhãn "xấu"	Tổng số dòng
0..5	2	0	2	4
6..10	0	1	2	3
11..20	2	1	1	4
>20	1	2	0	3

$$Gini_{0..5}(S) = 1 - \left(\frac{2}{4}\right)^2 - \left(\frac{0}{4}\right)^2 - \left(\frac{2}{4}\right)^2 = \frac{1}{2} = 0{,}5$$

$$Gini_{6..10}(S) = 1 - \left(\frac{0}{3}\right)^2 - \left(\frac{1}{3}\right)^2 - \left(\frac{2}{3}\right)^2 = \frac{4}{9} \approx 0{,}44$$

$$Gini_{11..20}(S) = 1 - \left(\frac{2}{4}\right)^2 - \left(\frac{1}{4}\right)^2 - \left(\frac{1}{4}\right)^2 = \frac{5}{8} = 0{,}625$$

$$Gini_{>20}(S) = 1 - \left(\frac{1}{3}\right)^2 - \left(\frac{2}{3}\right)^2 - \left(\frac{0}{3}\right)^2 = \frac{4}{9} \approx 0{,}44$$

$$Gini_{đi}(S) = \frac{4}{14} \times \frac{1}{2} + \frac{3}{14} \times \frac{4}{9} + \frac{4}{14} \times \frac{5}{8} + \frac{3}{14} \times \frac{4}{9} \approx 0{,}512$$

=> Chọn thuộc tính "đi" làm root node vì có chỉ số Gini thấp nhất.

Tập dữ liệu lúc này được chia làm hai phần tương ứng với hai nhánh cây theo giá trị của thuộc tính **"đi"**.

Phần có giá trị 0..5 gồm 4 dòng, phần có giá trị 6..10 gồm 3 dòng, phần có giá trị 11..20 gồm 4 dòng và phần có giá trị >20 gồm 3 dòng.

Với nhánh "0..5", xét lần lượt:

+ Thuộc tính "giảm" – đi = 0..5:

Giá trị phân biệt	Số dòng gán nhãn "tốt"	Số dòng gán nhãn "bình thường"	Số dòng gán nhãn "xấu"	Tổng số dòng
0..5	2	0	1	3
6..10	0	0	0	0
11..20	0	0	1	1
>20	0	0	0	0

$$Gini_{0..5}(S_{\text{đi}=0..5}) = 1 - \left(\frac{2}{3}\right)^2 - \left(\frac{0}{3}\right)^2 - \left(\frac{1}{3}\right)^2 = \frac{4}{9}$$

$$Gini_{giảm}(S_{\text{đi}=0..5}) = \frac{3}{4} \times \frac{4}{9} \approx 0{,}33$$

+ Thuộc tính "người" – đi = 0..5:

Giá trị phân biệt	Số dòng gán nhãn "tốt"	Số dòng gán nhãn "bình thường"	Số dòng gán nhãn "xấu"	Tổng số dòng
0..5	0	0	0	0
6..10	1	0	1	2
11..20	1	0	0	1
>20	0	0	1	1

$$Gini_{6..10}(S_{\text{đi}=0..5}) = 1 - \left(\frac{1}{2}\right)^2 - \left(\frac{0}{5}\right)^2 - \left(\frac{1}{2}\right)^2 = \frac{1}{2}$$

$$Gini_{người}(S_{\text{đi}=0..5}) = \frac{2}{4} \times \frac{1}{2} = 0{,}25$$

+ Thuộc tính "chuyển" – đi = 0..5 :

Giá trị phân biệt	Số dòng gán nhãn "tốt"	Số dòng gán nhãn "bình thường"	Số dòng gán nhãn "xấu"	Tổng số dòng
0..5	0	0	0	0
6..10	0	0	0	0
11..20	1	0	1	2
>20	1	0	1	2

$$Gini_{11..20}(S_{\text{đi=0..5}}) = 1 - \left(\frac{1}{2}\right)^2 - \left(\frac{0}{2}\right)^2 - \left(\frac{1}{2}\right)^2 = \frac{1}{2}$$

$$Gini_{>20}(S_{\text{đi=0..5}}) = 1 - \left(\frac{1}{2}\right)^2 - \left(\frac{0}{2}\right)^2 - \left(\frac{1}{2}\right)^2 = \frac{1}{2}$$

$$Gini_{chuyển}(S_{\text{đi=0..5}}) = \frac{2}{4} \times \frac{1}{2} + \frac{2}{4} \times \frac{1}{2} = 0{,}5$$

+ Thuộc tính "yêu" – đi = 0..5:

Giá trị phân biệt	Số dòng gán nhãn "tốt"	Số dòng gán nhãn "bình thường"	Số dòng gán nhãn "xấu"	Tổng số dòng
0..5	1	0	1	2
6..10	0	0	0	0
11..20	1	0	1	2
>20	0	0	0	0

$$Gini_{0..5}(S_{\text{đi=0..5}}) = 1 - \left(\frac{1}{2}\right)^2 - \left(\frac{0}{2}\right)^2 - \left(\frac{1}{2}\right)^2 = \frac{1}{2}$$

$$Gini_{11..20}(S_{\text{đi=0..5}}) = 1 - \left(\frac{1}{2}\right)^2 - \left(\frac{0}{2}\right)^2 - \left(\frac{1}{2}\right)^2 = \frac{1}{2}$$

$$Gini_{yêu}(S_{\text{đi=0..5}}) = \frac{2}{4} \times \frac{1}{2} + \frac{2}{4} \times \frac{1}{2} = 0{,}5$$

+ Thuộc tính "vừa" – đi = 0..5:

Giá trị phân biệt	Số dòng gán nhãn "tốt"	Số dòng gán nhãn "bình thường"	Số dòng gán nhãn "xấu"	Tổng số dòng
0..5	0	0	1	1
6..10	1	0	1	2
11..20	0	0	0	0
>20	1	0	0	1

$Gini_{0..5}(S_{\text{đi}=0..5}) = 0$

$$Gini_{6..10}(S_{\text{đi}=0..5}) = 1 - \left(\frac{1}{2}\right)^2 - \left(\frac{0}{2}\right)^2 - \left(\frac{1}{2}\right)^2 = \frac{1}{2}$$

$$Gini_{vừa}(S_{\text{đi}=0..5}) = \frac{2}{4} \times \frac{1}{2} = 0{,}25$$

Vì thuộc tính "người" và "vừa" có chỉ số Gini thấp bằng nhau, ta chọn ngẫu nhiên thuộc tính "người" để chia nhánh tiếp. Với giá trị "người" = 11..20, ta luôn có phân lớp tốt và giá trị "người" = >20 ta luôn có phân xấu, vì vậy nhánh này đi đến nút lá và không cần xét tiếp. Nhánh con tương ứng với giá trị còn lại là 6..10 sẽ tiếp tục được phát triển.

Với nhánh "người" tương ứng với giá trị "6..10", gồm 2 dòng dữ liệu, xét lần lượt:

+ Thuộc tính "giảm" – đi = 0..5 và người = 6..10:

Giá trị phân biệt	Số dòng gán nhãn "tốt"	Số dòng gán nhãn "bình thường"	Số dòng gán nhãn "xấu"	Tổng số dòng
0..5	1	0	0	1
6..10	0	0	0	0
11..20	0	0	1	1
>20	0	0	0	0

$Gini_{giảm}(S_{\text{người} = 6..10}) = 0$

+ Thuộc tính "chuyển" – đi = 0..5 và người = 6..10:

Giá trị phân biệt	Số dòng gán nhãn "tốt"	Số dòng gán nhãn "bình thường"	Số dòng gán nhãn "xấu"	Tổng số dòng
0..5	0	0	0	0
6..10	0	0	0	0
11..20	1	0	0	1
>20	0	0	1	1

$Gini_{chuyển}(S_{\text{người = 6..10}}) = 0$

+ Thuộc tính "yêu" – đi = 0..5 và người = 6..10:

Giá trị phân biệt	Số dòng gán nhãn "tốt"	Số dòng gán nhãn "bình thường"	Số dòng gán nhãn "xấu"	Tổng số dòng
0..5	1	0	0	1
6..10	0	0	0	0
11..20	0	0	1	1
>20	0	0	0	0

$Gini_{yêu}(S_{\text{người = 6..10}}) = 0$

+ Thuộc tính "vừa" – đi = 0..5 và người = 6..10:

Giá trị phân biệt	Số dòng gán nhãn "tốt"	Số dòng gán nhãn "bình thường"	Số dòng gán nhãn "xấu"	Tổng số dòng
0..5	1	0	0	1
6..10	0	0	1	1
11..20	0	0	0	0
>20	0	0	0	0

$Gini_{vừa}(S_{\text{người = 6..10}}) = 0$

Vì tất cả thuộc tính có chỉ số Gini thấp bằng nhau, ta chọn ngẫu nhiên thuộc tính "giảm" để chia nhánh tiếp. Với giá trị "giảm" = 11..20, ta luôn có phân lớp xấu và giá trị "giảm" = 0..5 ta luôn có phân tốt, vì vậy nhánh này đi đến nút lá và không cần xét tiếp. Kết thúc chia nhánh.

Trở lại với nhánh đi = "6..10" gồm 3 dòng dữ liệu, xét lần lượt:

+ Thuộc tính "giảm" – đi = 6..10:

Giá trị phân biệt	Số dòng gán nhãn "tốt"	Số dòng gán nhãn "bình thường"	Số dòng gán nhãn "xấu"	Tổng số dòng
0..5	0	0	0	0
6..10	0	1	1	2
11..20	0	0	0	0
>20	0	0	1	1

$$Gini_{6..10}(S_{\text{đi=6..10}}) = 1 - \left(\frac{1}{2}\right)^2 - \left(\frac{1}{2}\right)^2 = \frac{1}{2}$$

$$Gini_{giảm}(S_{\text{đi=6..10}}) = \frac{2}{3} \times \frac{1}{2} \approx 0{,}33$$

+ Thuộc tính "người" – đi = 6..10:

Giá trị phân biệt	Số dòng gán nhãn "tốt"	Số dòng gán nhãn "bình thường"	Số dòng gán nhãn "xấu"	Tổng số dòng
0..5	0	0	1	1
6..10	0	0	1	1
11..20	0	1	0	1
>20	0	0	0	0

$$Gini_{người}(S_{\text{đi=6..10}}) = 0$$

+ Thuộc tính "chuyển" – đi = 6..10:

Giá trị phân biệt	Số dòng gán nhãn "tốt"	Số dòng gán nhãn "bình thường"	Số dòng gán nhãn "xấu"	Tổng số dòng
0..5	0	0	1	1
6..10	0	1	1	2
11..20	0	0	0	0
>20	0	0	0	0

$$Gini_{6..10}(S_{\text{đi=6..10}}) = 1 - \left(\frac{0}{2}\right)^2 - \left(\frac{1}{2}\right)^2 - \left(\frac{1}{2}\right)^2 = \frac{1}{2}$$

$$Gini_{chuyển}(S_{\text{đi=6..10}}) = \frac{2}{3} \times \frac{1}{2} \approx 0{,}33$$

+ Thuộc tính "yêu" – đi = 6..10:

Giá trị phân biệt	Số dòng gán nhãn "tốt"	Số dòng gán nhãn "bình thường"	Số dòng gán nhãn "xấu"	Tổng số dòng
0..5	0	0	1	1
6..10	0	0	0	0
11..20	0	0	1	1
>20	0	1	0	1

$$Gini_{yêu}(S_{\text{đi=6..10}}) = 0$$

+ Thuộc tính "vừa" – đi = 6..10:

Giá trị phân biệt	Số dòng gán nhãn "tốt"	Số dòng gán nhãn "bình thường"	Số dòng gán nhãn "xấu"	Tổng số dòng
0..5	0	1	2	3
6..10	0	0	0	0
11..20	0	0	0	0

>20	0	0	0	0

$$Gini_{0..5}(S_{\text{đi=6..10}}) = 1 - \left(\frac{1}{3}\right)^2 - \left(\frac{2}{3}\right)^2 = \frac{4}{9}$$

$$Gini_{vừa}(S_{\text{đi=0..5}}) = \frac{3}{3} \times \frac{4}{9} \approx 0{,}44$$

Vì thuộc tính "người" và "yêu" có chỉ số Gini thấp bằng nhau, ta chọn ngẫu nhiên thuộc tính "người" để chia nhánh tiếp. Với giá trị "người" = 6..10 hoặc 0..5, ta luôn có phân lớp xấu, giá trị "người" = 11..20 ta luôn có phân lớp bình thường. Nhánh này đi đến tất cả nút lá và kết thúc chia nhánh.

Trở lại với nhánh đi = "11..20" gồm 4 dòng dữ liệu, xét lần lượt:

+ Thuộc tính "giảm" – đi = 11.20:

Giá trị phân biệt	Số dòng gán nhãn "tốt"	Số dòng gán nhãn "bình thường"	Số dòng gán nhãn "xấu"	Tổng số dòng
0..5	0	0	0	0
6..10	0	0	1	1
11..20	2	1	0	3
>20	0	0	0	0

$$Gini_{6..10}(S_{\text{đi=11..20}}) = 0$$

$$Gini_{11..20}(S_{\text{đi=11..20}}) = \frac{4}{9}$$

$$Gini_{giảm}(S_{\text{đi=11..20}}) = \frac{3}{4} \times \frac{4}{9} \approx 0{,}33$$

+ Thuộc tính "người" – đi = 11.20:

Giá trị phân biệt	Số dòng gán nhãn "tốt"	Số dòng gán nhãn "bình thường"	Số dòng gán nhãn "xấu"	Tổng số dòng
0..5	1	0	0	1

6..10	1	0	1	2
11..20	0	0	0	0
>20	0	1	0	1

$Gini_{6..10}(S_{\text{đi=11..20}}) = \frac{1}{2}$

$Gini_{người}(S_{\text{đi=11..20}}) = \frac{2}{4} \times \frac{1}{2} = 0{,}25$

+ Thuộc tính "chuyển" – đi = 11.20:

Giá trị phân biệt	Số dòng gán nhãn "tốt"	Số dòng gán nhãn "bình thường"	Số dòng gán nhãn "xấu"	Tổng số dòng
0..5	0	1	1	2
6..10	1	0	0	1
11..20	1	0	0	1
>20	0	0	0	0

$Gini_{0..5}(S_{\text{đi=11..20}}) = \frac{1}{2}$

$Gini_{chuyển}(S_{\text{đi=11..20}}) = \frac{2}{4} \times \frac{1}{2} = 0{,}25$

+ Thuộc tính "yêu" – đi = 11.20:

Giá trị phân biệt	Số dòng gán nhãn "tốt"	Số dòng gán nhãn "bình thường"	Số dòng gán nhãn "xấu"	Tổng số dòng
0..5	1	0	1	2
6..10	0	0	0	0
11..20	1	1	0	2
>20	0	0	0	0

$Gini_{0..5}(S_{\text{đi=11..20}}) = \frac{1}{2}$

$$Gini_{11..20}(S_{\text{đi}=11..20}) = \frac{1}{2}$$

$$Gini_{yêu}(S_{\text{đi}=11..20}) = \frac{2}{4} \times \frac{1}{2} + \frac{2}{4} \times \frac{1}{2} = 0{,}5$$

+ Thuộc tính "vừa" – đi = 11.20:

Giá trị phân biệt	Số dòng gán nhãn "tốt"	Số dòng gán nhãn "bình thường"	Số dòng gán nhãn "xấu"	Tổng số dòng
0..5	1	1	0	2
6..10	0	0	0	0
11..20	1	0	1	2
>20	0	0	0	0

$$Gini_{0..5}(S_{\text{đi}=11..20}) = \frac{1}{2}$$

$$Gini_{11..20}(S_{\text{đi}=11..20}) = \frac{1}{2}$$

$$Gini_{vừa}(S_{\text{đi}=11..20}) = \frac{2}{4} \times \frac{1}{2} + \frac{2}{4} \times \frac{1}{2} = 0{,}5$$

Vì thuộc tính "người" và "chuyển" có chỉ số Gini thấp bằng nhau, ta chọn ngẫu nhiên thuộc tính "người" để chia nhánh tiếp. Với giá trị "người" = 0..5, ta luôn có phân lớp tốt, giá trị "người" = >20 ta luôn có phân lớp bình thường. Nhánh con tương ứng với giá trị còn lại là 6..10 sẽ tiếp tục được phát triển.

Với nhánh "người" tương ứng với giá trị "6..10", gồm 2 dòng dữ liệu, xét lần lượt:

+ Thuộc tính "giảm" – đi = 11..20 và người = 6..10:

Giá trị phân biệt	Số dòng gán nhãn "tốt"	Số dòng gán nhãn "bình thường"	Số dòng gán nhãn "xấu"	Tổng số dòng
0..5	0	0	1	1
6..10	0	0	0	0
11..20	1	0	0	1

>20	0	0	0	0

$Gini_{giảm}(S_{\text{người = 6..10}}) = 0$

+ Thuộc tính "chuyển" – đi = 11..20 và người = 6..10:

Giá trị phân biệt	Số dòng gán nhãn "tốt"	Số dòng gán nhãn "bình thường"	Số dòng gán nhãn "xấu"	Tổng số dòng
0..5	0	0	1	1
6..10	1	0	0	1
11..20	0	0	0	0
>20	0	0	0	0

$Gini_{chuyển}(S_{\text{người = 6..10}}) = 0$

+ Thuộc tính "yêu" – đi = 0..5 và người = 6..10:

Giá trị phân biệt	Số dòng gán nhãn "tốt"	Số dòng gán nhãn "bình thường"	Số dòng gán nhãn "xấu"	Tổng số dòng
0..5	1	0	1	2
6..10	0	0	0	0
11..20	0	0	0	0
>20	0	0	0	0

$Gini_{0..5}(S_{\text{người = 6..10}}) = \frac{1}{2}$

$Gini_{yêu}(S_{\text{người = 6..10}}) = 0{,}5$

+ Thuộc tính "vừa" – đi = 0..5 và người = 6..10:

Giá trị phân biệt	Số dòng gán nhãn "tốt"	Số dòng gán nhãn "bình thường"	Số dòng gán nhãn "xấu"	Tổng số dòng
0..5	1	0	0	1
6..10	0	0	1	1
11..20	0	0	0	0
>20	0	0	0	0

$Gini_{11..20}(S_{\text{người = 6..10}}) = \frac{1}{2}$

$Gini_{vừa}(S_{\text{người = 6..10}}) = 0{,}5$

Giá trị phân biệt	Số dòng gán nhãn "tốt"	Số dòng gán nhãn "bình thường"	Số dòng gán nhãn "xấu"	Tổng số dòng
0..5	0	0	1	1
6..10	0	0	0	0
11..20	1	0	0	1
>20	0	0	0	0

Vì tất cả thuộc tính có chỉ số Gini thấp bằng nhau, ta chọn ngẫu nhiên thuộc tính "giảm" để chia nhánh tiếp. Với giá trị "giảm" = 0..5, ta luôn có phân lớp xấu và giá trị "giảm" = 11..20 ta luôn có phân tốt, vì vậy nhánh này đi đến nút lá và không cần xét tiếp. Kết thúc chia nhánh.

Trở lại với nhánh đi = ">20" gồm 3 dòng dữ liệu, xét lần lượt:

+ Thuộc tính "giảm" – đi = >20:

Giá trị phân biệt	Số dòng gán nhãn "tốt"	Số dòng gán nhãn "bình thường"	Số dòng gán nhãn "xấu"	Tổng số dòng
0..5	1	1	0	2

6..10	0	0	0	0
11..20	0	0	0	0
>20	0	1	0	1

$$Gini_{0..5}(S_{\text{đi= >20}}) = 1 - \left(\frac{1}{2}\right)^2 - \left(\frac{1}{2}\right)^2 = \frac{1}{2}$$

$$Gini_{giảm}(S_{\text{đi= >20}}) = \frac{2}{3} \times \frac{1}{2} \approx 0{,}33$$

+ Thuộc tính "người" – đi = >20:

Giá trị phân biệt	Số dòng gán nhãn "tốt"	Số dòng gán nhãn "bình thường"	Số dòng gán nhãn "xấu"	Tổng số dòng
0..5	1	2	0	3
6..10	0	0	0	0
11..20	0	0	0	0
>20	0	0	0	0

$$Gini_{0..5}(S_{\text{đi= >20}}) = 1 - \left(\frac{1}{3}\right)^2 - \left(\frac{2}{3}\right)^2 = \frac{4}{9}$$

$$Gini_{người}(S_{\text{đi= >20}}) = \frac{3}{4} \times \frac{4}{9} \approx 0{,}33$$

+ Thuộc tính "chuyển" – đi = >20:

Giá trị phân biệt	Số dòng gán nhãn "tốt"	Số dòng gán nhãn "bình thường"	Số dòng gán nhãn "xấu"	Tổng số dòng
0..5	0	0	0	0
6..10	1	0	0	1
11..20	0	2	0	2
>20	0	0	0	0

$$Gini_{chuyển}(S_{\text{đi= >20}}) = 0$$

+ Thuộc tính "yêu" – đi = >20:

Giá trị phân biệt	Số dòng gán nhãn "tốt"	Số dòng gán nhãn "bình thường"	Số dòng gán nhãn "xấu"	Tổng số dòng
0..5	0	1	0	1
6..10	1	1	0	2
11..20	0	0	0	0
>20	0	0	0	0

$Gini_{yêu}(S_{\text{đi= >20}}) = \frac{2}{3} \times \frac{1}{2} \approx 0{,}33$

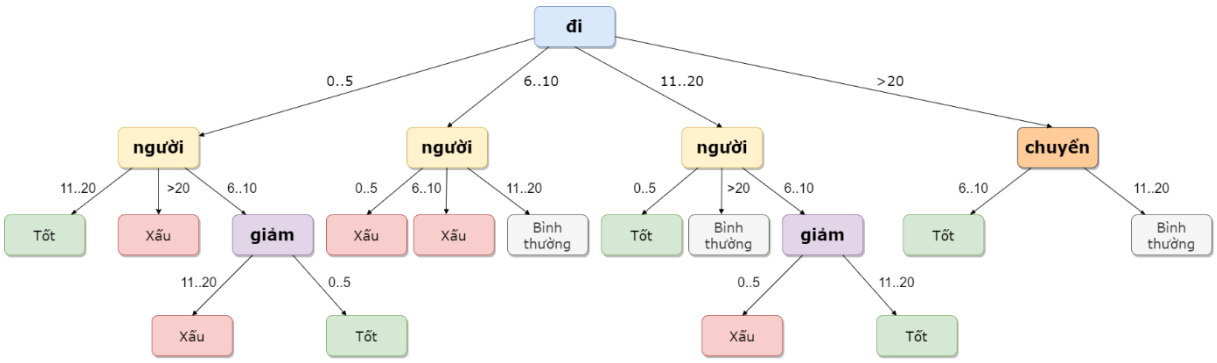
+ Thuộc tính "vừa" – đi = >20:

Giá trị phân biệt	Số dòng gán nhãn "tốt"	Số dòng gán nhãn "bình thường"	Số dòng gán nhãn "xấu"	Tổng số dòng
0..5	0	1	0	1
6..10	1	0	0	1
11..20	0	1	0	1
>20	0	0	0	0

$Gini_{vừa}(S_{\text{đi= >20}}) = 0$

Vì thuộc tính "chuyển" và "vừa" có chỉ số Gini thấp bằng nhau, ta chọn ngẫu nhiên thuộc tính "chuyển" để chia nhánh tiếp. Với giá trị "người" = 6..10, ta luôn có phân lớp tốt, giá trị "người" = 11..20 ta luôn có phân lớp bình thường. Nhánh này đi đến tất cả nút lá và kết thúc chia nhánh.

Thuật toán kết thúc, kết quả cây quyết định như sau:



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

+ Từ cây quyết định xây dựng ở câu b, ta có kết quả dự đoán:

giảm	người	chuyển	yêu	vừa	đi	Cảm xúc
0..5	6..10	0..5	11..20	6..10	0..5	Tốt
0..5	0..5	6..10	6..10	11..20	>20	Tốt
6..10	0..5	11..20	>20	6..10	6..10	Xấu
6..10	11..20	6..10	6..10	>20	0..5	Tốt

+ Dự đoán bằng thuật toán Naïve Bayes:

Xét lần lượt từng dòng (hồ sơ), dựa theo định lý Bayes để tính xác suất xảy ra của Cảm xúc và chọn giá trị xác suất cao nhất.

Với hồ sơ đầu tiên:

X = {giảm = 0..5, người = 6..10, chuyển = 0..5, yêu = 11..20, vừa = 6..10, đi = 0..5}

Áp dụng làm trơn Laplace, ta có:

$$p(Cảm\ xúc = Tốt) = \frac{5+1}{14+3} = \frac{6}{17}$$

$$p(\text{giảm} = 0..5 | Cảm\ xúc = Tốt) = \frac{3+1}{5+4} = \frac{4}{9}$$

$$p(\text{người} = 6..10 | Cảm\ xúc = Tốt) = \frac{2+1}{5+4} = \frac{3}{9}$$

$$p(\text{chuyển} = 0..5 | Cảm\ xúc = Tốt) = \frac{0+1}{5+4} = \frac{1}{9}$$

$$p(\text{yêu} = 11..20 | Cảm\ xúc = Tốt) = \frac{2+1}{5+4} = \frac{3}{9}$$

$$p(\text{vừa } = 6..10|\ Cảm\ xúc = Tốt) = \frac{2+1}{5+4} = \frac{3}{9}$$

$$p(\text{đi } = 0..5|\ Cảm\ xúc = Tốt) = \frac{2+1}{5+4} = \frac{3}{9}$$

$$p(\ X \mid Cảm\ xúc = Tốt) \times p(Cảm\ xúc = Tốt) \approx 2{,}15 \times 10^{-4}$$

$$p(Cảm\ xúc = Xấu) = \frac{5+1}{14+3} = \frac{6}{17}$$

$$p(\text{giảm } = 0..5 \mid Cảm\ xúc = Xấu) = \frac{2+1}{5+4} = \frac{3}{9}$$

$$p(\text{người } = 6..10 \mid Cảm\ xúc = Xấu) = \frac{3+1}{5+4} = \frac{4}{9}$$

$$p(\text{chuyển } = 0..5 \mid Cảm\ xúc = Xấu) = \frac{2+1}{5+4} = \frac{3}{9}$$

$$p(\text{yêu } = 11..20 \mid Cảm\ xúc = Xấu) = \frac{2+1}{5+4} = \frac{3}{9}$$

$$p(\text{vừa } = 6..10|\ Cảm\ xúc = Xấu) = \frac{1+1}{5+4} = \frac{2}{9}$$

$$p(\text{đi } = 0..5|\ Cảm\ xúc = Xấu) = \frac{2+1}{5+4} = \frac{3}{9}$$

$$p(\ X \mid Cảm\ xúc = Xấu) \times p(Cảm\ xúc = Xấu) \approx 4{,}3 \times 10^{-4}$$

$$p(Cảm\ xúc = Bình\ thường) = \frac{4+1}{14+3} = \frac{5}{17}$$

$$p(\text{giảm } = 0..5 \mid Cảm\ xúc = Bình\ thường) = \frac{1+1}{4+4} = \frac{2}{8}$$

$$p(\text{người } = 6..10 \mid Cảm\ xúc = Bình\ thường) = \frac{0+1}{4+4} = \frac{1}{8}$$

$$p(\text{chuyển } = 0..5 \mid Cảm\ xúc = Bình\ thường) = \frac{1+1}{4+4} = \frac{2}{8}$$

$$p(\text{yêu } = 11..20 \mid Cảm\ xúc = Bình\ thường) = \frac{1+1}{4+4} = \frac{2}{8}$$

$$p(\text{vừa } = 6..10|\ Cảm\ xúc = Bình\ thường) = \frac{0+1}{4+4} = \frac{1}{8}$$

$$p(\text{đi } = 0..5 |\, Cảm\ xúc = Bình\ thường) = \frac{0+1}{4+4} = \frac{1}{8}$$

$$p(\, X \,|Cảm\ xúc = Bình\ thường) \times p(Cảm\ xúc = Bình\ thường) \approx 8{,}98 \times 10^{-6}$$

Như vậy, ở hồ sơ đầu tiên có xác suất xảy ra cảm xúc xấu lớn nhất, vậy ta có thể kết luận dòng dữ liệu đầu tiên được dự đoán thuộc phân lớp Cảm xúc = Xấu.

Với hồ sơ thứ hai:

X = {giảm = 0..5, người = 0..5, chuyển = 6..10, yêu = 0..5, vừa = 11..20, đi = >20}

Áp dụng làm trơn Laplace, ta có:

$$p(Cảm\ xúc = Tốt) = \frac{5+1}{14+3} = \frac{6}{17}$$

$$p(\text{giảm } = 0..5 |\, Cảm\ xúc = Tốt) = \frac{3+1}{5+4} = \frac{4}{9}$$

$$p(\text{người } = 0..5 |\, Cảm\ xúc = Tốt) = \frac{2+1}{5+4} = \frac{3}{9}$$

$$p(\text{chuyển } = 6..10 |\, Cảm\ xúc = Tốt) = \frac{2+1}{5+4} = \frac{3}{9}$$

$$p(\text{yêu } = 0..5 |\, Cảm\ xúc = Tốt) = \frac{2+1}{5+4} = \frac{3}{9}$$

$$p(\text{vừa } = 11..20 |\, Cảm\ xúc = Tốt) = \frac{1+1}{5+4} = \frac{2}{9}$$

$$p(\text{đi } => 20 |\, Cảm\ xúc = Tốt) = \frac{1+1}{5+4} = \frac{2}{9}$$

$$p(\, X \mid Cảm\ xúc = Tốt) \times p(Cảm\ xúc = Tốt) \approx 2{,}87 \times 10^{-4}$$

$$p(Cảm\ xúc = Xấu) = \frac{5+1}{14+3} = \frac{6}{17}$$

$$p(\text{giảm } = 0..5 \mid Cảm\ xúc = Xấu) = \frac{2+1}{5+4} = \frac{3}{9}$$

$$p(\text{người } = 0..5 \mid Cảm\ xúc = Xấu) = \frac{1+1}{5+4} = \frac{2}{9}$$

$$p(\text{chuyển } = 6..10 \mid Cảm\ xúc = Xấu) = \frac{1+1}{5+4} = \frac{2}{9}$$

$$p(\text{yêu } = 0..5 \mid Cảm\ xúc = Xấu) = \frac{3+1}{5+4} = \frac{4}{9}$$

$$p(\text{vừa } = 11..20 \mid Cảm\ xúc = Xấu) = \frac{1+1}{5+4} = \frac{2}{9}$$

$$p(\text{đi } => 20 \mid Cảm\ xúc = Xấu) = \frac{0+1}{5+4} = \frac{1}{9}$$

$$p(X \mid Cảm\ xúc = Xấu) \times p(Cảm\ xúc = Xấu) \approx 6{,}37 \times 10^{-5}$$

$$p(Cảm\ xúc = Bình\ thường) = \frac{4+1}{14+3} = \frac{5}{17}$$

$$p(\text{giảm } = 0..5 \mid Cảm\ xúc = Bình\ thường) = \frac{1+1}{4+4} = \frac{2}{8}$$

$$p(\text{người } = 0..5 \mid Cảm\ xúc = Bình\ thường) = \frac{2+1}{4+4} = \frac{3}{8}$$

$$p(\text{chuyển } = 6..10 \mid Cảm\ xúc = Bình\ thường) = \frac{1+1}{4+4} = \frac{2}{8}$$

$$p(\text{yêu } = 0..5 \mid Cảm\ xúc = Bình\ thường) = \frac{1+1}{4+4} = \frac{2}{8}$$

$$p(\text{vừa } = 11..20 \mid Cảm\ xúc = Bình\ thường) = \frac{1+1}{4+4} = \frac{2}{8}$$

$$p(\text{đi } => 20 \mid Cảm\ xúc = Bình\ thường) = \frac{2+1}{4+4} = \frac{3}{8}$$

$$p(X \mid Cảm\ xúc = Bình\ thường) \times p(Cảm\ xúc = Bình\ thường) \approx 1{,}61 \times 10^{-4}$$

Như vậy, ở hồ sơ thứ hai có xác suất xảy ra cảm xúc tốt lớn nhất, vậy ta có thể kết luận dòng dữ liệu thứ hai được dự đoán thuộc phân lớp Cảm xúc = Tốt.

Với hồ sơ thứ ba:

X = {giảm = 6..10, người = 0..5, chuyển = 11..20, yêu = >20, vừa = 6..10, đi = 6..10}

Áp dụng làm trơn Laplace, ta có:

$$p(Cảm\ xúc = Tốt) = \frac{5+1}{14+3} = \frac{6}{17}$$

$$p(\text{giảm } = 6..10 \mid Cảm\ xúc = Tốt) = \frac{0+1}{5+4} = \frac{1}{9}$$

$p(\text{người } = 0..5|\ Cảm\ xúc = Tốt) = \frac{2+1}{5+4} = \frac{3}{9}$

$p(\text{chuyển } = 11..20|\ Cảm\ xúc = Tốt) = \frac{2+1}{5+4} = \frac{3}{9}$

$p(\text{yêu } => 20|\ Cảm\ xúc = Tốt) = \frac{0+1}{5+4} = \frac{1}{9}$

$p(\text{vừa } = 6..10|\ Cảm\ xúc = Tốt) = \frac{2+1}{5+4} = \frac{3}{9}$

$p(\text{đi } = 6..10|\ Cảm\ xúc = Tốt) = \frac{0+1}{5+4} = \frac{1}{9}$

$p(\ X \mid Cảm\ xúc = Tốt) \times p(Cảm\ xúc = Tốt) \approx 1{,}79 \times 10^{-5}$

$p(Cảm\ xúc = Xấu) = \frac{5+1}{14+3} = \frac{6}{17}$

$p(\text{giảm } = 6..10 \mid Cảm\ xúc = Xấu) = \frac{1+1}{5+4} = \frac{2}{9}$

$p(\text{người } = 0..5 \mid Cảm\ xúc = Xấu) = \frac{1+1}{5+4} = \frac{2}{9}$

$p(\text{chuyển } = 11..20 \mid Cảm\ xúc = Xấu) = \frac{1+1}{5+4} = \frac{2}{9}$

$p(\text{yêu } => 20 \mid Cảm\ xúc = Xấu) = \frac{0+1}{5+4} = \frac{1}{9}$

$p(\text{vừa } = 6..10 \mid Cảm\ xúc = Xấu) = \frac{1+1}{5+4} = \frac{2}{9}$

$p(\text{đi } = 6..10 \mid Cảm\ xúc = Xấu) = \frac{2+1}{5+4} = \frac{1}{9}$

$p(\ X \mid Cảm\ xúc = Xấu) \times p(Cảm\ xúc = Xấu) \approx 1{,}06 \times 10^{-5}$

$p(Cảm\ xúc = Bình\ thường) = \frac{4+1}{14+3} = \frac{5}{17}$

$p(\text{giảm } = 6..10 \mid Cảm\ xúc = Bình\ thường) = \frac{1+1}{4+4} = \frac{2}{8}$

$p(\text{người } = 0..5 \mid Cảm\ xúc = Bình\ thường) = \frac{2+1}{4+4} = \frac{3}{8}$

$$p(\text{chuyển } = 11..20 \mid Cảm\ xúc = Bình\ thường) = \frac{2+1}{4+4} = \frac{3}{8}$$

$$p(\text{yêu } => 20 \mid Cảm\ xúc = Bình\ thường) = \frac{1+1}{4+4} = \frac{2}{8}$$

$$p(\text{vừa } = 6..10 \mid Cảm\ xúc = Bình\ thường) = \frac{0+1}{4+4} = \frac{1}{8}$$

$$p(\text{đi } = 6..10 \mid Cảm\ xúc = Bình\ thường) = \frac{1+1}{4+4} = \frac{2}{8}$$

$$p(X \mid Cảm\ xúc = Bình\ thường) \times p(Cảm\ xúc = Bình\ thường) \approx 8{,}07 \times 10^{-5}$$

Như vậy, ở hồ sơ thứ ba có xác suất xảy ra cảm xúc bình thường lớn nhất, vậy ta có thể kết luận dòng dữ liệu thứ ba được dự đoán thuộc phân lớp Cảm xúc = Bình thường.

Với hồ sơ cuối cùng:

X = {giảm = 6..10, người = 11..20, chuyển = 6..10, yêu = 6..10, vừa = >20, đi = 0..5}

Áp dụng làm trơn Laplace, ta có:

$$p(Cảm\ xúc = Tốt) = \frac{5+1}{14+3} = \frac{6}{17}$$

$$p(\text{giảm } = 6..10 \mid Cảm\ xúc = Tốt) = \frac{0+1}{5+4} = \frac{1}{9}$$

$$p(\text{người } = 11..20 \mid Cảm\ xúc = Tốt) = \frac{1+1}{5+4} = \frac{2}{9}$$

$$p(\text{chuyển } = 6..10 \mid Cảm\ xúc = Tốt) = \frac{2+1}{5+4} = \frac{3}{9}$$

$$p(\text{yêu } = 6..10 \mid Cảm\ xúc = Tốt) = \frac{1+1}{5+4} = \frac{2}{9}$$

$$p(\text{vừa } => 20 \mid Cảm\ xúc = Tốt) = \frac{1+1}{5+4} = \frac{2}{9}$$

$$p(\text{đi } = 0..5 \mid Cảm\ xúc = Tốt) = \frac{2+1}{5+4} = \frac{3}{9}$$

$$p(X \mid Cảm\ xúc = Tốt) \times p(Cảm\ xúc = Tốt) \approx 4{,}78 \times 10^{-5}$$

$$p(Cảm\ xúc = Xấu) = \frac{5+1}{14+3} = \frac{6}{17}$$

$$p(\text{giảm } = 6..10 \mid Cảm\ xúc = Xấu) = \frac{1+1}{5+4} = \frac{2}{9}$$

$$p(\text{người } = 11..20 \mid Cảm\ xúc = Xấu) = \frac{0+1}{5+4} = \frac{1}{9}$$

$$p(\text{chuyển } = 6..10 \mid Cảm\ xúc = Xấu) = \frac{1+1}{5+4} = \frac{2}{9}$$

$$p(\text{yêu } = 6..10 \mid Cảm\ xúc = Xấu) = \frac{0+1}{5+4} = \frac{1}{9}$$

$$p(\text{vừa } => 20 \mid Cảm\ xúc = Xấu) = \frac{0+1}{5+4} = \frac{1}{9}$$

$$p(\text{đi } = 0..5 \mid Cảm\ xúc = Xấu) = \frac{2+1}{5+4} = \frac{1}{9}$$

$$p(X \mid Cảm\ xúc = Xấu) \times p(Cảm\ xúc = Xấu) \approx 2{,}66 \times 10^{-6}$$

$$p(Cảm\ xúc = Bình\ thường) = \frac{4+1}{14+3} = \frac{5}{17}$$

$$p(\text{giảm } = 6..10 \mid Cảm\ xúc = Bình\ thường) = \frac{1+1}{4+4} = \frac{2}{8}$$

$$p(\text{người } = 11..20 \mid Cảm\ xúc = Bình\ thường) = \frac{1+1}{4+4} = \frac{2}{8}$$

$$p(\text{chuyển } = 6..10 \mid Cảm\ xúc = Bình\ thường) = \frac{1+1}{4+4} = \frac{2}{8}$$

$$p(\text{yêu } = 6..10 \mid Cảm\ xúc = Bình\ thường) = \frac{1+1}{4+4} = \frac{2}{8}$$

$$p(\text{vừa } => 20 \mid Cảm\ xúc = Bình\ thường) = \frac{0+1}{4+4} = \frac{1}{8}$$

$$p(\text{đi } = 0..5 \mid Cảm\ xúc = Bình\ thường) = \frac{0+1}{4+4} = \frac{1}{8}$$

$$p(X \mid Cảm\ xúc = Bình\ thường) \times p(Cảm\ xúc = Bình\ thường) \approx 7{,}18 \times 10^{-5}$$

Như vậy, ở hồ sơ thứ ba có xác suất xảy ra cảm xúc bình thường lớn nhất, vậy ta có thể kết luận dòng dữ liệu thứ ba được dự đoán thuộc phân lớp Cảm xúc = Bình thường.

Với thuật toán Naïve Bayes, ta có kết quả dự đoán tổng hợp qua bảng sau:

giảm	người	chuyển	yêu	vừa	đi	Cảm xúc
0..5	6..10	0..5	11..20	6..10	0..5	Xấu

0..5	0..5	6..10	6..10	11..20	>20	Tốt
6..10	0..5	11..20	>20	6..10	6..10	Bình thường
6..10	11..20	6..10	6..10	>20	0..5	Bình thường

d) Trên thực tế những trạng thái này lần lượt có cảm xúc là: xấu, tốt, bình thường, tốt. Hãy lập ma trận nhầm lẫn, sau đó tính giá trị độ chính xác, độ phủ của cả hai phương pháp trên rồi so sánh chúng với nhau. Sinh viên có kết luận gì về kết quả này?

Kết quả dự đoán của cây quyết định như sau:

giảm	người	chuyển	yêu	vừa	đi	Cảm xúc
0..5	6..10	0..5	11..20	6..10	0..5	Tốt
0..5	0..5	6..10	6..10	11..20	>20	Tốt
6..10	0..5	11..20	>20	6..10	6..10	Xấu
6..10	11..20	6..10	6..10	>20	0..5	Tốt

Ma trận nhầm lẫn của cây quyết định như sau:

		Lớp dự đoán được từ mô hình		
Lớp trên thực tế		Xấu	Bình thường	Tốt
	Xấu	0	0	1
	Bình thường	1	0	0
	Tốt	0	0	2

$$precision(Xấu) = \frac{0}{0+1} = 0\%$$

$$precision(Bình\ thường) = \frac{0}{0+0} = 0\%$$

$$precision(Tốt) = \frac{2}{2+1} = 67\%$$

$$macro\_precision = \frac{\frac{2}{3}}{3} \approx 22\%$$

$$recall(Xấu) = \frac{0}{0+1} = 0\%$$

$$recall(Bình\ thường) = \frac{0}{0+1} = 0\%$$

$$recall(Tốt) = \frac{2}{2+0} = 100\%$$

$$macro\_recall(Xấu) = \frac{1}{3} \approx 33\%$$

Kết quả dự đoán của thuật toán Naïve Bayes qua bảng sau:

giảm	người	chuyển	yêu	vừa	đi	Cảm xúc
0..5	6..10	0..5	11..20	6..10	0..5	Xấu
0..5	0..5	6..10	6..10	11..20	>20	Tốt
6..10	0..5	11..20	>20	6..10	6..10	Bình thường
6..10	11..20	6..10	6..10	>20	0..5	Bình thường

Ma trận nhầm lẫn của thuật toán Naïve Bayes như sau:

		Lớp dự đoán được từ mô hình		
Lớp trên thực tế		Xấu	Bình thường	Tốt
	Xấu	1	0	0
	Bình thường	0	1	0
	Tốt	0	1	1

$$precision(Xấu) = \frac{1}{1+0} = 100\%$$

$$precision(Bình\ thường) = \frac{1}{1+1} = 50\%$$

$$precision(Tốt) = \frac{1}{1+0} = 100\%$$

$$macro\_precision = \frac{1+\frac{1}{2}+1}{3} \approx 83\%$$

$$recall(Xấu) = \frac{1}{1+0} = 100\%$$

$$recall(Bình\ thường) = \frac{1}{1+0} = 100\%$$

$$recall(Tốt) = \frac{1}{1+1} = 50\%$$

$$macro\_recall(Xấu) = \frac{1+\frac{1}{2}+1}{3} \approx 83\%$$

Kết luận:

Từ hai kết quả tính toán giá trị macro của precision và recall của hai thuật toán trên, ta nhận thấy thuật toán Naïve Bayes cho kết quả tốt hơn so với cây quyết định. Naïve Bayes có độ chính xác và độ phủ trung bình đều trên 80%, còn cây quyết định lại cho kết quả rất thấp dưới 40%.

e) Nếu nắm bắt được cảm xúc của người dùng mạng xã hội thì sinh viên sẽ sử dụng chúng như thế nào?

Nếu sinh viên có khả năng nắm bắt được cảm xúc của người dùng mạng xã hội, họ có thể sử dụng thông tin này để thực hiện nhiều mục đích có ích.

- Thứ nhất, việc hiểu rõ cảm xúc của người dùng mạng xã hội sẽ giúp sinh viên phát hiện sớm và can thiệp vào những tình huống có khả năng gây ra căng thẳng hoặc xung đột, từ đó giảm thiểu các hiện tượng bạo lực mạng và hỗ trợ những người gặp vấn đề về tâm lý.
- Thứ hai, nắm bắt được cảm xúc của cộng đồng mạng sẽ giúp sinh viên nhận diện được các xu hướng và mối quan tâm phổ biến, từ đó xây dựng và triển khai các chiến lược truyền thông và marketing hiệu quả, phù hợp với nhu cầu và mong muốn của người tiêu dùng.

Câu 4: (Lập trình)

Cho dữ liệu Red Wine Quality, liên quan đến các mẫu rượu vang Vinho Verde đỏ từ phía bắc Bồ Đào Nha. Mục tiêu của bài toán là mô hình hóa chất lượng rượu dựa trên các chỉ số hóa lý đo đạc được. Sử dụng câu lệnh sau để chia dữ liệu đầu vào thành hai phần huấn luyện 70% và kiểm thử 30%.

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=42)
```

Thuộc tính quyết định là “quanlity”.

(Bài giải ở file 21522005_Bai4_Lab4.ipynb)