TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN
KHOA HỆ THỐNG THÔNG TIN

TÀI LIỆU HƯỚNG DẪN THỰC HÀNH
Môn: Khai thác dữ liệu

BÀI THỰC HÀNH SỐ 4

I. Mục tiêu

1. Hiểu và vận dụng giải thuật cây quyết định để phân lớp dữ liệu.
2. Hiểu và vận dụng giải thuật Naïve Bayes để phân lớp dữ liệu.
3. Hiểu các phương pháp đánh giá kết quả phân lớp tìm được.

II. Thời gian

1. Thực hành: 4 tiết
2. Bài tập làm thêm: 8 tiết

III. Hướng dẫn chung

1. Yêu cầu cơ bản

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

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

Trong đó cột dữ liệu *Nguy cơ* là thuộc tính quyết định. Hãy thực hiện những yêu cầu sau:

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

2. 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.
3. 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.
4. Sử dụng một trong hai cây quyết định ở trên để tiên đoán giá trị *Nguy cơ* của những hồ sơ sau:

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

5. Nguy cơ trên thực tế của các hồ sơ ở Yêu cầu 4 lần lượt là *Rủi ro cao, Rủi ro thấp, Rủi ro thấp*. Hãy lập ma trận nhầm lẫn (confusion matrix), sau đó tính giá trị độ chính xác (precision), độ phủ (recall) của mô hình/cây đã xây dựng.
6. Xác suất không điều kiện của giá trị 'Truyền thông' trong tập dữ liệu là bao nhiêu?
7. Khi nguy cơ là 'Rủi ro cao', hãy tính xác suất đó là những người có thu nhập 'Thấp'.
8. Dựa theo định lý Bayes, hãy viết công thức tính xác suất Nguy cơ 'Rủi ro cao' của những người đã sở hữu nhà.
9. Sử dụng thuật toán Naïve Bayes và làm trơn Laplace để dự đoán giá trị *Nguy cơ* của những hồ sơ trong Yêu cầu 4.
10. Với kết quả thu được và nguy cơ thực tế (Yêu cầu 5), hãy lập ma trận nhầm lẫn (confusion matrix), sau đó tính giá trị độ chính xác (precision), độ phủ (recall) của thuật toán.
11. So sánh kết quả từ thuật toán cây quyết định và Naïve Bayes.

Hướng dẫn:

1. Sau khi tiền xử lý dữ liệu cho phù hợp với thuật toán phân lớp, người ta thường rà soát lại để tìm ra những mâu thuẫn. 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 bài tập trên không xảy ra trường hợp này.
2. 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:

$$G(S,A) = E(S) - \sum_{i=1}^{m} f_s(A_i) E\left(S_{A_i}\right)$$

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:

$$E(S) = -\sum_{j=1}^{n} f_s(A_j) \log_2 f_s(A_j)$$

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

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

$f_s(A_i)$ là tỷ lệ của số mẫu có thuộc tính A_i với S

S_{A_i} là một tập con của S chứa tất cả các mẫu có giá trị A_i

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

$$E(S) = -\frac{9}{15}\log_2\frac{9}{15} - \frac{6}{15}\log_2\frac{6}{15} \approx 0{,}97$$

Tính độ lợi thông tin ở lần lượt từng thuộc tính, với thuộc tính Thu nhập ta có:

- Thuộc tính Thu nhập có 3 giá trị phân biệt là: Thấp, Trung bình, Cao.
- Giá trị Thấp có 5 dòng trong đó 3 dòng được gán nhãn Rủi ro cao, 2 dòng được gán nhãn Rủi ro thấp.
- Giá trị Trung bình có 5 dòng trong đó 2 dòng được gán nhãn Rủi ro cao, 3 dòng được gán nhãn Rủi ro thấp.
- Giá trị Cao có 5 dòng trong đó 1 dòng được gán nhãn Rủi ro cao, 4 dòng được gán nhãn Rủi ro thấp.

Như vậy, ta tính được:

$$E\left(S_{Thấp}\right) = -\frac{3}{5}\log_2\frac{3}{5} - \frac{2}{5}\log_2\frac{2}{5} \approx 0{,}971$$

$$E\left(S_{Trung\ bình}\right) = -\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5} \approx 0{,}971$$

$$E(S_{Cao}) = -\frac{1}{5}\log_2\frac{1}{5} - \frac{4}{5}\log_2\frac{4}{5} \approx 0{,}722$$

$$G(S, \text{Thu nhập}) \approx 0{,}97 - \frac{5}{15} \times 0{,}971 - \frac{5}{15} \times 0{,}971 - \frac{5}{15} \times 0{,}722 \approx 0{,}083$$

Xét thuộc tính Nghề nghiệp ta có:

- Thuộc tính Nghề nghiệp có 3 giá trị phân biệt là: Vận tải, Truyền thông, Kinh doanh.
- Giá trị Vận tải có 4 dòng trong đó 2 dòng được gán nhãn Rủi ro cao, 2 dòng được gán nhãn Rủi ro thấp.
- Giá trị Truyền thông có 6 dòng trong đó 2 dòng được gán nhãn Rủi ro cao, 4 dòng được gán nhãn Rủi ro thấp.
- Giá trị Kinh doanh có 5 dòng trong đó 2 dòng được gán nhãn Rủi ro cao, 3 dòng được gán nhãn Rủi ro thấp.

Như vậy, ta tính được:

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

$$E(S_{Truyền\ thông}) = -\frac{2}{6}\log_2\frac{2}{6} - \frac{4}{6}\log_2\frac{4}{6} \approx 0{,}918$$

$$E(S_{Kinh\ doanh}) = -\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5} \approx 0{,}971$$

$$G(S, \text{Nghề nghiệp}) \approx 0{,}97 - \frac{4}{15} \times 1 - \frac{6}{15} \times 0{,}918 - \frac{5}{15} \times 0{,}971 \approx 0{,}013$$

Tiếp tục xét thuộc tính Tình trạng hôn nhân ta có:

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

Như vậy, ta tính được:

$$E(S_{\text{Độc thân}}) = -\frac{5}{8}\log_2\frac{5}{8} - \frac{3}{8}\log_2\frac{3}{8} \approx 0{,}954$$

$$E(S_{\text{Đã kết hôn}}) = -\frac{1}{7}\log_2\frac{1}{7} - \frac{6}{7}\log_2\frac{6}{7} \approx 0{,}6$$

$$G(S, \text{Tình trạng hôn nhân}) \approx 0{,}97 - \frac{8}{15} \times 0{,}954 - \frac{7}{15} \times 0{,}6 \approx 0{,}186$$

Cuối cùng, ta xét thuộc tính Sở hữu nhà:

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

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

Như vậy, ta tính được:

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

$$E(S_{\text{Rồi}}) = -\frac{3}{9}\log_2\frac{3}{9} - \frac{6}{9}\log_2\frac{6}{9} \approx 0{,}918$$

$$G(S, \text{Sở hữu nhà}) \approx 0{,}97 - \frac{6}{15} \times 1 - \frac{9}{15} \times 0{,}918 \approx 0{,}02$$

Trong 4 thuộc tính đã xem xét, Tình trạng hôn nhân 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 Tình trạng hôn nhân. Phần có giá trị Độc thân gồm 8 dòng, phần có giá trị Đã kết hôn gồm 7 dòng.

Với nhánh cây thứ nhất, nhánh Độc thân, xét thuộc tính Thu nhập, tính toán tương tự ta có:

$$E(S_{Thấp}) = 0;\ E(S_{Trung\ bình}) = 1;\ E(S_{Cao}) \approx 0{,}92;\ G(S_{\text{Độc thân}}, \text{Thu nhập}) \approx 0{,}36$$

Với hai thuộc tính Nghề nghiệp và Sở hữu nhà, ta có:

$$E(S_{Vận\ tải}) \approx 0{,}92;\ E(S_{Truyền\ thông}) = 1;\ E(S_{Kinh\ doanh}) = 0;$$

$$G(S_{\text{Độc thân}}, \text{Nghề nghiệp}) \approx 0{,}11$$

$$E(S_{Chưa}) \approx 0{,}92;\ E(S_{Rồi}) \approx 0{,}971;\ G(S_{\text{Độc thân}}, \text{Sở hữu nhà}) \approx 0{,}003$$

Vậy ở nhánh này ta chọn Thuộc tính Thu nhập làm phép chia nhánh. Với giá trị Thu nhập bằng Thấp, ta luôn có phân lớp Rủi ro 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 Trung bình và Cao sẽ tiếp tục được phát triển.

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

$$E(S_{Vận\ tải}) = 0;\ E(S_{Truyền\ thông}) = 0;\ E(S_{Kinh\ doanh}) = 0;$$

$$G(S_{Cao}, \text{Nghề nghiệp}) \approx 0{,}92$$

$$E(S_{Chưa}) = 0;\ E(S_{Rồi}) = 1;\ G(S_{Cao}, \text{Sở hữu nhà}) \approx 0{,}252$$

Với giá trị độ lợi thông tin lớn hơn, cây sẽ tiếp tục được phân nhánh bằng thuộc tính Nghề nghiệp. Đến đây, hai nhánh Vận tải và Truyền thông sẽ đi đến nút lá Rủi ro thấp, nhánh Kinh doanh sẽ đi đến nút lá Rủi ro cao.

Nhánh tương ứng với giá trị Trung bình gồm 2 dòng dữ liệu, tiếp tục tính độ lợi thông tin ta có:

$$E(S_{Truyền\ thông}) = 1;\ G(S_{Trung\ bình}, \text{Nghề nghiệp}) = 0$$

$$E(S_{Chưa}) = 0;\ E(S_{Rồi}) = 0;\ G(S_{Trung\ bình}, \text{Sở hữu nhà}) = 1$$

Nhánh này sẽ được phân chia bằng thuộc tính Sở hữu nhà, nhánh con Chưa sẽ đi đến nút lá Rủi ro cao, nhánh con Rồi sẽ đi đến nút là Rủi ro thấp.

Trở lại với nhánh Đã kết hôn, được phân chia từ nút gốc, ta có 7 dòng dữ liệu. Xét các thuộc tính Thu nhập, Nghề nghiệp và Sở hữu nhà ta được kết quả sau:

$$E(S_{Thấp}) = 0;\ E(S_{Trung\ bình}) \approx 0{,}92;\ E(S_{Cao}) = 0;\ G(S_{\text{Đã kết hôn}}, \text{Thu nhập}) \approx 0{,}2$$

$$E(S_{Vận\ tải}) = 0;\ E(S_{Truyền\ thông}) = 0;\ E(S_{Kinh\ doanh}) \approx 0{,}811;$$

$$G(S_{\text{Đã kết hôn}}, \text{Nghề nghiệp}) \approx 0{,}128$$

$$E(S_{Chưa}) \approx 0{,}92;\ E(S_{Rồi}) = 0;\ G(S_{\text{Đã kết hôn}}, \text{Sở hữu nhà}) \approx 0{,}2$$

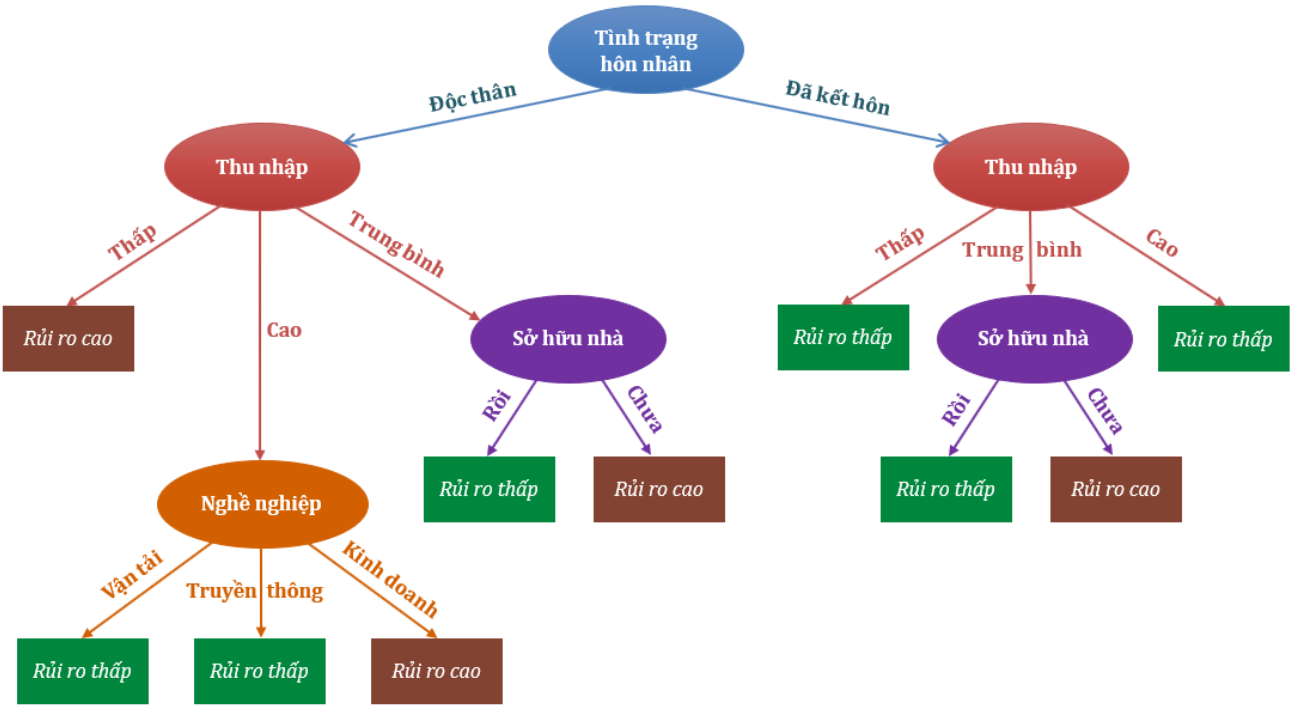
Ở nhánh này có hai thuộc tính có độ lợi thông tin cao bằng nhau đó là: Thu nhập và Sở hữu nhà. Ta chọn Thu nhập để tiếp tục, với giá trị Thấp và giá trị Cao, cây sẽ phát triển đến nút lá Rủi ro thấp. Chúng ta chỉ việc xem xét phát triển cây với giá trị Trung bình, ta có:

$$E(S_{Truyền\ thông}) = 0;\ E(S_{Kinh\ doanh}) = 1;\ G(S_{Trung\ bình}, \text{Nghề nghiệp}) \approx 0{,}252$$

$$E(S_{Chưa}) = 0;\ E(S_{Rồi}) = 0;\ G(S_{Trung\ bình}, \text{Sở hữu nhà}) \approx 0{,}92$$

Chọn Sở hữu nhà làm thuộc tình phân nhánh, lúc này với giá trị Chưa ta luôn có Rủi ro cao, với giá trị Rồi ta luôn có Rủi ro thấp, như vậy thuật toán kết thúc.

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



3. Chỉ số Gini dùng để đánh giá thuộc tính phân nhánh được tính theo công thức sau.

Chỉ số Gini của tập huấn luyện S:

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

Với $p(j|S)$ là tần suất của lớp j trong S.

Khi phân chia nút A thành k nhánh, chất lượng của phép chia được tính bằng công thức:

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

Trong đó:

n_i là số mẫu trong nút i

n là số mẫu trong nút A

Theo những thống kê từ câu 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:

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

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

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

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

Tiếp tục xét thuộc tính Nghề nghiệp, ta có:

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

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

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

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

Với thuộc tính Tình trạng hôn nhân, ta có:

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

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

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

Với thuộc tính Sở hữu nhà, ta có:

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

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

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

Chọn thuộc tính có chỉ số Gini thấp nhất là Tình trạng hôn nhân và phân nhánh theo đó.

Tương tự như trên, khi xét xuống nhánh Độc thân ta tính được giá trị chỉ số Gini như sau:

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

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

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

Chọn thuộc tính Thu nhập và tiếp tục phát triển cây, ta tính được các chỉ số Gini sau:

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

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

Đối với nhánh Cao ta chọn thuộc tính Nghề nghiệp để phân nhánh và đều đi đến các nút lá. Với nhánh Trung bình ta có:

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

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

Ở nhánh này ta chọn thuộc tính Sở hữu nhà làm thuộc tính phân nhánh và cây cũng đi đến các nút lá. Xét tiếp tục nhánh Đã kết hôn ta có:

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

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

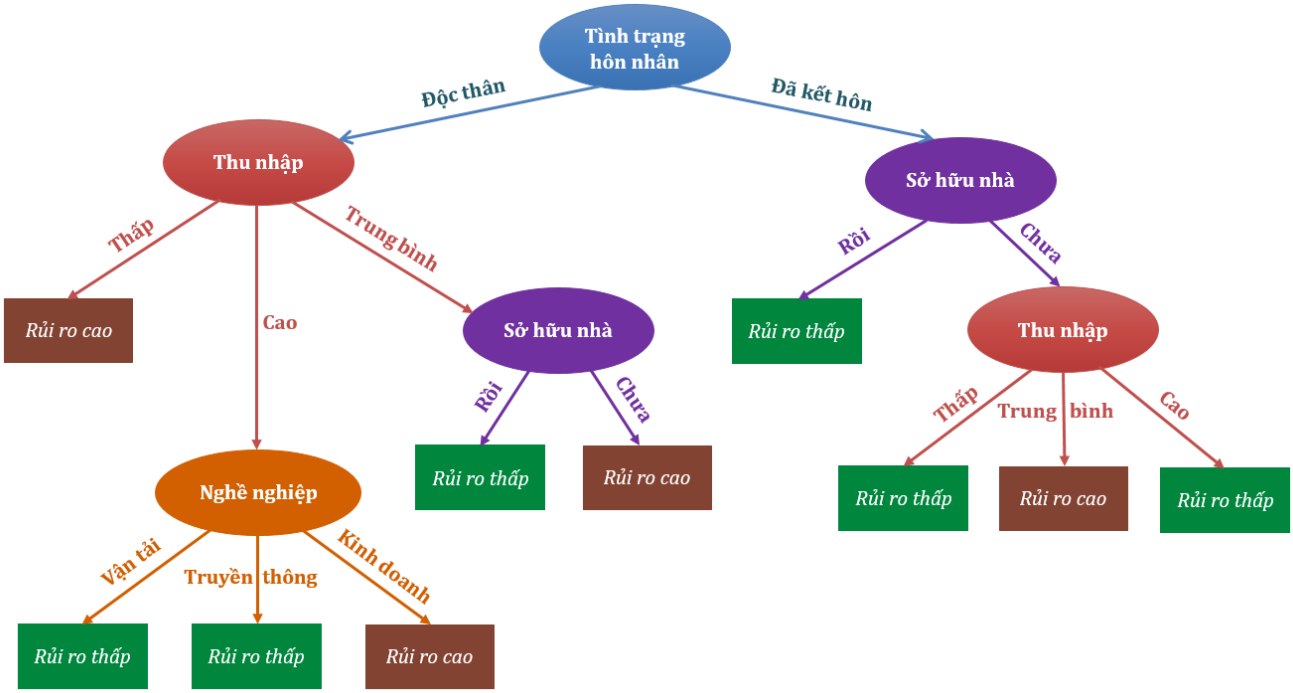
$$Gini_{Sở\ hữu\ nhà}(S_{\text{Đã kết hôn}}) = \frac{3}{7} \times 0{,}444 + \frac{4}{7} \times 0 \approx 0{,}19$$

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

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

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

Chọn thuộc tính Thu nhập để tiếp tục phân nhánh, đến đây các nhánh đều đi đến nút lá, kết quả ta thu được cây như sau:



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

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

5. Đầu tiên, ta hãy làm quen với ma trận nhầm lẫn (confusion matrix)

		Lớp dự đoán được từ mô hình	
		Lớp dương	*Lớp âm*
Lớp trên thực tế	*Lớp dương*	a	b
	Lớp âm	c	d

Ma trận nhầm lẫn - Confusion matrix

Quan sát ma trận nhầm lẫn, ta có các thông tin sau:

- **a**: *TP (true positive)* – mẫu mang nhãn dương được phân lớp **đúng** vào lớp **dương**.
- **b**: *FN (false negative)* – mẫu mang nhãn dương bị phân lớp **sai** vào lớp **âm**.
- **c**: *FP (false positive)* – mẫu mang nhãn âm bị phân lớp **sai** vào lớp **dương**.
- **d**: *TN (true negative)* – mẫu mang nhãn âm được phân lớp **đúng** vào lớp **âm**.

Từ đây, độ chính xác (precision), độ phủ (recall) của mô hình M được tính như sau:

$$precision(M) = \frac{a}{a + c}$$

$$recall(M) = \frac{a}{a+b}$$

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		*Rủi ro thấp*	*Rủi ro cao*
	Rủi ro thấp	2	0
	Rủi ro cao	1	0

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

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

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

6. Xác suất không điều kiện của giá trị 'Truyền thông' trong tập dữ liệu là:

$$p(\text{Nghề nghiệp} = \text{Truyền thông}) = \frac{6}{15} = 0{,}4$$

7. Xác suất hồ sơ là người thu nhập 'Thấp' khi có nguy cơ 'Rủi ro cao' xảy ra là:

$$p(\text{Thu nhập} = \text{Thấp}|\text{Nguy cơ} = \text{Rủi ro cao}) = \frac{3}{6} = 0{,}5$$

8. Dựa theo định lý Bayes, công thức tính xác suất Nguy cơ 'Rủi ro cao' của những người đã sở hữu nhà là:

$$p(\text{Nguy cơ} = \text{Rủi ro cao}|\text{Sở hữu nhà} = \text{Rồi})$$
$$= \frac{p(\text{Sở hữu nhà} = \text{Rồi}|\text{Nguy cơ} = \text{Rủi ro cao}) \times p(\text{Nguy cơ} = \text{Rủi ro cao})}{p(\text{Sở hữu nhà} = \text{Rồi})}$$

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

   Với hồ sơ đầu tiên

$$X = \{\text{Thu nhập} = \text{Trung bình}, \text{Nghề nghiệp} = \text{Vận tải}, \text{Tình trạng hôn nhân} = \text{Độc thân}, \text{Sở hữu nhà} = \text{Rồi}\}$$

   Ta cần tính được

$$p(\text{Nguy cơ} = \text{Rủi ro cao}|X)$$
$$= \frac{p(X|\text{Nguy cơ} = \text{Rủi ro cao}) \times p(\text{Nguy cơ} = \text{Rủi ro cao})}{p(X)}$$

   so sánh với

$$p(\text{Nguy cơ} = \text{Rủi ro thấp}|X)$$

$$= \frac{p(X|\text{Nguy cơ} = \text{Rủi ro thấp}) \times p(\text{Nguy cơ} = \text{Rủi ro thấp})}{p(X)}$$

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:

$$p(\text{Nguy cơ} = \text{Rủi ro cao}) = \frac{6+1}{15+2} = \frac{7}{17} \approx 0{,}412$$

$$p(\text{Thu nhập} = \text{Trung bình}|\text{Nguy cơ} = \text{Rủi ro cao}) = \frac{2+1}{6+3} = \frac{1}{3} \approx 0{,}333$$

$$p(\text{Nghề nghiệp} = \text{Vận tải}|\text{Nguy cơ} = \text{Rủi ro cao}) = \frac{2+1}{6+3} = \frac{1}{3} \approx 0{,}333$$

$$p(\text{Tình trạng hôn nhân} = \text{Độc thân}|\text{Nguy cơ} = \text{Rủi ro cao}) = \frac{5+1}{6+2} = \frac{3}{4} = 0{,}75$$

$$p(\text{Sở hữu nhà} = \text{Rồi}|\text{Nguy cơ} = \text{Rủi ro cao}) = \frac{3+1}{6+2} = \frac{1}{2} = 0{,}5$$

$$p(X|\text{Nguy cơ} = \text{Rủi ro cao}) \times p(\text{Nguy cơ} = \text{Rủi ro cao})$$

$$= p(\text{Thu nhập} = \text{Trung bình}|\text{Nguy cơ} = \text{Rủi ro cao})$$

$$\times p(\text{Nghề nghiệp} = \text{Vận tải}|\text{Nguy cơ} = \text{Rủi ro cao})$$

$$\times p(\text{Tình trạng hôn nhân} = \text{Độc thân}|\text{Nguy cơ} = \text{Rủi ro cao})$$

$$\times p(\text{Sở hữu nhà} = \text{Rồi}|\text{Nguy cơ} = \text{Rủi ro cao})$$

$$\times p(\text{Nguy cơ} = \text{Rủi ro cao}) = \frac{1}{3} \times \frac{1}{3} \times 0{,}75 \times 0{,}5 \times \frac{7}{17} \approx 0{,}017$$

$$p(\text{Nguy cơ} = \text{Rủi ro thấp}) = \frac{9+1}{15+2} = \frac{10}{17} \approx 0{,}588$$

$$p(\text{Thu nhập} = \text{Trung bình}|\text{Nguy cơ} = \text{Rủi ro thấp}) = \frac{3+1}{9+3} = \frac{1}{3} \approx 0{,}333$$

$$p(\text{Nghề nghiệp} = \text{Vận tải}|\text{Nguy cơ} = \text{Rủi ro thấp}) = \frac{2+1}{9+3} = \frac{1}{4} = 0{,}25$$

$$p(\text{Tình trạng hôn nhân} = \text{Độc thân}|\text{Nguy cơ} = \text{Rủi ro thấp}) = \frac{3+1}{9+2} = \frac{4}{11} \approx 0{,}364$$

$$p(\text{Sở hữu nhà} = \text{Rồi}|\text{Nguy cơ} = \text{Rủi ro thấp}) = \frac{6+1}{9+2} = \frac{7}{11} \approx 0{,}636$$

$$p(X|\text{Nguy cơ} = \text{Rủi ro thấp}) \times p(\text{Nguy cơ} = \text{Rủi ro thấp})$$
$$= p(\text{Thu nhập} = \text{Trung bình}|\text{Nguy cơ} = \text{Rủi ro thấp})$$
$$\times p(\text{Nghề nghiệp} = \text{Vận tải}|\text{Nguy cơ} = \text{Rủi ro thấp})$$
$$\times p(\text{Tình trạng hôn nhân} = \text{Độc thân}|\text{Nguy cơ} = \text{Rủi ro thấp})$$
$$\times p(\text{Sở hữu nhà} = \text{Rồi}|\text{Nguy cơ} = \text{Rủi ro thấp})$$
$$\times p(\text{Nguy cơ} = \text{Rủi ro thấp}) = \frac{1}{3} \times 0{,}25 \times \frac{4}{11} \times \frac{7}{11} \times \frac{10}{17} \approx 0{,}011$$

Như vậy hồ sơ đầu tiên có xác suất xảy ra Nguy cơ Rủi ro cao lớn hơn, vậy ta có thể kết luận dòng dữ liệu đầu tiên được dự đoán thuộc phân lớp Rủi ro cao.

Xét hồ sơ thứ hai

$$X = \{\text{Thu nhập} = \text{Cao}, \text{Nghề nghiệp} = \text{Truyền thông}, \text{Tình trạng hôn nhân} = \text{Độc thân}, \text{Sở hữu nhà} = \text{Rồi}\}$$

Ta cũng có những tính toán sau:

$$p(\text{Thu nhập} = \text{Cao}|\text{Nguy cơ} = \text{Rủi ro cao}) = \frac{1+1}{6+3} = \frac{2}{9} \approx 0{,}222$$

$$p(\text{Nghề nghiệp} = \text{Truyền thông}|\text{Nguy cơ} = \text{Rủi ro cao}) = \frac{2+1}{6+3} = \frac{1}{3} \approx 0{,}333$$

$$p(\text{Tình trạng hôn nhân} = \text{Độc thân}|\text{Nguy cơ} = \text{Rủi ro cao}) = \frac{5+1}{6+2} = \frac{3}{4} = 0{,}75$$

$$p(\text{Sở hữu nhà} = \text{Rồi}|\text{Nguy cơ} = \text{Rủi ro cao}) = \frac{3+1}{6+2} = \frac{1}{2} = 0{,}5$$

$$p(X|\text{Nguy cơ} = \text{Rủi ro cao}) \times p(\text{Nguy cơ} = \text{Rủi ro cao})$$
$$= p(\text{Thu nhập} = \text{Cao}|\text{Nguy cơ} = \text{Rủi ro cao})$$
$$\times p(\text{Nghề nghiệp} = \text{Truyền thông}|\text{Nguy cơ} = \text{Rủi ro cao})$$
$$\times p(\text{Tình trạng hôn nhân} = \text{Độc thân}|\text{Nguy cơ} = \text{Rủi ro cao})$$
$$\times p(\text{Sở hữu nhà} = \text{Rồi}|\text{Nguy cơ} = \text{Rủi ro cao})$$
$$\times p(\text{Nguy cơ} = \text{Rủi ro cao}) = \frac{2}{9} \times \frac{1}{3} \times 0{,}75 \times 0{,}5 \times \frac{7}{17} \approx 0{,}011$$

$$p(\text{Thu nhập} = \text{Cao}|\text{Nguy cơ} = \text{Rủi ro thấp}) = \frac{4+1}{9+3} = \frac{5}{12} \approx 0{,}417$$

$$p(\text{Nghề nghiệp} = \text{Truyền thông}|\text{Nguy cơ} = \text{Rủi ro thấp}) = \frac{4+1}{9+3} = \frac{5}{12} = 0{,}417$$

$$p(\text{Tình trạng hôn nhân} = \text{Độc thân}|\text{Nguy cơ} = \text{Rủi ro thấp}) = \frac{3+1}{9+2} = \frac{4}{11} \approx 0{,}364$$

$$p(\text{Sở hữu nhà} = \text{Rồi}|\text{Nguy cơ} = \text{Rủi ro thấp}) = \frac{6+1}{9+2} = \frac{7}{11} \approx 0{,}636$$

$$p(X|\text{Nguy cơ} = \text{Rủi ro thấp}) \times p(\text{Nguy cơ} = \text{Rủi ro thấp})$$
$$= p(\text{Thu nhập} = \text{Cao}|\text{Nguy cơ} = \text{Rủi ro thấp})$$
$$\times p(\text{Nghề nghiệp} = \text{Truyền thông}|\text{Nguy cơ} = \text{Rủi ro thấp})$$
$$\times p(\text{Tình trạng hôn nhân} = \text{Độc thân}|\text{Nguy cơ} = \text{Rủi ro thấp})$$
$$\times p(\text{Sở hữu nhà} = \text{Rồi}|\text{Nguy cơ} = \text{Rủi ro thấp})$$
$$\times p(\text{Nguy cơ} = \text{Rủi ro thấp}) = \frac{5}{12} \times \frac{5}{12} \times \frac{4}{11} \times \frac{7}{11} \times \frac{10}{17} \approx 0{,}024$$

Như vậy hồ sơ thứ hai có xác suất xảy ra Nguy cơ Rủi ro thấp lớn hơn, vậy ta có thể kết luận dòng dữ liệu thứ hai được dự đoán thuộc phân lớp Rủi ro thấp.

Xét hồ sơ thứ ba:

$$\text{X} = \{\text{Thu nhập} = \text{Thấp}, \text{Nghề nghiệp} = \text{Vận tải}, \text{Tình trạng hôn nhân} = \text{Đã kết hôn}, \text{Sở hữu nhà} = \text{Rồi}\}$$

Ta cũng có những tính toán sau:

$$p(\text{Thu nhập} = \text{Thấp}|\text{Nguy cơ} = \text{Rủi ro cao}) = \frac{3+1}{6+3} = \frac{4}{9} \approx 0{,}444$$

$$p(\text{Nghề nghiệp} = \text{Vận tải}|\text{Nguy cơ} = \text{Rủi ro cao}) = \frac{2+1}{6+3} = \frac{1}{3} \approx 0{,}333$$

$$p(\text{Tình trạng hôn nhân} = \text{Đã kết hôn}|\text{Nguy cơ} = \text{Rủi ro cao}) = \frac{1+1}{6+2} = \frac{1}{4} = 0{,}25$$

$$p(\text{Sở hữu nhà} = \text{Rồi}|\text{Nguy cơ} = \text{Rủi ro cao}) = \frac{3+1}{6+2} = \frac{1}{2} = 0{,}5$$

$$p(X|\text{Nguy cơ} = \text{Rủi ro cao}) \times p(\text{Nguy cơ} = \text{Rủi ro cao})$$
$$= p(\text{Thu nhập} = \text{Thấp}|\text{Nguy cơ} = \text{Rủi ro cao})$$
$$\times p(\text{Nghề nghiệp} = \text{Vận tải}|\text{Nguy cơ} = \text{Rủi ro cao})$$
$$\times p(\text{Tình trạng hôn nhân} = \text{Đã kết hôn}|\text{Nguy cơ} = \text{Rủi ro cao})$$
$$\times p(\text{Sở hữu nhà} = \text{Rồi}|\text{Nguy cơ} = \text{Rủi ro cao})$$
$$\times p(\text{Nguy cơ} = \text{Rủi ro cao}) = \frac{4}{9} \times \frac{1}{3} \times 0{,}25 \times 0{,}5 \times \frac{7}{17} \approx 0{,}007$$

$$p(\text{Thu nhập} = \text{Thấp}|\text{Nguy cơ} = \text{Rủi ro thấp}) = \frac{2+1}{9+3} = \frac{1}{4} = 0{,}25$$

$$p(\text{Nghề nghiệp} = \text{Vận tải}|\text{Nguy cơ} = \text{Rủi ro thấp}) = \frac{2+1}{9+3} = \frac{1}{4} = 0{,}25$$

$$p(\text{Tình trạng hôn nhân} = \text{Đã kết hôn}|\text{Nguy cơ} = \text{Rủi ro thấp}) = \frac{6+1}{9+2} = \frac{7}{11}$$

$$\approx 0{,}636$$

$$p(\text{Sở hữu nhà} = \text{Rồi}|\text{Nguy cơ} = \text{Rủi ro thấp}) = \frac{6+1}{9+2} = \frac{7}{11} \approx 0{,}636$$

$$p(X|\text{Nguy cơ} = \text{Rủi ro thấp}) \times p(\text{Nguy cơ} = \text{Rủi ro thấp})$$

$$= p(\text{Thu nhập} = \text{Thấp}|\text{Nguy cơ} = \text{Rủi ro thấp})$$

$$\times p(\text{Nghề nghiệp} = \text{Vận tải}|\text{Nguy cơ} = \text{Rủi ro thấp})$$

$$\times p(\text{Tình trạng hôn nhân} = \text{Đã kết hôn}|\text{Nguy cơ} = \text{Rủi ro thấp})$$

$$\times p(\text{Sở hữu nhà} = \text{Rồi}|\text{Nguy cơ} = \text{Rủi ro thấp})$$

$$\times p(\text{Nguy cơ} = \text{Rủi ro thấp}) = \frac{1}{4} \times \frac{1}{4} \times \frac{7}{11} \times \frac{7}{11} \times \frac{10}{17} \approx 0{,}014$$

Như vậy dòng dữ liệu thứ ba cũng được dự đoán thuộc phân lớp Rủi ro thấp.

10. Chọn Rủi ro thấp là lớp dương, lớp còn lại – Rủi ro cao sẽ là lớp âm, ta có ma trận nhầm lẫn của thuật toán Naïve Bayes như sau:

		Lớp dự đoán được từ mô hình	
Lớp trên thực tế		*Rủi ro thấp*	*Rủi ro cao*
	Rủi ro thấp	2	0
	Rủi ro cao	0	1

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

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

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

11. Từ ma trận nhầm lẫn và giá trị của độ chính xác, độ phủ trên dữ liệu kiểm thử được cho ở câu 4, ta có thể kết luận được mô hình được xây dựng bởi thuật toán Naïve Bayes có độ chính xác cao hơn mô hình cây quyết định theo thuật toán CART.

2. Yêu cầu lập trình

Cho Adult Income[1] là dữ liệu về thu nhập của những người trẻ tuổi ở Hoa Kỳ với thuộc tính quyết định 'Income'. Thuộc tính này chứa hai giá trị là '>50K' (thu nhập lớn hơn

50.000 USD/năm) và '<=50K' (thu nhập bé hơn hoặc bằng 50.000 USD/năm). Sinh viên hãy thực hiện những yêu cầu sau đây:

1. Nhập dữ liệu đầu vào và cho biết số lượng dữ liệu huấn luyện (tương ứng với file adult.data.csv) và số lượng dữ liệu kiểm thử (tương ứng với file adult.test.csv)
2. Tiến hành tiền xử lý dữ liệu:
   - Xóa những dòng có chứa dữ liệu trống, biết rằng dữ liệu trống được ký hiệu bằng dấu '?'.
   - Xóa cột *final weight* 'fnlwgt' trong dữ liệu huấn luyện vì cột này không có trong dữ liệu kiểm thử.
   - Nối dữ liệu huấn luyện và kiểm thử lại với nhau để phục vụ các bước tiếp theo.
3. Khảo sát độ tương đồng giữa các cột với nhau bằng công thức Pearson. Những cột nào có giá trị tương đồng cao thì hãy loại bỏ.
4. Tách các cột dữ liệu thành hai phần, một phần chứa các thuộc tính bình thường, một phần chứa riêng thuộc tính quyết định.
5. Chuyển đổi các cột không phải dạng số về dạng one-hot vector để phù hợp với đầu vào của thư viện.
6. Tách các dòng dữ liệu ra thành hai phần huấn luyện và kiểm thử như ban đầu.
7. Xây dựng cây ID3 dựa trên dữ liệu huấn luyện và sau đó tiến hành kiểm thử kết quả của cây bằng ma trận nhầm lẫn. Biểu diễn cây vào trong kết quả thực hiện.
8. Lặp lại yêu cầu 7 đối với cây CART.
9. Xây dựng mô hình phân lớp bằng thuật toán Naïve Bayes và kiểm tra kết quả đạt được.
10. So sánh kết quả của các mô hình trên.

Hướng dẫn

Để thực hiện phân lớp dữ liệu trên ngôn ngữ Python, sinh viên cần cài đặt thư viện *sklearn*. Sau đó, thực hiện import các thư viện sau vào bài làm.

```
%matplotlib inline
import matplotlib.pyplot as plt
import numpy as np
import pandas as pd
import seaborn as sns

from sklearn.model_selection import train_test_split
from sklearn import tree
from sklearn.naive_bayes import GaussianNB
from sklearn import metrics
```

Sinh viên tự thực hiện yêu cầu 1, 2.

Số lượng dữ liệu huấn luyện và kiểm thử là

```
Number of training data: 30162
Number of test data: 15060
```

Sau khi tiền xử lý nối hai phần dữ liệu vào biến data sẽ nhận được khối dữ liệu có thông tin như sau

```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 45222 entries, 0 to 45221
Data columns (total 14 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   age             45222 non-null  int64
 1   workclass       45222 non-null  object
 2   education       45222 non-null  object
 3   education-num   45222 non-null  int64
 4   marital-status  45222 non-null  object
 5   occupation      45222 non-null  object
 6   relationship    45222 non-null  object
 7   race            45222 non-null  object
 8   sex             45222 non-null  object
 9   capital-gain    45222 non-null  int64
 10  capital-loss    45222 non-null  int64
 11  hours-per-week  45222 non-null  int64
 12  native-country  45222 non-null  object
 13  income          45222 non-null  object
dtypes: int64(5), object(9)
memory usage: 4.8+ MB
```

3. Khảo sát độ tương đồng giữa các cột với nhau bằng công thức Pearson.

   Kết hợp với thư viện biểu diễn đồ thị seaborn để trực quan hóa kết quả bằng đồ thị *heatmap*.

```
plt.figure(figsize=(16,9))
sns.heatmap(data.corr(method='pearson'), annot=True)
```

Lưu ý, biến *annot=True* giúp hiển thị các giá trị độ tương đồng trong đồ thị kết quả.

Trong đồ thị này, màu càng đậm chứng tỏ mức độ tương đồng càng thấp.

Nhận thấy, không có thuộc tính nào tương đồng cao với nhau nên không cần thiết phải loại bỏ cột nào.

4. Tách các cột dữ liệu thành hai phần, một phần chứa các thuộc tính bình thường, một phần chứa riêng thuộc tính quyết định.

   Thực hiện việc tách các cột dữ liệu vào hai biến *features* (chứa các thuộc tính bình thường) và biến *labels* (chứa riêng thuộc tính quyết định 'Income') bằng câu lệnh sau.

```
features = data.drop('income', axis=1)
labels = data['income']
```

   Với *data* là *DataFrame* chứa toàn bộ dữ liệu Adult Income đã gộp lại ở yêu cầu tiền xử lý.

5. Chuyển đổi các cột không phải dạng số về dạng one-hot vector để phù hợp với đầu vào của thư viện.

   Xác định các thuộc tính không phải dạng số trong biến *features* bằng câu lệnh sau.

```
features.select_dtypes(exclude=['int64']).columns
```

```
Index(['workclass', 'education', 'marital-status', 'occupation',
       'relationship', 'race', 'sex', 'native-country'],
      dtype='object')
```

Kết hợp câu lệnh *get_dummies* của thư viện *pandas* để chuyển đổi các cột này về dạng one-hot vector.

```
features_onehot = pd.get_dummies(features, columns=features.select_dtypes(exclude=['int64']).columns)
features_onehot
```

	age	education-num	capital-gain	capital-loss	hours-per-week	workclass_ Federal-gov	workclass_ Local-gov	workclass_ Private	workclass_ Self-emp-inc	workclass_ Self-emp-not-inc	...	native-country_ Portugal	native-country_ Puerto-Rico	native-country_ Scotland	native-country_ South	cc
0	39	13	2174	0	40	0	0	0	0	0	...	0	0	0	0	
1	50	13	0	0	13	0	0	0	0	1	...	0	0	0	0	
2	38	9	0	0	40	0	0	1	0	0	...	0	0	0	0	
3	53	7	0	0	40	0	0	1	0	0	...	0	0	0	0	
4	28	13	0	0	40	0	0	1	0	0	...	0	0	0	0	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
45217	33	13	0	0	40	0	0	1	0	0	...	0	0	0	0	
45218	39	13	0	0	36	0	0	1	0	0	...	0	0	0	0	
45219	38	13	0	0	50	0	0	1	0	0	...	0	0	0	0	
45220	44	13	5455	0	40	0	0	1	0	0	...	0	0	0	0	
45221	35	13	0	0	60	0	0	0	1	0	...	0	0	0	0	

45222 rows × 103 columns

6. Tách các dòng dữ liệu ra thành hai phần huấn luyện và kiểm thử như ban đầu.
   Tiến hành tách dữ liệu huấn luyện và kiểm thử như thông tin thu được ban đầu.

```
X_train = features_onehot[:30162]
X_test = features_onehot[30162:]
y_train = labels[:30162]
y_test = labels[30162:]
```

7. Xây dựng cây ID3 dựa trên dữ liệu huấn luyện và sau đó tiến hành kiểm thử kết quả của cây bằng ma trận nhầm lẫn. Biểu diễn cây vào trong kết quả thực hiện.
   Để xây dựng cây ID3, sinh viên thực hiện câu lệnh sau.

```
clf = tree.DecisionTreeClassifier(criterion="entropy", random_state=0)
#Train Decision Tree Classifer
clf.fit(X_train, y_train)
```

```
DecisionTreeClassifier(ccp_alpha=0.0, class_weight=None, criterion='entropy',
                       max_depth=None, max_features=None, max_leaf_nodes=None,
                       min_impurity_decrease=0.0, min_impurity_split=None,
                       min_samples_leaf=1, min_samples_split=2,
                       min_weight_fraction_leaf=0.0, presort='deprecated',
                       random_state=0, splitter='best')
```

Với biến *criterion='entropy'* để yêu cầu thư viện thực hiện phân nhánh theo *information gain*. Sinh viên có thể tham khảo thêm các cài đặt ở tài liệu của thư viện sklearn[4].

Sau khi xây dựng xong cây ID3, tiến hành áp dụng mô hình trên dữ liệu kiểm thử. Thư viện này cũng sẽ hỗ trợ việc tính toán các thông tin như độ chính xác, độ phủ của mô hình đã xây dựng.

```
#Predict the response for test dataset
tree_pred = clf.predict(X_test)

# Model Accuracy, how often is the classifier correct?
tree_score = metrics.accuracy_score(y_test, tree_pred)
print("Accuracy:", tree_score)
print("Report:", metrics.classification_report(y_test, tree_pred))
```

```
Accuracy: 0.8175298804780876
Report:               precision    recall  f1-score   support

       <=50K       0.88      0.88      0.88     11360
        >50K       0.63      0.62      0.62      3700

    accuracy                           0.82     15060
   macro avg       0.75      0.75      0.75     15060
weighted avg       0.82      0.82      0.82     15060
```

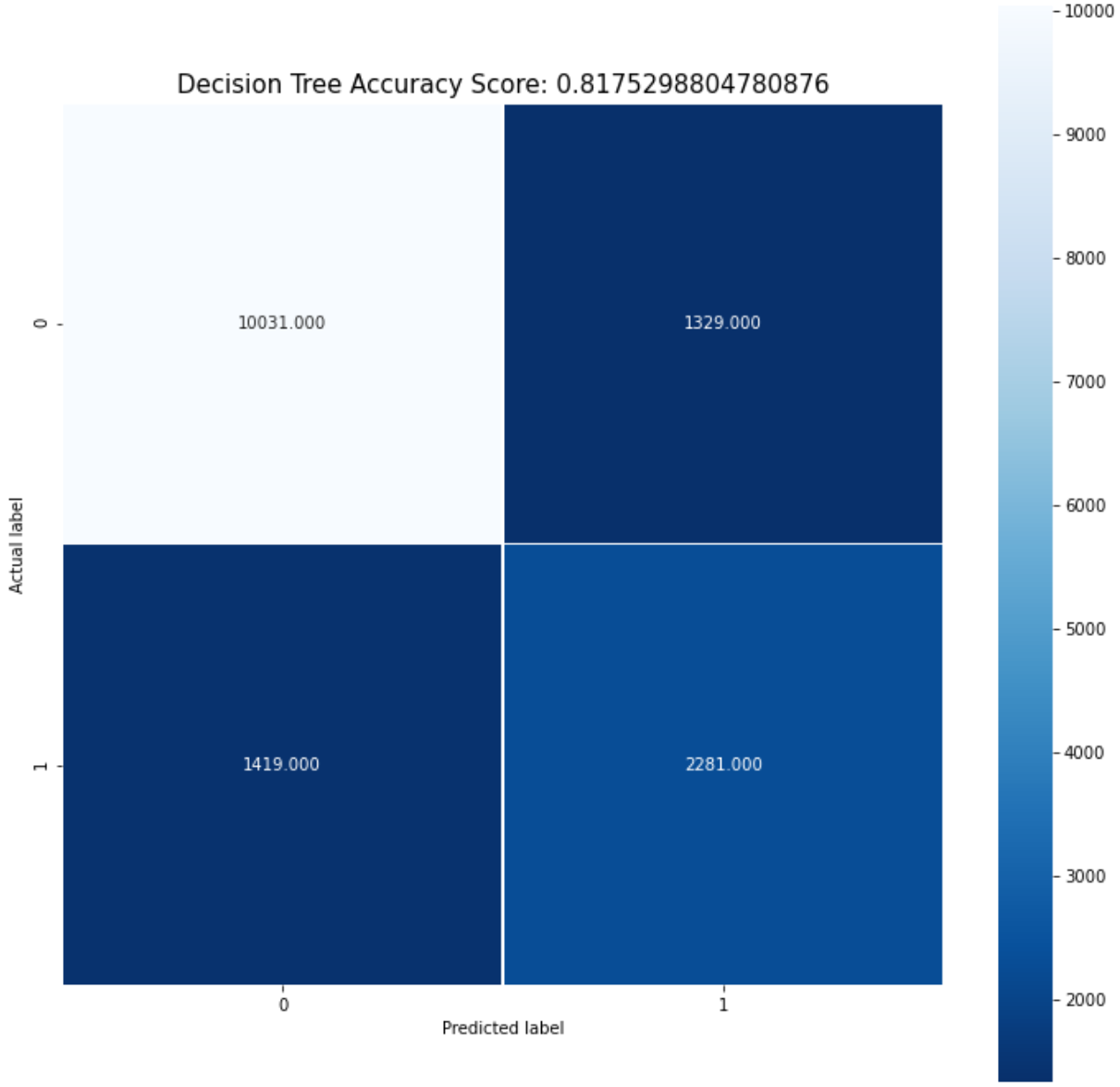
Tính toán ma trận nhầm lẫn

```
tree_cm = metrics.confusion_matrix(y_test, tree_pred)
```

Và biểu diễn nó lên đồ thị *heatmap*

```
plt.figure(figsize=(12,12))
sns.heatmap(tree_cm, annot=True, fmt=".3f", linewidths=.5, square = True, cmap = 'Blues_r');
plt.ylabel('Actual label');
plt.xlabel('Predicted label');
title = 'Decision Tree Accuracy Score: {0}'.format(tree_score)
plt.title(title, size = 15);
```

Đồ thị thu được sẽ có hình dạng như sau

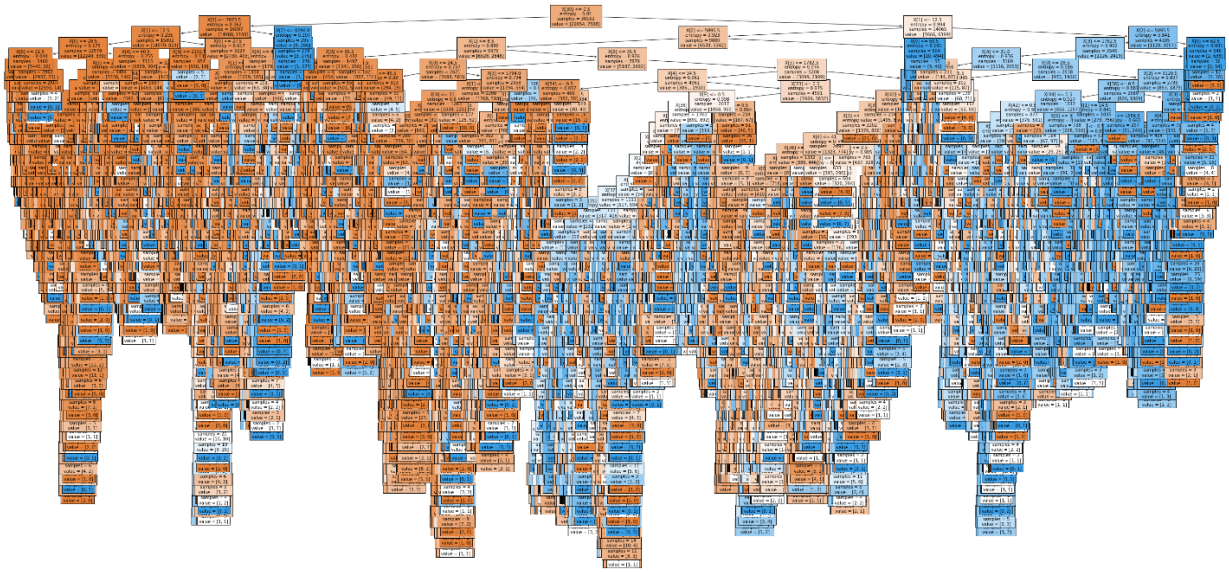


Sinh viên có thể biểu diễn cây ID3 bằng câu lệnh.

```
fig, ax = plt.subplots(figsize=(50, 24))
tree.plot_tree(clf, filled=True, fontsize=10)
plt.savefig('decision_tree', dpi=100)
plt.show()
```

Câu lệnh *savefig* với *dpi=100* giúp lưu lại cây vào file *'decision_tree.png'* với chất lượng tốt.

Cây thu được có hình dạng như sau



8. Sinh viên tự thực hiện yêu cầu này tương tự như trên. Lưu ý, thay thế giá trị *criterion='gini'*.
9. Xây dựng mô hình phân lớp bằng thuật toán Naïve Bayes và kiểm tra kết quả đạt được.

   Đối với thuật toán Naïve Bayes, sinh viên cũng làm tương tự với câu lệnh sau

```
gnb = GaussianNB()
bayes_pred = gnb.fit(X_train, y_train).predict(X_test)
```

   Tính toán các giá trị của mô hình

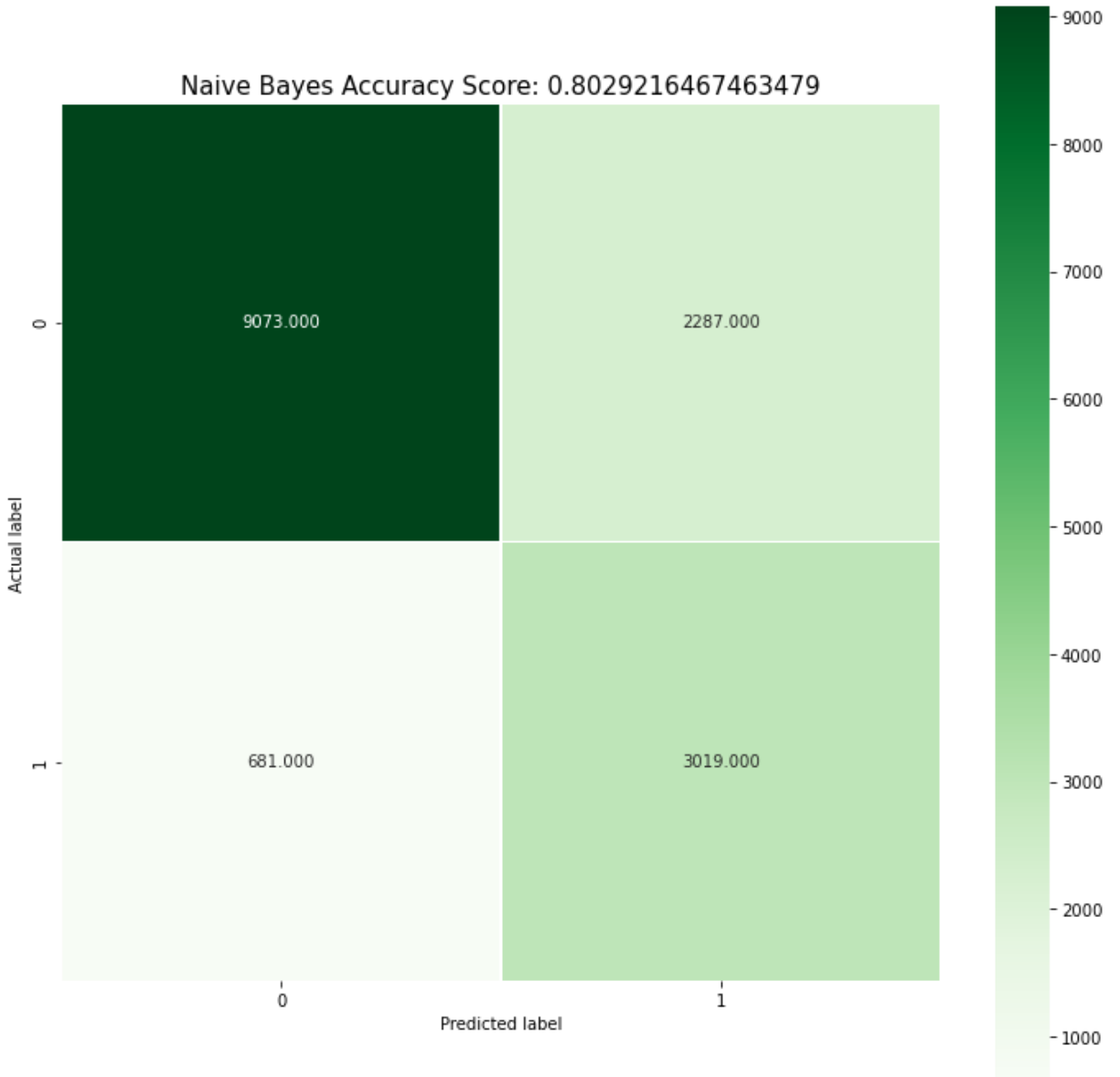
```
# Model Accuracy, how often is the classifier correct?
bayes_score = metrics.accuracy_score(y_test, bayes_pred)
print("Accuracy:", bayes_score)
print("Report:", metrics.classification_report(y_test, bayes_pred))
```

```
Accuracy: 0.8029216467463479
Report:               precision    recall  f1-score   support

       <=50K       0.93      0.80      0.86     11360
        >50K       0.57      0.82      0.67      3700

    accuracy                           0.80     15060
   macro avg       0.75      0.81      0.76     15060
weighted avg       0.84      0.80      0.81     15060
```

   Và biểu diễn ma trận nhầm lẫn bằng đồ thị *heatmap*.

```
bayes_cm = metrics.confusion_matrix(y_test, bayes_pred)

plt.figure(figsize=(12,12))
sns.heatmap(bayes_cm, annot=True, fmt=".3f", linewidths=.5, square = True, cmap = 'Greens');
plt.ylabel('Actual label');
plt.xlabel('Predicted label');
title = 'Naive Bayes Accuracy Score: {0}'.format(bayes_score)
plt.title(title, size = 15);
```

Đồ thị thu được có dạng



10. Dựa vào các thông tin đã thu được, sinh viên tự đưa ra nhận xét, so sánh kết quả của các mô hình trên.

IV. Thực hành

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.

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.

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.

d) Sử dụng một trong hai cây quyết định ở trên để tiên đoán giá trị *Doanh số bán* của những sản phẩm sau:

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

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.

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.

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.

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

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.

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

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
0..5	6..10	0..5	11..20	6..10	0..5
0..5	0..5	6..10	0..5	11..20	>20
6..10	0..5	11..20	>20	6..10	6..10
6..10	11..20	6..10	6..10	>20	0..5

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?

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

3. (Lập trình) Sinh viên hãy thực hiện lại bài tập trong phần *Yêu cầu lập trình* một cách hoàn chỉnh.
4. (Lập trình) Cho dữ liệu Red Wine Quality[2], liên quan đến các mẫu rượu vang Vinho Verde đỏ từ phía bắc Bồ Đào Nha. Mục tiêu của bài toán là mô hình hóa chất lượng rượu dựa trên các chỉ số hóa lý đo đạc được.

   Sử dụng câu lệnh sau để chia dữ liệu đầu vào thành hai phần huấn luyện 70% và kiểm thử 30%.

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

   Sinh viên hãy thực hiện lại từ yêu cầu 3 trở đi trong phần *Yêu cầu lập trình*, áp dụng với dữ liệu trên. Lưu ý, thuộc tính quyết định là *'quality'*.
5. (Lập trình) Cho dữ liệu Heart Disease[3] chứa 14 thuộc tính thu thập ở Cleveland, Mỹ. Thuộc tính *'target'* là số nguyên có giá trị 0 (không có sự hiện diện) hoặc 1 (có bệnh tim ở bệnh nhân).

   a) Nhập dữ liệu đầu vào và tiến hành tiền xử lý xóa các dòng dữ liệu trống, sửa tên các cột dữ liệu lại như sau để dễ theo dõi: *'age', 'sex', 'chest_pain_type', 'resting_blood_pressure', 'cholesterol', 'fasting_blood_sugar', 'rest_ecg', 'max_heart_rate_achieved', 'exercise_induced_angina', 'st_depression', 'st_slope', 'num_major_vessels', 'thalassemia', 'target'*.

b) Thực hiện lại từ yêu cầu 3 trở đi trong phần *Yêu cầu lập trình*, áp dụng với dữ liệu trên. Lưu ý, thuộc tính quyết định là *'target'*.

6. (Lập trình) Cho dữ liệu Mushroom[4], chứa các thông tin đặc điểm của nhiều loại nấm cùng với phân loại *'class'* là nấm độc *(poisonous=p)* hay ăn được *(edible=e)*.
   Sử dụng câu lệnh sau để chia dữ liệu đầu vào thành hai phần huấn luyện 70% và kiểm thử 30%.

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

   Sinh viên hãy thực hiện lại từ yêu cầu 4 trở đi trong phần *Yêu cầu lập trình*, áp dụng với dữ liệu trên. Lưu ý, thuộc tính quyết định là *'class'*.

7. (Lập trình) Xây dựng ứng dụng nhận dạng số viết tay dựa trên dữ liệu MNIST[5]. Đây là thư viện chữ số viết tay được gán nhãn sẵn và được hỗ trợ trực tiếp trong sklearn. Sinh viên có thể viết ứng dụng nhận dạng theo các bước sau.
   Chuẩn bị các thư viện cần thiết

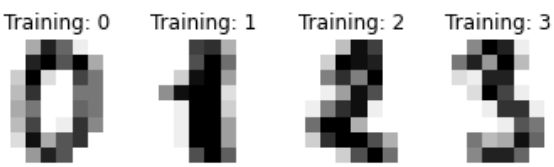
```
%matplotlib inline
import matplotlib
import matplotlib.pyplot as plt

from sklearn import datasets, tree, metrics
from sklearn.model_selection import train_test_split
```

   Download dữ liệu có sẵn trong thư viện sklearn

```
# The digits dataset
digits = datasets.load_digits()
```

   Biểu diễn 4 mẫu dữ liệu huấn luyện tương ứng với 4 lớp 0, 1, 2, 3

```
# The data that we are interested in is made of 8x8 images of digits, let's
# have a look at the first 4 images, stored in the `images` attribute of the
# dataset.  If we were working from image files, we could load them using
# matplotlib.pyplot.imread.  Note that each image must have the same size. For these
# images, we know which digit they represent: it is given in the 'target' of
# the dataset.
_, axes = plt.subplots(1, 4)
images_and_labels = list(zip(digits.images, digits.target))
for ax, (image, label) in zip(axes, images_and_labels[:4]):
    ax.set_axis_off()
    ax.imshow(image, cmap=plt.cm.gray_r, interpolation='nearest')
    ax.set_title('Training: %i' % label)
plt.show()
```



Chuyển đổi mỗi đối tượng hình ảnh – ma trận 8x8 thành ma trận 1x64 để thỏa yêu cầu đầu vào của thuật toán.

```
# To apply a classifier on this data, we need to flatten the image, to
# turn the data in a (samples, feature) matrix:
n_samples = len(digits.images)
data = digits.images.reshape((n_samples, -1))
```

Sử dụng thuật toán cây ID3

```
# Create a classifier: a decision tree classifier
classifier = tree.DecisionTreeClassifier(criterion="entropy", random_state=33)
```

Chia dữ liệu thành hai phần: huấn luyện 80%, kiểm thử 20%

```
# Split data into train and test subsets
X_train, X_test, y_train, y_test = train_test_split(
    data, digits.target, test_size=0.2, shuffle=False)
```

Tiến hành xây dựng cây ID3

```
# We learn the digits on the first part of the digits
classifier.fit(X_train, y_train)
```

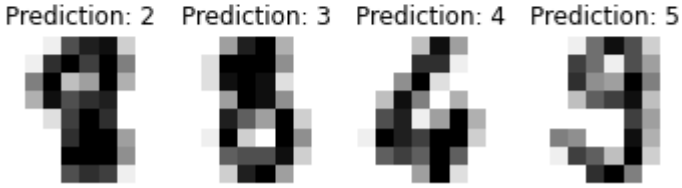
```
DecisionTreeClassifier(ccp_alpha=0.0, class_weight=None, criterion='entropy',
                       max_depth=None, max_features=None, max_leaf_nodes=None,
                       min_impurity_decrease=0.0, min_impurity_split=None,
                       min_samples_leaf=1, min_samples_split=2,
                       min_weight_fraction_leaf=0.0, presort='deprecated',
                       random_state=33, splitter='best')
```

Dự đoán nhãn dữ liệu cho tập kiểm thử

```
# Now predict the value of the digit on the second part:
predicted = classifier.predict(X_test)
```

Biểu diễn một vài kết quả dự đoán

```
_, axes = plt.subplots(1, 4)
images_and_predictions = list(zip(digits.images[n_samples // 2:], predicted))
for ax, (image, prediction) in zip(axes, images_and_predictions[:4]):
    ax.set_axis_off()
    ax.imshow(image, cmap=plt.cm.gray_r, interpolation='nearest')
    ax.set_title('Prediction: %i' % prediction)
plt.show()
```

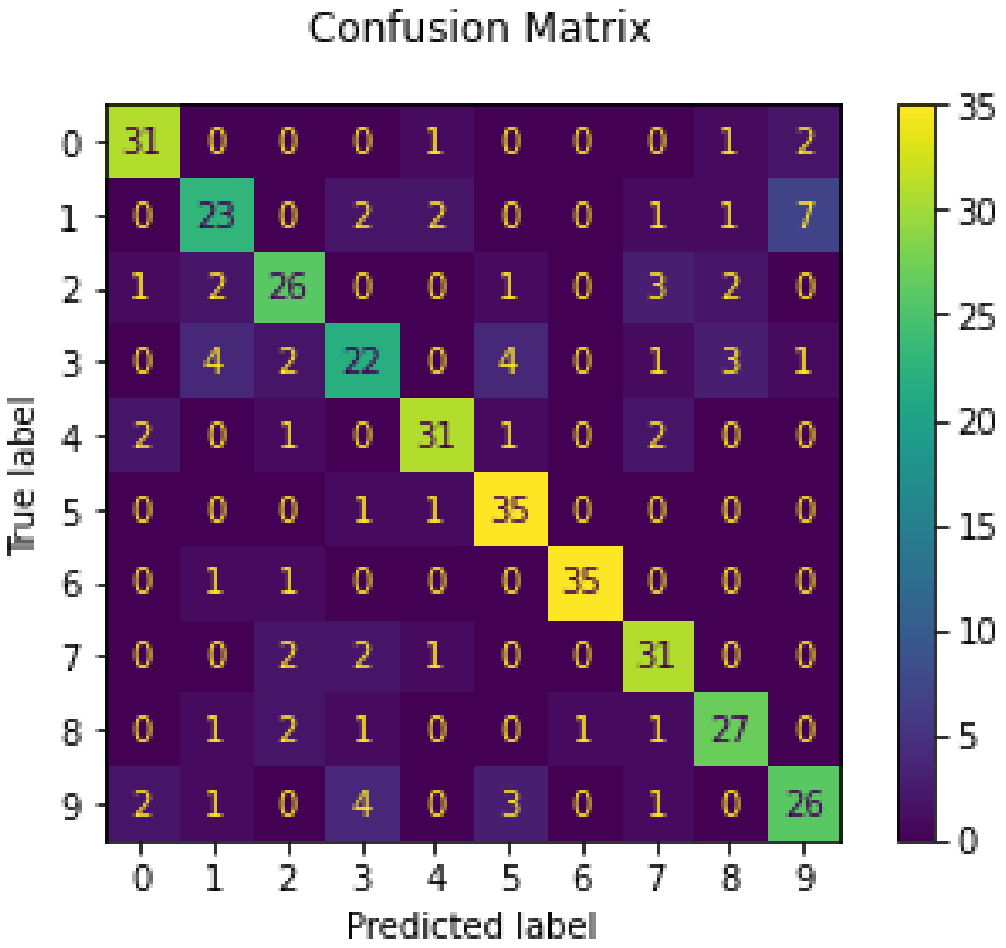


Thông số đạt được của cây ID3

```
print("Classification report for classifier %s:\n%s\n"
      % (classifier, metrics.classification_report(y_test, predicted)))
disp = metrics.plot_confusion_matrix(classifier, X_test, y_test)
disp.figure_.suptitle("Confusion Matrix")
print("Confusion matrix:\n%s" % disp.confusion_matrix)
```

```
Classification report for classifier DecisionTreeClassifier(ccp_alpha=0.0, class_weight=None, criterion='entropy',
                       max_depth=None, max_features=None, max_leaf_nodes=None,
                       min_impurity_decrease=0.0, min_impurity_split=None,
                       min_samples_leaf=1, min_samples_split=2,
                       min_weight_fraction_leaf=0.0, presort='deprecated',
                       random_state=33, splitter='best'):
              precision    recall  f1-score   support

           0       0.86      0.89      0.87        35
           1       0.72      0.64      0.68        36
           2       0.76      0.74      0.75        35
           3       0.69      0.59      0.64        37
           4       0.86      0.84      0.85        37
           5       0.80      0.95      0.86        37
           6       0.97      0.95      0.96        37
           7       0.78      0.86      0.82        36
           8       0.79      0.82      0.81        33
           9       0.72      0.70      0.71        37

    accuracy                           0.80       360
   macro avg       0.80      0.80      0.79       360
weighted avg       0.80      0.80      0.79       360
```

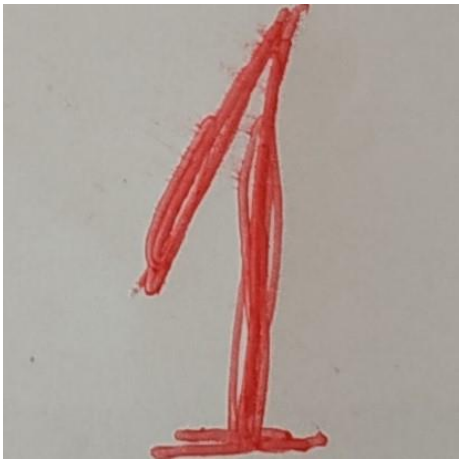
Ma trận nhầm lẫn



Cài đặt thư viện xử lý hình ảnh pillow cho Python và import vào

```
from PIL import Image, ImageOps
import numpy as np
```

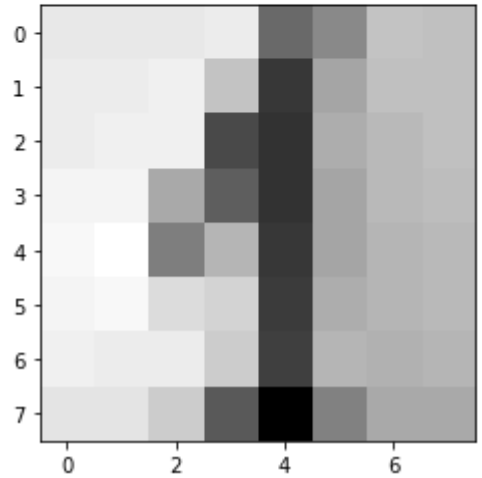
Chụp ảnh một chữ số viết tay bất kỳ, cắt theo hình vuông sát với ký tự như hình minh họa dưới đây



Đọc hình trên vào phần mềm

```
#Open image
img = Image.open('sample_1.jpg').convert("L").resize((8,8))
#Invert image
img = ImageOps.invert(img)
#Convert to numpy array
im2arr = np.array(img)
#Show
plt.imshow(im2arr, cmap=plt.cm.gray_r, interpolation='nearest')
```

```
<matplotlib.image.AxesImage at 0x1fd5070d490>
```



Làm rõ nét lại bức ảnh bằng cách phân biệt rõ những giá trị đen, trắng. Sinh viên có thể điều chỉnh giá trị phân biệt này (ở đây là 110) tùy thuộc vào điều kiện ảnh chụp.

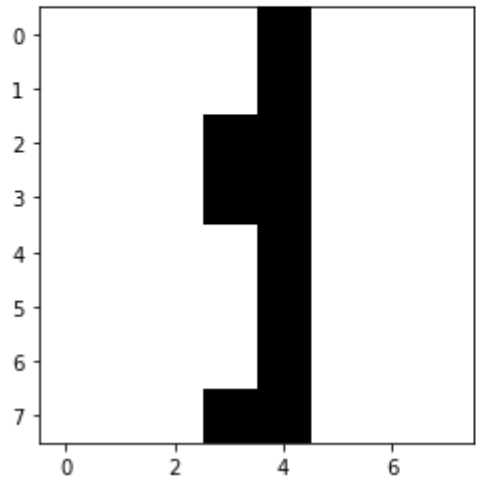
```
#Change from 2d array (8x8) to 1d array (1x64)
img1d=im2arr.reshape([1,64])
#normailize black & white image array
img1d[img1d > 109] = 155
img1d[img1d < 110] = 0
img1d
```

```
array([[  0,   0,   0,   0, 155,   0,   0,   0,   0,   0,   0,   0, 155,
          0,   0,   0,   0,   0,   0, 155, 155,   0,   0,   0,   0,   0,
          0, 155, 155,   0,   0,   0,   0,   0,   0,   0, 155,   0,   0,
          0,   0,   0,   0,   0, 155,   0,   0,   0,   0,   0,   0,   0,
        155,   0,   0,   0,   0,   0,   0, 155, 155,   0,   0,   0]],
      dtype=uint8)
```

Hình ảnh chữ số sau khi làm rõ

```
#After normalize
plt.imshow(im2arr, cmap=plt.cm.gray_r, interpolation='nearest')
```

```
<matplotlib.image.AxesImage at 0x1fd50a7b1f0>
```



Tiến hành nhận diện bằng cây ID3 và in ra kết quả

```
#Predict label of number
y_pred = classifier.predict(img1d)
print(y_pred)
```

```
[1]
```

V. Bài tập thêm

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

   a) Sinh viên hãy thiết kế một mô hình để áp dụng có hiệu quả thuật toán Naïve Bayes trên dữ liệu này.

b) Người ta muốn theo dõi, so sánh sự thay đổi của mô hình phân lớp theo thời gian (ví dụ: mô hình phân lớp của tuần trước so với hiện tại…). Sinh viên hãy gợi ý phương pháp thực hiện điều này.

2. Thông thường để dễ so sánh hai giá trị độ chính xác (precision) và độ phủ (recall) của những phương pháp với nhau người ta sử dụng một giá trị gọi là F-measure. F-measure là trung bình điều hòa (harmonic mean) của hai giá trị độ chính xác và độ phủ được tính bằng công thức:

$$F - measure = 2 \times \frac{precision \times recall}{precision + recall}$$

Hãy tính giá trị F-measure của các bài tập cơ bản.

VI. Tài liệu tham khảo

1. Adult Dataset, UCI Machine Learning Repository.
2. Wine Quality Dataset, UCI Machine Learning Repository.
3. Heart Disease UCI, Kaggle Datasets.
4. Mushroom Classification, Kaggle Datasets.
5. Recognizing hand-written digits, scikit-learn Documentation.
6. Slide bài giảng lý thuyết môn Khai thác dữ liệu.
7. Han, J., Kamber, M. & Pei, J. (2012). Data mining concepts and techniques, third edition Morgan Kaufmann Publishers.
8. scikit-learn Tutorials