

ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH

**TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN**

KHAI THÁC DỮ LIỆU

**BÀI TẬP LAP**

**NGUYỄN THỊ NGỌC HÀ: 17520421**

MỤC LỤC

[I. BÀI TẬP 1: 4](#_Toc41518641)

[a. Xác định tất cả những mâu thuẫn có thể có trong dữ liệu 4](#_Toc41518642)

[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. 4](#_Toc41518643)

[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. 11](#_Toc41518644)

[d. Sử dụng một trong hai cây quyết định ở trên để tiên đoán giá trị DOANH SỐ BÁN của những sản phẩm sau: 16](#_Toc41518645)

[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. 16](#_Toc41518646)

[f. Xác suất không điều kiện của giá trị ‘Xếp Hình’ trong tập dữ liệu là bao nhiêu? 17](#_Toc41518647)

[g. Khi DOANH SỐ BÁN là ‘THẤP’, hãy tính xác suất đó là những sản phẩm có chất liệu là ‘CAO SU’ 18](#_Toc41518648)

[h. Dựa theo định lý Bayes, hãy viết công thức tính xác suất DOANH SỐ BÁN ‘CAO’ của những sản phẩm thuộc loại ‘ĐIỀU KHIỂN’ 18](#_Toc41518649)

[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. 18](#_Toc41518650)

[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. 23](#_Toc41518651)

[k. So sánh kết quả từ thuật toán cây quyết định và Naïve Bayes. 24](#_Toc41518652)

[l. Sản phẩm mới của doanh nghiệp dự định tung ra thị trường như sau. Sử dụng mô hình đã xây dựng được để dự đoán DOANH SỐ của công ty với sản phẩm này. 24](#_Toc41518653)

[II. BÀI TẬP 2: 26](#_Toc41518654)

[a. Xác định tất cả nhưng mâu thuẫn có thể có trong dữ liệu: 26](#_Toc41518655)

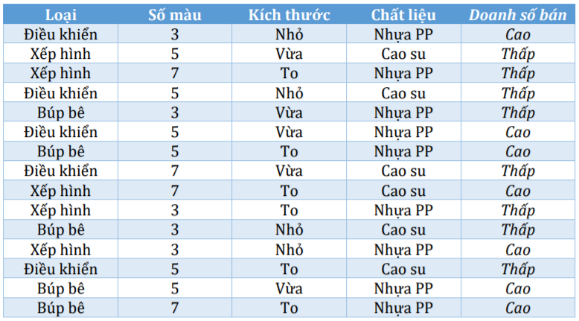
[b. Tính giá trị chỉ số Gini (Gini index) của các thuộc tính và vẽ cây quyết định theo thuật toán CART cho dữ liệu trên. 26](#_Toc41518656)

[c. Sử dụng cây quyết định và thuật toán Naïve Bayes để dự đoán cảm xúc của những trạng thái sau: 32](#_Toc41518657)

[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? 42](#_Toc41518658)

[e. Nếu nắm bắt được cảm xúc của người dùng mạng xã hội thì sinh viên sẽ sử dụng chúng như thế nào? 43](#_Toc41518659)

# BÀI TẬP 1:



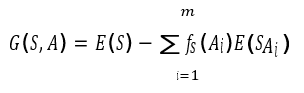
Một doanh nghiệp sản xuất đồ chơi cho trẻ em muốn dự đoán doanh số của các sản phẩm sắp đưa ra thị trường, do đó họ đã thu thập các dữu liệu trên.

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

Mâu thuẫn trong tập dữ liệu phân lớp là những dòng dữ liệu có giá trị thuộc tính giống nhau nhưng lại thuộc các phân lớp khác nhau. Trong dữ liệu trên, không có mâu thuẫn dữ liệu xảy ra.

## 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.

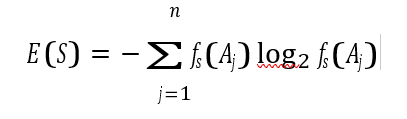
Đầu tiên, người ta tính giá trị độ lợi thông tin ở tất cả các thuộc tính để chọn ra thuộc tính ở nút gốc, theo công thức sau:



Trong đó:

𝐺(𝑆, 𝐴) là độ lợi thông tin của tập S khi phân chia theo thuộc tính A.

𝐸(𝑆) là độ bất định (Entropy) của tập S được tính theo công thức sau:



Với:

𝑚 là số giá trị khác nhau của thuộc tính A đang xét

𝐴i là số mẫu tương ứng với mỗi giá trị i của thuộc tính A

𝑓s (𝐴i) là tỷ lệ của số mẫu có thuộc tính 𝐴i với S

𝑆Ai là một tập con của S chứa tất cả các mẫu có giá trị 𝐴i

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

E(S) = = 1

Tính độ lợi thông tin ở lần lượt từng thuộc tính

* **Tính Gain (S, LOẠI)**

Với thuộc tính LOẠI ta có 2 giá trị phân biệt: Điều khiển (ĐK), Búp bê (BB), Xếp hình (XH).

Có: 5 Trường hợp ĐK với 2 CAO 3 THẤP

5 Trường hợp BB với 3 CAO 2 THẤP

5 Trường hợp ĐK với 2 CAO 3 THẤP

Như vậy, ta tính được:

E (SĐK) = = 0,97

E (SBB) = = 0,97

E (SXH) = = 0,97

* Gain (S, LOẠI) = = 0,03
* **Tính Gain (S, SỐ MÀU)**

Với thuộc tính SỐ MÀU ta có 3 giá trị phân biệt: 3 Màu (3), 5 Màu (5), 7 Màu (7).

Có: 5 Trường hợp 3 với 2 CAO 3 THẤP

6 Trường hợp 5 với 3 CAO 3 THẤP

4 Trường hợp 7 với 2 CAO 2 THẤP

Như vậy, ta tính được:

E (S3) = = 0,97

E (S5) = = 1

E (S7) = = 1

* Gain (S, SỐ MÀU) = = 0,01
* **Tính Gain (S, KÍCH THƯỚC)**

Với thuộc tính KÍCH THƯỚC ta có 3 giá trị phân biệt: Nhỏ (S), Vừa (M), To (L).

Có: 4 Trường hợp S với 2 CAO 2 THẤP

5 Trường hợp M với 2 CAO 3 THẤP

6 Trường hợp L với 3 CAO 3 THẤP

Như vậy, ta tính được:

E (SS) = = 1

E (SM) = = 0,97

E (SL) = = 1

* Gain (S, KÍCH THƯỚC) = = 0,01
* **Tính Gain (S, CHẤT LIỆU)**

Với thuộc tính CHẤT LIỆU ta có 2 giá trị phân biệt: Nhựa PP (PP) và Cao su (CS).

Có: 9 Trường hợp PP với 6 CAO 3 THẤP

6 Trường hợp CS với 1 CAO 5 THẤP

Như vậy, ta tính được:

E (SPP) = = 0,92

E (SCS) = = 0,65

* Gain (S, CHẤT LIỆU) = = 0,188

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 (PP) gồm 9 dòng, phần có giá trị Cao su (CS) gồm 6 dòng.

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

E (SĐK) = 0; E (SBB) = 0,81; E (SXH) = 0,92

* Gain( SPP, LOẠI) = = 0,25

Với 2 thuốc tính KÍCH THƯỚC và SỐ MÀU, ta có:

E (S3) = 0; E (S5) = 0; E (S7) = 1

* Gain( SPP, SỐ MÀU) = = 0,7

E (SS) = 0; E (SM) = 0,92; E (SL) = 0,81

* Gain( SPP, KÍCH THƯỚC) = = 0,25

Vậy ở nhánh này ta chọn Thuộc tính SỐ MÀU làm phép chia nhánh. Với giá trị 5 Màu (5), ta luôn có phân lớp DOANH SỐ BÁN là CAO. Vì vậy nhánh này đi đến nút lá và không cần xét tiếp.

Hai nhánh con tương ứng với hai giá trị còn lại là 3 Màu (3) và 7 Màu (7).

Nhánh tương ứng với giá trị 3 Màu (3) gồm 4 dòng, xét các thuộc tính còn lại ta có:

E (SĐK) = 0; E (SBB) = 0; E (SXH) = 1

* Gain( S3, LOẠI) = = 0,47

E (SS) = 0; E (SM) = 0; E (SL) = 0

* Gain( S3, KÍCH THƯỚC) = = 0,97

Với giá trị độ lợi thông tin lớn hơn, cây sẽ tiếp tục được phân nhánh bằng thuộc tính KÍCH THƯỚC. Đến đây, nhánh Nhỏ (S) sẽ đi đến nút lá CAO, nhánh Vừa (M) và To (L) sẽ đi đến nút lá THẤP.

Nhánh tương ứng với giá trị 7 Màu (7) gồm 2 dòng dữ liệu, tiếp tục tính độ lợi thông tin ta có:

E (SBB) = 0; E (SXH) = 0

* Gain( S7, LOẠI) = = 1

E (SL) = 1

* Gain( S7, KÍCH THƯỚC) = = 0

Với giá trị độ lợi thông tin lớn hơn, cây sẽ tiếp tục được phân nhánh bằng thuộc tính LOẠI. Đến đây, nhánh Búp bê (BB) sẽ đi đến nút lá CAO, nhánh Xếp hình (XH) sẽ đi đến nút lá THẤP.

Trở lại với nhánh cây thứ hai, nhánh Cao su (CS), xét thuộc tính LOẠI, tính toán tương tự, ta có:

E (SĐK) = 0; E (SBB) = 0; E (SXH) = 1

* Gain( SCS, LOẠI) = = 0,32

Với 2 thuốc tính KÍCH THƯỚC và SỐ MÀU, ta có:

E (S3) = 0; E (S5) = 0; E (S7) = 1

* Gain( SCS, SỐ MÀU) = = 0,32

E (SS) = 0; E (SM) = 0; E (SL) = 1

* Gain( SPP, KÍCH THƯỚC) = = 0,32

Ở nhánh này có ba thuộc tính có độ lợi thông tin cao bằng nhau đó là: LOẠI, KÍCH THƯỚC và SỐ MÀU. Ta chọn LOẠI để tiếp tục, với giá trị Điều khiển (ĐK) và Búp bê (BB) cây sẽ phát triển đến nút lá THẤP. Chúng ta chỉ việc xem xét phát triển cây với giá trị Xếp hình (XH), ta có:

E (SM) = 0; E (SL) = 0

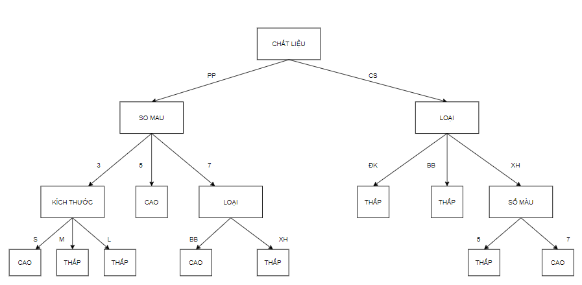
* Gain( SXH, KÍCH THƯỚC) = = 0,97

E (S5) = 0; E (S7) = 0

* Gain( SXH, SỐ MÀU) = = 0,97

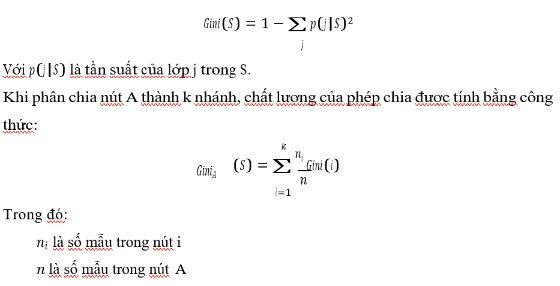
Ở nhánh này cả 2 thuộc tính có độ lợi thông tin cao bằng nhau đó là: KÍCH THƯỚC và SỐ MÀU. Ta chọn SỐ MÀU để tiếp tục, với giá trị 5 Màu (5) cây sẽ phát triển đến nút lá THẤP, với giá trị 7 Màu (7) cây sẽ phát triển đến nút lá CAO, như vậy thuật toán kết thúc.

Kết quả cây quyết định được trình bày như hình bên dưới:



## 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.

Ta có chỉ số Gini của tập huấn luyện S:



* **Tính GiniLOẠI(S)**

Gini(SĐK) = = 0,48

Gini(SBB) = = 0,48

Gini(SXH) = = 0,48

* GiniLOẠI(S) = = 0,48
* **Tính GiniSỐ MÀU(S)**

Gini(S3) = = 0,48

Gini(S5) = = 0,5

Gini(S7) = = 0,5

* GiniLOẠI(S) = = 0,49
* **Tính GiniKÍCH THƯỚC(S)**

Gini(SS) = = 0,5

Gini(SM) = = 0,48

Gini(SL) = = 0,5

* GiniKÍCH THƯỚC(S) = = 0,49
* **Tính GiniCHẤT LIỆU(S)**

Gini(SPP) = = 0,44

Gini(SCS) = = 0,28

* GiniCHẤT LIỆU(S) = = 0,376

Chọn thuộc tính có chỉ số Gini thấp nhất là CHẤT LIỆU và phân nhánh theo.

Tương tự như câu b, khi xét xuống nhánh Nhựa PP (PP), ta tính được giá trị chỉ số Gini như sau:

GiniLOẠI(SPP) = = 0,31

GiniSỐ MÀU(SPP) = = 0,33

GiniKÍCH THƯỚC(SPP) = = 0,37

Chọn thuộc tính LOẠI và tiếp tục phát triển cây. Với giá trị Điều khiển (ĐK) ta luôn có phân lớp CAO. Do đó, ta chỉ cần tính các chỉ số Gini của giá trị Búp bê (BB) và Xếp hình (XH) như sau:

Với nhánh Búp bê, ta có:

GiniSỐ MÀU(SBB) = = 0

GiniKÍCH THƯỚC(SBB) = = 0,25

Chọn thuộc tính SỐ MÀU và tiếp tục phát triển cây. Với giá trị 3 Màu (3) ta luôn có phân lớp THẤP, giá trị 7 Màu (7), ta luôn có phân lớp CAO và giá trị 5 Màu (5), ta luôn có phân lớp CAO.

Với nhánh Xếp hình (XH), ta có:

GiniSỐ MÀU(SXH) = = 0,33

GiniKÍCH THƯỚC(SXH) = = 0

Chọn thuộc tính KÍCH THƯỚC và tiếp tục phát triển cây. Với giá trị Nhỏ (S) ta luôn có phân lớp CAO và giá trị To (L), ta luôn có phân lớp THẤP.

Trở lại với nhánh cây thứ hai, nhánh Cao su (CS), ta tính được giá trị chỉ số Gini như sau:

GiniLOẠI(SCS) = = 0,166

GiniSỐ MÀU(SCS) = = 0,166

GiniKÍCH THƯỚC(SCS) = = 0,166

Ở nhánh này cả 3 thuộc tính có độ lợi thông tin cao bằng nhau đó là: LOẠI, KÍCH THƯỚC và SỐ MÀU. Ta chọn LOẠI để tiếp tục, với giá trị Điều khiển (ĐK) cây sẽ phát triển đến nút lá THẤP, với giá trị Búp bê (BB) cây sẽ phát triển đến nút lá THẤP.

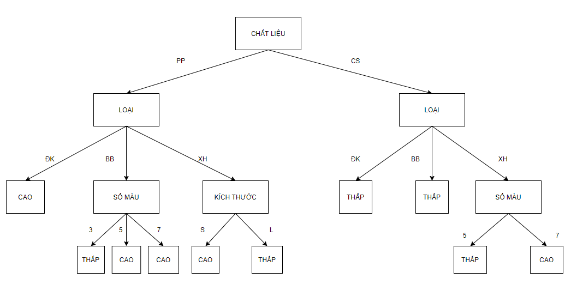
Với nhánh Xếp hình (XH), ta có:

GiniSỐ MÀU(SXH) = = 0

GiniKÍCH THƯỚC(SXH) = = 0

Ở nhánh này cả 2 thuộc tính có độ lợi thông tin cao bằng nhau đó là: KÍCH THƯỚC và SỐ MÀU. Ta chọn SỐ MÀU để tiếp tục, với giá trị 5 Màu (5) cây sẽ phát triển đến nút lá THẤP và với giá trị 7 Màu (7) cây sẽ phát triển đến nút lá CAO, như vậy thuật toán kết thúc.

Kết quả cây quyết định được trình bày như hình bên dưới:



## Sử dụng một trong hai cây quyết định ở trên để tiên đoán giá trị DOANH SỐ BÁN của những sản phẩm sau:



Dựa theo cây quyết định CART ở câu b, cả 3 sản phẩm này đều có tiên đoán DOANH SỐ BÁN là THẤP.

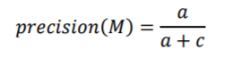
## Doanh số bán trên thực tế của các sản phẩm ở Yêu cầu d lần lượt là Thấp, Thấp, Cao. Hãy lập ma trận nhầm lẫn, sau đó tính giá trị độ chính xác, độ phủ của mô hình/ cây đã xây dựng.

Ma trận nhầm lẫn hiển thị các thông tin sau:

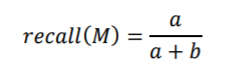


Từ đó ta có công thức sau:

* Công thức tính Độ chính xác (Precision):



* Công thức tính Độ phủ (Recall):



Ta chọn phân lớp THẤP là lớp dương, phân lớp CAO là lớp âm, ta có ma trận nhầm lẫn sau:

|  |  |  |  |
| --- | --- | --- | --- |
| LỚP DỰ ĐOÁN ĐƯỢC TỪ MÔ HÌNH | | | |
| LỚP TRÊN  THỰC TẾ |  | THẤP | CAO |
| THẤP | 2 |  |
| CAO | 1 |  |

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

Precision(M) = ≈ 0,67 tức xấp xỉ 67%

Recall(M) = = 1 tức bằng 100%

## Xác suất không điều kiện của giá trị ‘Xếp Hình’ trong tập dữ liệu là bao nhiêu?

P(LOẠI = XH) = = 0,33

## Khi DOANH SỐ BÁN là ‘THẤP’, hãy tính xác suất đó là những sản phẩm có chất liệu là ‘CAO SU’

P(CHẤT LIỆU = CS | DOANH SỐ BÁN = THẤP) = = 0,55

## Dựa theo định lý Bayes, hãy viết công thức tính xác suất DOANH SỐ BÁN ‘CAO’ của những sản phẩm thuộc loại ‘ĐIỀU KHIỂN’

P(DOANH SỐ BÁN = CAO | LOẠI = ĐK)

=

## 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.

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

Với hồ sơ đầu tiên

X = { LOẠI = BB, SỐ MÀU = 3, KÍCH THƯỚC = L, CHẤT LIỆU = CS }

Ta cần tính được:

P(DOANH SỐ = CAO|X)

=

So sánh với:

P(DOANH SỐ = THẤP|X)

=

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

Áp dụng lằm trơn Laplace, ta có:

\*\* P(DOANH SỐ = CAO) = = = 0,47

P( LOẠI = BB | DOANH SỐ = CAO) = = = 0,4

P( SỐ MÀU = 3 | DOANH SỐ = CAO) = = = 0,3

P( KÍCH THƯỚC = L | DOANH SỐ = CAO) = = = 0,4

P( CHẤT LIỆU = CS | DOANH SỐ = CAO) = = = 0,22

* P( X | DOANH SỐ = CAO) x P(DOANH SỐ = CAO)

= P( LOẠI = BB | DOANH SỐ = CAO) x P( SỐ MÀU = 3 | DOANH SỐ = CAO) x P( KÍCH THƯỚC = L | DOANH SỐ = CAO) x P( CHẤT LIỆU = CS | DOANH SỐ = CAO) x P(DOANH SỐ = CAO)

= = 0,005

\*\* P(DOANH SỐ = THẤP) = = = 0,47

P( LOẠI = BB | DOANH SỐ = THẤP) = = = 0,27

P( SỐ MÀU = 3 | DOANH SỐ = THẤP) = = = 0,36

P( KÍCH THƯỚC = L | DOANH SỐ = THẤP) = = = 0,36

P( CHẤT LIỆU = CS | DOANH SỐ = THẤP) = = = 0,6

* P( X | DOANH SỐ = THẤP) x P(DOANH SỐ = THẤP)

= P( LOẠI = BB | DOANH SỐ = THẤP) x P( SỐ MÀU = 3 | DOANH SỐ = THẤP) x P( KÍCH THƯỚC = L | DOANH SỐ = THẤP) x P( CHẤT LIỆU = CS | DOANH SỐ = THẤP) x P(DOANH SỐ = THẤP)

= = 0,1113

Như vậy, hồ sơ đầu tiên có xác suất xảy ra DOANH SỐ là THẤP lớn hơn, vậy ta có thể kết luận dòng dữ liệu đầu tiên được dự đoán thuộc phân lớp THẤP.

Xét hồ sơ thứ hai:

X = { LOẠI = XH, SỐ MÀU = 5, KÍCH THƯỚC = L, CHẤT LIỆU = PP }

Ta cũng có những tính toán sau:

\*\* P(DOANH SỐ = CAO) = = = 0,47

P( LOẠI = XH | DOANH SỐ = CAO) = = = 0,3

P( SỐ MÀU = 5 | DOANH SỐ = CAO) = = = 0,4

P( KÍCH THƯỚC = L | DOANH SỐ = CAO) = = = 0,4

P( CHẤT LIỆU = PP | DOANH SỐ = CAO) = = = 0,78

* P( X | DOANH SỐ = CAO) x P(DOANH SỐ = CAO)

= P( LOẠI = XH | DOANH SỐ = CAO) x P( SỐ MÀU = 5 | DOANH SỐ = CAO) x P( KÍCH THƯỚC = L | DOANH SỐ = CAO) x P( CHẤT LIỆU = PP | DOANH SỐ = CAO) x P(DOANH SỐ = CAO)

= = 0,0176

\*\* P(DOANH SỐ = THẤP) = = = 0,47

P( LOẠI = XH | DOANH SỐ = THẤP) = = = 0,36

P( SỐ MÀU = 5 | DOANH SỐ = THẤP) = = = 0,36

P( KÍCH THƯỚC = L | DOANH SỐ = THẤP) = = = 0,36

P( CHẤT LIỆU = PP | DOANH SỐ = THẤP) = = = 0,4

* P( X | DOANH SỐ = THẤP) x P(DOANH SỐ = THẤP)

= P( LOẠI = XH | DOANH SỐ = THẤP) x P( SỐ MÀU = 5 | DOANH SỐ = THẤP) x P( KÍCH THƯỚC = L | DOANH SỐ = THẤP) x P( CHẤT LIỆU = PP | DOANH SỐ = THẤP) x P(DOANH SỐ = THẤP)

= = 0,01

Như vậy, hồ sơ thứ hai có xác suất xảy ra DOANH SỐ là CAO lớn hơn, vậy ta có thể kết luận dòng dữ liệu thứ hai này được dự đoán thuộc phân lớp CAO.

Xét hồ sơ thứ ba:

X = { LOẠI = ĐK, SỐ MÀU = 3, KÍCH THƯỚC = M, CHẤT LIỆU = CS }

Ta cũng có những tính toán sau:

\*\* P(DOANH SỐ = CAO) = = = 0,47

P( LOẠI = ĐK | DOANH SỐ = CAO) = = = 0,3

P( SỐ MÀU = 3 | DOANH SỐ = CAO) = = = 0,3

P( KÍCH THƯỚC = M | DOANH SỐ = CAO) = = = 0,3

P( CHẤT LIỆU = CS | DOANH SỐ = CAO) = = = 0,22

* P( X | DOANH SỐ = CAO) x P(DOANH SỐ = CAO)

= P( LOẠI = ĐK | DOANH SỐ = CAO) x P( SỐ MÀU = 3 | DOANH SỐ = CAO) x P( KÍCH THƯỚC = M | DOANH SỐ = CAO) x P( CHẤT LIỆU = CS | DOANH SỐ = CAO) x P(DOANH SỐ = CAO)

= = 0,0028

\*\* P(DOANH SỐ = THẤP) = = = 0,47

P( LOẠI = ĐK | DOANH SỐ = THẤP) = = = 0,36

P( SỐ MÀU = 3 | DOANH SỐ = THẤP) = = = 0,36

P( KÍCH THƯỚC = M | DOANH SỐ = THẤP) = = = 0,36

P( CHẤT LIỆU = CS | DOANH SỐ = THẤP) = = = 0,6

* P( X | DOANH SỐ = THẤP) x P(DOANH SỐ = THẤP)

= P( LOẠI = ĐK | DOANH SỐ = THẤP) x P( SỐ MÀU = 3 | DOANH SỐ = THẤP) x P( KÍCH THƯỚC = M | DOANH SỐ = THẤP) x P( CHẤT LIỆU = CS | DOANH SỐ = THẤP) x P(DOANH SỐ = THẤP)

= = 0,015

Như vậy, hồ sơ thứ ba có xác suất xảy ra DOANH SỐ là THẤP lớn hơn, vậy ta có thể kết luận dòng dữ liệu thứ ba này được dự đoán thuộc phân lớp THẤP.

## 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.

Kết quả thu được từ câu I cho 3 hồ sơ lần lượt là THẤP, CAO, THẤP

So với câu e là THẤP, THẤP, CAO, ta có được ma trận nhầm lẫn như sau.

Chọn phân lớp THẤP là lớp Dương, ta có:

|  |  |  |  |
| --- | --- | --- | --- |
| LỚP DỰ ĐOÁN ĐƯỢC TỪ MÔ HÌNH | | | |
| LỚP TRÊN  THỰC TẾ |  | THẤP | CAO |
| THẤP | 1 | 1 |
| CAO | 1 |  |

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

Precision(M) = = 0,5 tức 50%

Recall(M) = = 0,5 tức 50%

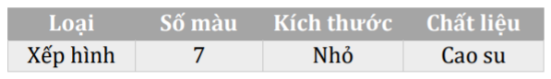
## So sánh kết quả từ thuật toán cây quyết định và Naïve Bayes.

Ta có độ chính xác từ thuật toán cây quyết định là xấp xỉ 67% và độ chính xác từ thuật toán Naïve Bayes là 50%.

Độ phủ của thuật toán cây quyết định là 100% và độ phủ từ thuật toán Naïve Bayes là 50%.

* Sử dụng thuật toán cây quyết định ta có được kết quả chính xác hơn.

## Sản phẩm mới của doanh nghiệp dự định tung ra thị trường như sau. Sử dụng mô hình đã xây dựng được để dự đoán DOANH SỐ của công ty với sản phẩm này.

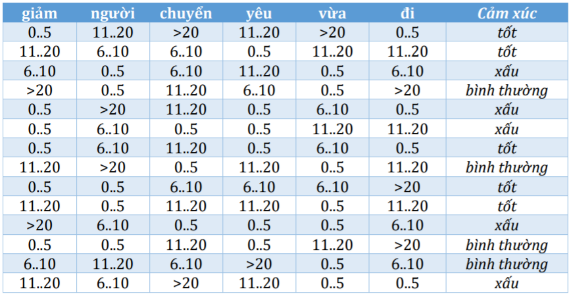


Dựa theo thuật toán cây quyết định CART, sản phẩm sắp tung ra của công ty có phân lớp CAO.

Dựa theo thuật toán Naïve Bayes, sản phẩm sắp tung ra của công ty có phân lớp THẤP.

# BÀI TẬP 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.



## Xác định tất cả nhưng mâu thuẫn có thể có trong dữ liệu:

Mâu thuẫn trong tập dữ liệu phân lớp là những dòng dữ liệu có giá trị thuộc tính giống nhau nhưng lại thuộc các phân lớp khác nhau. Trong dữ liệu trên, không có mâu thuẫn dữ liệu xảy ra.

## 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.

* **Tính Gini (S, GIẢM)**

Gini(S0-5) = = 0,61

Gini(S6-10) = = 0,5

Gini(S11-20) = = 0,625

Gini(S>20) = = 0,5

* GiniGIẢM(S) = = 0,58
* **Tính Gini (S, NGƯỜI)**

Gini(S0-5) = = 0,64

Gini(S6-10) = = 0,48

Gini(S11-20) = = 0,5

Gini(S>20) = = 0,5

* GiniNGƯỜI(S) = = 0,543
* **Tính Gini (S, CHUYỂN)**

Gini(S0-5) = = 0,44

Gini(S6-10) = = 0,625

Gini(S11-20) = = 0,64

Gini(S>20) = = 0,5

* GiniCHUYỂN(S) = = 1,12
* **Tính Gini (S, YÊU)**

Gini(S0-5) = = 0,61

Gini(S6-10) = = 0,5

Gini(S11-20) = = 0,64

Gini(S>20) = = 0

* GiniYÊU(S) = = 0,56
* **Tính Gini (S, VỪA)**

Gini(S0-5) = = 0,61

Gini(S6-10) = = 0,44

Gini(S11-20) = = 0,66

Gini(S>20) = = 0

* GiniVỪA(S) = = 0,541
* **Tính Gini (S, ĐI)**

Gini(S0-5) = = 0,5

Gini(S6-10) = = 0,44

Gini(S11-20) = = 0,625

Gini(S>20) = = 0,44

* GiniĐI(S) = = 0,51

Trong 6 thuộc tính đã xem xét, ta chọn thuộc tính có chỉ số Gini thấp nhất là ĐI làm phép chia nhánh cho cây tại nút gốc.

Tập dữ liệu lúc này được chia làm 4 phần tương ứng với 4 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, phần có giá trị >20 gồm 3 dòng.

Với nhánh cây thứ nhất, nhánh 0-5, tương tự như trên, ta có:

GiniGIẢM(S0-5) = = 0,33

GiniNGƯỜI(S0-5) = = 0,25

GiniCHUYỂN(S0-5) = = 0,5

GiniYÊU(S0-5) = = 0,5

GiniVỪA(S0-5) = = 0,25

Chọn thuộc tính VỪA và tiếp tục phát triển cây. Với giá trị 0-5 ta luôn có phân lớp XẤU, với giá trị >20 ta luôn có phân lớp TỐT. Do đó, ta chỉ cần tính các chỉ số Gini của giá trị 6-10 như sau:

GiniGIẢM(S6-10) = = 0,5

GiniNGƯỜI(S6-10) = = 0

GiniCHUYỂN(S6-10) = = 0,5

GiniYÊU(S6-10) = = 0,5

Chọn thuộc tính NGƯỜI và tiếp tục phát triển cây. Với giá trị 6-10 ta luôn có phân lớp TỐT, với giá trị >20 ta luôn có phân lớp XẤU.

Trở lại với nhánh 6-10:

GiniGIẢM(S6-10) = = 0,33

GiniNGƯỜI(S6-10) = = 0

GiniCHUYỂN(S6-10) = = 0

GiniYÊU(S6-10) = = 0

GiniVỪA(S6-10) = = 0,44

Ở nhánh này có 3 thuộc tính có độ lợi thông tin cao bằng nhau đó là: NGƯỜI, CHUYỂN và YÊU. Ta chọn NGƯỜI để tiếp tục, với giá trị 0-5 và giá trị 6-10 cây sẽ phát triển đến nút lá XẤU, với giá trị 11-20 cây sẽ phát triển đến nút lá BÌNH THƯỜNG.

Trở lại với nhánh 11-20:

GiniGIẢM(S11-20) = = 0,33

GiniNGƯỜI(S11-20) = = 0,25

GiniCHUYỂN(S11-20) = = 0,25

GiniYÊU(S11-20) = = 0,5

GiniVỪA(S11-20) = = 0,5

Ở nhánh này có 2 thuộc tính có độ lợi thông tin cao bằng nhau đó là: NGƯỜI, CHUYỂN. Ta chọn NGƯỜI để tiếp tục, với giá trị 0-5 cây sẽ phát triển đến nút lá TỐT, với giá trị >20 cây sẽ phát triển đến nút lá BÌNH THƯỜNG. Do đó, ta chỉ cần tính các chỉ số Gini của giá trị 6-10 như sau:

GiniGIẢM(S6-10) = = 0

GiniCHUYỂN(S6-10) = = 0

GiniYÊU(S6-10) = = 0,5

Ở nhánh này có 2 thuộc tính có độ lợi thông tin cao bằng nhau đó là: GIẢM và CHUYỂN. Ta chọn GIẢM để tiếp tục, với giá trị 0-5 cây sẽ phát triển đến nút lá XẤU, với giá trị 11-20 cây sẽ phát triển đến nút lá TỐT.

Trở lại với nhánh >20:

GiniGIẢM(S>20) = = 0,33

GiniNGƯỜI(S>20) = = 0,44

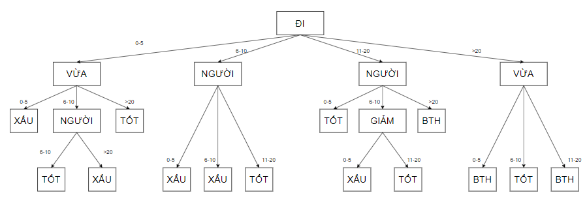
GiniCHUYỂN(S>20) = = 0

GiniYÊU(S>20) = = 0,33

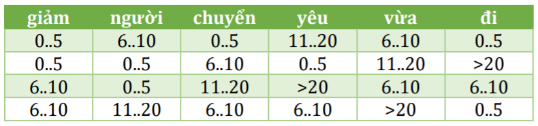
GiniVỪA(S>20) = = 0

Ở nhánh này có 2 thuộc tính có độ lợi thông tin cao bằng nhau đó là: VỪA và CHUYỂN. Ta chọn VỪA để tiếp tục, với giá trị 0-5 và giá trị 11-20 cây sẽ phát triển đến nút lá BÌNH THƯỜNG, với giá trị 6-10 cây sẽ phát triển đến nút lá TỐT. như vậy thuật toán kết thúc.

Kết quả cây quyết định được trình bày như hình bên dưới:



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



Xét tập hồ sơ thứ nhất.

Áp dụng lằm trơn Laplace, ta có:

\*\* P(CẢM XÚC = TỐT) = 0,35

P( GIẢM = 0-5 | CẢM XÚC = TỐT) = 0,44

P( NGƯỜI = 6-10 | CẢM XÚC = TỐT) = 0,33

P( CHUYỂN = 0-5 | CẢM XÚC = TỐT) = 0,11

P( YÊU = 11-20 | CẢM XÚC = TỐT) = 0,33

P( VỪA = 6-10 | CẢM XÚC = TỐT) = 0,11

P( ĐI = 0-5 | CẢM XÚC = TỐT) = 0,33

* P( X | CẢM XÚC = TỐT) x P(CẢM XÚC = TỐT)

= P( GIẢM = 0-5 | CẢM XÚC = TỐT) x P( NGƯỜI = 6-10 | CẢM XÚC = TỐT) x P( CHUYỂN = 0-5 | CẢM XÚC = TỐT) x P( YÊU = 11-20 | CẢM XÚC = TỐT) x P( VỪA = 6-10 | CẢM XÚC = TỐT) x P( ĐI = 0-5 | CẢM XÚC = TỐT) x P(CẢM XÚC = TỐT)

= 0,000067

\*\* P(CẢM XÚC = XẤU) = 0,35

P( GIẢM = 0-5 | CẢM XÚC = XẤU) = 0,33

P( NGƯỜI = 6-10 | CẢM XÚC = XẤU) = 0,44

P( CHUYỂN = 0-5 | CẢM XÚC = XẤU) = 0,33

P( YÊU = 11-20 | CẢM XÚC = XẤU) = 0,33

P( VỪA = 6-10 | CẢM XÚC = XẤU) = 0,22

P( ĐI = 0-5 | CẢM XÚC = XẤU) = 0,33

* P( X | CẢM XÚC = XẤU) x P( CẢM XÚC = XẤU)

= P( GIẢM = 0-5 | CẢM XÚC = XẤU) x P( NGƯỜI = 6-10 | CẢM XÚC = XẤU) x P( CHUYỂN = 0-5 | CẢM XÚC = XẤU) x P( YÊU = 11-20 | CẢM XÚC = XẤU) x P( VỪA = 6-10 | CẢM XÚC = XẤU) x P( ĐI = 0-5 | CẢM XÚC = XẤU) x P( CẢM XÚC = XẤU)

= 0,0004

\*\* P(CẢM XÚC = BT) = 0,294

P( GIẢM = 0-5 | CẢM XÚC = BT) = 0,25

P( NGƯỜI = 6-10 | CẢM XÚC = BT) = 0,125

P( CHUYỂN = 0-5 | CẢM XÚC = BT) = 0,25

P( YÊU = 11-20 | CẢM XÚC = BT) = 0,25

P( VỪA = 6-10 | CẢM XÚC = BT) = 0,125

P( ĐI = 0-5 | CẢM XÚC = BT) = 0,125

* P( X | CẢM XÚC = BT) x P( CẢM XÚC = BT)

= P( GIẢM = 0-5 | CẢM XÚC = BT) x P( NGƯỜI = 6-10 | CẢM XÚC = BT) x P( CHUYỂN = 0-5 | CẢM XÚC = BT) x P( YÊU = 11-20 | CẢM XÚC = BT) x P( VỪA = 6-10 | CẢM XÚC = BT) x P( ĐI = 0-5 | CẢM XÚC = BT) x P( CẢM XÚC = BT)

= 0,00000897

Như vậy, hồ sơ đầu tiên có xác suất xảy ra CẢM XÚC là XẤU lớn hơn, vậy ta có thể kết luận dòng dữ liệu đầu tiên được dự đoán thuộc phân lớp XẤU.

Xét hồ sơ thứ hai:

\*\* P(CẢM XÚC = TỐT) = 0,35

P( GIẢM = 0-5 | CẢM XÚC = TỐT) = 0,44

P( NGƯỜI = 0-5 | CẢM XÚC = TỐT) = 0,33

P( CHUYỂN = 6-10 | CẢM XÚC = TỐT) = 0,33

P( YÊU = 0-5 | CẢM XÚC = TỐT) = 0,33

P( VỪA = 11-20 | CẢM XÚC = TỐT) = 0,22

P( ĐI = >20 | CẢM XÚC = TỐT) = 0,22

* P( X | CẢM XÚC = TỐT) x P(CẢM XÚC = TỐT)

= P( GIẢM = 0-5 | CẢM XÚC = TỐT) x P( NGƯỜI = 0-5 | CẢM XÚC = TỐT) x P( CHUYỂN = 6-10 | CẢM XÚC = TỐT) x P( YÊU = 0-5 | CẢM XÚC = TỐT) x P( VỪA = 11-20 | CẢM XÚC = TỐT) x P( ĐI = >20 | CẢM XÚC = TỐT) x P(CẢM XÚC = TỐT)

= 0,000268

\*\* P(CẢM XÚC = XẤU) = 0,35

P( GIẢM = 0-5 | CẢM XÚC = XẤU) = 0,33

P( NGƯỜI = 0-5 | CẢM XÚC = XẤU) = 0,22

P( CHUYỂN = 6-10 | CẢM XÚC = XẤU) = 0,22

P( YÊU = 0-5 | CẢM XÚC = XẤU) = 0,44

P( VỪA = 11-20 | CẢM XÚC = XẤU) = 0,22

P( ĐI = >20 | CẢM XÚC = XẤU) = 0,11

* P( X | CẢM XÚC = XẤU) x P( CẢM XÚC = XẤU)

= P( GIẢM = 0-5 | CẢM XÚC = XẤU) x P( NGƯỜI = 0-5 | CẢM XÚC = XẤU) x P( CHUYỂN = 6-10 | CẢM XÚC = XẤU) x P( YÊU = 0-5 | CẢM XÚC = XẤU) x P( VỪA = 11-20 | CẢM XÚC = XẤU) x P( ĐI = >20 | CẢM XÚC = XẤU) x P( CẢM XÚC = XẤU)

= 0,00006

\*\* P(CẢM XÚC = BT) = 0,294

P( GIẢM = 0-5 | CẢM XÚC = BT) = 0,25

P( NGƯỜI = 0-5 | CẢM XÚC = BT) = 0,375

P( CHUYỂN = 6-10 | CẢM XÚC = BT) = 0,25

P( YÊU = 0-5 | CẢM XÚC = BT) = 0,25

P( VỪA = 11-20 | CẢM XÚC = BT) = 0,25

P( ĐI = >20 | CẢM XÚC = BT) = 0,375

* P( X | CẢM XÚC = BT) x P( CẢM XÚC = BT)

= P( GIẢM = 0-5 | CẢM XÚC = BT) x P( NGƯỜI = 0-5 | CẢM XÚC = BT) x P( CHUYỂN = 6-10 | CẢM XÚC = BT) x P( YÊU = 0-5 | CẢM XÚC = BT) x P( VỪA = 11-20 | CẢM XÚC = BT) x P( ĐI = >20 | CẢM XÚC = BT) x P( CẢM XÚC = BT)

= 0,00016

Như vậy, hồ sơ thứ hai có xác suất xảy ra CẢM XÚC là TỐT lớn hơn, vậy ta có thể kết luận dòng dữ liệu thứ hai được dự đoán thuộc phân lớp TỐT.

Xét hồ sơ thứ ba:

\*\* P(CẢM XÚC = TỐT) = 0,35

P( GIẢM = 6-10 | CẢM XÚC = TỐT) = 0,11

P( NGƯỜI = 0-5 | CẢM XÚC = TỐT) = 0,33

P( CHUYỂN = 11-20 | CẢM XÚC = TỐT) = 0,33

P( YÊU = >20 | CẢM XÚC = TỐT) = 0,11

P( VỪA = 6-10 | CẢM XÚC = TỐT) = 0,33

P( ĐI = 6-10 | CẢM XÚC = TỐT) = 0,11

* P( X | CẢM XÚC = TỐT) x P(CẢM XÚC = TỐT)

= P( GIẢM = 6-10 | CẢM XÚC = TỐT) x P( NGƯỜI = 0-5 | CẢM XÚC = TỐT) x P( CHUYỂN = 11-20 | CẢM XÚC = TỐT) x P( YÊU = >20 | CẢM XÚC = TỐT) x P( VỪA = 6-10 | CẢM XÚC = TỐT) x P( ĐI = 6-10 | CẢM XÚC = TỐT) x P(CẢM XÚC = TỐT)

= 0,0000167

\*\* P(CẢM XÚC = XẤU) = 0,35

P( GIẢM = 6-10 | CẢM XÚC = XẤU) = 0,22

P( NGƯỜI = 0-5 | CẢM XÚC = XẤU) = 0,22

P( CHUYỂN = 11-20 | CẢM XÚC = XẤU) = 0,22

P( YÊU = >20 | CẢM XÚC = XẤU) = 0,11

P( VỪA = 6-10 | CẢM XÚC = XẤU) = 0,22

P( ĐI = 6-10 | CẢM XÚC = XẤU) = 0,33

* P( X | CẢM XÚC = XẤU) x P(CẢM XÚC = XẤU)

= P( GIẢM = 6-10 | CẢM XÚC = XẤU) x P( NGƯỜI = 0-5 | CẢM XÚC = XẤU) x P( CHUYỂN = 11-20 | CẢM XÚC = XẤU) x P( YÊU = >20 | CẢM XÚC = XẤU) x P( VỪA = 6-10 | CẢM XÚC = XẤU) x P( ĐI = 6-10 | CẢM XÚC = XẤU) x P(CẢM XÚC = XẤU)

= 0,00003

\*\* P(CẢM XÚC = BT) = 0,294

P( GIẢM = 6-10 | CẢM XÚC = BT) = 0,25

P( NGƯỜI = 0-5 | CẢM XÚC = BT) = 0,375

P( CHUYỂN = 11-20 | CẢM XÚC = BT) = 0,375

P( YÊU = >20 | CẢM XÚC = BT) = 0,25

P( VỪA = 6-10 | CẢM XÚC = BT) = 0,125

P( ĐI = 6-10 | CẢM XÚC = BT) = 0,25

* P( X | CẢM XÚC = BT) x P(CẢM XÚC = BT)

= P( GIẢM = 6-10 | CẢM XÚC = BT) x P( NGƯỜI = 0-5 | CẢM XÚC = BT) x P( CHUYỂN = 11-20 | CẢM XÚC = BT) x P( YÊU = >20 | CẢM XÚC = BT) x P( VỪA = 6-10 | CẢM XÚC = BT) x P( ĐI = 6-10 | CẢM XÚC = BT) x P(CẢM XÚC = BT)

= 0,00008

Như vậy, hồ sơ thứ ba có xác suất xảy ra CẢM XÚC là BT (Bình Thường) lớn hơn, vậy ta có thể kết luận dòng dữ liệu thứ ba được dự đoán thuộc phân lớp BT (Bình Thường).

Xét hồ sơ thứ tư:

\*\* P(CẢM XÚC = TỐT) = 0,35

P( GIẢM = 6-10 | CẢM XÚC = TỐT) = 0,11

P( NGƯỜI = 11-20 | CẢM XÚC = TỐT) = 0,22

P( CHUYỂN = 6-10 | CẢM XÚC = TỐT) = 0,33

P( YÊU = 6-10 | CẢM XÚC = TỐT) = 0,22

P( VỪA = >20 | CẢM XÚC = TỐT) = 0,22

P( ĐI = 0-5 | CẢM XÚC = TỐT) = 0,33

* P( X | CẢM XÚC = TỐT) x P(CẢM XÚC = TỐT)

= P( GIẢM = 6-10 | CẢM XÚC = TỐT) x P( NGƯỜI = 11-20 | CẢM XÚC = TỐT) x P( CHUYỂN = 6-10 | CẢM XÚC = TỐT) x P( YÊU = 6-10 | CẢM XÚC = TỐT) x P( VỪA = >20 | CẢM XÚC = TỐT) x P( ĐI = 0-5 | CẢM XÚC = TỐT) x P(CẢM XÚC = TỐT)

= 0,0000446

\*\* P(CẢM XÚC = XẤU) = 0,35

P( GIẢM = 6-10 | CẢM XÚC = XẤU) = 0,22

P( NGƯỜI = 11-20 | CẢM XÚC = XẤU) = 0,11

P( CHUYỂN = 6-10 | CẢM XÚC = XẤU) = 0,22

P( YÊU = 6-10 | CẢM XÚC = XẤU) = 0,11

P( VỪA = >20 | CẢM XÚC = XẤU) = 0,11

P( ĐI = 0-5 | CẢM XÚC = XẤU) = 0,33

* P( X | CẢM XÚC = XẤU) x P(CẢM XÚC = XẤU)

= P( GIẢM = 6-10 | CẢM XÚC = XẤU) x P( NGƯỜI = 11-20 | CẢM XÚC = XẤU) x P( CHUYỂN = 6-10 | CẢM XÚC = XẤU) x P( YÊU = 6-10 | CẢM XÚC = XẤU) x P( VỪA = >20 | CẢM XÚC = XẤU) x P( ĐI = 0-5 | CẢM XÚC = XẤU) x P(CẢM XÚC = XẤU)

= 0,00000744

\*\* P(CẢM XÚC = BT) = 0,294

P( GIẢM = 6-10 | CẢM XÚC = BT) = 0,25

P( NGƯỜI = 11-20 | CẢM XÚC = BT) = 0,25

P( CHUYỂN = 6-10 | CẢM XÚC = BT) = 0,25

P( YÊU = 6-10 | CẢM XÚC = BT) = 0,25

P( VỪA = >20 | CẢM XÚC = BT) = 0,125

P( ĐI = 0-5 | CẢM XÚC = BT) = 0,125

* P( X | CẢM XÚC = BT) x P(CẢM XÚC = BT)

= P( GIẢM = 6-10 | CẢM XÚC = BT) x P( NGƯỜI = 11-20 | CẢM XÚC = BT) x P( CHUYỂN = 6-10 | CẢM XÚC = BT) x P( YÊU = 6-10 | CẢM XÚC = BT) x P( VỪA = >20 | CẢM XÚC = BT) x P( ĐI = 0-5 | CẢM XÚC = BT) x P(CẢM XÚC = BT)

= 0,0000179

Như vậy, hồ sơ thứ tư có xác suất xảy ra CẢM XÚC là TỐT lớn hơn, vậy ta có thể kết luận dòng dữ liệu thứ tư được dự đoán thuộc phân lớp TỐT.

Như vậy:

Dùng thuật toán cây quyết định, ta có kết quả cho 4 tập hồ sơ trên lần lượt là: TỐT, TỐT, XẤU, TỐT.

Dùng thuật toán Naïve Bayes, ta có kết quả cho 4 tập hồ sơ trên lần lượt là: XẤU, TỐT, BT, TỐT.

## 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 phân lớp TỐT là lớp dương, ta có ma trận nhầm lẫn cho thuật toán cây quyết định như sau:

|  |  |  |  |
| --- | --- | --- | --- |
| LỚP DỰ ĐOÁN ĐƯỢC TỪ MÔ HÌNH | | | |
| LỚP TRÊN  THỰC TẾ |  | TỐT | XẤU+BT |
| TỐT | 2 | 0 |
| XẤU+BT | 1 | 1 |

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

Precision(M) = = 0,67 tức 67%

Recall(M) = = 1 tức 100%

Chọn phân lớp TỐT là lớp dương, ta có ma trận nhầm lẫn cho thuật toán cây quyết định như sau:

|  |  |  |  |
| --- | --- | --- | --- |
| LỚP DỰ ĐOÁN ĐƯỢC TỪ MÔ HÌNH | | | |
| LỚP TRÊN  THỰC TẾ |  | TỐT | XẤU+BT |
| TỐT | 2 | 0 |
| XẤU+BT | 0 | 2 |

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

Precision(M) = = 1 tức 100%

Recall(M) = = 1 tức 100%

Như vậy, đối với tập hồ sơ này, thuật toán Naïve Bayes cho kết quả độ chính xác và độ phủ lớn hơn so với thuật toán cây quyết định, tức sử dụng thuật toán Naïve Bayes trong trường hợp này sẽ cho kết quả đúng hơn.

Điều này còn cho thấy độ chính xác và độ phủ phụ thuộc vào giá trị thực tế mà bài toán đạt được. Ngoài ra, còn cho thấy muốn xác định độ chính xác và độ phủ của từng tập dữ liệu, ta cần tính giá trị này đối với nhiều thuật toán khác nhau để chọn ra thuật toán phù hợp nhất cho tập dữ liệu.

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