TRƯỜNG ĐẠI HỌC XÂY DỰNG HÀ NỘI
KHOA CÔNG NGHỆ THÔNG TIN

-----□□🕮□□-----

BÁO CÁO BÀI TẬP NHÓM

ÁP DỤNG CÁC THUẬT TOÁN PHÂN LOẠI CHUẨN ĐOÁN BỆNH UNG THƯ

Sinh viên thực hiện	MSSV
Nguyễn Xuân Cường	1507764
Nguyễn Đỗ Lân	118664
Đặng Đình Thanh	182164
Ngô Đức Thịnh	189464
Đỗ Đức Tiến	1660364
Giảng viên hướng dẫn: ThS TS Phạm Hồng Phong	

Hà Nội, 10-2022

Breast Cancer - 64CS3 - Group 4

Nhóm 4 - 64CS3

19 tháng 10 2022

Mục lục

Lời nói đầu

- Trong thời đại phát triển ngày nay, nhiều bài toán yêu cầu đưa ra trong việc giải quyết vấn đề phân loại.
- Phân loại (Classification) là một trong những bài toán phạm vi nghiên cứu phổ biến nhất trong lĩnh vực Học máy. Ngày nay, bạn có thể thấy ứng dụng của phân loại học máy ở nhiều nơi. Ví dụ: khi bạn đăng một hình ảnh lên Facebook, nó có thể nhận ra khuôn mặt của bạn và bạn bè bạn; hoặc khi bạn truy cập Internet, bạn có thể thấy có rất nhiều quảng cáo hiển thị dựa trên sở thích và những gì bạn đã tìm kiếm trên Google trước đó. Hơn nữa, có thể bạn đã nghe nói rằng phân loại học máy có thể giúp dự đoán xem một bệnh nhân có bị bệnh hay không.
- Và sau đây nhóm sẽ áp dụng các thuật toán Bayes, Navie Bayes, Decision Tree, vào chuẩn đoán dữ liệu bệnh ung thư vú. Trong quá trình thực hiện đề tài không tránh khỏi những sai sót, nhóm em mong sẽ nhận được sự góp ý và đánh giá của thầy Phạm Hồng Phong.

Xin chân thành cảm ơn!

Chương 1

Các thuật toán phân loại

1.1 Cây quyết định

1.1.1 Tổng quan

Trong lý thuyết quyết định (chẳng hạn quản lý rủi ro), một **cây quyết định** (Tiếng Anh: Decision Tree) là một đồ thị của các quyết định và các hậu quả có thể của nó (bao gồm rủi ro và hao phí tài nguyên). Cây quyết định được sử dụng để xây dựng một kế hoạch nhằm đạt được mục tiêu mong muốn. Các cây quyết định được dùng để hỗ trợ quá trình ra quyết định. Cây quyết định là một dạng đặc biệt của cấu trúc cây.

1.1.2 Ý tưởng

- Trong lĩnh vực học máy, **cây quyết định** là một kiểu mô hình dự báo (Predictive Model), nghĩa là một ánh xạ từ các quan sát về một sự vật/hiện tượng tới các kết luận về giá trị mục tiêu của sự vật - hiện tượng. Mỗi một nút trong (internal node) tương ứng với một biến; đường nối giữa nó với nút con của nó thể hiện một giá trị cụ thể cho biến đó. Mỗi nút lá đại diện cho giá trị dự đoán của biến mục tiêu, cho trước các giá trị của các biến được biểu diễn bởi đường đi từ nút gốc tới nút lá đó. Kỹ thuật học máy dùng trong cây quyết định được gọi là học bằng cây quyết định, hay chỉ gọi với cái tên ngắn gọn là **cây quyết định**.
- Học bằng cây quyết định cũng là một phương pháp thông dụng trong khai phá dữ liệu. Khi đó, cây quyết định mô tả một cấu trúc cây, trong đó, các lá đại diện cho các phân loại còn cành đại diện cho các kết hợp của các thuộc tính dẫn tới phân loại đó. Một cây quyết định có thể được học bằng cách chia tập hợp nguồn thành các tập con dựa theo một kiểm tra giá trị thuộc tính. Quá trình này được lặp lại một cách đệ quy cho mỗi tập con dẫn xuất. Quá trình đệ quy hoàn thành khi không thể tiếp tục thực hiện việc chia tách được nữa, hay khi một phân loại đơn có thể áp dụng cho từng phần tử của tập con dẫn xuất. Một bộ phân loại rừng ngẫu nhiên (Tiếng Anh: Random Forest) sử dụng một số cây quyết định để có thể cải thiện tỉ lệ phân loại.

1.1.3 Phân loại

Cây quyết định còn có 2 tên khác:

- **Cây hồi quy** (Tiếng Anh: Regression Tree) ước lượng các hàm giá có giá trị là số thực thay vì được sử dụng cho các nhiệm vụ phân loại. (ví dụ: ước tính giá một ngôi nhà hoặc khoảng thời gian một bệnh nhân nằm viện)
- **Cây phân loại** (Tiếng Anh: Classification Tree), nếu y là một biến phân loại như: giới tính (nam hay nữ), kết quả của một trận đấu (thắng hay thua).

1.1.4 Ví dụ thực hành

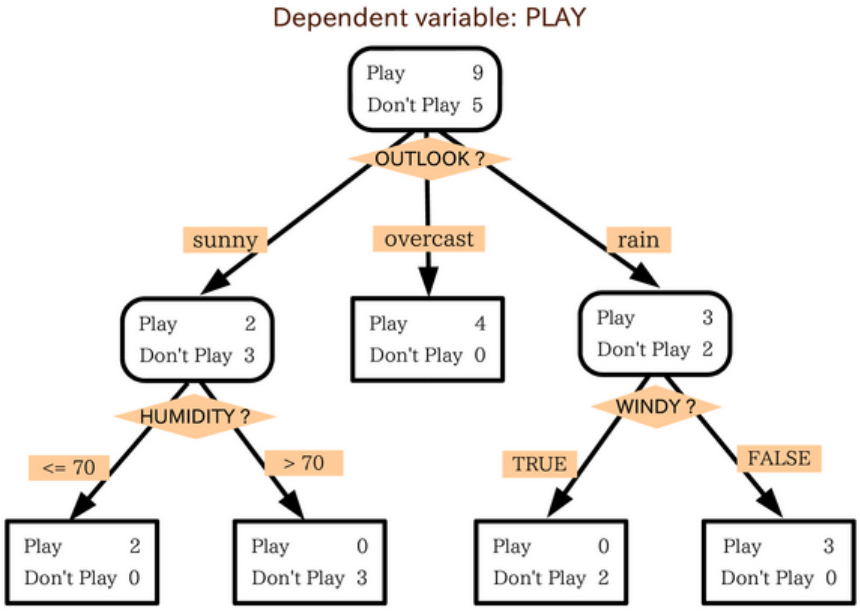
Ta sẽ dùng một ví dụ để giải thích về cây quyết định

- David là quản lý của một câu lạc bộ đánh golf nổi tiếng. Anh ta đang có rắc rối chuyện các thành viên đến hay không đến. Có ngày ai cũng muốn chơi golf nhưng số nhân viên câu lạc bộ lại không đủ phục vụ. Có hôm, không hiểu vì lý do gì mà chẳng ai đến chơi, và câu lạc bộ lại thừa nhân viên.
- Mục tiêu của David là tối ưu hóa số nhân viên phục vụ mỗi ngày bằng cách dựa theo thông tin dự báo thời tiết để đoán xem khi nào người ta sẽ đến chơi golf. Để thực hiện điều đó, anh cần hiểu được tại sao khách hàng quyết định chơi và tìm hiểu xem có cách giải thích nào cho việc đó hay không.
- Vậy là trong hai tuần, anh ta thu thập thông tin về:
  - Trời (outlook) - nắng (sunny)
  - Và tất nhiên là số người đến chơi golf vào hôm đó. David thu được một bộ dữ liệu gồm 14 dòng và 5 cột.

Independent Variables				Dep .Var
OUTLOOK	TEMPERATURE	HUMIDITY	WINDY	PLAY
sunny	85	85	FALSE	Don't Play
sunny	80	90	TRUE	Don't Play
overcast	83	78	FALSE	Play
rain	70	96	FALSE	Play
rain	68	80	FALSE	Play
rain	65	70	TRUE	Don't Play
overcast	64	65	TRUE	Play
sunny	72	95	FALSE	Don't Play
sunny	69	70	FALSE	Play
rain	75	80	FALSE	Play
sunny	75	70	TRUE	Play
overcast	72	90	TRUE	Play
overcast	81	75	FALSE	Play
rain	71	80	TRUE	Don't Play

Bảng 1.1: Tập dữ liệu về chơi Golf - Play Golf Dataset

  - Sau đó, để giải quyết bài toán của David, người ta đã đưa ra một mô hình cây quyết định.



- Nhóm người chơi golf khi trời nắng, nhóm chơi khi trời nhiều mây, và nhóm chơi khi trời mưa.
- Kết luận thứ nhất: nếu trời nhiều mây, người ta luôn luôn chơi golf. Và có một số người ham mê đến mức chơi golf cả khi trời mưa. Tiếp theo, ta lại chia nhóm trời nắng thành hai nhóm con. Ta thấy rằng khách hàng không muốn chơi golf nếu độ ẩm lên quá 70
- Cuối cùng, ta chia nhóm trời mưa thành hai và thấy rằng khách hàng sẽ không chơi golf nếu trời nhiều gió.
- Và đây là lời giải ngắn gọn cho bài toán mô tả bởi cây phân loại. David cho phần lớn nhân viên nghỉ vào những ngày trời nắng và ẩm, hoặc những ngày mưa gió. Vì hầu như sẽ chẳng có ai chơi golf trong những ngày đó. Vào những hôm khác, khi nhiều người sẽ đến chơi golf, anh ta có thể thuê thêm nhân viên thời vụ để phụ giúp công việc.
- Kết luận là cây quyết định giúp ta biến một biểu diễn dữ liệu phức tạp thành một cấu trúc đơn giản hơn rất nhiều.

1.1.5 Các công thức

1. **Entropy**:

   - **Entropy** đo độ hỗn độn hay độ không chắc chắn của một hệ thống. Trong bài toán phân lớp, mỗi một miền sẽ có entropy càng thấp nếu nó càng tinh khiết, nghĩa là chứa hầu hết các điểm thuộc về cùng một lớp.

$$H(D) = \sum_{i=1}^{k} P(c_i|D)\log_2(c_i|D)$$

   - Trong đó, là xác suất phân lớp ci xuất hiện trong D; k là số lượng phân lớp. Nếu một miền hoàn toàn tinh khiết thì sẽ có entropy bằng 0 . Một miền hoàn toàn hỗn độn sẽ có entropy bằng $\log_2 k$.

- Giả sử một điểm chia thực hiện tách D thành DY và DN . Ta định nghĩa split entropy là trung bình có trọng số entropy của các nửa nhận được:

$$H(D_Y|D_N) = \frac{n_Y}{n}H(D_Y) + \frac{n_N}{n}H(D_N)$$

  Trong đó: $n = |D|$, $n_Y = |D_Y|$, $n_N = |D_N|$

- Tỷ lệ tăng thông tin càng cao thì điểm chia càng tốt. Do đó, ta có thể chọn điểm chia sao cho tỷ lệ tăng thông tin lớn nhất.
- Áp dụng giải ví dụ:
  - Tính chỉ số Entropy của thuộc tính outlook:

$$E\,(Parent) = \left(-\frac{9}{14}\right)\log_2\frac{9}{14} + \left(-\frac{5}{14}\right)\log_2\frac{5}{14} = 0.94\,(1)$$

  - Để tìm điểm chia tốt nhất, tiến hành lặp qua tất cả thuộc tính. Giả sử xét thuộc tính "outlook" gồm 3 giá trị: sunny, overcast, rain

$$E\,(Parent|outlook = sunny) = \left(-\frac{2}{5}\right)\log_2\frac{2}{5} + \left(-\frac{3}{5}\right)\log_2\frac{3}{5} = 0.97\,(2)$$

$$E\,(Parent|outlook = overcast) = \left(-\frac{4}{4}\right)\log_2\frac{4}{4} + \left(-\frac{0}{4}\right)\log_2\frac{0}{4} = 0\,(3)$$

$$E\,(Parent|outlook = rain) = \left(-\frac{2}{5}\right)\log_2\frac{2}{5} + \left(-\frac{3}{5}\right)\log_2\frac{3}{5} = 0.97\,(4)$$

  - Từ (1), (2), (3), (4) ta tính được Entropy của Outlook

$$E\,(Parent|outlook) = 0.97 \times \frac{5}{14} + 0.97 \times \frac{5}{14} = 0.7$$

  - Ta tính được chỉ số Information Gain của tập outlook:
    Information Gain = 0.94 – 0.7 = 0.24
  - Tương tự các thuộc tính còn lại, humidity có Information Gain có = 0.02, windy có Information Gain = 0.23. Ta sẽ chọn điểm chia tiếp theo là thuộc tính outlook vì có chỉ số Information Gain lớn nhất

2. **Gini impurity**

- Chỉ số Gini (gini index): chỉ số gini đước sử dụng trong thuật toán CART. Trái ngược với độ đo Gain, chỉ số gini là độ đo về tính "không trong suốt" của tập dữ liệu. Chỉ số gini của 1 tập dữ liệu D được tính như sau:

$$G(D) = 1 - \sum_{i=1}^{k} P(c_i|D)^2$$

- Với k là tổng số nhãn lớp, ci|D là xác suất để 1 bộ bất kì trong D thuộc về 1 nhãn c_i
- Nếu một miền hoàn toàn tinh khiết thì sẽ có chỉ số GINI bằng 0. Một miền hoàn toàn hỗn độn sẽ có chỉ số GINI bằng $\frac{k-1}{k}$

- Chỉ số GINI có trọng số của điểm chia được định nghĩa như sau:

$$H(G_Y|G_N) = \frac{n_Y}{n}G(D_Y) + \frac{n_N}{n}G(D_N)$$

Trong đó: $n = |D|$, $n_Y = |D_Y|$, $n_N = |D_N|$

- Một độ đo khác có thể được dùng để thay thế cho entropy và chỉ số GINI là CART (Classification And Regression Trees measure):

$$CART(D_Y, D_N) = 2\frac{n_Y}{n}\frac{n_N}{n}\sum_{i=1}^{k}|P(c_i|D_Y)| - |P(c_i|D_N)|$$

Độ đo CART càng cao thì điểm chia càng tốt.

- Áp dụng giải ví dụ:
Tính chỉ số gini của thuộc tính outlook:

$$Gini(Parent) = 1 - \left(\frac{9}{14}\right)^2 - \left(\frac{5}{14}\right)^2 = 0.459(1)$$

Để tìm điểm chia tốt nhất, tiến hành lặp qua tất cả thuộc tính. Giả sử xét thuộc tính "outlook" gồm 3 giá trị: sunny, overcast, rain

$$Gini(Parent|outlook = sunny) = 1 - \left(\frac{2}{5}\right)^2 - \left(\frac{3}{5}\right)^2 = 0.48(2)$$

$$Gini(Parent|outlook = overcast) = 1 - \left(\frac{4}{4}\right)^2 - \left(\frac{0}{4}\right)^2 = 0(3)$$

$$Gini(Parent|outlook = rain) = 1 - \left(\frac{2}{5}\right)^2 - \left(\frac{3}{5}\right)^2 = 0.48(4)$$

Tương tự các thuộc tính còn lại, humidity có gini = 0.45, windy có gini = 0.41. Ta sẽ chọn điểm chia tiếp theo là thuộc tính outlook vì có chỉ số gini nhỏ nhất

1.2 Bayes

1.2.1 Định lý

Định lý Bayes (Tiếng Anh: Bayes' Theorem) là một định lý toán học để tính xác suất xảy ra của một sự kiện ngẫu nhiên A khi biết sự kiện liên quan B đã xảy ra.

1.2.2 Công thức

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Kí hiệu $\neg A$ là không A (hay bù A). Ta có: $P(A) + P(\neg A) = 1$
Từ đó:

$$P(B) = P(B, A) + P(B, \neg A) = P(B|A)P(A) + P(B|\neg A)P(\neg A)$$

Định lý Bayes được viết dưới dạng biến thể như sau:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|\neg A)P(\neg A)}$$

1.2.3 Ví dụ

Một gia đình có 2 đứa trẻ. Biết có ít nhất có một đứa trẻ là con gái và sinh vào thứ 3. Hỏi xác suất 2 đứa trẻ đều là con gái là bao nhiêu? Biết rằng:

- Xác suất để một đứa trẻ sinh vào một ngày nhất định trong tuần là 1/7
- Giới tính của đứa trẻ và ngày sinh của nó là 2 sự kiện không liên quan đến nhau

Ký hiệu	Sự kiện	Xác suất
B	Ít nhất 1 đứa trẻ là con gái sinh vào ngày thứ 3	?
A	Cả 2 đứa trẻ đều là con gái	1/4
A_1	Chỉ 1 trong 2 đứa trẻ là con gái	1/2
C	Đứa trẻ sinh ra vào ngày thứ 3	1/7
$\neg C$	Đứa trẻ không sinh ra vào ngày thứ 3	6/7

- Để sử dụng định lý Bayes tính $P(A|B)$ ta cần tính được $P(B|A)$ và $P(B)$
  - P(B|A) được hiểu là xác suất ít nhất 1 đứa trẻ là con gái sinh ra vào thứ 3 nếu biết trước 2 đứa trẻ là con gái
  - Ta sẽ tính xác suất phần bù $P(\neg B|A)$ là xác suất để không có đứa trẻ nào sinh ra vào thứ 3
$$P(\neg B|A) = P(\neg C)P(\neg C) = \frac{6}{7} \times \frac{6}{7} = \frac{36}{49}$$
  - Vậy ta có:
$$P(B|A) = 1 - P(\neg B|A) = \frac{13}{49}$$
  - $P(B)$ là xác suất sự ít nhất 1 đứa trẻ là con gái sinh ra vào thứ 3.
  - Sự kiện này bao gồm 2 khả năng:
    * Cả 2 đứa trẻ đều là con gái (A)
    * Chỉ 1 đứa trẻ là con gái (A_1)
    * Ta có:
$$P(B) = P(B, A) + P(B, A_1)$$
$$P(B) = P(B|A)P(A) + P(B|A_1)P(A_1)$$
$$P(B) = \frac{13}{49} \times \frac{1}{4} + \frac{1}{7} \times \frac{1}{2}$$
$$P(B) = \frac{27}{196}$$

1.3 Naïve Bayes

1.3.1 Ý tưởng

- Thuật toán Naïve Bayes là thuật toán học máy nâng cao hiệu quả, Mặc dù mới ra mắt cơ mà nó đã thể hiện khả năng không chỉ đơn giản, hiệu năng của thuật toán naïve bayes

cũng có tốc độ xử lý khá nhanh, chính xác và đáng tin cậy. Nó đã được ứng dụng thành công cho rất nhiều mục đích, nhưng đặt biệt đốt khi xử lý các vấn đề về xử lý ngôn ngữ tự nhiên Natural Language Processing (NLP).

- Thuật toán Naïve Bayes được xây dựng dựa trên "Định lý Bayes"- một định lý được xử dụng khá rộng rãi, phổ biến trong các công việc, phần việc phân lớp. Bây giờ chúng ta sẽ bắt đầu đi tìm hiểu thuật toán Naïve Bayes và một vài định nghĩa thiết yếu để hiểu một cách rõ rang và vững vàng hơn về định nghĩa của thuật toán này.

1.3.2 Định lý Bayes

- **Định lý Bayes** đó chỉ là một công thức toán học đơn giản được sử dụng cho việc tính toán xác suất có điều kiện.
- Xác suất có điều kiện là một thước đo xác suất của một sự kiện xảy ra khi các sự kiện khác xảy ra (theo giả định, khẳng định và đã xác thực).

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{1.1}$$

- Chú thích công thức 1.1:
  - Tần suất của A xảy ra khi B xảy ra, được viết bằng **P(A|B)** còn được gọi là xác suất hậu nghiệm
  - Tần suất B xảy ra khi A xảy ra được viết bằng **P(B|A)**
  - Tần suất của riêng A được viết bằng **P(A)**
  - Tần suất của riêng B được viết bằng **P(B)**
- Nói một cách đơn giản hơn, định lý Bayes là một cách để tìm khả năng xác suất khi bạn biết một số xác suất khác
- Dưới đây là một ví dụ thống kê data về các vụ trộm ô tô :

Example No.	Color	Type	Origin	Stolen?
1	Red	Sports	Domestic	Yes
2	Red	Sports	Domestic	No
3	Red	Sports	Domestic	Yes
4	Yellow	Sports	Domestic	No
5	Yellow	Sports	Imported	Yes
6	Yellow	SUV	Imported	No
7	Yellow	SUV	Imported	Yes
8	Yellow	SUV	Domestic	No
9	Red	SUV	Imported	No
10	Red	Sports	Imported	Yes

- Các quan hệ, liên quan giữa các đặc tính có thể được hiểu như sau:

  - Chúng ta giả định rằng không có đặc tính, tính năng nào phụ thuộc vào nhau. Ví dụ màu sắc đỏ của xe không liên quan gì tới nguồn gốc của xe. Do đó các tính năng được coi là độc lập
  - Thứ hai, mỗi đặc trưng của data sẽ cho chúng ta một độ ảnh hưởng và tầm quan trọng giống nhau giống nhau. Ví dụ như khi biết mỗi color hoặc type thì đều sẽ không thể đoán chính xác được kết quả. Vì vậy có cặp thuộc tính nào không liên quan tới nhau, và nó đều có độ ảnh hưởng tới chính xác dự đoán là như nhau

- Các giả thuyết được đưa ra bởi Naïve Bayes đều không đúng ở các tình huống thực tế. Các giả thuyết độc lập thường sai sót, nhưng thường hoạt động tốt ở thực tế

1.3.3 Ví dụ

- Trên là tất cả tập dữ liệu của chúng ta, chúng ta cần phải phân loại xem chiếc xe có bị đánh cắp hay không dựa trên các đặc trưng của chiếc xe. Các cột đại diện cho các đặc trưng và các hàng đại diện cho các đối tượng riêng lẻ. Ví dụ nếu chúng ta nhìn vào hàng đầu tiên của tập dữ liệu, chúng ta có thể quan sát rằng chiếc xe bị đánh cắp có color là red, type là sport, và origin là Domestic. Nhưng giờ chúng ta muốn dự đoán một chiếc Red Domestic SUV có bị đánh cắp hay không. Chú ý rằng là không có data Red Domestic SUV trong tập data của chúng ta, theo ví dụ này định lý Bayes có thể được viết lại là:

$$P(y|X) = \frac{P(X|y)P(y)}{P(X)}$$

- Y là lớp biến mang giá trị là có bị đánh cắp hay không?, đại diện cho chiếc xe có bị đánh cắp hay không dựa trên các điều kiện, biến X đại diện cho tham số hoặc tính năng:

$$X = (x_1, x_2, x_3, ..., x_n)$$

- $x_1, x_2, x_3, x_4, x_5, ..., x_n$ đại diện cho các đặc trưng, chúng có thể ánh xạ tới Color, Type và Origin. Bằng cách thay thế cho x và mở rộng bằng cách sử dụng quy tắc chuỗi chúng ta nhận được.

$$P(y|x_1, ..., x_n) = \frac{P(x_1|y)P(x_2|y)...P(x_n|y)P(y)}{P(x_1)P(x_2)...P(x_n)}$$

- Giờ đây bạn có thể lấy được giá trị cho một tập dữ liệu và ghép chúng vào phương trình. Đối với tất cả các mục của tập dữ liệu, mẫu số không thay đổi, nó vẫn tĩnh. Do đó, mẫu số có thể được lược bỏ và tỷ lệ có thể được tiêm

$$P(y|x_1, ..., x_n) \propto P(y) \prod_{i=1}^{n} P(x_i|y)$$

- Ở trường hợp này của chúng ta, lớp biến (y) chỉ có hai kết quả đầu ra, có thể có các trường hợp phân loại là đa biến. Do đó chúng ta cần phải tìm biến lớp (y) có xác suất lớn nhất

$$y = argmax_y P(y) \prod_{i=1}^{n} P(x_i|y)$$

- Sau khi sử dụng hàm trên chúng ta có thể nhận được lớp với các yếu tố/ đặc trưng dự đoán. Trước tiên có thể tính xác suất $P(y|X)$, tạo bảng xác suất:

Bảng tần suất		Bị trộm?	
		Yes	No
Màu sắc	Đỏ	3/5	2/5
	Vàng	2/5	3/5
Bảng khả năng		Bị trộm?	
		P(Yes)	P(No)
Màu sắc	Đỏ	3/5	2/5
	Vàng	2/5	3/5

Bảng 1.2: Bảng tần suất và khả năng của Màu sắc - Colors:

Bảng tần suất		Bị trộm?	
		Yes	No
Loại xe	Thể thao	4	2
	SUV	1	3
Bảng khả năng		Bị trộm?	
		P(Yes)	P(No)
Loại xe	Thể thao	4/5	2/5
	SUV	1/5	3/5

Bảng 1.3: Bảng tuần suất và khả năng của Type: Phân loại

Bảng tần suất		Bị trộm?	
		Yes	No
Xuất xứ	Nội địa	2	3
	Xuất khẩu	3	2
Bảng khả năng		Bị trộm?	
		P(Yes)	P(No)
Xuất xứ	Nội địa	2/5	3/5
	Xuất khẩu	3/5	2/5

Bảng 1.4: Bảng tần suất và khả năng của Origin : Nguồn gốc

Màu sắc	Loại xe	Xuất xứ	Bị trộm
Đỏ	SUV	Nội địa	?

Bảng 1.5: Bảng dự đoán

- Theo các phương trình được thảo luận ở trên cùng với các dữ kiện đã cho ở các bảng ở trên (Bảng 1.2, Bảng 1.3 và Bảng 1.4), chúng ta có thể tính xác suất P (Yes|X) và P (No|X) là:

$$P(Yes|X) = P(\text{Red}|Yes) \times P(SUV|Yes) \times P(domestic|\text{yes}) = \frac{3}{5} \times \frac{1}{5} \times \frac{2}{5} \times 1 = 0.048$$

$$P(No|X) = P(No|Yes) \times P(No|Yes) \times P(domestic|\text{yes}) = \frac{2}{5} \times \frac{3}{5} \times \frac{3}{5} \times 1 = 0.144$$

→ Vì 0,144 > 0,048, có nghĩa là với các xe có các đặc trưng là loại xe SUV và màu đỏ có nguồn gốc nội địa, ví dụ trên được phân loại là 'KHÔNG', vậy chúng ta kết luận rằng chiếc xe không bị đánh cắp.

1.3.4 Phân loại Naïve Bayes

- Multinomial: Nó được sử dụng cho các số đếm rời rạc. Ví dụ: giả sử, về vấn đề về phân loại văn bản. Ở đây chúng ta có thể coi thử nghiệm Bernoulli là một bước tiến xa hơn và thay vì "từ xuất hiện trong tài liệu", thì bộ phân loại multinomial "đếm tần suất từ xuất hiện trong tài liệu", bạn có thể coi nó là "số lần số kết quả xi được quan sát thấy qua n lần thử nghiệm.
- Bernoulli: Mô hình nhị thức hữu ích nếu vectơ đặc trưng của bạn là nhị phân (0 và 1). Một ứng dụng sẽ là phân loại văn bản với mô hình "bag of words"trong đó số 1 và số 0 lần lượt là "từ xuất hiện trong tài liệu"và "từ không xuất hiện trong tài liệu".
- Gaussian: Nó được sử dụng trong phân loại và nó giả định rằng các đối tượng địa lý tuân theo phân phối chuẩn.

Chương 2

Thực nghiệm

2.1 Tập dữ liệu

Tập dữ liệu nhóm lấy từ tập dữ liệu UCI (Truy cập: https://archive.ics.uci.edu/ml/datasets/Breast+Cancer). Tập dữ liệu gồm các thuộc tính sau:

	Class	Age	menopause	tumor-size	inv-nodes	node-caps	deg-malig	breast	breast-quad	irradiat
0	no-recurrence-events	30-39	premeno	30-34	0-2	no	3	left	left_low	no
1	no-recurrence-events	40-49	premeno	20-24	0-2	no	2	right	right_up	no
2	no-recurrence-events	40-49	premeno	20-24	0-2	no	2	left	left_low	no
3	no-recurrence-events	60-69	ge40	15-19	0-2	no	2	right	left_up	no
4	no-recurrence-events	40-49	premeno	0-4	0-2	no	2	right	right_low	no
...	...	...	...	...	...	...	...	...	...	...
281	recurrence-events	30-39	premeno	30-34	0-2	no	2	left	left_up	no
282	recurrence-events	30-39	premeno	20-24	0-2	no	3	left	left_up	yes
283	recurrence-events	60-69	ge40	20-24	0-2	no	1	right	left_up	no
284	recurrence-events	40-49	ge40	30-34	3-5	no	3	left	left_low	no
285	recurrence-events	50-59	ge40	30-34	3-5	no	3	left	left_low	no

286 rows × 10 columns

Hình 2.1: Hình ảnh thông tin trong tập dữ liệu

- Class: gồm 2 lớp tái phát và không tái phát.
- Age: tuổi của bệnh nhân tại thời điểm chẩn đoán.
- Menopause: cho dù bệnh nhân là tiền hoặc sau mãn kinh tại thời điểm chẩn đoán.
- tumor-size: đường kính lớn nhất (tính bằng mm) của khối u đã cắt.
- breast: ung thư vú rõ ràng có thể xảy ra ở một trong hai bên vú.
- irradiate: xạ trị là phương pháp điều trị sử dụng tia X năng lượng cao để tiêu diệt các tế bào ung thư.
- inv-nodes: số lượng (phạm vi 0 - 39) các hạch bạch huyết ở nách chứa ung thư vú di căn có thể nhìn thấy khi xét nghiệm mô học

- node-caps: nếu ung thư di căn đến một hạch bạch huyết, mặc dù bên ngoài vị trí ban đầu của khối u, nó có thể vẫn bị ”chứa”bởi nang của hạch bạch huyết. Tuy nhiên, theo thời gian và khi bệnh phát triển mạnh hơn, khối u có thể thay thế hạch bạch huyết và sau đó xâm nhập vào nang, cho phép nó xâm lấn các mô xung quanh.
- deg-malig: mức độ mô học (khoảng 1-3) của khối u. Các khối u lớp 1 chủ yếu bao gồm các tế bào, mặc dù là khối u tân sinh, vẫn giữ được nhiều đặc điểm bình thường của chúng.
- breast-quad: vú có thể được chia thành bốn góc phần tư, lấy núm vú được lấy làm điểm trung tâm.

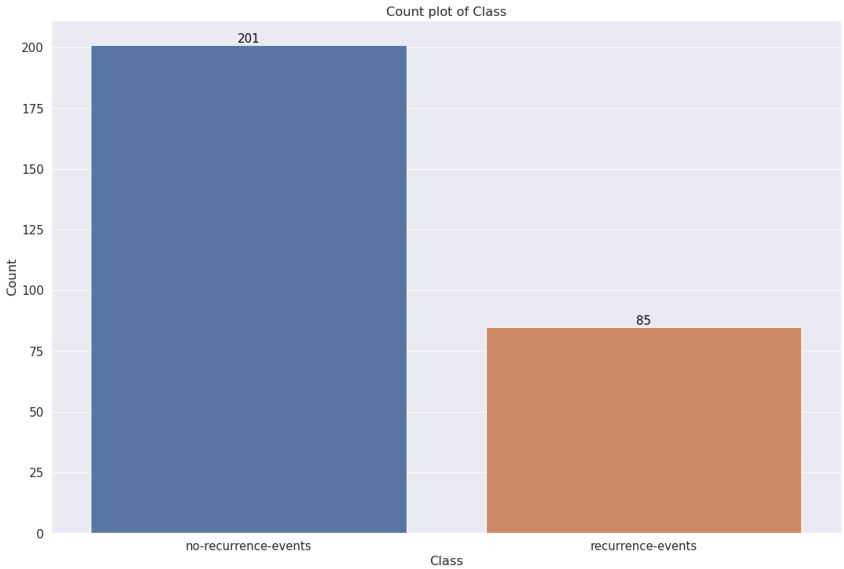
2.2 Tiến hành thực nghiệm

Sau đây chúng ta sẽ đi phân tích dữ liệu các cột trong đó thông qua các biểu đồ, để làm rõ tập dữ liệu bệnh ung thư.

2.2.1 Phân tích dữ liệu

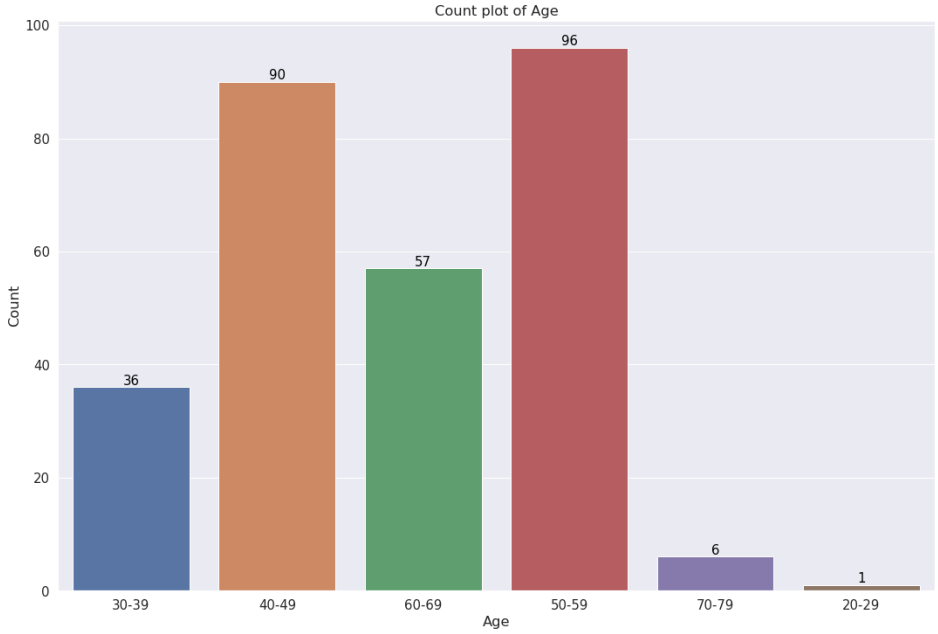
Trước tiên thế nào là ung thư?

1. Ung thư xảy ra khi tế bào của cơ thể tăng trưởng theo một cách không bình thường. Các tế bào ung thư này lan vào các mô khỏe mạnh của cơ thể. Đôi khi, ung thư còn được gọi là khối u.
2. Vậy ung thư vú là u ung thư phát triển ở vú. Có nhiều loại ung thư vú khác nhau.
3. Sau đây chúng ta sẽ đi phân tích dữ liệu qua các biểu đồ, để làm rõ tập dữ liệu bệnh ung thư vú của các bệnh nhân ung thư giai đoạn đầu, phát hiện ung thư sớm và loại bỏ nhưng vẫn được theo dõi tình trạng sức khỏe một thời gian sau khi điều trị.
    - Tập dữ liệu chia làm hai lớp: lớp tái nhiễm và lớp không tái nhiễm



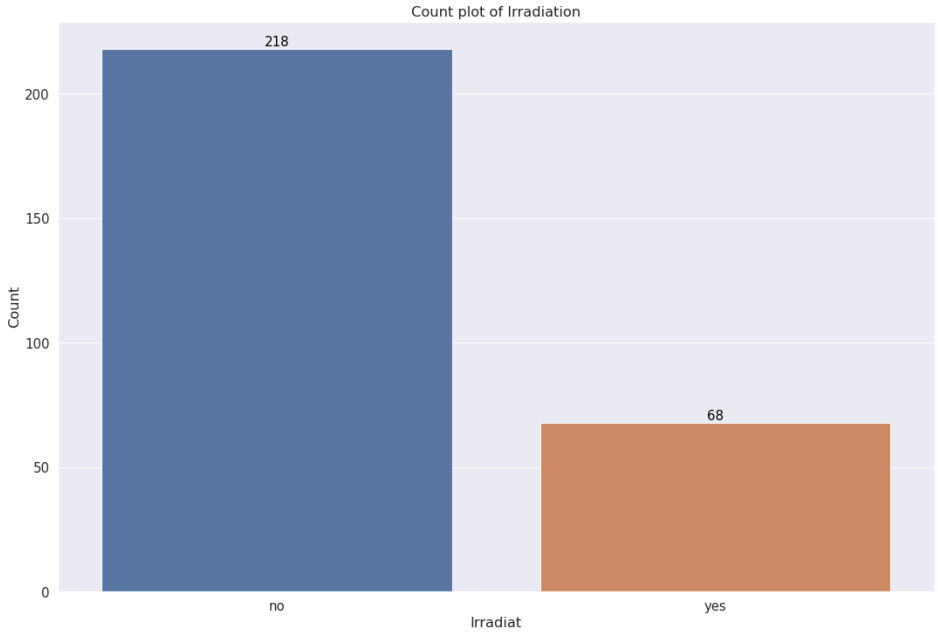
Hình 2.2: Biểu đồ số lớp tái nhiễm và lớp không tái nhiễm

- Tập dữ liệu chia làm hai lớp: lớp tái nhiễm và lớp không tái nhiễm. Với biểu đồ trên của tập dữ liệu gốc biểu hiện cho thấy số ca tái nhiễm vẫn xuất hiện nhiều sau khi bệnh nhân loại bỏ tế bào ung thư. Vậy nguyên nhân do đâu?



Hình 2.3: Biểu đồ phân bố độ tuổi

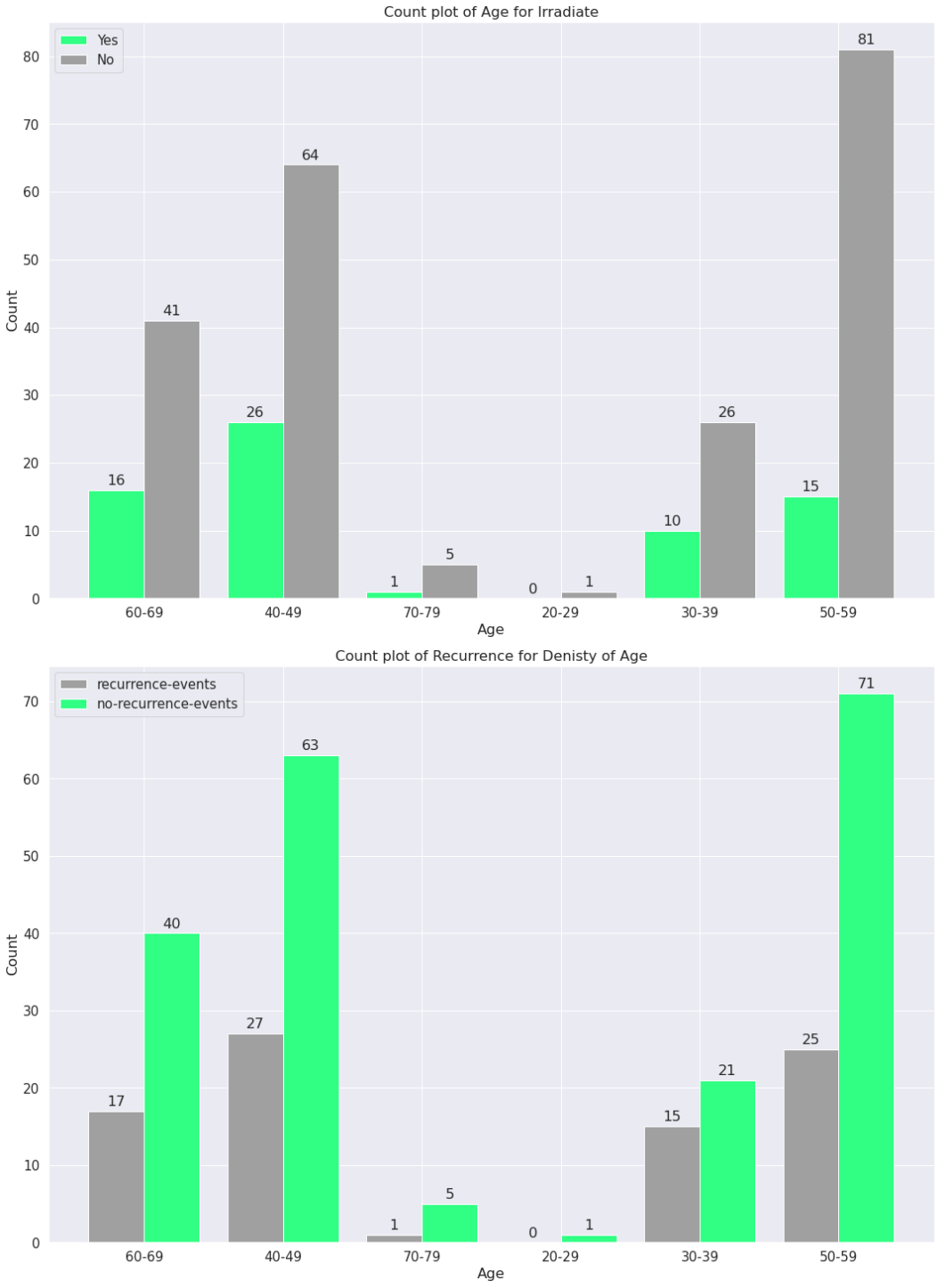
- Ung thư vú thường hay xuất hiện ở phụ nữ ở độ tuổi trung niên (45-65). Nhưng hiện giờ có thể thấy ung thư vú đang trẻ hóa dần như biểu đồ thể hiện trên. Đây là một vấn đề đáng quan ngoại với các bác sĩ nói riêng, với phụ nữ nói chung.



Hình 2.4: Biểu đồ bệnh nhân áp dụng phương pháp xạ trị

- Số lượng bệnh nhân áp dụng xạ trị vẫn còn thấp. Có thể do tâm lý người bệnh hay do phương pháp xạ trị đắt nên việc điều trị bằng phương pháp thấp.

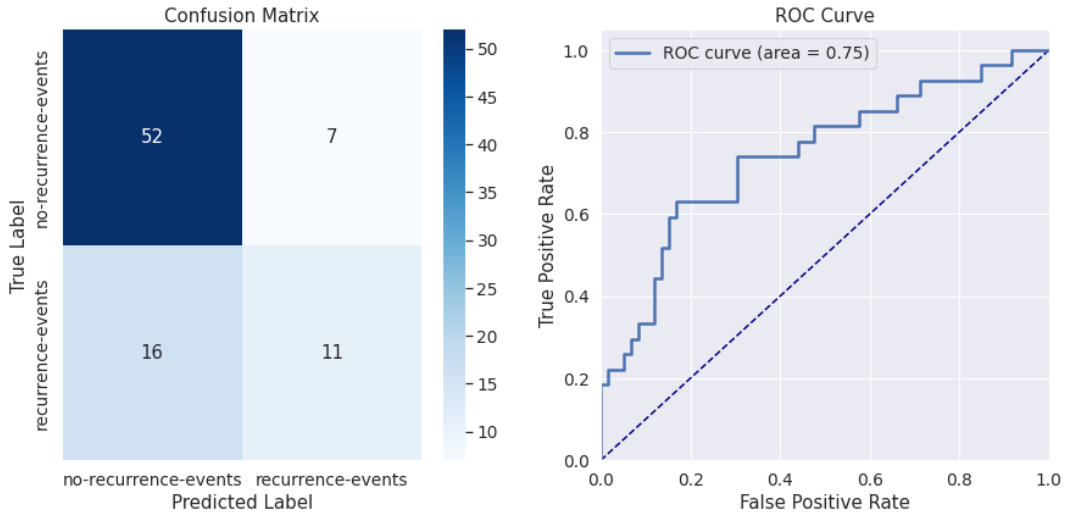
- Từ hai biểu đồ thì việc áp dụng phương pháp xạ trị để loại bỏ khối ung thư xong thì bệnh nhân vẫn tái nhiễm vẫn còn cao. Vậy điều trị bằng xạ trị có tốt không?



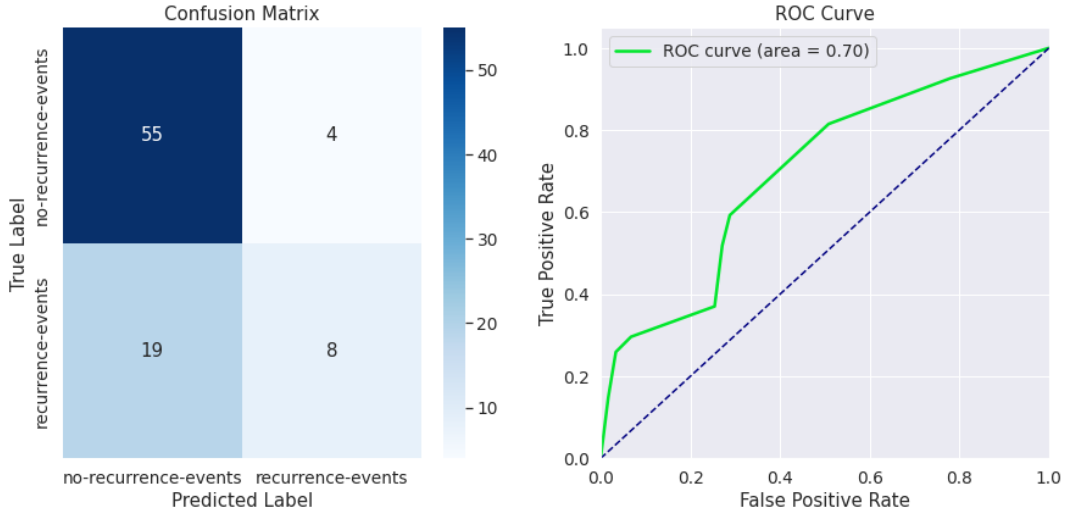
Hình 2.5: Biểu đồ phân bố độ tuổi, xạ trị và tái nhiễm

2.2.2 Đánh giá ROC

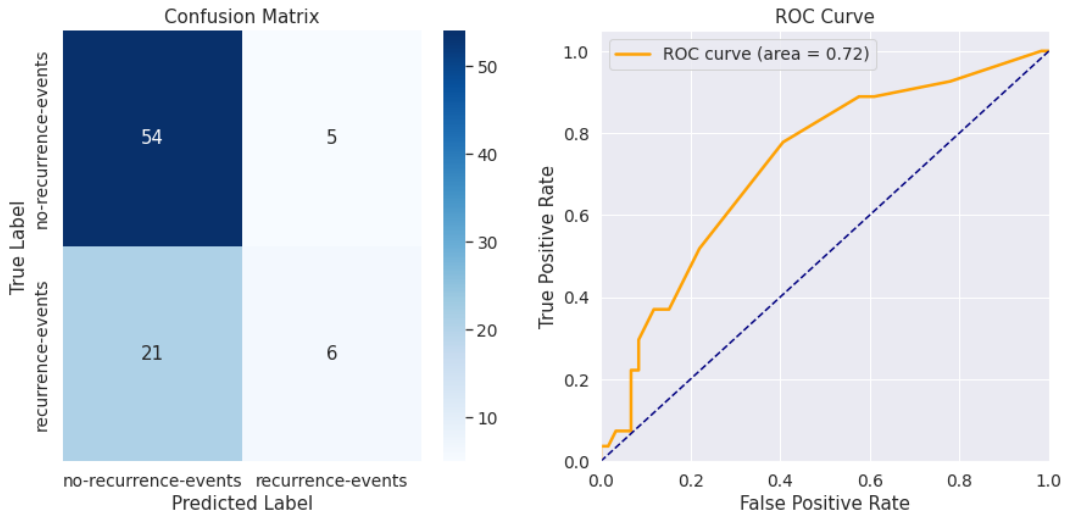
Đánh giá thuật toán



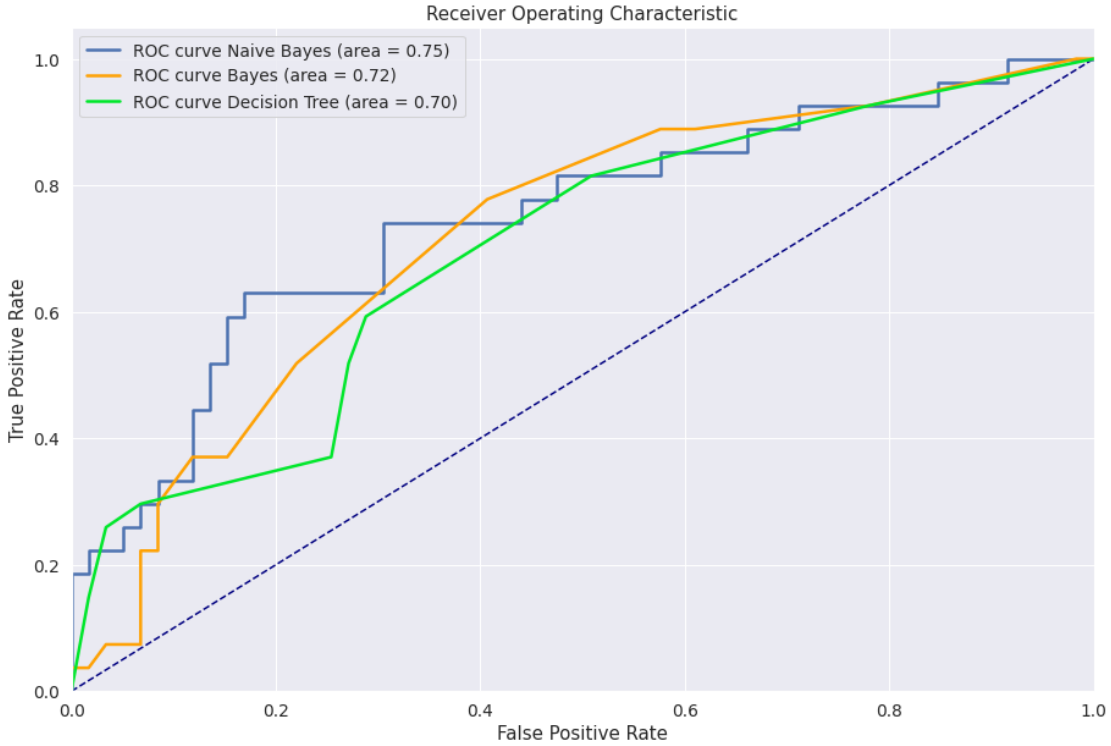
Hình 2.6: ROC của Naive Bayes



Hình 2.7: ROC của Cây quyết định



Hình 2.8: ROC của Bayes



Hình 2.9: ROC của các thuật toán

Chương 3

Nhận xét và biện pháp

3.1 Nhận xét

1. Qua dữ liệu phân tích và các biểu đồ, bệnh ung thư là một bệnh rất nguy hiểm và ảnh hưởng đến đời sống rất nhiều đến đời sống của người phụ nữ (có thể là đàn ông)

2. Với việc tìm hiểu và phân tích dữ liệu của nhóm, bệnh ung thư vú xảy ra không biết nguyên nhân do đâu, nó thường xảy ra đối với phụ nữ ở độ tuổi trung niên, có thể một phần nào cũng do hoàn cảnh hay tâm lý thời kỳ mãn kinh của phụ nữ. Và số lượng người nhiễm đang trẻ hoá là một điều đáng quan ngại.

3. Số lượng người điều trị bằng phương pháp xạ trị tái nhiễm còn cao, cho thấy việc tái nhiễm có thể do sức khoẻ của bệnh nhân hay lối sống mà ảnh hưởng đến bệnh tái nhiễm.

3.2 Biện pháp

Các biện pháp làm nhằm giảm thiểu nguy cơ bị nhiễm

1. Chế độ ăn uống phù hợp, ngủ đủ giấc để tăng sức đề kháng

2. Khám sức khoẻ định kỳ để kiểm tra và phát hiện sớm để loại bỏ sớm

3. Tập thể dục hay các bài tập aerobics để giúp thư giãn tinh thần

4. Tránh suy nghĩ tiêu cực, căng thẳng gây áp lực tâm lý