Tăng Việt Diện – 20521185

LAB04

1. (Cơ bản) 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 hãy giúp doanh nghiệp bằng cách thực hiện những yêu cầu sau:

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

Không có mâu thuẫn

b) 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.

Ban đầu tập S bao gồm toàn bộ 15 dòng dữ liệu đã cho, trong đó có 7 dòng Doanh số bán cao, 8 dòng Doanh số bán thấp. Vậy độ bất định của tập S lúc này là:

$$E(S) = -\frac{7}{15}\log_2\frac{7}{15} - \frac{8}{15}\log_2\frac{8}{15} \approx 0.99$$

Xét thuộc tính Loại ta có:

$$E(S_{\text{Điều khiển}}) = -\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5} \approx 0.97$$

$$E(S_{\text{Xếp hình}}) = -\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5} \approx 0.97$$

$$E(S_{\text{Búp bê}}) = -\frac{3}{5}\log_2\frac{3}{5} - \frac{2}{5}\log_2\frac{2}{5} \approx 0.97$$

$$G(S, Loại) \approx 0.99 - \frac{5}{15} \times 0.97 - \frac{5}{15} \times 0.97 - \frac{5}{15} \times 0.97 \approx 0.02$$

Xét thuộc tính Số màu ta có:

$$E(S_3) = -\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5} \approx 0.97$$

$$E(S_5) = -\frac{3}{6}\log_2\frac{3}{6} - \frac{3}{6}\log_2\frac{3}{6} = 1$$

$$E(S_7) = -\frac{2}{4}\log_2\frac{2}{4} - \frac{2}{4}\log_2\frac{2}{4} = 1$$

$$G(S, Số\ màu) \approx 0.99 - \frac{5}{15} \times 0.97 - \frac{6}{15} \times 1 - \frac{4}{15} \times 1 \approx 0$$

Xét thuộc tính Kích thước ta có:

$$E(S_{nhỏ}) = -\frac{2}{4}\log_2\frac{2}{4} - \frac{2}{4}\log_2\frac{2}{4} = 1$$

$$E(S_{vừa}) = -\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5} = 0.97$$

$$E(S_{to}) = -\frac{3}{6}\log_2\frac{3}{6} - \frac{3}{6}\log_2\frac{3}{6} = 1$$

$$G(S, Kích\ thước) \approx 0.99 - \frac{4}{15} \times 1 - \frac{5}{15} \times 0.97 - \frac{6}{15} \times 1 \approx 0$$

Xét thuộc tính Chất liệu ta có:

$$E(S_{Nhựa\ PP}) = -\frac{6}{9}\log_2\frac{6}{9} - \frac{3}{9}\log_2\frac{3}{9} \approx 0.92$$

$$E(S_{Cao\ su}) = -\frac{1}{6}\log_2\frac{1}{6} - \frac{5}{6}\log_2\frac{5}{6} \approx 0.65$$

$$G(S, Chất\ liệu) \approx 0.99 - \frac{9}{15} \times 0.92 - \frac{6}{15} \times 0.65 \approx 0.178$$

Trong 4 thuộc tính đã xem xét, Chất liệu có độ lợi thông tin lớn nhất. Do đó, ta chọn thuộc tính này làm phép chia nhánh cho cây tại nút gốc.

Tập dữ liệu lúc này được chia làm hai phần tương ứng với hai nhánh cây theo giá trị của thuộc tính Chất liệu. Phần có giá trị Nhựa PP gồm 9 dòng, phần có giá trị Cao su gồm 6 dòng.

Với nhánh cây thứ nhất, nhánh Nhựa PP, xét thuộc tính Loại, tính toán tương tự ta có:

$E(S_{Điều\ khiển}) = 0$; $E(S_{Xếp\ hình}) \approx 0.918$; $E(S_{Búp\ bê}) \approx 0.811$; $G(S_{Nhựa\ PP}, Loại) \approx 0.254$

Xét thuộc tính Số màu:

$E(S_3) = 1$; $E(S_5) = 0$; $E(S_7) = 1$; $G(S_{Nhựa\ PP}, Số\ màu) \approx 0.253$

Xét thuộc tính Kích thước:

$E(S_{Nhỏ}) = 0$; $E(S_{Vừa}) \approx 0.918$; $E(S_{To}) = 1$; $G(S_{Nhựa\ PP}, Kích\ thước) \approx 0.17$

Vậy ở nhánh này ta chọn Thuộc tính Loại làm phép chia nhánh. Với Loại Điều khiển, ta luôn có phân lớp Doanh số bán cao, vì vậy nhánh này đi đến nút lá và không cần xét tiếp. Hai nhánh con tương ứng với hai Loại còn lại là Xếp hình và Búp bê sẽ tiếp tục được phát triển.

Nhánh tương ứng với giá trị Xếp hình bao gồm 3 dòng dữ liệu, xét các thuộc tính còn lại ta có:

$$E(S_3) = 1; E(S_5) = 0; E(S_7) = 0; G(S_{Xếp\ hình}, Số\ màu) \approx 0.251$$

$$E(S_{Nhỏ}) = 0; E(S_{Vừa}) = 0; E(S_{To}) = 0; G(S_{Xếp\ hình}, Kích\ thước) \approx 0.918$$

Với giá trị độ lợi thông tin lớn hơn, cây sẽ tiếp tục được phân nhánh bằng thuộc tính Kích thước. Đến đây, nhánh To sẽ đi đến nút lá Doanh số bán thấp, nhánh Nhỏ sẽ đi đến nút lá Doanh số bán cao.

Nhánh tương ứng với giá trị Búp bê gồm 4 dòng dữ liệu, tiếp tục tính độ lợi thông tin ta có:

$$E(S_3) = 0; E(S_5) = 0; E(S_7) = 0; G(S_{Búp\ bê}, Số\ màu) \approx 0.811$$

$$E(S_{Nhỏ}) = 0; E(S_{Vừa}) = 1; E(S_{To}) = 0; G(S_{Búp\ bê}, Kích\ thước) \approx 0.3.11$$

Nhánh này sẽ được phân chia bằng thuộc tính Số màu, nhánh con Số màu 3 sẽ đi đến nút lá Doanh số bán thấp, nhánh con Số màu 5 và 7 sẽ đi đến nút là Doanh số bán cao.

Trở lại với nhánh Cao su, được phân chia từ nút gốc, ta có 6 dòng dữ liệu. Xét các thuộc tính Loại, Số màu và Kích thước ta được kết quả sau:

$$E(S_{Điều\ khiển}) = 0; E(S_{Xếp\ hình}) = 1; E(S_{Búp\ bê}) \approx 0; G(S_{Cao\ su}, Loại) \approx 0.32$$

$$E(S_3) = 0; E(S_5) = 0; E(S_7) = 1; G(S_{Cao\ su}, Số\ màu) \approx 0.32$$

$$E(S_{Nhỏ}) = 0; E(S_{Vừa}) = 0; E(S_{To}) = 1; G(S_{Cao\ su}, Kích\ thước) \approx 0.32$$

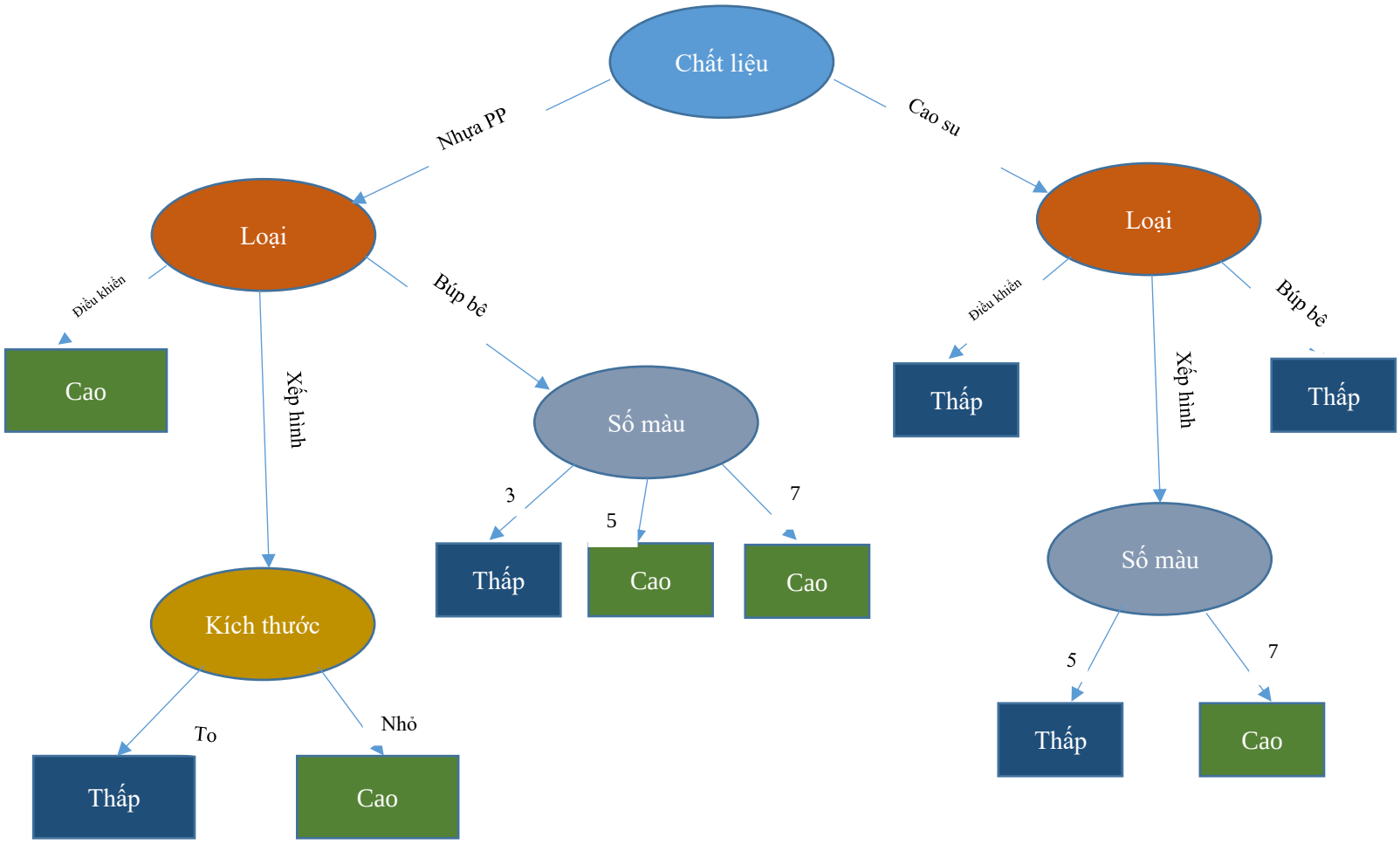
Ở nhánh này cả ba thuộc tính đều có độ lợi thông tin cao bằng nhau. Ta chọn Loại để tiếp tục, với loại Điều khiển và Búp bê cây sẽ phát triển đến nút Doanh số bán thấp. Chúng ta chỉ việc xem xét phát triển cây với Loại Xếp hình, ta có:

$$E(S_3) = 0; E(S_5) = 0; E(S_7) = 0; G(S_{Xếp\ hình}, Số\ màu) \approx 1$$

$$E(S_{Nhỏ}) = 0; E(S_{Vừa}) = 0; E(S_{To}) = 0; G(S_{Xếp\ hình}, Kích\ thước) \approx 1$$

Ở nhánh này cả hai thuộc tính đều có độ lợi thông tin cao bằng nhau. Ta chọn Số màu để tiếp tục, với Số màu 5 cây sẽ phát triển đến nút Doanh số bán thấp và Số màu 7 cây sẽ phát triển đến Doanh số bán cao.

Như vậy thuật toán kết thúc. Kết quả cây quyết định được trình bày như sau:



c) 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.

Theo những thống kê từ câu 2, ta tính chỉ số Gini của lần lượt từng thuộc tính để tìm ra thuộc tính phân nhánh có lợi nhất. Xét thuộc tính Thu nhập, ta có:

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

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

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

$$Gini_{Loại}(S) = \frac{5}{15} \times 0{,}48 + \frac{5}{15} \times 0{,}48 + \frac{5}{15} \times 0{,}48 = 0{,}48$$

Tiếp tục xét thuộc tính Số màu, ta có:

$$Gini(S_3) = 1 - \left(\frac{2}{5}\right)^2 - \left(\frac{3}{5}\right)^2 = 0{,}48$$

$$Gini(S_5) = 1 - \left(\frac{3}{6}\right)^2 - \left(\frac{3}{6}\right)^2 = 0{,}5$$

$$Gini(S_7) = 1 - \left(\frac{2}{4}\right)^2 - \left(\frac{2}{4}\right)^2 = 0{,}5$$

$$Gini_{Số\ màu}(S) = \frac{5}{15} \times 0{,}48 + \frac{6}{15} \times 0{,}5 + \frac{4}{15} \times 0{,}5 \approx 0{,}49$$

Với thuộc tính Kích thước, ta có:

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

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

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

$$Gini_{Số\ màu}(S) = \frac{4}{15} \times 0{,}5 + \frac{5}{15} \times 0{,}48 + \frac{6}{15} \times 0{,}5 \approx 0{,}49$$

Với thuộc tính Kích thước, ta có:

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

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

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

$$Gini_{Số\ màu}(S) = \frac{4}{15} \times 0{,}5 + \frac{5}{15} \times 0{,}48 + \frac{6}{15} \times 0{,}5 \approx 0{,}49$$

Với thuộc tính Chất liệu, ta có:

$$Gini(S_{Nhựa\ PP}) = 1 - \left(\frac{6}{9}\right)^2 - \left(\frac{3}{9}\right)^2 \approx 0{,}44$$

$$Gini(S_{Cao\ su}) = 1 - \left(\frac{1}{6}\right)^2 - \left(\frac{5}{6}\right)^2 \approx 0{,}28$$

$$Gini_{Số\ màu}(S) = \frac{9}{15} \times 0{,}44 + \frac{6}{15} \times 0{,}28 \approx 0{,}376$$

Chọn thuộc tính có chỉ số Gini thấp nhất là Chất liệu và phân nhánh.

Xét nhánh Nhựa PP, tính giá trị chỉ số Gini tương tự như trên:

$$Gini_{Loại}(S_{Nhựa\ PP}) = \frac{2}{9} \times 0 + \frac{3}{9} \times 0{,}44 + \frac{4}{9} \times 0{,}375 \approx 0{,}313$$

$$Gini_{Số\ màu}(S_{Nhựa\ PP}) = \frac{4}{9} \times 0{,}5 + \frac{3}{9} \times 0 + \frac{2}{9} \times 0{,}5 \approx 0{,}333$$

$$Gini_{Kích\ thước}(S_{Nhựa\ PP}) = \frac{2}{9} \times 0 + \frac{3}{9} \times 0{,}44 + \frac{4}{9} \times 0{,}375 \approx 0{,}368$$

Chọn thuộc tính Loại và tiếp tục phát triển cây, ta tính các chỉ số Gini:

$$Gini_{Số\ màu}(S_{Điều\ khiển}) = \frac{1}{2} \times 0 + \frac{1}{2} \times 0 = 0$$

$$Gini_{Kích\ thước}(S_{Điều\ khiển}) = \frac{1}{2} \times 0 + \frac{1}{2} \times 0 = 0$$

Đối với nhánh Điều khiển các thuộc tính đều có chỉ số Gini bằng 0 nên đi đến nút lá Cao. Với nhánh Xếp hình ta có:

$$Gini_{Số\ màu}(S_{Xếp\ hình}) = \frac{2}{3} \times 0{,}5 + \frac{1}{3} \times 0 = 0{,}333$$

$$Gini_{Kích\ thước}(S_{Xếp\ hình}) = \frac{1}{3} \times 0 + \frac{2}{3} \times 0 = 0$$

Ở nhánh này ta chọn thuộc tính Kích thước làm thuộc tính phân nhánh và cây cũng đi đến các nút lá. Xét tiếp tục nhánh Búp bê ta có:

$$Gini_{Số\ màu}(S_{Búp\ bê}) = \frac{1}{4} \times 0 + \frac{2}{4} \times 0 + \frac{1}{4} \times 0 = 0$$

$$Gini_{Kích\ thước}(S_{Búp\ bê}) = \frac{2}{4} \times 0{,}5 + \frac{2}{4} \times 0 = 0{,}25$$

Ở nhánh này ta chọn thuộc tính Số màu làm thuộc tính phân nhánh và cây cũng đi đến các nút lá. Ta quay lại xét nhánh Cao su:

$$Gini_{Loại}(S_{Cao\ su}) = \frac{3}{6} \times 0 + \frac{2}{6} \times 0{,}5 + \frac{1}{6} \times 0 \approx 0{,}167$$

$$Gini_{Số\ màu}(S_{Cao\ su}) = \frac{1}{6} \times 0 + \frac{3}{6} \times 0{,}44 + \frac{2}{6} \times 0{,}5 \approx 0{,}388$$

$$Gini_{Kích\ thước}(S_{Cao\ su}) = \frac{2}{6} \times 0 + \frac{2}{6} \times 0 + \frac{2}{6} \times 0{,}5 \approx 0{,}167$$

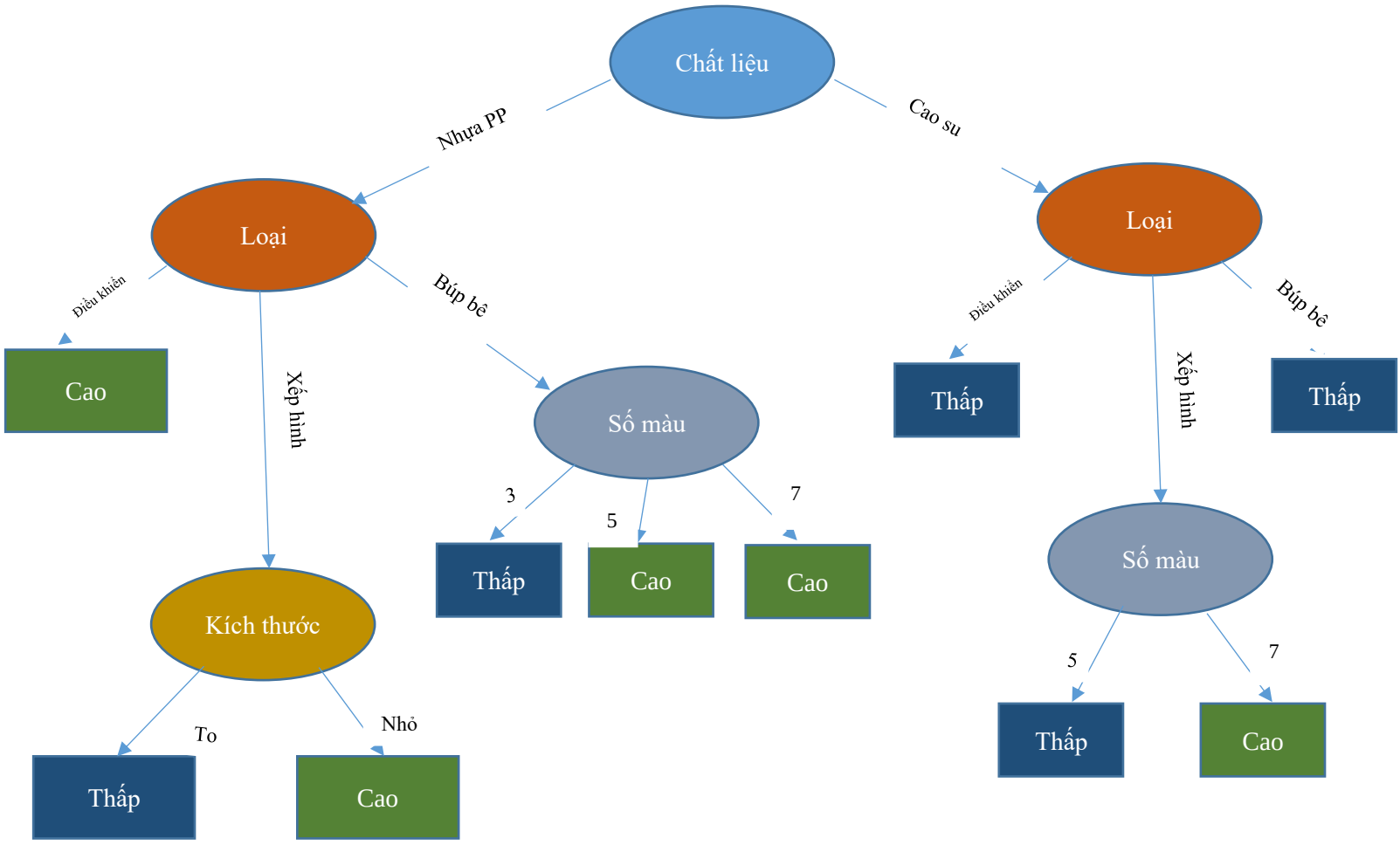
Ở nhánh này thuộc tính Loại và Kích thước có chỉ số Gini nhỏ bằng nhau, ta lựa chọn Loại để chia nhánh. Nhánh với giá trị Điều khiển và Búp bê nối đến nút lá Thấp, Xét nhánh Xếp hình:

$$Gini_{Số\ màu}(S_{Xếp\ hình}) = \frac{1}{2} \times 0 + \frac{1}{2} \times 0 = 0$$

$$Gini_{Kích\ thước}(S_{Xếp\ hình}) = \frac{1}{2} \times 0 + \frac{1}{2} \times 0 = 0$$

Ở nhánh này thuộc tính Số màu và Kích thước có chỉ số Gini nhỏ bằng nhau, ta chọn Số màu để chia nhánh và các nhánh đều đi đến nút lá.

Kết thúc thuật toán. Ta được cây như sau:



d) Dựa theo cây CART xây dựng được ở câu trên, kết quả dự đoán như sau:

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

e) Doanh số bán trên thực tế của các sản phẩm ở Yêu cầu d lần lượt là *Thấp, Thấp, Cao*. Hãy lập ma trận nhầm lẫn, sau đó tính giá trị độ chính xác, độ phủ của mô hình/cây đã xây dựng.

Chọn Doanh số bán thấp là lớp dương, lớp còn lại – Doanh số bán cao sẽ là lớp âm, ta có ma trận nhầm lẫn của cây quyết định như sau:

		Lớp dự đoán được từ mô hình	
Lớp trên thực tế		*Doanh số bán thấp*	*Doanh số bán cao*
	Doanh số bán thấp	2	0
	Doanh số bán cao	1	0

Theo công thức tính độ chính xác và độ phủ ta có:

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

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

f) Xác suất không điều kiện của giá trị 'Xếp hình' trong tập dữ liệu là bao nhiêu?

$$p(Loại = Xếp\ hình) = \frac{5}{15} = 0{,}333$$

g) Khi doanh số bán là 'Thấp', hãy tính xác suất đó là những sản phẩm có chất liệu là 'Cao su'.

$$p(Chất\ liệu = Cao\ su \mid Doanh\ số\ bán = Thấp) = \frac{5}{8} = 0{,}625$$

h) Dựa theo định lý Bayes, hãy viết công thức tính xác suất Doanh số 'Cao' của những sản phẩm thuộc loại 'Điều khiển'.

$$p(Doanh\ số = Cao \mid Loại = Điều\ khiển)$$

$$= \frac{p(Loại{=}Điều\ khiển \mid Doanh\ số{=}Cao) \times p(Doanh\ số{=}Cao)}{p(Loại{=}Điều\ khiển)}$$

$$= \frac{\frac{2}{7} \times \frac{7}{15}}{\frac{5}{15}} = 0{,}4$$

i) 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 d.

Với hồ sơ đầu tiên

X = {Loại = Búp bê, Số màu = 3, Kích thước = To, Chất liệu = Cao su}

Ta cần tính được

$$p(Doanh\ số\ bán = Cao|X)$$

$$= \frac{p\left(X|Doanh\ số\ bán = Cao\right) \times p(Doanh\ số\ bán = Cao)}{p(X)}$$

so sánh với

$$p(Doanh\ số\ bán = Thấp|X)$$

$$= \frac{p\left(X|Doanh\ số\ bán = Thấp\right) \times p(Doanh\ số\ bán = Thấp)}{p(X)}$$

Vì mẫu số bằng nhau nên chỉ cần tính toán và so sánh hai tử số chúng ta sẽ có kết quả.

Áp dụng làm trơn Laplace, ta có:

$p(Doanh\ số\ bán\ =\ Cao) = \frac{7+1}{15+2} = \frac{8}{17} \approx 0{,}47$

$p(Loại = Búp\ bê|Doanh\ số\ bán = Cao) = \frac{3+1}{7+3} = 0{,}4$

$p(Số\ màu = 3|Doanh\ số\ bán = Cao) = \frac{2+1}{7+3} = 0{,}3$

$p(Kích\ thước = To|Doanh\ số\ bán = Cao) = \frac{3+1}{7+3} = 0{,}4$

$p(Chất\ liệu = Cao\ su|Doanh\ số\ bán = Cao) = \frac{1+1}{7+2} = \frac{2}{9} \approx 0{,}222$

$p(X|Doanh\ số\ bán = Cao)\ \times\ p(Doanh\ số\ bán\ =\ Cao)$

$= \frac{8}{17} \times 0{,}4 \times 0{,}3 \times 0{,}4 \times \frac{2}{9} \approx 0{,}005$

$p(Doanh\ số\ bán\ =\ Thấp) = \frac{8+1}{15+2} = \frac{9}{17} \approx 0{,}529$

$p(Loại = Búp\ bê | Doanh\ số\ bán = Thấp) = \frac{2+1}{8+3} = \frac{3}{11} \approx 0{,}273$

$p(Số\ màu = 3 | Doanh\ số\ bán = Thấp) = \frac{3+1}{8+3} = \frac{4}{11} \approx 0{,}364$

$p(Kích\ thước = To | Doanh\ số\ bán = Thấp) = \frac{3+1}{8+3} = \frac{4}{11} \approx 0{,}364$

$p(Chất\ liệu = Cao\ su | Doanh\ số\ bán = Thấp) = \frac{5+1}{8+2} = 0{,}6$

$p(X | Doanh\ số\ bán = Thấp) \times p(Doanh\ số\ bán\ =\ Thấp)$

$= \frac{3}{11} \times \frac{4}{11} \times \frac{4}{11} \times 0{,}6 \times \frac{9}{17} \approx 0{,}011$

Như vậy hồ sơ đầu tiên có xác suất xảy ra Doanh số bán thấp lớn hơn, vậy ta có thể kết luận dòng dữ liệu đầu tiên được dự đoán thuộc phân lớp Doanh số bán thấp.

Xét hồ sơ thứ hai

X = {Loại = Xếp hình, Số màu = 5, Kích thước = To, Chất liệu = Nhựa PP}

Ta cũng có những tính toán sau:

$p(Loại = Xếp\ hình | Doanh\ số\ bán = Cao) = \frac{2+1}{7+3} = 0{,}3$

$p(Số\ màu = 5 | Doanh\ số\ bán = Cao) = \frac{3+1}{7+3} = 0{,}4$

$p(Kích\ thước = To | Doanh\ số\ bán = Cao) = \frac{3+1}{7+3} = 0{,}4$

$p(Chất\ liệu = Nhựa\ PP | Doanh\ số\ bán = Cao) = \frac{6+1}{7+2} = \frac{7}{9} \approx 0{,}778$

$p(X | Doanh\ số\ bán = Cao) \times p(Doanh\ số\ bán\ =\ Cao)$

$= 0{,}3 \times 0{,}4 \times 0{,}4 \times \frac{7}{9} \times \frac{8}{17} \approx 0{,}018$

$p(Loại = Xếp\ hình | Doanh\ số\ bán = Thấp) = \frac{3+1}{8+3} = \frac{4}{11} \approx 0{,}364$

$p(Số\ màu = 5 | Doanh\ số\ bán = Thấp) = \frac{3+1}{8+3} = \frac{4}{11} \approx 0{,}364$

$$p(Kích\ thước = To|Doanh\ số\ bán = Thấp) = \frac{3+1}{8+3} = \frac{4}{11} \approx 0{,}364$$

$$p(Chất\ liệu = Nhựa\ PP|Doanh\ số\ bán = Thấp) = \frac{3+1}{8+2} = 0{,}4$$

$$p(X|Doanh\ số\ bán = Thấp) \times p(Doanh\ số\ bán = Thấp)$$

$$= \frac{4}{11} \times \frac{4}{11} \times \frac{4}{11} \times 0{,}4 \times \frac{9}{17} \approx 0{,}01$$

Như vậy hồ sơ thứ hai có xác suất xảy ra Doanh số bán cao lớn hơn, vậy ta có thể kết luận dòng dữ liệu thứ hai được dự đoán thuộc phân lớp Doanh số bán cao.

Xét hồ sơ thứ ba:

X = {Loại = Điều khiển, Số màu = 3, Kích thước = Vừa, Chất liệu = Cao su}

Ta cũng có những tính toán sau:

$$p(Loại = Điều\ khiển|Doanh\ số\ bán = Cao) = \frac{2+1}{7+3} = 0{,}3$$

$$p(Số\ màu = 3|Doanh\ số\ bán = Cao) = \frac{2+1}{7+3} = 0{,}3$$

$$p(Kích\ thước = Vừa|Doanh\ số\ bán = Cao) = \frac{2+1}{7+3} = 0{,}3$$

$$p(Chất\ liệu = Cao\ su|Doanh\ số\ bán = Cao) = \frac{1+1}{7+2} = \frac{2}{9} \approx 0{,}222$$

$$p(X|Doanh\ số\ bán = Cao) \times p(Doanh\ số\ bán = Cao)$$

$$= 0{,}3 \times 0{,}3 \times 0{,}3 \times \frac{2}{9} \times \frac{8}{17} \approx 0{,}003$$

$$p(Loại = Điều\ khiển|Doanh\ số\ bán = Thấp) = \frac{3+1}{8+3} = \frac{4}{11} \approx 0{,}364$$

$$p(Số\ màu = 3|Doanh\ số\ bán = Thấp) = \frac{3+1}{8+3} = \frac{4}{11} \approx 0{,}364$$

$$p(Kích\ thước = Vừa|Doanh\ số\ bán = Thấp) = \frac{3+1}{8+3} = \frac{4}{11} \approx 0{,}364$$

$$p(Chất\ liệu = Cao\ su|Doanh\ số\ bán = Thấp) = \frac{5+1}{8+2} = 0{,}6$$

$$p(X|Doanh\ số\ bán = Thấp) \times p(Doanh\ số\ bán = Thấp)$$

$$= \frac{4}{11} \times \frac{4}{11} \times \frac{4}{11} \times 0{,}6 \times \frac{9}{17} \approx 0{,}015$$

Như vậy dòng dữ liệu thứ ba được dự đoán thuộc phân lớp Doanh số bán thấp.

j) Với kết quả thu được và doanh số trên thực tế (Yêu cầu e), hãy lập ma trận nhầm lẫn, sau đó tính giá trị độ chính xác, độ phủ của thuật toán.

		Lớp dự đoán được từ mô hình	
Lớp trên thực tế		*Doanh số bán thấp*	*Doanh số bán cao*
	Doanh số bán thấp	1	1
	Doanh số bán cao	1	0

Theo công thức tính độ chính xác và độ phủ ta có:

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

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

k) So sánh kết quả từ thuật toán cây quyết định và Naïve Bayes.
Ở bài toán này thuật toán cây quyết định có độ chính xác và độ phủ cao hơn thuật toán Naïve Bayes

l) Sản phẩm mới của doanh nghiệp dự định tung ra thị trường có thông tin như sau:

Loại	Số màu	Kích thước	Chất liệu
Xếp hình	7	Nhỏ	Cao su

Hãy sử dụng các mô hình đã xây dựng được để dự đoán Doanh số bán của công ty với sản phẩm này

.

Theo thuật toán CART thì dự đoán Doanh số bán của công ty với sản phẩm này là Cao.

Thuật toán Naïve Bayes:

X = {Loại = Xếp hình, Số màu = 7, Kích thước = Nhỏ, Chất liệu = Cao su}

Ta cần tính được

$$p(Doanh\ số\ bán = Cao|X)$$

$$= \frac{p(X|Doanh\ số\ bán = Cao) \times p(Doanh\ số\ bán = Cao)}{p(X)}$$

so sánh với

$$p(Doanh\ số\ bán = Thấp|X)$$

$$= \frac{p(X|Doanh\ số\ bán = Thấp) \times p(Doanh\ số\ bán = Thấp)}{p(X)}$$

Vì mẫu số bằng nhau nên chỉ cần tính toán và so sánh hai tử số chúng ta sẽ có kết quả.

Áp dụng làm trơn Laplace, ta có:

$$p(Doanh\ số\ bán\ =\ Cao) = \frac{7+1}{15+2} = \frac{8}{17} \approx 0{,}47$$

$$p(Loại = Xếp\ hình|Doanh\ số\ bán = Cao) = \frac{2+1}{7+3} = 0{,}3$$

$$p(Số\ màu = 7|Doanh\ số\ bán = Cao) = \frac{2+1}{7+3} = 0{,}3$$

$$p(Kích\ thước = Nhỏ|Doanh\ số\ bán = Cao) = \frac{2+1}{7+3} = 0{,}3$$

$p(Chất\ liệu = Cao\ su|Doanh\ số\ bán = Cao) = \frac{1+1}{7+2} = \frac{2}{9} \approx 0{,}222$

$p(X|Doanh\ số\ bán = Cao) \times p(Doanh\ số\ bán = Cao)$

$= \frac{8}{17} \times 0{,}3 \times 0{,}3 \times 0{,}3 \times \frac{2}{9} \approx 0{,}003$

$p(Doanh\ số\ bán = Thấp) = \frac{8+1}{15+2} = \frac{9}{17} \approx 0{,}529$

$p(Loại = Xếp\ hình|Doanh\ số\ bán = Thấp) = \frac{3+1}{8+3} = \frac{4}{11} \approx 0{,}273$

$p(Số\ màu = 7|Doanh\ số\ bán = Thấp) = \frac{2+1}{8+3} = \frac{3}{11} \approx 0{,}364$

$p(Kích\ thước = Nhỏ|Doanh\ số\ bán = Thấp) = \frac{2+1}{8+3} = \frac{3}{11} \approx 0{,}364$

$p(Chất\ liệu = Cao\ su|Doanh\ số\ bán = Thấp) = \frac{5+1}{8+2} = 0{,}6$

$p(X|Doanh\ số\ bán = Thấp) \times p(Doanh\ số\ bán = Thấp)$

$= \frac{4}{11} \times \frac{3}{11} \times \frac{3}{11} \times 0{,}6 \times \frac{9}{17} \approx 0{,}009$

Như vậy dự đoán Doanh số bán thấp

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*

Sinh viên hãy thực hiện những yêu cầu sau:

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

Không có mâu thuẫn.

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

Xét thuộc tính Giảm, ta có:

$$Gini(S_{0..5}) = 1 - \left(\frac{3}{6}\right)^2 - \left(\frac{1}{6}\right)^2 - \left(\frac{2}{6}\right)^2 = \frac{11}{18}$$

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

$$Gini(S_{11..20}) = 1 - \left(\frac{2}{4}\right)^2 - \left(\frac{1}{4}\right)^2 - \left(\frac{1}{4}\right)^2 = 0{,}625$$

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

$$Gini_{Giảm}(S) = \frac{6}{14} \times \frac{11}{18} + \frac{2}{14} \times 0{,}5 + \frac{4}{14} \times 0{,}625 + \frac{2}{14} \times 0{,}5 \approx 0{,}583$$

Tiếp tục xét thuộc tính Người, ta có:

$$Gini(S_{0..5}) = 1 - \left(\frac{2}{5}\right)^2 - \left(\frac{2}{5}\right)^2 - \left(\frac{1}{5}\right)^2 = 0{,}64$$

$$Gini(S_{6..10}) = 1 - \left(\frac{2}{5}\right)^2 - \left(\frac{3}{5}\right)^2 = 0{,}48$$

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

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

$$Gini_{Người}(S) = \frac{5}{14} \times 0{,}64 + \frac{5}{14} \times 0{,}48 + \frac{2}{14} \times 0{,}5 + \frac{2}{14} \times 0{,}5 \approx 0{,}543$$

Với thuộc tính Chuyển, ta có:

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

$$Gini(S_{6..10}) = 1 - \left(\frac{2}{4}\right)^2 - \left(\frac{1}{4}\right)^2 - \left(\frac{1}{4}\right)^2 = 0{,}625$$

$$Gini(S_{11..20}) = 1 - \left(\frac{2}{5}\right)^2 - \left(\frac{2}{5}\right)^2 - \left(\frac{1}{5}\right)^2 = 0{,}64$$

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

$$Gini_{Chuyển}(S) = \frac{3}{14} \times \frac{4}{9} + \frac{4}{14} \times 0{,}625 + \frac{5}{14} \times 0{,}64 + \frac{2}{14} \times 0{,}5 \approx 0{,}574$$

Với thuộc tính Yêu, ta có:

$$Gini(S_{0..5}) = 1 - \left(\frac{2}{6}\right)^2 - \left(\frac{1}{6}\right)^2 - \left(\frac{3}{6}\right)^2 = \frac{11}{18}$$

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

$$Gini(S_{11..20}) = 1 - \left(\frac{2}{5}\right)^2 - \left(\frac{1}{5}\right)^2 - \left(\frac{2}{5}\right)^2 = 0{,}64$$

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

$$Gini_{Yêu}(S) = \frac{6}{14} \times \frac{11}{18} + \frac{2}{14} \times 0{,}5 + \frac{5}{14} \times 0{,}64 + \frac{1}{14} \times 0 \approx 0{,}562$$

Với thuộc tính Vừa, ta có:

$$Gini(S_{0..5}) = 1 - \left(\frac{1}{7}\right)^2 - \left(\frac{3}{7}\right)^2 - \left(\frac{3}{7}\right)^2 = \frac{30}{49}$$

$$Gini(S_{6..10}) = 1 - \left(\frac{2}{3}\right)^2 - \left(\frac{1}{3}\right)^2 = \frac{4}{9}$$

$$Gini(S_{11..20}) = 1 - \left(\frac{1}{3}\right)^2 - \left(\frac{1}{3}\right)^2 - \left(\frac{1}{3}\right)^2 = \frac{2}{9}$$

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

$$Gini_{Vừa}(S) = \frac{7}{14} \times \frac{30}{49} + \frac{3}{14} \times \frac{4}{9} + \frac{3}{14} \times \frac{2}{9} + \frac{1}{14} \times 0 \approx 0{,}449$$

Với thuộc tính Đi, ta có:

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

$$Gini(S_{6..10}) = 1 - \left(\frac{1}{3}\right)^2 - \left(\frac{2}{3}\right)^2 = \frac{4}{9}$$

$$Gini(S_{11..20}) = 1 - \left(\frac{2}{4}\right)^2 - \left(\frac{1}{4}\right)^2 - \left(\frac{1}{4}\right)^2 = 0{,}625$$

$$Gini(S_{>20}) = 1 - \left(\frac{1}{3}\right)^2 - \left(\frac{2}{3}\right)^2 = \frac{4}{9}$$

$$Gini_{Đi}(S) = \frac{4}{14} \times 0{,}5 + \frac{3}{14} \times \frac{4}{9} + \frac{4}{14} \times 0{,}625 + \frac{3}{14} \times \frac{4}{9} \approx 0{,}512$$

Chọn thuộc tính có chỉ số Gini thấp nhất là Vừa và phân nhánh.

Xét nhánh 0..5, tính giá trị chỉ số Gini tương tự như trên:

$$Gini_{giảm}(S_{0..5}) = \frac{2}{7} \times 0{,}5 + \frac{3}{7} \times \frac{2}{9} + \frac{2}{7} \times 0{,}5 \approx 0{,}38$$

$$Gini_{người}(S_{0..5}) = \frac{3}{7} \times \frac{2}{9} + \frac{2}{7} \times 0 + \frac{1}{7} \times 0 + \frac{1}{7} \times 0 \approx 0{,}095$$

$$Gini_{chuyển}(S_{0..5}) = \frac{2}{7} \times 0{,}5 + \frac{2}{7} \times 0{,}5 + \frac{2}{7} \times 0{,}5 + \frac{1}{7} \times 0 \approx 0{,}429$$

$$Gini_{yêu}(S_{0..5}) = \frac{1}{7} \times 0{,}5 + \frac{1}{7} \times 0{,}5 + \frac{4}{7} \times 0{,}625 + \frac{1}{7} \times 0 = 0{,}5$$

$$Gini_{đi}(S_{0..5}) = \frac{1}{7} \times 0{,}5 + \frac{3}{7} \times \frac{4}{9} + \frac{2}{7} \times 0{,}5 + \frac{1}{7} \times 0 \approx 0{,}405$$

- Chọn thuộc tính Người và tiếp tục phát triển cây, ta tính các chỉ số Gini, xét nhánh 0..5:

$$Gini_{giảm}(S_{0..5}) = \frac{1}{3} \times 0 + \frac{1}{3} \times 0 + \frac{1}{3} \times 0 = 0$$

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

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

$$Gini_{đi}(S_{0..5}) = \frac{1}{3} \times 0 + \frac{1}{3} \times 0 + \frac{1}{3} \times 0 = 0$$

Chỉ số Gini của thuộc tính Giảm và Đi nhỏ bằng nhau, Chọn thuộc tính Giảm để phân nhánh, đến đây các nhánh 6..11 đến lá Xấu, nhánh 11.20 đến lá Tốt, nhánh >20 đến lá Bình thường.

- Xét nhánh 6..10, cây đi đến lá Xấu
- Xét nhánh 11..20 và >20, cây đều đi đến lá Bình thường:

Quay lại xét 6..10, Ta có:

$$Gini_{giảm}(S_{6..10}) = \frac{3}{3} \times \frac{4}{9} \approx 0{,}444$$

$$Gini_{người}(S_{6..10}) = \frac{1}{3} \times 0 + \frac{1}{3} \times 0 + \frac{1}{3} \times 0 = 0$$

$$Gini_{chuyển}(S_{6..10}) = \frac{1}{3} \times 0 + \frac{2}{3} \times 0{,}5 \approx 0{,}333$$

$$Gini_{yêu}(S_{6..10}) = \frac{2}{3} \times 0{,}5 + \frac{1}{3} \times 0 \approx 0{,}333$$

$$Gini_{đi}(S_{6..10}) = \frac{2}{3} \times 0{,}5 + \frac{1}{3} \times 0 \approx 0{,}333$$

Chọn thuộc tính Người để phân nhánh, nhánh 0..5 và 6..10 đến lá Tốt, nhánh >20 đến lá Xấu

Tiếp túc xét nhánh 11.20, ta có:

$$Gini_{giảm}(S_{11.20}) = \frac{2}{3} \times 0{,}5 + \frac{1}{3} \times 0 \approx 0{,}333$$

$$Gini_{người}(S_{11.20}) = \frac{2}{3} \times 0{,}5 + \frac{1}{3} \times 0 \approx 0{,}333$$

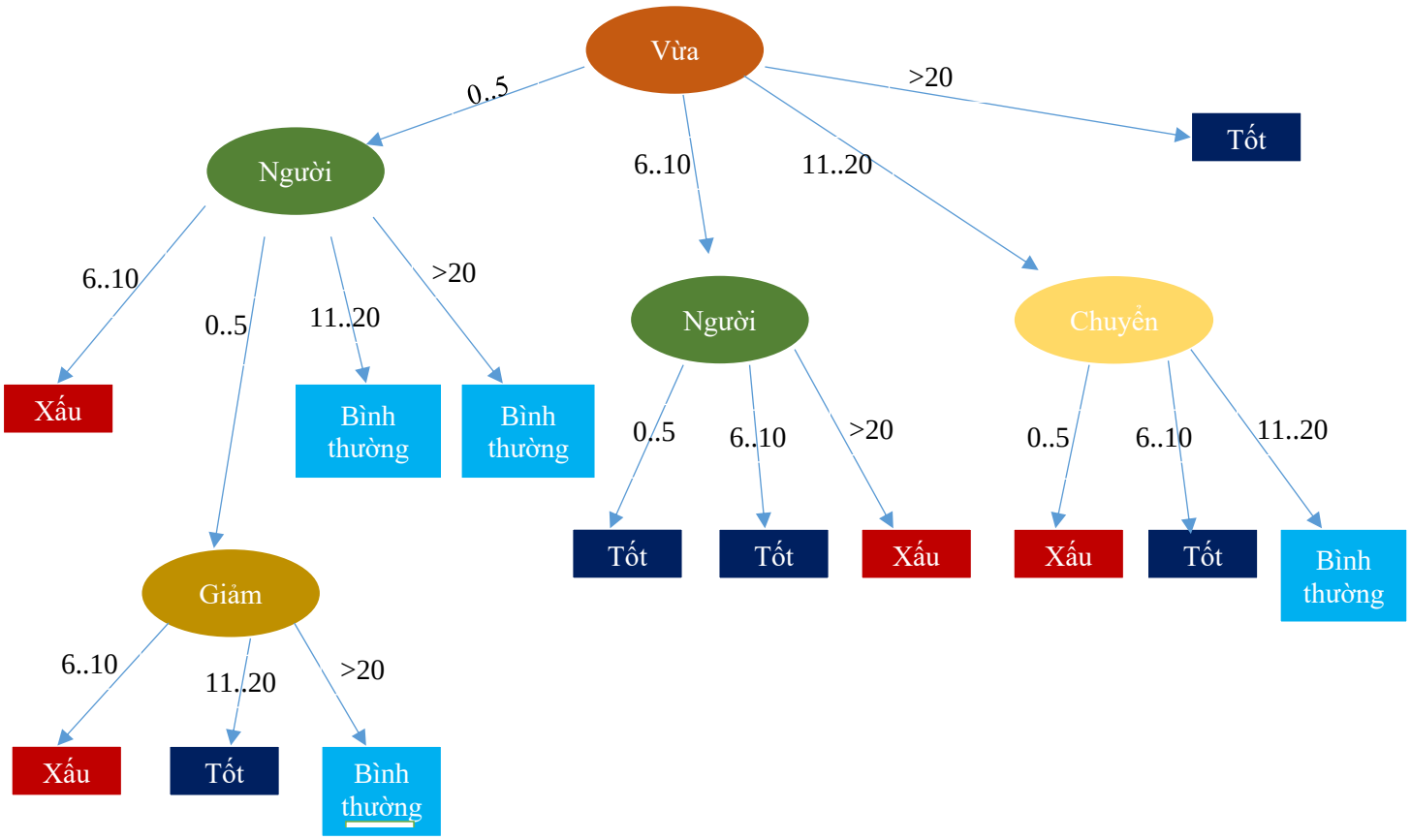
$$Gini_{chuyển}(S_{11.20}) = \frac{1}{3} \times 0 + \frac{1}{3} \times 0 + \frac{1}{3} \times 0 = 0$$

$$Gini_{yêu}(S_{11.20}) = \frac{3}{3} \times \frac{4}{9} \approx 0{,}444$$

$$Gini_{đi}(S_{11.20}) = \frac{2}{3} \times 0{,}5 + \frac{1}{3} \times 0 \approx 0{,}333$$

Chọn thuộc tính Chuyển để phân nhánh, nhánh 0..5 đến lá Xấu, nhánh 6..10 đến lá Tốt, và nhánh 11..20 đến lá Bình thường

Với nhánh >20, cây phân đến lá Tốt



c) 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	Cảm xúc
0..5	6..10	0..5	11..20	6..10	0..5	Tốt
0..5	0..5	6..10	0..5	11..20	>20	Tốt
6..10	0..5	11..20	>20	6..10	6..10	Tốt
6..10	11..20	6..10	6..10	>20	0..5	Tốt

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?

Chọn Cảm xúc tốt là lớp dương, lớp còn lại – Cảm xúc xấu và bình thường sẽ là lớp âm, ta có ma trận nhầm lẫn của cây quyết định như sau:

Lớp dự đoán được từ mô hình				
Lớp trên thực tế		*Cảm xúc tốt*	*Cảm xúc bình thường*	*Cảm xúc xấu*
	Cảm xúc tốt	2	0	0
	Cảm xúc bình thường	1	0	0
	Cảm xúc xấu	1	0	0

Theo công thức tính độ chính xác và độ phủ ta có:

$$precision(M) = \frac{2}{2+1+1} = 50\%$$

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

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

Sẽ sử dụng để có các chiến lược quảng cáo và marketing hiệu quả