**TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN BÁO CÁO THỰC HÀNH**

KHOA HỆ THỐNG THÔNG TIN Môn: Khai thác dữ liệu

**BÀI THỰC HÀNH SỐ 4**

Giảng viên hướng dẫn: ThS. Nguyễn Hồ Duy Trí

Sinh viên thực hiện: Phạm Thiện Bảo

MSSV: 20521107

Lớp: IS252.N22.HTCL

**Bài làm**

**IV. Thực hành:**

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

A picture containing text, screenshot, number, font

Description automatically generated

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

Tập dữ liệu trên **không xảy ra mâu thuẫn** trong tập dữ liệu phân lớp vì không tồn tại các dòng dữ liệu có giá trị thuộc tính giống nhau nhưng lại thuộc phân lớp khác nhau => không thể tìm ra 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

Ta có các công thức sau:

- Giá trị độ lợi thông tin (Information Gain):

- Độ bất định (Entropy) của tập S:

Trong đó:

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

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.

là một tập con của S chứa tất cả các mẫu có giá trị 𝐴.

**\*Bước 1: Tính độ bất định của tập S**

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

**\*Bước 2: Tính độ lợi thông tin ở lần lượt từng thuộc tính**

***+ Thuộc tính “Loại” :***

|  |  |  |  |
| --- | --- | --- | --- |
| Giá trị phân biệt | Số dòng gán nhãn “Cao” | Số dòng gán nhãn “Thấp” | Tổng số dòng |
| Điều khiển | 2 | 3 | 5 |
| Xếp hình | 2 | 3 | 5 |
| Búp bê | 3 | 2 | 5 |

***+ Thuộc tính “Số màu” :***

|  |  |  |  |
| --- | --- | --- | --- |
| Giá trị phân biệt | Số dòng gán nhãn “Cao” | Số dòng gán nhãn “Thấp” | Tổng số dòng |
| 3 | 2 | 3 | 5 |
| 5 | 3 | 3 | 6 |
| 7 | 2 | 2 | 4 |

***+ Thuộc tính “Kích thước” :***

|  |  |  |  |
| --- | --- | --- | --- |
| Giá trị phân biệt | Số dòng gán nhãn “Cao” | Số dòng gán nhãn “Thấp” | Tổng số dòng |
| Nhỏ | 2 | 2 | 4 |
| Vừa | 3 | 2 | 5 |
| To | 3 | 3 | 6 |

***+ Thuộc tính “Chất liệu” :***

|  |  |  |  |
| --- | --- | --- | --- |
| Giá trị phân biệt | Số dòng gán nhãn “Cao” | Số dòng gán nhãn “Thấp” | Tổng số dòng |
| Nhựa PP | 6 | 3 | 9 |
| Cao su | 1 | 5 | 6 |

**\*Bước 3: Chia nhánh cho cây**

Trong 4 thuộc tính đã xem xét, **“Chất liệu”** có độ lợi thông tin lớn nhất => Root node là “Chất liệu”.

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 “Nhựa PP”, xét lần lượt:

***+ Thuộc tính “Loại” – Chất liệu = “Nhựa PP” :***

|  |  |  |  |
| --- | --- | --- | --- |
| Giá trị phân biệt | Số dòng gán nhãn “Cao” | Số dòng gán nhãn “Thấp” | Tổng số dòng |
| Điều khiển | 2 | 0 | 2 |
| Xếp hình | 1 | 2 | 3 |
| Búp bê | 3 | 1 | 4 |

***+ Thuộc tính “Số màu” – Chất liệu = “Nhựa PP”:***

|  |  |  |  |
| --- | --- | --- | --- |
| Giá trị phân biệt | Số dòng gán nhãn “Cao” | Số dòng gán nhãn “Thấp” | Tổng số dòng |
| 3 | 2 | 2 | 4 |
| 5 | 3 | 0 | 3 |
| 7 | 1 | 1 | 2 |

***+ Thuộc tính “Kích thước” – Chất liệu = “Nhựa PP”:***

|  |  |  |  |
| --- | --- | --- | --- |
| Giá trị phân biệt | Số dòng gán nhãn “Cao” | Số dòng gán nhãn “Thấp” | Tổng số dòng |
| Nhỏ | 2 | 0 | 2 |
| Vừa | 1 | 2 | 3 |
| To | 2 | 2 | 4 |

Vậy ở nhánh này ta chọn Thuộc tính “Loại” làm phép chia nhánh. Với giá trị “Loại” = Điều khiển, ta luôn có phân lớp 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 Xếp hình và Búp bê sẽ tiếp tục được phát triển.

Với nhánh tương ứng với giá trị “Xếp hình”, gồm 3 dòng dữ liệu, xét lần lượt:

***+ Thuộc tính “Số màu” – Loại = “Xếp hình” và Chất liệu = “Nhựa PP” :***

|  |  |  |  |
| --- | --- | --- | --- |
| Giá trị phân biệt | Số dòng gán nhãn “Cao” | Số dòng gán nhãn “Thấp” | Tổng số dòng |
| 3 | 1 | 1 | 2 |
| 5 | 0 | 0 | 0 |
| 7 | 0 | 1 | 1 |

***+ Thuộc tính “Kích thước” – Loại = “Xếp hình” và Chất liệu = “Nhựa PP”:***

|  |  |  |  |
| --- | --- | --- | --- |
| Giá trị phân biệt | Số dòng gán nhãn “Cao” | Số dòng gán nhãn “Thấp” | Tổng số dòng |
| Nhỏ | 1 | 0 | 1 |
| Vừa | 0 | 0 | 0 |
| To | 0 | 2 | 0 |

Với giá trị độ lợi thông tin bằng nhau, ta chọn ngẫu nhiên một thuộc tính, ở đây ta cho cây sẽ tiếp tục được phân nhánh bằng thuộc tính “Kích thước”. Đến đây, nhánh Nhỏ sẽ cho nút lá thuộc phân lớp cao và nhánh To sẽ cho nút lá thuộc phân lớp thấp. Kết thúc chia nhánh.

Với nhánh tương ứng với giá trị “Búp bê”, gồm 4 dòng dữ liệu, xét lần lượt:

***+ Thuộc tính “Số màu” – Loại = “Búp bê” và Chất liệu = “Nhựa PP”:***

|  |  |  |  |
| --- | --- | --- | --- |
| Giá trị phân biệt | Số dòng gán nhãn “Cao” | Số dòng gán nhãn “Thấp” | Tổng số dòng |
| 3 | 0 | 1 | 1 |
| 5 | 2 | 0 | 2 |
| 7 | 1 | 0 | 1 |

***+ Thuộc tính “Kích thước” – Loại = “Búp bê” và Chất liệu = “Nhựa PP”:***

|  |  |  |  |
| --- | --- | --- | --- |
| Giá trị phân biệt | Số dòng gán nhãn “Cao” | Số dòng gán nhãn “Thấp” | Tổng số dòng |
| Nhỏ | 0 | 0 | 0 |
| Vừa | 1 | 1 | 2 |
| To | 2 | 0 | 2 |

Vậy ở nhánh “Búp bê” ta chọn thuộc tính “Số màu” làm phép chia nhánh. Với giá trị “Số màu” = 3, ta luôn có phân lớp Thấp; “Số màu” = 5 hoặc “Số màu” = 7 thì ta luôn có phân lớp cao. Vì vậy nhánh này đi đến nút lá và không cần xét tiếp.

Trở lại nhánh “Cao su” được phân chia từ nút gốc gồm 6 dòng dữ liệu, xét lần lượt:

***+ Thuộc tính “Loại” – Chất liệu = “Cao su”:***

|  |  |  |  |
| --- | --- | --- | --- |
| Giá trị phân biệt | Số dòng gán nhãn “Cao” | Số dòng gán nhãn “Thấp” | Tổng số dòng |
| Điều khiển | 0 | 3 | 3 |
| Xếp hình | 1 | 1 | 2 |
| Búp bê | 0 | 1 | 1 |

***+ Thuộc tính “Số màu” – Chất liệu = “Cao su”:***

|  |  |  |  |
| --- | --- | --- | --- |
| Giá trị phân biệt | Số dòng gán nhãn “Cao” | Số dòng gán nhãn “Thấp” | Tổng số dòng |
| 3 | 0 | 1 | 1 |
| 5 | 0 | 3 | 3 |
| 7 | 1 | 1 | 2 |

***+ Thuộc tính “Kích thước” – Chất liệu = “Cao su”:***

|  |  |  |  |
| --- | --- | --- | --- |
| Giá trị phân biệt | Số dòng gán nhãn “Cao” | Số dòng gán nhãn “Thấp” | Tổng số dòng |
| Nhỏ | 0 | 2 | 2 |
| Vừa | 0 | 2 | 2 |
| To | 1 | 1 | 2 |

Với giá trị độ lợi thông tin bằng nhau, ta chọn ngẫu nhiên một thuộc tính, ở đây ta cho cây sẽ tiếp tục được phân nhánh bằng thuộc tính “Loại” có 3 giá trị “Điều khiển”, “Xếp hình” và “Búp bê”. Đến đây, nhánh Điều khiển và nhánh Búp bê sẽ cho nút lá thuộc phân lớp thấp. Ta chỉ xem xét phát triển cây với giá trị Xếp hình của thuộc tính.

Với nhánh tương ứng với giá trị “Xếp hình”, gồm 2 dòng dữ liệu, xét lần lượt:

***+ Thuộc tính “Kích thước” – Loại = “Xếp hình” và Chất liệu = “Cao su”:***

|  |  |  |  |
| --- | --- | --- | --- |
| Giá trị phân biệt | Số dòng gán nhãn “Cao” | Số dòng gán nhãn “Thấp” | Tổng số dòng |
| Nhỏ | 0 | 0 | 0 |
| Vừa | 0 | 1 | 1 |
| To | 1 | 0 | 1 |

***+ Thuộc tính “Số màu” – Loại = “Xếp hình” và Chất liệu = “Cao su”:***

|  |  |  |  |
| --- | --- | --- | --- |
| Giá trị phân biệt | Số dòng gán nhãn “Cao” | Số dòng gán nhãn “Thấp” | Tổng số dòng |
| 3 | 0 | 0 | 0 |
| 5 | 0 | 1 | 1 |
| 7 | 1 | 0 | 1 |

Với giá trị độ lợi thông tin bằng nhau, ta chọn ngẫu nhiên một thuộc tính, ở đây ta cho cây sẽ tiếp tục được phân nhánh bằng thuộc tính “Số màu”. Nhánh 5 sẽ cho nút lá thuộc phân lớp thấp và nhánh 7 sẽ cho nút lá thuộc phân lớp cao.

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

A picture containing diagram, plan, text, design

Description automatically generated

## 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.

***+ Thuộc tính “Loại” :***

|  |  |  |  |
| --- | --- | --- | --- |
| Giá trị phân biệt | Số dòng gán nhãn “Cao” | Số dòng gán nhãn “Thấp” | Tổng số dòng |
| Điều khiển | 2 | 3 | 5 |
| Xếp hình | 2 | 3 | 5 |
| Búp bê | 3 | 2 | 5 |

***+ Thuộc tính “Số màu” :***

|  |  |  |  |
| --- | --- | --- | --- |
| Giá trị phân biệt | Số dòng gán nhãn “Cao” | Số dòng gán nhãn “Thấp” | Tổng số dòng |
| 3 | 2 | 3 | 5 |
| 5 | 3 | 3 | 6 |
| 7 | 2 | 2 | 4 |

***+ Thuộc tính “Kích thước” :***

|  |  |  |  |
| --- | --- | --- | --- |
| Giá trị phân biệt | Số dòng gán nhãn “Cao” | Số dòng gán nhãn “Thấp” | Tổng số dòng |
| Nhỏ | 2 | 2 | 4 |
| Vừa | 3 | 2 | 5 |
| To | 3 | 3 | 6 |

***+ Thuộc tính “Chất liệu” :***

|  |  |  |  |
| --- | --- | --- | --- |
| Giá trị phân biệt | Số dòng gán nhãn “Cao” | Số dòng gán nhãn “Thấp” | Tổng số dòng |
| Nhựa PP | 6 | 3 | 9 |
| Cao su | 1 | 5 | 6 |

Chọn thuộc tính “Chất liệu” 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 **“Chất liệu”**. Phần có giá trị Nhựa PP gồm 9 dòng, phần có giá trị Cao su gồm 6 dòng.

Với nhánh “Nhựa PP”, xét lần lượt:

***+ Thuộc tính “Loại” – Chất liệu = “Nhựa PP” :***

|  |  |  |  |
| --- | --- | --- | --- |
| Giá trị phân biệt | Số dòng gán nhãn “Cao” | Số dòng gán nhãn “Thấp” | Tổng số dòng |
| Điều khiển | 2 | 0 | 2 |
| Xếp hình | 1 | 2 | 3 |
| Búp bê | 3 | 1 | 4 |

***+ Thuộc tính “Số màu” – Chất liệu = “Nhựa PP”:***

|  |  |  |  |
| --- | --- | --- | --- |
| Giá trị phân biệt | Số dòng gán nhãn “Cao” | Số dòng gán nhãn “Thấp” | Tổng số dòng |
| 3 | 2 | 2 | 4 |
| 5 | 3 | 0 | 3 |
| 7 | 1 | 1 | 2 |

***+ Thuộc tính “Kích thước” – Chất liệu = “Nhựa PP”:***

|  |  |  |  |
| --- | --- | --- | --- |
| Giá trị phân biệt | Số dòng gán nhãn “Cao” | Số dòng gán nhãn “Thấp” | Tổng số dòng |
| Nhỏ | 2 | 0 | 2 |
| Vừa | 1 | 2 | 3 |
| To | 2 | 2 | 4 |

Vậy ở nhánh này ta chọn Thuộc tính “Loại” làm phép chia nhánh. Với giá trị “Loại” = Điều khiển, ta luôn có phân lớp 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 Xếp hình và Búp bê sẽ tiếp tục được phát triển.

Với nhánh tương ứng với giá trị “Xếp hình”, gồm 3 dòng dữ liệu, xét lần lượt:

***+ Thuộc tính “Số màu” – Loại = “Xếp hình” và Chất liệu = “Nhựa PP”:***

|  |  |  |  |
| --- | --- | --- | --- |
| Giá trị phân biệt | Số dòng gán nhãn “Cao” | Số dòng gán nhãn “Thấp” | Tổng số dòng |
| 3 | 1 | 1 | 2 |
| 5 | 0 | 0 | 0 |
| 7 | 0 | 1 | 1 |

***+ Thuộc tính “Kích thước” – Loại = “Xếp hình” và Chất liệu = “Nhựa PP”:***

|  |  |  |  |
| --- | --- | --- | --- |
| Giá trị phân biệt | Số dòng gán nhãn “Cao” | Số dòng gán nhãn “Thấp” | Tổng số dòng |
| Nhỏ | 1 | 0 | 1 |
| Vừa | 0 | 2 | 2 |
| To | 0 | 0 | 0 |

Ta cho cây sẽ tiếp tục được phân nhánh bằng thuộc tính “Kích thước” vì Gini thấp nhất. Đến đây, nhánh Nhỏ sẽ cho nút lá thuộc phân lớp cao và nhánh vừa sẽ cho nút lá thuộc phân lớp thấp. Kết thúc chia nhánh.

Với nhánh tương ứng với giá trị “Búp bê”, gồm 4 dòng dữ liệu, xét lần lượt:

***+ Thuộc tính “Số màu” – Loại = “Xếp hình” và Chất liệu = “Nhựa PP”:***

|  |  |  |  |
| --- | --- | --- | --- |
| Giá trị phân biệt | Số dòng gán nhãn “Cao” | Số dòng gán nhãn “Thấp” | Tổng số dòng |
| 3 | 0 | 1 | 1 |
| 5 | 2 | 0 | 2 |
| 7 | 1 | 0 | 1 |

***+ Thuộc tính “Kích thước” – Loại = “Xếp hình” và Chất liệu = “Nhựa PP”:***

|  |  |  |  |
| --- | --- | --- | --- |
| Giá trị phân biệt | Số dòng gán nhãn “Cao” | Số dòng gán nhãn “Thấp” | Tổng số dòng |
| Nhỏ | 0 | 0 | 0 |
| Vừa | 1 | 1 | 2 |
| To | 2 | 0 | 2 |

Vậy ở nhánh “Búp bê” ta chọn thuộc tính “Số màu” làm phép chia nhánh. Với giá trị “Số màu” = 3, ta luôn có phân lớp Thấp; “Số màu” = 5 hoặc “Số màu” = 7 thì ta luôn có phân lớp cao. Vì vậy nhánh này đi đến nút lá và không cần xét tiếp.

Trở lại nhánh “Cao su” được phân chia từ nút gốc gồm 6 dòng dữ liệu, xét lần lượt:

***+ Thuộc tính “Loại” – Chất liệu = “Cao su”:***

|  |  |  |  |
| --- | --- | --- | --- |
| Giá trị phân biệt | Số dòng gán nhãn “Cao” | Số dòng gán nhãn “Thấp” | Tổng số dòng |
| Điều khiển | 0 | 3 | 3 |
| Xếp hình | 1 | 1 | 2 |
| Búp bê | 0 | 1 | 1 |

***+ Thuộc tính “Số màu” – Chất liệu = “Cao su” :***

|  |  |  |  |
| --- | --- | --- | --- |
| Giá trị phân biệt | Số dòng gán nhãn “Cao” | Số dòng gán nhãn “Thấp” | Tổng số dòng |
| 3 | 0 | 1 | 1 |
| 5 | 0 | 3 | 3 |
| 7 | 1 | 1 | 2 |

***+ Thuộc tính “Kích thước” – Chất liệu = “Cao su”:***

|  |  |  |  |
| --- | --- | --- | --- |
| Giá trị phân biệt | Số dòng gán nhãn “Cao” | Số dòng gán nhãn “Thấp” | Tổng số dòng |
| Nhỏ | 0 | 2 | 2 |
| Vừa | 0 | 2 | 2 |
| To | 1 | 1 | 2 |

Với giá trị Gini bằng nhau, ta chọn ngẫu nhiên một thuộc tính, ở đây ta cho cây sẽ tiếp tục được phân nhánh bằng thuộc tính “Loại” có 3 giá trị “Điều khiển”, “Xếp hình” và “Búp bê”. Đến đây, nhánh Điều khiển và nhánh Búp bê sẽ cho nút lá thuộc phân lớp thấp. Ta chỉ xem xét phát triển cây với giá trị Xếp hình của thuộc tính.

Với nhánh tương ứng với giá trị “Xếp hình”, gồm 2 dòng dữ liệu, xét lần lượt:

***+ Thuộc tính “Kích thước” – Loại = “Xếp hình” và Chất liệu = “Cao su”:***

|  |  |  |  |
| --- | --- | --- | --- |
| Giá trị phân biệt | Số dòng gán nhãn “Cao” | Số dòng gán nhãn “Thấp” | Tổng số dòng |
| Nhỏ | 0 | 0 | 0 |
| Vừa | 0 | 1 | 1 |
| To | 1 | 0 | 1 |

***+ Thuộc tính “Số màu” – Loại = “Xếp hình” và Chất liệu = “Cao su”:***

|  |  |  |  |
| --- | --- | --- | --- |
| Giá trị phân biệt | Số dòng gán nhãn “Cao” | Số dòng gán nhãn “Thấp” | Tổng số dòng |
| 3 | 0 | 0 | 0 |
| 5 | 0 | 1 | 1 |
| 7 | 1 | 0 | 1 |

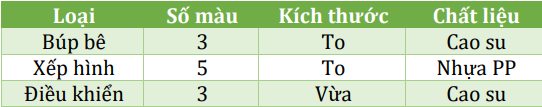
Với giá trị độ lợi thông tin bằng nhau, ta chọn ngẫu nhiên một thuộc tính, ở đây ta cho cây sẽ tiếp tục được phân nhánh bằng thuộc tính “Số màu”. Nhánh 5 sẽ cho nút lá thuộc phân lớp thấp và nhánh 7 sẽ cho nút lá thuộc phân lớp cao.

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

A picture containing diagram, plan, text, design

Description automatically generated

## d) 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 CART xây dựng được ở câu c, 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 Rủi ro thấp là lớp dương, lớp còn lại – Rủi ro 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 |  | Thấp | Cao |
| Thấp | 2 | 0 |
| Cao | 1 | 0 |

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

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

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

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

## 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.

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

Với hồ sơ đầu tiên:

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

Áp dụng làm trơn Laplace, ta có:

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.

Với hồ sơ thứ 2:

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

Áp dụng làm trơn Laplace, ta có:

Như vậy hồ sơ thứ 2 có xác suất xảy ra Doanh số bán cao lớn hơn, vậy ta có thể kết luận dòng dữ liệu đầu tiên được dự đoán thuộc phân lớp Doanh số bán = Cao.

Với hồ sơ thứ 3:

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

Áp dụng làm trơn Laplace, ta có:

Như vậy hồ sơ thứ 3 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.

Kết quả dự đoán:

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

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

Doanh số thực tế : Thấp – Thấp – Cao

Chọn Rủi ro thấp là lớp dương, lớp còn lại – Rủi ro 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 |  | Thấp | Cao |
| Thấp | 1 | 1 |
| Cao | 1 | 0 |

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

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

Từ ma trận nhầm lẫn và giá trị của độ chính xác, độ phủ trên dữ liệu doanh số thực tế, ta có thể kết luận được mô hình được xây dựng bởi thuật toán Naïve Bayes có độ chính xác thấp hơn mô hình cây quyết định theo thuật toán CART.

## l) Dự đoán Doanh số bán:

Vì thuật toán CART cho độ chính xác tốt, nên ta chọn thuật toán này để dự đoán:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Loại | Số màu | Kích Thước | Chất liệu | Doanh số bán |
| Búp bê | 3 | To | Cao su | Thấp |

# Câu 2: Bảng dữ liệu phân tích cảm xúc:

A picture containing text, screenshot, number, font

Description automatically generated

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

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

***+ 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 |

***+ 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 |

***+ 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 |

***+ 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 |

***+ 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 |

***+ 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 |

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 |

***+ 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 |

***+ 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 |

***+ 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 |

***+ 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 |

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 |

***+ 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 |

***+ 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 |

***+ 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 |

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 |

***+ 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 |

***+ 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 |

***+ 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 |

***+ 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 |

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 |

***+ 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 |

***+ 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 |

***+ 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 |

***+ 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 |

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 |

***+ 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 |

***+ 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 |

***+ 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 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 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 |

***+ 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 |

***+ 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 |

***+ 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 |

***+ 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 |

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:

A picture containing text, diagram, screenshot, line

Description automatically generated

## 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:

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:

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:

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:

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 |

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 |

**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?

- Việc nắm bắt được cảm xúc người dùng mạng xã hội giúp ta ngăn chặn và giải quyết các cảm xúc tiêu cực, giảm thiểu bạo lực mạng và vấn nạn trầm cảm của giới trẻ hiện nay.

- Ngoài ra, nắm bắt cảm xúc còn giúp ta tìm ra vấn đề đang được quan tâm của xã hội, bắt kịp được xu hướng để triển khai hoạt động marketing trong tương lai.