ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC BÁCH KHOA
KHOA KHOA HỌC VÀ KỸ THUẬT MÁY TÍNH

XÁC SUẤT THỐNG KÊ (MT2013)

Bài tập lớn

"Phân tích thống kê dữ liệu ung thư vú thực tế"

Giáo viên hướng dẫn: Nguyễn Đình Huy (Lớp TN01)

Sinh viên: Nguyễn Thái Học - 2311100 *(Nhóm 1, **Trưởng nhóm**)*
Lê Chí Đại - 2310621 *(Nhóm 1)*
Phạm Lê Tiến Đạt - 2310687 *(Nhóm 1)*
Nguyễn Quốc Huy - 2311209 *(Nhóm 1)*
Huỳnh Đức Nhân - 2312420 *(Nhóm 1)*
Phạm Trần Minh Trí - 2313622 *(Nhóm 1)*

THÀNH PHỐ HỒ CHÍ MINH, THÁNG 5 NĂM 2025

Mục lục

Danh sách các ảnh

Danh sách các bảng

Danh sách mã nguồn

Bảng phân công công việc

No.	Họ tên	MSSV	Công việc	Đánh giá
1	Nguyễn Thái Học	2311100	Thống kê suy diễn Kiểm định giả thuyết thống kê	100%
2	Lê Chí Đại	2310621	Thống kê suy diễn Hồi quy Logistic nhị phân đa biến	100%
3	Phạm Lê Tiến Đạt	2310687	Thống kê suy diễn Cây quyết định	100%
4	Nguyễn Quốc Huy	2311209	Thống kê mô tả Trực quan hóa dữ liệu	100%
5	Huỳnh Đức Nhân	2312420	Tiền xử lý dữ liệu Tổng quan dữ liệu	100%
6	Phạm Trần Minh Trí	2313622	Thống kê mô tả Trực quan hóa dữ liệu	100%

Bảng 1: Bảng phân công công việc

1 Tổng quan dữ liệu

Trong bối cảnh ô nhiễm môi trường đang ngày một trầm trọng, số người mắc bệnh ung thư ngày càng tăng. Theo thống kê của GLOBOCAN 2022, Việt Nam ghi nhận trên 180.000 ca mắc mới và 120.000 ca tử vong do ung thư. Trong số những ca mắc ung thư ở nữ, có đến 28.9% số ca là bị ung thư vú. Ung thư vú là dạng u vú ác tính. Chúng thường bắt đầu từ các ống dẫn sữa, sau đó phát triển lớn lên gây chèn ép các cơ quan xung quanh và cuối cùng là di căn vào các bộ phận khác gây đau đớn tột cùng. Ung thư vú luôn nằm trong nhóm có tỷ lệ tử vong cao nhất dù biện pháp chữa trị đơn giản hơn các loại ung thư khác. Nguyên nhân chủ yếu là do tâm lý chủ quan của người bệnh, chỉ tiến hành chữa trị khi đã quá muộn. Vì vậy, ta cần có một thống kê về tỉ lệ sống sót ở các giai đoạn của bệnh ung thư vú, để người bệnh có cái nhìn chính xác nhất về căn bệnh và tiến hành chữa trị kịp thời. Đó cũng chính là mục tiêu của bài báo cáo này.

Để tiến hành phân tích dữ liệu về bệnh ung thư vú, nhóm sử dụng bộ dữ liệu được thu thập từ hơn 300 người bệnh về các thông tin như tuổi, giới tính, giai đoạn của khối u, phân loại mô,... được cung cấp tại Kaggle[1]. Sau đây là một vài thông tin cơ bản của bộ dữ liệu:

- Tiêu đề: Real Breast Cancer Data.
- Nguồn: Nghiên cứu ung thư của Đại học Queen Belfast.
- Thời gian: 2021.
- Cỡ mẫu: 341 mẫu.
- Số biến: 16 biến.

Mô tả chi tiết của 16 biến trong bộ dữ diệu như sau:

- **Patient_ID**: Mã định danh của bệnh nhân (Kiểu chr). Mã này được sử dụng để theo dõi thông tin y tế mà không làm lộ thông tin cá nhân của bệnh nhân.
- **Age**: Tuổi của bệnh nhân (Kiểu int). Dữ liệu này rất cần thiết cho việc xác định tỉ lệ mắc, khởi bệnh ở các độ tuổi khác nhau.
- **Gender**: Giới tính của bệnh nhân: Nam, nữ hoặc không rõ (Kiểu chr). Giới tính của bệnh nhân cũng rất cần thiết cho việc xác định tỉ lệ sống sót khi mắc ung thư vú ở các giới tính khác nhau.
- **Protein1, 2, 3, 4**: Mức biểu hiện của các loại protein (Kiểu num). Dữ liệu này có thể biểu hiện các hoạt động sinh học bên trong các tế bào khối u và có thể được sử dụng để xác định các mục tiêu điều trị tiềm năng.
- **Tumour_stage**: Giai đoạn của khối u: I, II, III (Kiểu chr). Đây là dữ liệu quan trọng để xác định chiến lược điều trị phù hợp nhất cho từng bệnh nhân ở từng giai đoạn khác nhau.
- **Histology**: Loại mô học của khối u (Kiểu chr). Trong bộ dữ liệu này có 3 loại mô học: Infiltrating Ductal Carcinoma (Ung thư ống xâm lấn), Infiltrating Lobular Carcinoma (Ung thư tiêu thùy xâm lấn) và Mucinous Carcinoma (Ung thư tuyến nhầy). Với mỗi loại tế bào ung thư khác nhau, bác sĩ sẽ có các biện pháp chữa trị khác nhau cho bệnh nhân.

[1] https://www.kaggle.com/datasets/amandam1/breastcancerdataset/data

- **ER.status**: Tình trạng thụ thể Estrogen (Kiểu chr): Positive (Dương tính) hoặc Negative (Âm tính). Dữ liệu này thể hiện liệu khối u có phản ứng với estrogen không. Tùy thuộc vào kết quả mà sẽ có những liệu pháp nội tiết khác nhau.
- **PR.status**: Tình trạng thụ thể Progesterone (Kiểu chr): Positive (Dương tính) hoặc Negative (Âm tính). Tương tự như ER.status, dữ liệu này thể hiện độ nhạy hormone của khối u.
- **HER2.status**: Tình trạng HER2 (Kiểu chr): Positive (Dương tính) hoặc Negative (Âm tính). Dữ liệu này thể hiện kết quả xét nghiệm hàm lượng protein HER2 của bệnh nhân, qua đó xác định khả năng điều trị bằng liệu pháp nhắm mục tiêu HER2.
- **Surgery_type**: Loại phẫu thuật đã thực hiện (Kiểu chr): Lumpectomy (Cắt bỏ khối u), Simple Mastectomy (Cắt toàn bộ tuyến vú đơn giản), Modified Radical Mastectomy (Cắt toàn bộ tuyến vú có kèm nạo hạch) hoặc Other (Khác). Thông tin này rất quan trọng để ghi nhận phương pháp phẫu thuật được áp dụng.
- **Date_of_Surgery, Date_of_Last_Visit**: Ngày phẫu thuật và ngày tái khám cuối cùng (Kiểu chr). Dữ liệu ngày được nhập dưới dạng chr và định dạng theo kiểu: "DD-MON-YY". Ngoài ra, Date_of_Last_Visit có thể trống nếu bệnh nhân không quay lại sau phẫu thuật.
- **Patient_Status**: Tình trạng của bệnh nhân (Kiểu chr): Alive (Còn sống) hoặc Dead (Đã mất). Giá trị này có thể là rỗng nếu bệnh nhân không quay trở lại tái khám. Đây cũng là thông tin quan trọng nhất, thông tin mà chúng ta cần dự đoán.

Sau đây là bảng mô tả thông tin về 9 biến xuất hiện trong bộ dữ liệu:

STT	Tên biến	Loại biến	Đơn vị	Ý nghĩa
1	Patient_ID	phân loại (chr)	-	Số định danh phân biệt các bệnh nhân khác nhau
2	Age	định lượng (int)	năm	Tuổi của bệnh nhân
3	Gender	phân loại (char)	-	Giới tính của bệnh nhân
4-7	Protein1-4	định lượng (num)	-	Mức biểu hiện của Protein
8	Tumour_stage	phân loại (char)	-	Giai đoạn của khối u
9	Histology	phân loại (char)	-	Loại mô học của khối u
10	ER.status	phân loại (char)	-	Tình trạng thụ thể Estrogen
11	PR.status	phân loại (char)	-	Tình trạng thụ thể Progesterone

STT	Tên biến	Loại biến	Đơn vị	Ý nghĩa
12	HER2.status	phân loại (char)	-	Tình trạng HER2
13	Surgery_type	phân loại (char)	-	Loại phẫu thuật thực hiện
14	Date_of_Surgery	phân loại (char)	-	Ngày phẫu thuật
15	Date_of_Last_Visit	phân loại (char)	-	Ngày tái khám cuối cùng
16	Patient_Status	phân loại (char)	-	Tình trạng của bệnh nhân

Bảng 2: Bảng mô tả biến trong bộ dữ liệu

Mục tiêu bài báo cáo: Với bộ dữ liệu nêu trên, nhóm sẽ tiến hành các bước tiền xử lý, thống kê mô tả và kiểm định giả thuyết thống kê. Sau đó, nhóm sẽ xây dựng các mô hình phân tích nhằm phân biệt và dự đoán trạng thái sống hoặc tử vong của bệnh nhân mắc bệnh ung thư vú. Cụ thể như sau:

1. Tiền xử lý dữ liệu:
   - Đọc và làm sạch dữ liệu: Loại bỏ dữ liệu rỗng, điều chỉnh định dạng dữ liệu.
   - Xử lý dữ liệu: mã hóa biến phân loại, điều chỉnh biến dữ liệu.
2. Thống kê mô tả
   - Phân tích đặc điểm của các biến định lượng, định tính.
   - Trực quan hóa dữ liệu bằng các loại biểu đồ: biểu đồ cột, histogram, boxplot.
3. Thống kê suy diễn
   - Kiểm định giả thuyết thống kê.
   - Đánh giá hoạt động mô hình Hồi quy Logistic nhị phân đa biến.
   - Khởi tạo mô hình Cây quyết định.

2 Tổng quan kiến thức nền

2.1 Về thống kê mô tả

Thống kê mô tả giúp cung cấp cái nhìn tổng quát các khía cạnh của dữ liệu, diễn tả những đặc trưng cơ bản của mẫu. Trong thống kê mô tả, những bộ phận chính được sử dụng bao gồm các đại lượng trung bình, độ lệch, tần suất, cùng các dạng đồ thị giúp trực quan hóa dữ liệu.

1. **Các đại lượng trung tâm**

   Những đại lượng trung tâm thông dụng trong thống kê bao gồm: kỳ vọng (arithmetic mean), trung vị (median), và mốt (mode).

   - Kỳ vọng bằng tổng tất cả các giá trị trong mẫu chia cho số giá trị của mẫu. Nếu ta có bảng phân phối xác suất của đại lượng ngẫu nhiên X

X	x_1	x_2	...	x_n
P	p_1	p_2	...	p_n

     thì kỳ vọng của X là $E(X) = x_1p_1 + x_2p_2 + \ldots + x_np_n$. Giá trị kỳ vọng thường được sử dụng làm ước lượng cho trung bình, tuy nhiên nó dễ bị ảnh hưởng từ các giá trị ngoại lai.
   - Trung vị là giá trị nằm ngay chính giữa của mẫu, chia mẫu ra làm hai nửa, một nửa nhỏ hơn và một nửa lớn hơn trung vị. Thông thường, với bộ dữ liệu có số điểm dữ liệu là lẻ, điểm nằm ở trung tâm được chọn làm trung vị, còn khi số điểm dữ liệu là chẵn, thì trung bình của 2 điểm trung tâm được chọn làm trung vị.

     Một đặc trưng của trung vị là nó ít chịu ảnh hưởng bởi các điểm ngoại lai, do đó trong nhiều trường hợp trung vị là ước lượng trung tâm tốt hơn so với kỳ vọng. Khi kỳ vọng nhỏ hơn trung vị, đồ thị sẽ lệch trái (lefted-skew), và ngược lại.
   - Mode là giá trị có tần suất lớn nhất trong mẫu. Mode hữu dụng khi giá trị các quan sát không có thứ tự dễ thấy (chẳng hạn như phân loại cam, táo, xoài,...). Trong phân phối chuẩn, các giá trị kỳ vọng, trung vị, mode đều trùng nhau.

   Ngoài ra, ta còn thường xét các điểm tứ phân vị: tứ phân vị thứ nhất Q_1, thứ hai Q_2 (trung vị), thứ ba Q_3 chia tập dữ liệu thành 4 phần có số lượng bằng nhau. Từ đó ta tính khoảng tứ phân vị $\Delta Q = Q_3 - Q_1$. Từ đó ta có 1 cách xác định điểm ngoại lai, đó la các điểm nằm ngoài khoảng $(Q_1 - 1.5\Delta Q, Q_3 + 1.5\Delta Q)$.

2. **Độ lệch chuẩn**

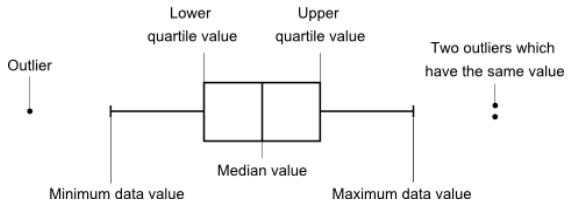
   Các giá trị trung tâm cho ta biết vị trí tập trung của các điểm dữ liệu, nhưng để biết độ phân tán của dữ liệu quanh trung tâm như thế nào, ta cần các đại lượng phương sai và độ lệch chuẩn.

   Phương sai được tính bằng trung bình các bình phương chênh lệch giữa giá trị cụ thể và kỳ vọng. Độ lệch chuẩn được tính bằng căn bậc hai của phương sai, cho ta biết khoảng cách trung bình giữa một điểm dữ liệu đến kỳ vọng.

3. **Các đồ thị thường dùng**

   Trong thống kê mô tả, các biểu đồ thường được sử dụng để trực quan hóa các đặc trưng, tính chất, làm rõ hơn những biểu hiện, quan hệ của các khía cạnh trong tập dữ liệu.

- **Histogram:** Là dạng biểu đồ thể hiện sự phân bố tần suất của dữ liệu bằng các cột. Khoảng giá trị của tập dữ liệu được chia thành các "bucket", sau đó đếm số giá trị thuộc vào mỗi bucket trên. Histogram giúp ta ước lượng hàm mật độ xác suất của tập dữ liệu.
- **Box plot:** Là biểu đồ biểu diễn dữ liệu dưới dạng các hình chữ nhật, trong đó đoạn thẳng ở giữa hộp là trung vị; hai cạnh hai bên là hai tứ phân vị thứ nhất và thứ ba; hai đầu mút của râu (whisker) là giá trị nhỏ nhất và giá trị lớn nhất không tính điểm ngoại lai; và các điểm ngoại lai được biểu diễn bởi các dấu chấm bên ngoài hộp. Hình dạng của box plot thể hiện sự phân bố của dữ liệu và các điểm ngoại lai. Box plot hữu ích khi cần so sánh giữa các khoảng khác nhau trong tập dữ liệu.



Hình 1: Biểu diễn tập dữ liệu bằng Boxplot

- **Scatter plot:** Là loại biểu đồ thường dùng để biểu diễn mối quan hệ giữa 2 biến trong tập dữ liệu, với một biến được biểu diễn bởi trục Ox, biến còn lại trên trục Oy. Dạng biểu đồ này được sử dụng khi cần xem xét sự tương quan giữa hai biến trong cùng tập dữ liệu, đánh giá mức độ tương quan mạnh yếu, thuận nghịch giữa chúng.

2.2 Về thống kê suy diễn

2.2.1 Khoảng tin cậy

Định nghĩa: Cho tham số θ cần khảo sát của tổng thể và $X_1, X_2, ..., X_n$ là các quan sát ngẫu nhiên. Với độ tin cậy γ khoảng (a, b) được gọi là khoảng ước lượng (hay khoảng tin cậy) của tham số θ nếu:

$$P(\theta \in (a, b)) = \gamma$$

Xác suất để tham số θ nằm trong trong khoảng (a, b) là γ. Ngược lại, xác suất để ước lượng tham số θ sai được gọi là mức ý nghĩa, ký hiệu là α. Ta có $\alpha + \gamma = 1$. Công thức tổng quát cho mọi khoảng tin cậy là:

Ước lượng điểm $\pm$ (nhân tố độ tin cậy)(sai số chuẩn)

Bảng tóm tắt các bài toán tìm khoảng tin cậy đối xứng (trung bình, tỷ lệ):

Dạng	Giả định	Độ chính xác	Khoảng tin cậy
Tỷ lệ	$n > 30$	$\varepsilon = z_{\alpha/2} \cdot \sqrt{\frac{f(1-f)}{n}}$	$f - \varepsilon < p < f + \varepsilon$
Trung bình	(1)	$\varepsilon = z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$	$\overline{x} - \varepsilon < \mu < \overline{x} + \varepsilon$
	(2)	$\varepsilon = t_{\alpha/2;n-1} \cdot \frac{s}{\sqrt{n}}$	$\overline{x} - \varepsilon < \mu < \overline{x} + \varepsilon$
	(3)	$\varepsilon = z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$	$\overline{x} - \varepsilon < \mu < \overline{x} + \varepsilon$

Lưu ý đối với bài toán ước lượng trung bình dạng (3), trường hợp chưa biết σ thì thay bằng s.
Lưu ý khác:

Giả định (1): $X_i \sim N(\mu, \sigma^2)$, đã biết σ^2
Giả định (2): $X_i \sim N(\mu, \sigma^2)$, chưa biết σ^2
Giả định (3): Phân phối tuỳ ý, mẫu lớn ($n \geq 30$)

2.2.2 Kiểm định giả thuyết thống kê

Giả thuyết thống kê là một phát biểu về các đặc trưng, tham số của một hoặc nhiều tổng thể. Trong một bài toán kiểm định, ta sẽ ra hai giả thuyết: giả thuyết không H_0 và đối giả thuyết H_1 . Áp dụng quy trình kiểm định phù hợp sẽ đưa ra kết luận tốt dựa trên các quan sát từ mẫu. Quy trình kiểm định được diễn ra như sau.

1. Xây dựng giả thuyết thống kê.
   (a) Giả thuyết không H_0: không có tác động, hay sự khác biệt giữa các tham số từ tổng thể đang xét với tham số mẫu, hoặc giữa các tổng thể.
   (b) Đối giả thuyết H_1: Có sự khác biệt.
2. Thu thập mẫu dựa trên quy trình.
3. Sử dụng quy trình kiểm định phù hợp dựa trên tính chất của biến quan sát. Từ đó tính toán các tham số tương ứng như là miền bác bỏ (RR: Reject Region), p-value, ...
4. So sánh kết quả từ tổng thể với phân phối từ mô hình. Ở đây, ta có thể gặp một số sai lầm như sau:
   (a) Sai lầm loại I: Bác bỏ H_0 đúng
   (b) Sai lầm loại II: Chấp nhận H_0 sai
5. Phân tích sâu cho bộ dữ liệu bằng các quy trình khác (nếu cần).

Mức ý nghĩa α được định nghĩa là

$$\alpha = \text{P(sai lầm loại I)} = \text{P(bác bỏ } H_0 \text{ đúng)}$$

2.2.2.1 Kiểm định nhị phân một mẫu - Binomial Testing Quy trình kiểm định được sử dụng khi xác định tỷ lệ phần trăm p_0 của tổng thể X mang đặc tính A nào đó với mức ý nghĩa α. Quy trình này tính toán chính xác các giá trị đặc trưng hơn là quy trình kiểm định tỷ lệ một mẫu.

Dựa trên quan sát từ mẫu, ta thấy phần tử mang đặc tính A hay không, trong n phép thử giống với phép thử Bernoulli. Gọi Y là số lần xuất hiện biến cố trên ,thì $Y \sim \text{B}(n, p)$ và $\hat{P} = \frac{Y}{n}$ là một ước lượng không chệch cho p.

Ở đây, ta có hai giả thuyết như sau:

1. Giả thuyết H_0: $\pi = \pi_0$
2. Giả thuyết H_1: $\pi < \pi_0$

Để kiểm định, ta sẽ đi tính p-value cho trường hợp này như sau:

$$p = \sum_{i=0}^{k} \Pr(X = i) = \sum_{i=0}^{k} \binom{n}{i} \pi_0^i (1 - \pi_0)^{n-i}$$

Sau đó, ta so sánh với p-value kỳ vọng (thường là 0.05). Nếu p-value của mẫu lớn hơn thì ta kết luận rằng, dữ liệu từ mẫu không có đủ bằng chứng để bác bỏ giả thuyết H_0. Ngược lại, ta sẽ bác bỏ giả thuyết H_0.

2.2.2.2 Phân tích phương sai một nhân tố Quy trình được sử dụng để so sánh trung bình của nhiều tổng thể, đánh giá mức độ tương quan, ảnh hưởng của một biến định lượng lên môt biến định tính.

Tổng thể được giả định là tuân theo phân phối chuẩn, phương sai của các tổng thể bằng nhau, và các quan sát là độc lập.

Gọi $\mu_1, \mu_2, \mu_3, ..., \mu_n$ là trung bình của các tổng thể. Các giả thuyết được phát biểu như sau:

$$H_0 : \mu_1 = \mu_2 = \mu_3 = ... = \mu_n$$
$$H_1 : \exists i, \exists j : \mu_i \neq \mu_j$$

Sau đó, ta tính toán các thông số đặc trưng của quy trình và đưa quyết định như sau.

Bảng mô hình phân tích phương sai một nhân tố:

Nguồn của sự biến thiên	**SS**	**df**	**MS**	**F**
Giữa các nhóm	SSB	$k-1$	MSB	$F = \frac{MSB}{MSW}$
Trong từng nhóm	SSW	$N-k$	MSW	
Toàn bộ	SST	$N-1$		

Xác định miền bác bỏ của bài toán và đưa ra kết luận

$$RR = (F_{\alpha;k-1;N-k}; +\infty) \text{ hay } F > F_{\alpha;k-1;N-k}$$

1. Nếu $F > F_{\alpha;k-1;N-k} \Leftrightarrow F \in RR \Rightarrow$ Bác bỏ H_0, chấp nhận H_1
2. Nếu $F < F_{\alpha;k-1;N-k} \Leftrightarrow F \notin RR \Rightarrow$ Không bác bỏ H_0

Phân tích bội là một quy trình nhằm so sánh các giá trị trung bình của tất cả các cặp với giả thuyết tương ứng như sau

$$H_0 : \mu_i = \mu_j$$
$$H_1 : \mu_i \neq \mu_j, \text{ với } i \neq j$$

Thực chất, phân tích bội thực hiện kiểm định hai mẫu cho tất cả các cặp trong bộ dữ liệu. Ở bài báo cáo này, nhóm sẽ sử dụng thư viện có sẵn t-test để thực hiện chức năng này.

2.2.2.3 Kiểm định về tính độc lập của hai nhân tố Quy trình kiểm định chi bình phương được sử dụng cho bài toán xác định tính độc lập của hai nhân tố định tính , tức là xác định mức độ tương quan của hai biến này.

Quy trình này sẽ thống kê dữ liệu theo dạng bảng, trong đó các hàng, các cột tương ứng với các đặc tính quan sát của hai biến đó thông qua bảng tương quan (Contingency table). Bộ dữ liệu sẽ được chia thành nhiều nhóm thông qua các đặc tính trên, và số lượng của chúng được gọi là tần suất quan sát (Observed frequency - O).

Sau đó, mô hình sẽ tính tần suất kỳ vọng (Expected frequency - E) cho các nhóm đặc tính khác nhau thông qua xác suất lý thuyết. Công thức xác định cụ thể như sau:

$$E = \frac{(Row\_total * Col\_total)}{Num\_observations}$$

chỉ số thống kê Chi-square (X^2) là mọt đặc tính quan trọng, thể hiện mức độ khác biệt của tần suất quan sát và tần suất kỳ vọng của các nhóm.Quy trình sẽ ước lượng chỉ số thống kê Chi-square (X^2) bằng công thức sau:

$$X^2 = \sum \frac{(O-E)^2}{E}$$

Dưới đây là ví dụ mình họa cho quy trình.

	Y	$\overline{Y}$	
X	O_1 (E_1)	O_2 (E_2)	$O_1 + O_2$
$\overline{X}$	O_3 (E_3)	O_4 (E_4)	$O_3 + O_4$
	$O_1 + O_3$	$O_2 + O_4$	

Hình 2: Minh họa bảng tương quan cho kiểm định Chi Bình Phương

Sau khi có được các thông số tương ứng, ta sẽ so sánh chỉ số thống kê Chi-Square với giá trị cực hạn (critical value) thông qua việc xác định mức ý nghĩa α và bậc tự do df. Ta có thể xem qua các giá trị cực hạn thông qua bảng dưới đây. Nếu $X_O^2 > X_C^2$ (chi-square value greater than the critical value), thể hiện rằng có sự khác biệt lớn giữa tần suất ghi nhận và tần suất kỳ vọng trong một nhóm nào đó, ta có đủ bằng chứng để bác bỏ giả thuyết H_0.

	Significance level (α)							
Degrees of freedom (*df*)	.99	.975	.95	.9	.1	.05	.025	.01
1	-------	0.001	0.004	0.016	2.706	3.841	5.024	6.635
2	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210
3	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345
4	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277
5	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086
6	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812
7	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475
8	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090
9	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666
10	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209
11	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725
12	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217
13	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688
14	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141
15	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578
16	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000
17	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409
18	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805
19	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191
20	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566
21	8.897	10.283	11.591	13.240	29.615	32.671	35.479	38.932
22	9.542	10.982	12.338	14.041	30.813	33.924	36.781	40.289
23	10.196	11.689	13.091	14.848	32.007	35.172	38.076	41.638
24	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.980
25	11.524	13.120	14.611	16.473	34.382	37.652	40.646	44.314
26	12.198	13.844	15.379	17.292	35.563	38.885	41.923	45.642
27	12.879	14.573	16.151	18.114	36.741	40.113	43.195	46.963
28	13.565	15.308	16.928	18.939	37.916	41.337	44.461	48.278
29	14.256	16.047	17.708	19.768	39.087	42.557	45.722	49.588
30	14.953	16.791	18.493	20.599	40.256	43.773	46.979	50.892
40	22.164	24.433	26.509	29.051	51.805	55.758	59.342	63.691
50	29.707	32.357	34.764	37.689	63.167	67.505	71.420	76.154
60	37.485	40.482	43.188	46.459	74.397	79.082	83.298	88.379
70	45.442	48.758	51.739	55.329	85.527	90.531	95.023	100.425
80	53.540	57.153	60.391	64.278	96.578	101.879	106.629	112.329
100	61.754	65.647	69.126	73.291	107.565	113.145	118.136	124.116
1000	70.065	74.222	77.929	82.358	118.498	124.342	129.561	135.807

Hình 3: Điểm tới hạn của phân phối chi-square dựa trên bậc tự do

2.2.3 Hồi quy logistic nhị phân đa biến

Hồi quy logistic nhị phân là một phương pháp thống kê được sử dụng để mô hình hóa mối quan hệ giữa một biến phụ thuộc nhị phân (có hai giá trị 0 hoặc 1) và một hoặc nhiều biến độc lập[5]. Phương pháp này giúp ước lượng xác suất xảy ra một sự kiện cụ thể dựa trên các biến giải thích, đồng thời khắc phục các hạn chế của hồi quy tuyến tính khi biến phụ thuộc không phải là biến liên tục mà chỉ có hai trạng thái.

Trong bài toán này, biến phụ thuộc là `Patient_Status` với hai trạng thái "Alive" (0) và "Dead" (1), nhằm dự đoán xác suất bệnh nhân ung thư vú tử vong dựa trên các đặc trưng như tuổi, các chỉ số protein, giai đoạn khối u, loại mô học, trạng thái HER2, loại phẫu thuật và khoảng thời gian khác biệt giữa ngày phẫu thuật và ngày tái khám.

Mô hình hồi quy logistic nhị phân được biểu diễn bằng hàm logit, là logarit của tỉ số chênh lệch (odds) giữa xác suất xảy ra sự kiện và xác suất không xảy ra sự kiện:

$$logit(p) = \log\left(\frac{1-p}{p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k$$

trong đó:

- $p = P(Y = 1|X)$ là xác suất biến phụ thuộc Y nhận giá trị 1 (ví dụ: bệnh nhân tử vong) với các biến độc lập $X = (X_1, X_2, \ldots, X_k)$.
- β_0 là hệ số chặn (intercept).
- β_i là hệ số hồi quy tương ứng với biến độc lập X_i.

Hàm logit là hàm chuyển đổi giúp biến đổi xác suất p từ khoảng a sang khoảng giá trị thực $(-\infty, +\infty)$, giúp mô hình hóa mối quan hệ tuyến tính giữa biến độc lập và log-odds của biến phụ thuộc.

Ngược lại, xác suất p có thể được tính từ logit thông qua hàm logistic (hàm sigmoid):

$$p = logit^{-1}(\eta) = \frac{1}{1+e^{-\eta}} = \frac{e^{\eta}}{1+e^{\eta}}$$

,với $\eta = \beta_0 + \Sigma_{i=1}^{k} \beta_i X_i$.

Đây là hàm đặc trưng của hồi quy logistic, đảm bảo kết quả dự đoán luôn nằm trong khoảng từ 0 đến 1, phù hợp với xác suất.

2.2.3.1 Ước lượng tham số Trong hồi quy logistic thường được thực hiện bằng phương pháp *maximum likelihood estimation* (MLE), tối đa hóa hàm khả năng (likelihood function) của mô hình dựa trên dữ liệu quan sát.

Trong hồi quy logistic, việc ước lượng các tham số $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_k)$ không được thực hiện bằng phương pháp bình phương tối thiểu như hồi quy tuyến tính, mà thay vào đó sử dụng phương pháp *Ước lượng hợp lý cực đại* (Maximum Likelihood Estimation - MLE) [6].

Ý tưởng cơ bản của MLE là tìm bộ tham số $\boldsymbol{\beta}$ sao cho mô hình logistic dự đoán xác suất xảy ra các kết quả quan sát trong dữ liệu là cao nhất. Cụ thể, với mỗi quan sát i có biến phụ thuộc nhị phân $y_i \in 0, 1$ và biến độc lập $X_i = (X_{i1}, \ldots, X_{ik})$, xác suất dự đoán được bởi mô hình là:

$$p_i = P(Y_i = 1|X_i) = \frac{1}{1 + e^{-(\beta_0 + \Sigma_{j=1}^{k} \beta_j \cdot X_{ij})}}$$

Do các quan sát được giả định độc lập, hàm hợp lý (likelihood function) của toàn bộ dữ liệu là tích xác suất của từng quan sát:

$$L(\beta) = \prod_{i=1}^{n} p_i^{y_i} \cdot (1 - p_i)^{1-y_i}$$

Để thuận tiện trong tính toán, ta thường sử dụng log-likelihood:

$$\ell(\beta) = \log\left(L(\beta)\right) = \sum_{i=1}^{n} \left[y_i \cdot \log(p_i) + (1 - y_i) \cdot \log(1 - p_i)\right]$$

Phương pháp MLE sẽ tìm bộ tham số $\boldsymbol{\beta}$ sao cho $\ell(\boldsymbol{\beta})$ đạt giá trị cực đại, tức là làm cho dữ liệu quan sát được có xác suất cao nhất dưới mô hình logistic.

$$\hat{\boldsymbol{\beta}} = \arg\max_{\boldsymbol{\beta}} \ell(\boldsymbol{\beta})$$

Quá trình tối ưu này thường được thực hiện bằng các thuật toán số như *gradient descent* hoặc *Newton-Raphson*.

Trong phần mềm R, khi sử dụng hàm `glm(..., family = "binomial")`, quá trình ước lượng tham số logistic được thực hiện tự động dựa trên MLE. Kết quả trả về là các ước lượng tham số $\hat{\boldsymbol{\beta}}$ tối đa hóa log-likelihood, giúp mô hình có khả năng dự đoán tốt nhất trên dữ liệu huấn luyện.

2.2.4 Mô hình cây quyết định - Decision Tree [8]

2.2.4.1 Giới thiệu Mô hình cây quyết định là một trong những mô hình học máy phổ biến, được sử dụng trong thống kê và khai phá dữ liệu. Mô hình này có thể được áp dụng trong cả bài toán phân loại và hồi quy. Bài toán được sử dụng trong Bài tập lớn này là phân loại.

Mô hình cây quyết định biểu diễn dưới dạng cấu trúc cây phân cấp, trong đó mỗi nút nội bộ là một điều kiện kiểm tra thuộc tính, mỗi nhánh đại diện cho một giá trị của thuộc tính và mỗi nút lá tương ứng với kết quả dự đoán (nhãn đối với phân loại hoặc giá trị đối với hồi quy). Khi dự đoán, ta bắt đầu từ nút gốc và lần lượt kiểm tra các điều kiện tại mỗi nút theo chiều xuống dưới cho đến khi đến một nút lá và lấy kết quả cuối cùng. Quá trình xây dựng cây thường nhằm chia tập dữ liệu thành các vùng con càng "thuần nhất" (hầu hết thuộc cùng một lớp) càng tốt.

Cây quyết định có thể được xây dựng bằng cách sử dụng các thuật toán như ID3, C4.5, CART, v.v. Mỗi thuật toán có cách tiếp cận riêng để chọn thuộc tính phân chia và điều kiện dừng xây dựng cây [2].

2.2.4.2 Thuật toán xây dựng cây quyết định

Các thuật toán xây dựng cây

- **Thuật toán ID3 (Iterative Dichotomiser 3)**: Thuật toán do Quinlan đề xuất, thuật toán này xây dựng cây bằng cách chọn lần lượt thuộc tính có lợi ích thông tin (Information Gain) cao nhất để tách dữ liệu. Cụ thể, ID3 tính entropy của tập dữ liệu khi tách theo mỗi thuộc tính, chọn thuộc tính có entropy con thấp nhất (tức IG lớn nhất).
- **Thuật toán C4.5**: Là phiên bản mở rộng của ID3 cũng do Quinlan phát triển, C4.5 khắc phục những hạn chế của ID3. Thuật toán C4.5 xây dựng cây giống ID3, nhưng dùng gain ratio làm tiêu chí để chọn thuộc tính phân chia.

- **Thuật toán CART (Classification and Regression Trees)**: Thuật toán do Breiman và cộng sự phát triển, được sử dụng cho cả phân loại và hồi quy, sử dụng độ Gini (phân loại) hoặc độ lệch bình phương (hồi quy) làm tiêu chí phân chia. Và đây cũng là thuật toán được sử dụng để xây dựng cây quyết định trong Bài tập lớn.

Tiêu chí chia nhánh: để lựa chọn thuộc tính và ngưỡng tách tối ưu tại mỗi nút.

- **Entropy**: Entropy của tập dữ liệu S có C lớp được định nghĩa như sau:

$$Entropy(S) = -\sum_{i=1}^{C} p_i \log_2 p_i \tag{1}$$

  Trong đó, p_i là xác suất của lớp i trong tập dữ liệu S. Entropy đo lường độ hỗn loạn của một tập hợp. Nếu tất cả các mẫu đều thuộc cùng một lớp, entropy sẽ bằng 0 (tập hợp thuần nhất). Ngược lại, nếu các mẫu phân bố đều giữa các lớp, entropy sẽ cao hơn.

- **Information Gain (IG):** được tính bằng hiệu entropy trước và sau khi tách tập cha S thành các tập con:

$$IG(S, A) = Entropy(S) - \sum_{v \in A} \frac{|S_v|}{|S|} Entropy(S_v) \tag{2}$$

  Trong đó, A là một thuộc tính, v là một giá trị cụ thể của thuộc tính A

- **Gain Ratio**: là tỷ lệ giữa IG và entropy của thuộc tính. Gain Ratio giúp điều chỉnh bias của IG đối với thuộc tính có nhiều giá trị.

$$GainRatio(S, A) = \frac{IG(S, A)}{Entropy(A)} \tag{3}$$

- **Gini Impurity**: được sử dụng trong thuật toán CART, đo lường mức độ không thuần nhất của một tập hợp. Gini của một tập S được tính bằng:

$$Gini(S) = 1 - \sum_{i=1}^{C} p_i^2 \tag{4}$$

  Trong đó, p_i là tỉ lệ mẫu trong tập dữ liệu S thuộc về lớp thứ i, trong tổng số C lớp. Gini Impurity càng thấp, tập hợp càng thuần nhất.

2.2.5 Kiểm định và đánh giá

Akaike Information Criterion (AIC) là một chỉ số dùng để so sánh các mô hình thống kê khác nhau dựa trên dữ liệu cùng một tập hợp[1]. AIC được tính theo công thức:

$$AIC = 2k - 2ln(\hat{L})$$

trong đó:

- k là số lượng tham số trong mô hình,
- $\hat{L}$ là giá trị hàm khả năng tối đa của mô hình.

AIC đánh đổi giữa độ phù hợp của mô hình với dữ liệu và độ phức tạp của mô hình (số tham số). Mô hình có giá trị AIC thấp hơn được coi là mô hình tốt hơn vì nó mất ít thông tin hơn khi mô hình hóa dữ liệu, đồng thời tránh hiện tượng quá khớp (overfitting).

Ma trận nhầm lẫn (Confusion Matrix) là bảng tóm tắt kết quả dự đoán của mô hình phân loại, thể hiện số lượng dự đoán đúng và sai theo từng lớp [7]. Với bài toán nhị phân, ma trận có dạng:

	Dự đoán sống	**Dự đoán chết**
Thực tế sống	TP	FN
Thực tế chết	FP	TN

Bảng 3: Ma trận nhầm lẫn

Từ ma trận này, ta tính được các chỉ số quan trọng như:

- **Độ chính xác (Accuracy)**: tỉ lệ dự đoán đúng trên tổng số mẫu:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

- **Độ nhạy (Sensitivity)**: khả năng phát hiện đúng các trường hợp sống:

$$Sensitivity = \frac{TP}{TP + FN}$$

- **Độ đặc hiệu (Specificity)**: khả năng phát hiện đúng các trường hợp chết:

$$Specificity = \frac{TN}{TN + FP}$$

Đường cong ROC (Receiver Operating Characteristic Curve) là đồ thị biểu diễn mối quan hệ giữa Tỉ lệ dương tính thực (True Positive Rate - TPR) và Tỉ lệ dương tính giả (False Positive Rate - FPR) ở các ngưỡng phân loại khác nhau. Cụ thể:

$$TPR = \frac{TP}{TP + FN}, \quad FPR = \frac{FP}{FP + TN}$$

Đường cong ROC giúp đánh giá khả năng phân biệt giữa hai lớp của mô hình phân loại[4]. Diện tích dưới đường cong (AUC - Area Under Curve) là một chỉ số tổng quát thể hiện độ tốt của mô hình, với giá trị nằm trong khoảng từ 0.5 (mô hình ngẫu nhiên) đến 1 (mô hình hoàn hảo).

3 Tiền xử lý dữ liệu

Trước khi tiến hành xử lý dữ liệu, ta cài đặt một số thư viện cần thiết như sau: Thư viện "questionr" cho các xử lý cơ bản và "fastDummies" để sử dụng One-hot-encoding, thư viện "psych" để sử dụng describe cho thống kê mô tả, "ggplot2" và "corrplot" để vẽ các đồ thị.

```
install.packages(c("questionr", "fastDummies", "psych", "ggplot2", "corrplot", "
    caret", "caTools", "pROC", "rpart", "rpart.plot"))
library(questionr)     # Used for freq.na, freq
library(fastDummies)   # Used to create dummy variables
library(psych)         # Used for the describe() function for descriptive
    statistics
library(ggplot2)       # Used for plotting graphs
library(corrplot)      # Used for plotting correlation matrix heatmaps
```

Mã nguồn 1: Cài đặt thư viện

Ta tiến hành đọc file dữ liệu đầu vào "BRCA.csv" vào data tên "brc" trong rstudio:

```
brc <- read.csv("~/BRCA/BRCA.csv")
View(brc)
str(brc)
```

Mã nguồn 2: Đọc dữ liệu

Tổng quan về bộ dữ liệu vừa đọc thu được như sau:

```
'data.frame':   341 obs. of  16 variables:
 $ Patient_ID        : chr  "TCGA-D8-A1XD" "TCGA-EW-A1OX" "TCGA-A8-A079" "TCGA-D8-
    A1XR" ...
 $ Age               : int  36 43 69 56 56 84 53 50 77 40 ...
 $ Gender            : chr  "FEMALE" "FEMALE" "FEMALE" "FEMALE" ...
 $ Protein1          : num  0.0804 -0.4203 0.214 0.3451 0.2215 ...
 $ Protein2          : num  0.426 0.578 1.311 -0.211 1.907 ...
 $ Protein3          : num  0.547 0.614 -0.327 -0.193 0.52 ...
 $ Protein4          : num  0.2737 -0.0315 -0.2343 0.1243 -0.312 ...
 $ Tumour_Stage      : chr  "III" "II" "III" "II" ...
 $ Histology         : chr  "Infiltrating Ductal Carcinoma" "Mucinous Carcinoma" "
    Infiltrating Ductal Carcinoma" "Infiltrating Ductal Carcinoma" ...
 $ ER.status         : chr  "Positive" "Positive" "Positive" "Positive" ...
 $ PR.status         : chr  "Positive" "Positive" "Positive" "Positive" ...
 $ HER2.status       : chr  "Negative" "Negative" "Negative" "Negative" ...
 $ Surgery_type      : chr  "Modified Radical Mastectomy" "Lumpectomy" "Other" "
    Modified Radical Mastectomy" ...
 $ Date_of_Surgery   : chr  "15-Jan-17" "26-Apr-17" "08-Sep-17" "25-Jan-17" ...
 $ Date_of_Last_Visit: chr  "19-Jun-17" "09-Nov-18" "09-Jun-18" "12-Jul-17" ...
 $ Patient_Status    : chr  "Alive" "Dead" "Alive" "Alive" ...
```

Mã nguồn 3: Kết quả đọc dữ liệu

Tiến hành xem xét bộ dữ liệu, ta nhận thấy rằng 7 hàng cuối của dữ liệu là dữ liệu rỗng, vì vậy ta sẽ xóa 7 hàng này ra khỏi bộ dữ liệu. Đồng thời, một số ô trong cột có kiểu dữ liệu nhận giá trị là **""** thay vì **NA**, vì vậy, ta cần điền NA cho các ô này rồi mới kiểu tra tỉ lệ thiếu (NA) của bộ dữ liệu.

```
                   missing %    missing %
Age                        7 2          7 2
Protein1                   7 2          7 2
Protein2                   7 2          7 2
Protein3                   7 2          7 2
```

```
Protein4                 7 2          7 2
Patient_ID               0 0          7 2
Gender                   0 0          7 2
Tumour_Stage             0 0          7 2
Histology                0 0          7 2
ER.status                0 0          7 2
PR.status                0 0          7 2
HER2.status              0 0          7 2
Surgery_type             0 0          7 2
Date_of_Surgery          0 0          7 2
Date_of_Last_Visit       0 0         24 7
Patient_Status           0 0         20 6
```

Mã nguồn 4: Tỉ lệ khuyết của bộ dữ liệu trước (trái) và sau (phải) khi điền NA

Ta thấy rằng bộ dữ liệu có tỷ lệ khuyết thấp (cao nhất là 7%), nên ta sẽ tiến hành xử lý khuyết bằng cách xóa khỏi bộ dữ liệu. Mặt khác, ta nhận thấy một điểm khác lạ trong tỷ lệ khuyết như sau: tỷ lệ khuyết của `Date_of_Last_Visit` lớn hơn `Patient_Status`. Trong phần Tổng quan, nhóm đã mô tả rằng biến `Patient_Status` có thể rỗng nếu `Date_of_Last_Visit` rỗng (vì bệnh nhân vẫn còn sống tại thời điểm phẫu thuật nhưng không quay lại tái khám nên không có thông tin về tình trạng bệnh nhân). Vì vậy, theo lẽ thường thì tỷ lệ khuyết của `Patient_Status` phải lớn hơn hoặc bằng `Date_of_Last_Visit`. Thông qua kiểm tra bộ dữ liệu, nhóm nhận thấy rằng, có 4 vị trí mà bệnh nhân không tái khám nhưng vẫn biết tình trạng bệnh nhân (đều là đã mất). Vì không biết được nguyên nhân khuyết `Date_of_Last_Visit` tại vị trí đó (bệnh nhân đã mất lúc phẫu thuật nhưng điền thiếu hay biết được tình trạng bệnh nhân qua cách khác) nên nhóm không khôi phục dữ liệu mà vẫn xóa chung với các dữ liệu khuyết khác.

```r
freq.na(brc)
brc[brc == ""] <- NA
freq.na(brc)
# Xoa du lieu khuyet
brc <- na.omit(brc)
# Kiem tra lai ty le khuyet va ty le trung Patient_ID
freq.na(brc)
any(duplicated(brc$Patient_ID))
```

Mã nguồn 5: Xử lý dữ liệu khuyết

Sau khi xóa dữ liệu khuyết, ta tiến hành kiểm tra lại tỷ lệ khuyết và đồng thời kiểm tra xem có hai bệnh nhân nào trùng `Patient_ID` không. Kết quả thu được là không. Vậy ta đã xử lý được vấn đề khuyết và trùng lắp của bộ dữ liệu.

Ta tiến hành kiểm tra tỷ lệ giá trị của các biến bằng cách sử dụng lệnh `freq(brc$Ten_bien)`, theo đó, ta nhận thấy vấn đề ở các biến `Gender`, `ER.status`, `PR.status` như sau:

```
> freq(brc$Gender)
         n    % val%
FEMALE 313 98.7 98.7
MALE     4  1.3  1.3
> freq(brc$ER.status)
           n   % val%
Positive 317 100  100
> freq(brc$PR.status)
           n   % val%
Positive 317 100  100
```

Mã nguồn 6: Tỉ lệ giá trị của Gender, ER.status, PR.status

Có thể thấy rằng, tỷ lệ nhận giá trị là Positive của `ER.status` và `PR.status` là 100%, tỷ lệ nhận giá trị FEMALE của `Gender` cũng đạt tới 98.7%. Vì tỉ lệ đều ở mức tuyệt đối nên 3 biến trên sẽ không mang lại nhiều ý nghĩa trong quá trình dự đoán tỷ lệ sống sót. Vì vậy, nhóm quyết định xóa 3 biến trên cùng với biến `Patient_ID` (vì chỉ có ý nghĩa định danh) ra khỏi bộ dữ liệu.

```
brc <- subset(brc, select = -ER.status)
brc <- subset(brc, select = -PR.status)
brc <- subset(brc, select = -Gender)
brc <- subset(brc, select = -Patient_ID)
```

Mã nguồn 7: Cài đặt thư viện

Như vậy, sau quá trình xử lý dữ liệu khuyết, loại bỏ các biến không cần thiết, nhóm đã thu được bộ dữ liệu mới gồm 317 mẫu và 12 biến (5 biến định lượng, 7 biến phân loại). Tiếp theo đây nhóm sẽ tiến hành kiểm tra, xử lý và mã hóa các biến định lượng và biến phân loại.

Với biến định lượng, sau khi tiến hành kiểm tra giá trị ngoại lai, ta thu được kết quả gồm 7 giá trị ngoại lai ở biến `Protein1`, 5 giá trị ở `Protein3` và 4 ở `Protein4`. Số lượng là không đáng kể nên nhóm sẽ không xóa các giá trị này ra khỏi bộ dữ liệu.

Với biến phân loại, trước hết ta xét đến 2 biến chỉ có 2 loại giá trị là `HER2.status` và `Patient_Status`. Vì chỉ có 2 loại giá trị nên ta chỉ cần xử lý một cách đơn giản là Label Encoding chúng về dạng 0 hoặc 1. Với biến `HER2.status`, giá trị 0 tương ứng với Negative và 1 là Positive. Với biến `Patient_Status`, 0 là Dead và 1 là Alive.

```
brc$Patient_Status <- match(brc$Patient_Status, c("Dead", "Alive"))
brc$Patient_Status <- brc$Patient_Status - 1
brc$HER2.status    <- match(brc$HER2.status, c("Negative", "Positive"))
brc$HER2.status    <- brc$HER2.status - 1
```

Mã nguồn 8: Label Encoding biến HER2.status và Patient_Status

Với 3 biến phân loại có 3-4 loại giá trị là `Tumour_Stage`, `Histology` và `Surgery_type`, ta sẽ xử lý chúng bằng One-hot-encoding. Rất may mắn là thư viện fastDummies đã cung cấp sẵn câu lệnh cho chúng ta sử dụng. Ở đây nhóm chọn cờ `remove_first_dummy` để xóa bớt 1 biến giá trị được tạo ra, giúp làm giảm số lượng biến cho bộ dữ liệu.

```
brc <- dummy_cols(brc, select_columns = c("Tumour_Stage", "Histology", "Surgery_
    type"), remove_first_dummy = TRUE)
# Xoa cac bien cu sau khi da one-hot-encoding
brc <- subset(brc, select = -Tumour_Stage)
brc <- subset(brc, select = -Histology)
brc <- subset(brc, select = -Surgery_type)
```

Mã nguồn 9: One-hot-encoding biến Tumour_Stage, Histology và Surgery_type

Ta còn lại hai biến phân loại cuối cùng chưa xử lý là `Date_of_Surgery` và `Date_of_Last_Visit`. Đây đều là hai biến biểu thị ngày tháng năm nên ngoài việc tách chúng thành 3 biến ngày, tháng và năm, ta sẽ tính thêm khoảng cách thời gian giữa lần phẫu thuật và lần khám cuối cùng.

```
# Chuyen du lieu tu chr -> Date
brc$Date_of_Last_Visit <- as.Date(brc$Date_of_Last_Visit, format = "%d-%b-%y")
brc$Date_of_Surgery    <- as.Date(brc$Date_of_Surgery, format = "%d-%b-%y")
# Tach ra ngay, thang, nam
brc$Surgery_Day        <- as.numeric(format(brc$Date_of_Surgery, "%d"))
brc$Surgery_Month      <- as.numeric(format(brc$Date_of_Surgery, "%m"))
brc$Surgery_Year       <- as.numeric(format(brc$Date_of_Surgery, "%y"))
```

```r
brc$Last_Visit_Day      <- as.numeric(format(brc$Date_of_Last_Visit, "%d"))
brc$Last_Visit_Month    <- as.numeric(format(brc$Date_of_Last_Visit, "%m"))
brc$Last_Visit_Year     <- as.numeric(format(brc$Date_of_Last_Visit, "%y"))
# Tinh khoang cach giua Date_of_Surgery va Date_of_Last_Visit
brc$Difference_Days     <- as.numeric(brc$Date_of_Last_Visit - brc$Date_of_Surgery)
# Xoa cac bien cu sau khi da xu ly
brc <- subset(brc, select = -Date_of_Surgery)
brc <- subset(brc, select = -Date_of_Last_Visit)
```

Mã nguồn 10: Xử lý biến ngày tháng Date_of_Surgery và Date_of_Last_Visit

Đến đây, nhóm đã hoàn tất các bước tiền xử lý dữ liệu, bước cuối cùng sẽ dành cho việc chỉnh sửa tên biến và thay đổi thứ tự các biến về với thứ tự ban đầu để thuận tiện cho việc quan sát và xử lý dữ liệu sau này.

```r
# Rut gon ten bien do one-hot-encoding tao ra
colnames(brc)[colnames(brc) == "Histology_Infiltrating Lobular Carcinoma"] <- "
    Histology_L"
colnames(brc)[colnames(brc) == "Histology_Mucinous Carcinoma"] <- "Histology_M"
colnames(brc)[colnames(brc) == "Surgery_type_Modified Radical Mastectomy"] <- "
    Surgery_type_M"
colnames(brc)[colnames(brc) == "Surgery_type_Simple Mastectomy"] <- "Surgery_type_
    S"
colnames(brc)[colnames(brc) == "Surgery_type_Other"] <- "Surgery_type_O"
# Doi vi tri cac bien
brc <- brc[, c(1, 2, 3, 4, 5, 8, 9, 10, 11, 6, 12, 13, 14, 15, 16, 17, 18, 19, 20,
     21, 7)]
```

Mã nguồn 11: Chỉnh sửa tên biến, thứ tự biến

Cuối cùng, bộ dữ liệu Breast Cancer Data ban đầu đã hoàn thành tiền xử lý dữ liệu. Từ bộ dữ liệu với 341 mẫu và 16 biến, nhóm đã tiến hành xử lý và đưa ra bộ dữ liệu gồm 317 mẫu và 21 biến. Sau đây, nhóm sẽ sử dụng bộ dữ liệu đã qua xử lý để tiến hành thống kê.

4 Thống kê mô tả

4.1 Các đại lượng thống kê mô tả

4.1.1 Thống kê mô tả các biến định lượng

Thực hiện tính toán các thống kê mô tả bao gồm trung bình mẫu, độ lệch chuẩn, các phân vị, giá trị nhỏ nhất, giá trị lớn nhất cho các biến liên tục.

```r
# 1. Lay cac cot can thiet
continuous_vars <- c("Age", "Protein1", "Protein2", "Protein3", "Protein4", "
    Difference_Days")
brc_continuous <- brc[, continuous_vars]

# 2. Dinh nghia ham tinh toan thong ke
calculate_custom_stats <- function(x) {
  q <- quantile(x, probs = c(0.25, 0.75), na.rm = TRUE, type = 7)
  stats <- c(
    Mean = mean(x, na.rm = TRUE),
    SD = sd(x, na.rm = TRUE),
    Min = min(x, na.rm = TRUE),
    Q1 = q[1],
    Median = median(x, na.rm = TRUE),
    Q3 = q[2],
    Max = max(x, na.rm = TRUE)
  )
  return(stats)
}

# 3. Ap dung ham cho tung cot
stats_list <- lapply(brc_continuous, calculate_custom_stats)

# 4. Ket hop ket qua thanh bang
summary_table <- do.call(rbind, stats_list)

# 5. Chuyen thanh dataframe
summary_df <- as.data.frame(summary_table)

# 6. In ket qua
print(round(summary_df, 3))
```

Mã nguồn 12: Tính toán giá trị thống kê cho các biến định lượng

	Mean	SD	Min	Q1.25%	Median	Q3.75%	Max
Age	58.726	12.827	29.000	49.000	58.000	67.000	90.000
Protein1	-0.027	0.544	-2.145	-0.351	0.006	0.336	1.594
Protein2	0.950	0.906	-0.979	0.369	0.997	1.612	3.402
Protein3	-0.095	0.589	-1.627	-0.531	-0.193	0.251	2.193
Protein4	0.007	0.626	-2.026	-0.382	0.039	0.436	1.630
Difference_Days	447.776	386.279	0.000	189.000	372.000	595.000	3019.000

Hình 4: Kết quả TKMT các biến định lượng

Dưới đây là phân tích chi tiết về đặc điểm phân bố của các biến định lượng chính trong tập dữ liệu:

- **Age (Tuổi)**:

  - **Khoảng giá trị**: Tuổi của bệnh nhân dao động từ 29 đến 90 tuổi. Giá trị trung bình là 58.7 tuổi, rất gần với giá trị trung vị (58.0 tuổi).
  - **Phân phối**: Sự tương đồng giữa giá trị trung bình và trung vị cho thấy phân bố tuổi của bệnh nhân trong mẫu dữ liệu này là **tương đối đối xứng**. Độ lệch chuẩn (SD = 12.8 tuổi) cho thấy có sự biến động vừa phải về độ tuổi quanh giá trị trung tâm. 50% số bệnh nhân trung tâm có độ tuổi nằm trong khoảng từ 49 đến 67 tuổi (Khoảng tứ phân vị - IQR).
  - **Kết luận**: Biến Age có phân bố khá cân đối, tập trung chủ yếu quanh độ tuổi cuối 50, với phần lớn bệnh nhân nằm trong nhóm tuổi từ cuối 40 đến cuối 60.

- **Protein1**:
  - **Khoảng giá trị**: Giá trị của Protein1 biến động trong khoảng từ -2.145 đến 1.594. Giá trị trung bình (-0.027) và trung vị (0.006) đều rất gần với 0 và gần nhau.
  - **Phân phối**: Giá trị trung bình và trung vị gần bằng nhau và gần 0 cho thấy phân bố của Protein1 **khá đối xứng quanh giá trị 0**. Độ lệch chuẩn (SD = 0.544) tương đối lớn so với giá trị trung bình (gần 0), thể hiện mức độ phân tán đáng kể của dữ liệu. Khoảng tứ phân vị (IQR) từ -0.351 đến 0.336 cho thấy 50% giá trị trung tâm tập trung trong một khoảng khá hẹp quanh 0.
  - **Kết luận**: Dữ liệu Protein1 phân bố tương đối đối xứng, tập trung quanh giá trị 0. Mặc dù giá trị trung tâm gần 0, dữ liệu có độ phân tán đáng kể.

- **Protein2**:
  - **Khoảng giá trị**: Protein2 có giá trị nằm trong khoảng từ -0.979 đến 3.402. Giá trị trung bình (0.950) và trung vị (0.997) khá gần nhau.
  - **Phân phối**: Sự chênh lệch nhỏ giữa trung bình và trung vị (trung bình hơi thấp hơn) gợi ý một độ lệch trái nhẹ, nhưng nhìn chung phân bố có thể xem là **tương đối đối xứng** quanh giá trị 1. Độ lệch chuẩn (SD = 0.906) khá lớn, gần bằng giá trị trung bình, cho thấy sự biến thiên cao của giá trị Protein2. Khoảng tứ phân vị (IQR) từ 0.369 đến 1.612 cho thấy một nửa số quan sát nằm trong khoảng này.
  - **Kết luận**: Biến Protein2 có độ phân tán cao, với các giá trị tập trung quanh 1.0 và phân bố tương đối cân đối. Phần lớn giá trị nằm giữa 0.369 và 1.612.

- **Protein3**:
  - **Khoảng giá trị**: Giá trị của Protein3 dao động từ -1.627 đến 2.193. Giá trị trung bình (-0.095) lớn hơn một chút so với giá trị trung vị (-0.193).
  - **Phân phối**: Giá trị trung bình cao hơn trung vị cho thấy phân bố của Protein3 có xu hướng **lệch phải** (đuôi phân bố kéo dài về phía các giá trị dương lớn hơn). Độ lệch chuẩn (SD = 0.589) cho thấy mức độ phân tán đáng kể quanh giá trị trung tâm (hơi âm). Khoảng cách từ Q3 (0.251) đến Max (2.193) lớn hơn nhiều so với khoảng cách từ Min (-1.627) đến Q1 (-0.531), càng khẳng định độ lệch phải.
  - **Kết luận**: Phân bố của Protein3 bị lệch về bên phải, với trung tâm dữ liệu hơi thấp hơn 0. Phần lớn giá trị tập trung trong khoảng -0.531 đến 0.251, tuy nhiên có sự xuất hiện của các giá trị dương lớn hơn làm kéo dài đuôi phân phối.

- **Protein4**:

- **Khoảng giá trị**: Protein4 có giá trị trong khoảng từ -2.026 đến 1.630. Giá trị trung bình (0.007) và trung vị (0.039) rất gần nhau và gần 0.
- **Phân phối**: Sự tương đồng giữa trung bình và trung vị cho thấy phân bố của Protein4 là **khá đối xứng quanh giá trị 0**. Độ lệch chuẩn (SD = 0.626) cho thấy độ phân tán đáng kể so với giá trị trung tâm gần 0. Khoảng tứ phân vị (IQR) từ -0.382 đến 0.436 bao gồm 50% giá trị dữ liệu trung tâm.
- **Kết luận**: Biến Protein4 có phân bố tương đối đối xứng, tập trung quanh giá trị 0. Dữ liệu có độ phân tán đáng kể, với khoảng giá trị điển hình nằm giữa -0.382 và 0.436.

- **Difference_Days (Số ngày giữa lần khám cuối và ngày phẫu thuật)**:
  - **Khoảng giá trị**: Khoảng thời gian này dao động rất rộng, từ 0 ngày đến 3019 ngày. Giá trị trung bình (447.8 ngày) cao hơn đáng kể so với giá trị trung vị (372.0 ngày).
  - **Phân phối**: Sự chênh lệch lớn giữa trung bình và trung vị là dấu hiệu rõ ràng cho thấy phân bố của biến này **lệch phải mạnh**. Điều này có nghĩa là phần lớn bệnh nhân có khoảng thời gian theo dõi/sống sót sau phẫu thuật tương đối ngắn, nhưng có một số ít trường hợp với khoảng thời gian rất dài, kéo giá trị trung bình lên cao. Độ lệch chuẩn (SD = 386.3 ngày) rất lớn, gần bằng giá trị trung bình, cũng phản ánh sự biến động rất cao và phân bố trải rộng. Khoảng tứ phân vị (IQR) từ 189 đến 595 ngày cho thấy 50% bệnh nhân có thời gian theo dõi trong khoảng này. Giá trị Max (3019) rất xa Q3, xác nhận đuôi phân phối dài về bên phải.
  - **Kết luận**: Biến Difference_Days có độ biến thiên rất lớn và phân bố lệch phải rõ rệt. Mặc dù một nửa số bệnh nhân có thời gian theo dõi từ khoảng 6 tháng (189 ngày) đến gần 1 năm 8 tháng (595 ngày), sự tồn tại của các trường hợp theo dõi rất lâu (lên đến hơn 8 năm) làm tăng đáng kể giá trị trung bình và thể hiện tính không đồng đều của dữ liệu này.

4.1.2 Thống kê mô tả các biến định tính

Thực hiện lập bảng thống kê số lượng đối với các biến định tính gốc (`HER2.status`, `Tumour_Stage`, `Histology`, `Surgery_type`, `Patient_Status`) từ tập dữ liệu `dummy_dataset`. Kết quả tần số thu được từ R được trình bày dưới đây:

```
> # Table of HER2.status
> print(table(dummy_dataset$HER2.status))

Negative Positive
     288       29
> # Table of Tumour_Stage
> print(table(dummy_dataset$Tumour_Stage))

  I  II III
 60 180  77
>
> # Table of Histology
> print(table(dummy_dataset$Histology))

 Infiltrating Ductal Carcinoma   Infiltrating Lobular Carcinoma  Mucinou Carcinoma
                           224                               81                 12
>
> # Table of Surgery_type
> print(table(dummy_dataset$Surgery_type))

                 Lumpectomy        Modified Radical Mastectomy         Other
```

```
                        66                                      89             97
          Simple Mastectomy
                        65
>
> # Table of Patient_Status
> print(table(dummy_dataset$Patient_Status))

Alive  Dead
  255    62
```

Mã nguồn 13: Kết quả bảng tần số cho các biến định tính.

Nhận xét kết quả:
Dựa trên kết quả bảng tần số được trình bày trong ở trên, ta có nhận xét về phân bố của các biến định tính như sau:

- **HER2.status (Tình trạng HER2)**:
  - Tình trạng HER2 được phân loại thành hai nhóm:
    - Negative: 288 bệnh nhân
    - Positive: 29 bệnh nhân
  - Tình trạng "Negative" chiếm tỉ lệ áp đảo so với "Positive" trong tập dữ liệu đã xử lý.
- **Tumour_Stage (Giai đoạn khối u)**:
  - Số lượng bệnh nhân được phân loại theo 3 giai đoạn chính:
    - Giai đoạn I: 60 bệnh nhân
    - Giai đoạn II: 180 bệnh nhân
    - Giai đoạn III: 77 bệnh nhân
  - Giai đoạn II là phổ biến nhất trong tập dữ liệu này, chiếm số lượng lớn nhất. Giai đoạn I có số lượng ít nhất.
- **Histology (Mô bệnh học)**:
  - Các loại mô bệnh học chính được ghi nhận:
    - Infiltrating Ductal Carcinoma: 224 trường hợp
    - Infiltrating Lobular Carcinoma: 81 trường hợp
    - Mucinous Carcinoma: 12 trường hợp
  - Loại *Infiltrating Ductal Carcinoma* chiếm đa số tuyệt đối trong mẫu dữ liệu. Loại *Mucinous Carcinoma* xuất hiện rất ít.
- **Surgery_type (Loại phẫu thuật)**:
  - Các loại phẫu thuật được thực hiện bao gồm:
    - Lumpectomy: 66 ca
    - Modified Radical Mastectomy: 89 ca
    - Other (Loại khác): 97 ca
    - Simple Mastectomy: 65 ca

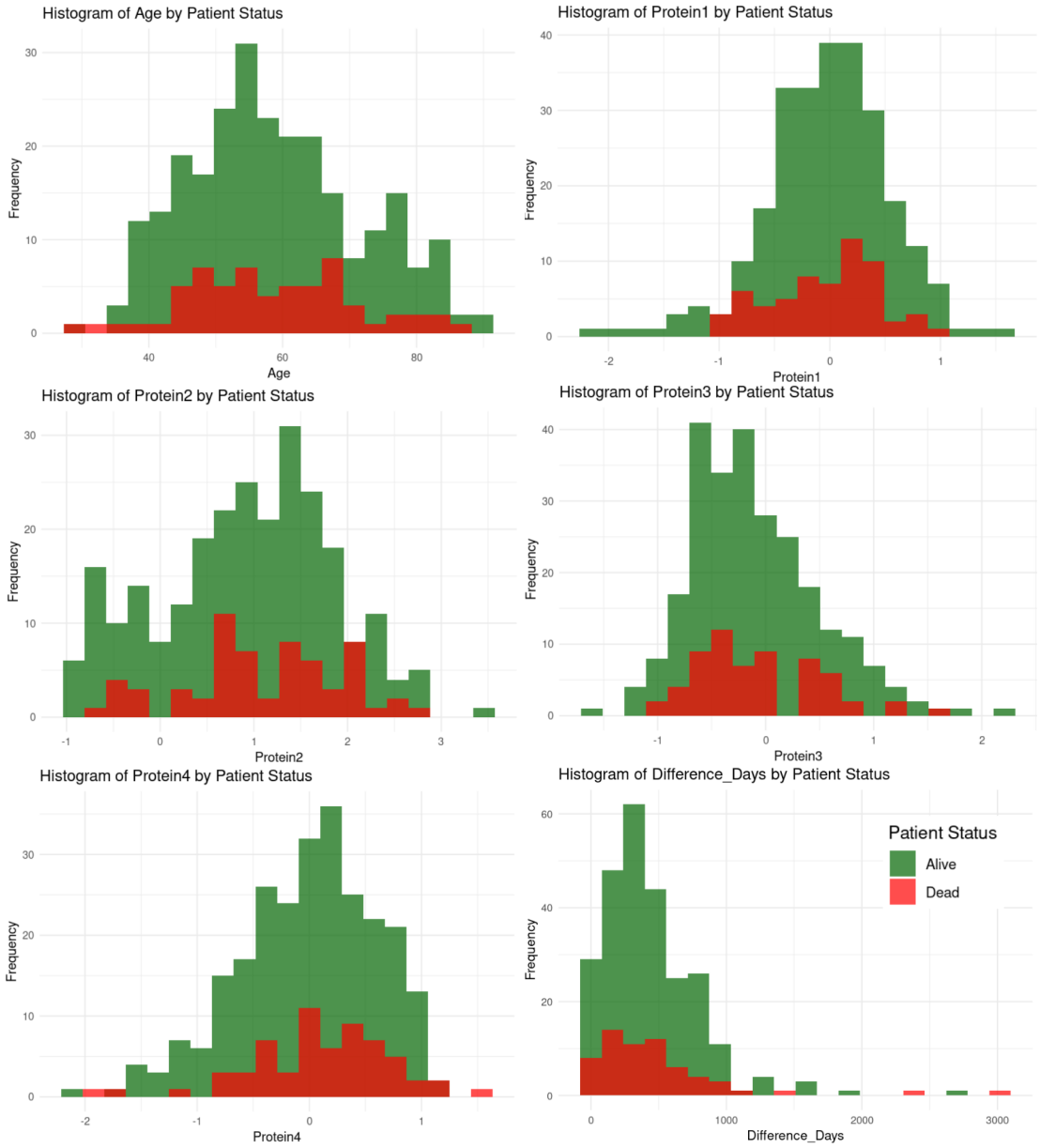
  - Số lượng các ca phẫu thuật được phân bố tương đối đồng đều giữa các loại, tuy nhiên nhóm "Other" (các loại phẫu thuật khác không thuộc 3 nhóm chính còn lại) có số lượng cao nhất. *Simple Mastectomy* và *Lumpectomy* có số lượng ít hơn một chút.

- **Patient_Status (Tình trạng bệnh nhân)**:
  - Tình trạng của bệnh nhân tại thời điểm ghi nhận lần cuối:
    * Alive (Còn sống): 255 bệnh nhân
    * Dead (Đã mất): 62 bệnh nhân
  - Số lượng bệnh nhân còn sống cao hơn đáng kể so với số lượng bệnh nhân đã mất trong tập dữ liệu đã qua xử lý (loại bỏ NA).

4.2 Trực quan hóa dữ liệu bằng đồ thị

Để có cái nhìn tổng quan, toàn diện về các biến cũng như mối liên hệ giữa chúng, nhóm đã quyết định sử dụng các kiểu đồ thị sau để trực quan hóa bộ dữ liệu, cụ thể:

4.2.1 Biểu đồ historgram cho các biến liên tục



Hình 5: Histogram các biến liên tục cùng trạng thái bệnh nhân

Ta sử dụng package `ggplot2` để vẽ biểu đồ histogram. Dùng vòng lặp để chạy qua các biến

liên tục, với mỗi biến ta tạo một histogram biểu diễn biến đó cùng với trạng thái của bệnh nhân (Dead/Alive).

```
print("--- Plotting Histograms ---")
# Loop through continuous variables and plot histograms colored by Patient_Status
for (v in continuous_vars) {
  # Create clear labels for Patient_Status on the plot
  plot_data <- brc
  plot_data$Patient_Status_Label <- factor(plot_data$Patient_Status, levels = c(1,
      0), labels = c("Alive", "Dead"))

  p <- ggplot(plot_data, aes_string(x = v, fill = "Patient_Status_Label")) +
    # geom_histogram(binwidth = 1, alpha = 0.7, position = "identity") + #
        binwidth = 1 might be too small for some variables
    geom_histogram(alpha = 0.7, position = "identity", bins = 20) + # Use number
        of bins instead of fixed binwidth
    labs(title = paste("Histogram of", v, "by Patient Status"),
         x = v,
         y = "Frequency",
         fill = "Patient Status") + # Set legend title
    scale_fill_manual(values = c("Alive" = "darkgreen", "Dead" = "red")) + #
        Ensure colors match labels
    theme_minimal() # Use a cleaner theme
  print(p) # Display the plot
}
```

Mã nguồn 14: Mã R vẽ histogram

Nhận xét:

- **Histogram độ tuổi với tình trạng bệnh nhân**
  - Đồ thị trông như lệch về phía người cao tuổi, nhưng nhìn chung thì đồ thị tương đối tuân theo phân phối chuẩn.
  - Tỉ lệ người chết có vẻ cao hơn đối với nhóm người cao tuổi, nhưng không thể hiện quá rõ ràng trên đồ thị.

  Đồ thị có vẻ phần nào cho biết người cao tuổi có khả năng thuộc nhóm Dead cao hơn.

- **Histogram protein 1 với tình trạng bệnh nhân**

  Cả nhóm Alive và Dead đều phân phối xung quanh mức 0, với nhóm Dead hơi lệch trái. Có khả năng các bệnh nhân với chỉ số protein 1 lớn hơn 0 có khả năng tử vong cao hơn.

- **Histogram protein 2 với tình trạng bệnh nhân**

  Protein 2 có phân phối xung quanh giá trị 1 và gần giống hình chuông, với lượng người bị ung thư có chỉ số Protein 2 nhỏ hơn 0 có vẻ lớn hơn so với lượng người bị ung thư có chỉ số Protein 2 lớn hơn 2. Tương tự, tỉ lệ người chết trong nhóm >2 có vẻ cao hơn.

- **Histogram protein 3 với tình trạng bệnh nhân**

  Phân phối protein 3 trông tương đối lệch phải; và ta cũng thấy lượng người chết trong nhóm <0 có vẻ lớn hơn nhóm >0. Điều này gợi ý rằng người có chỉ số protein 3 thấp có khả năng ung thư cao hơn và tử vong nhiều hơn.

- **Histogram protein 4 với tình trạng bệnh nhân**

  Nhóm Alive và Dead đều phân phối khoảng xung quanh điểm 0, với đồ thị tương đối lệch trái. Điều này có thể cho thấy người có chỉ số protein 4 cao có khả năng bị ung thư vú cao hơn.

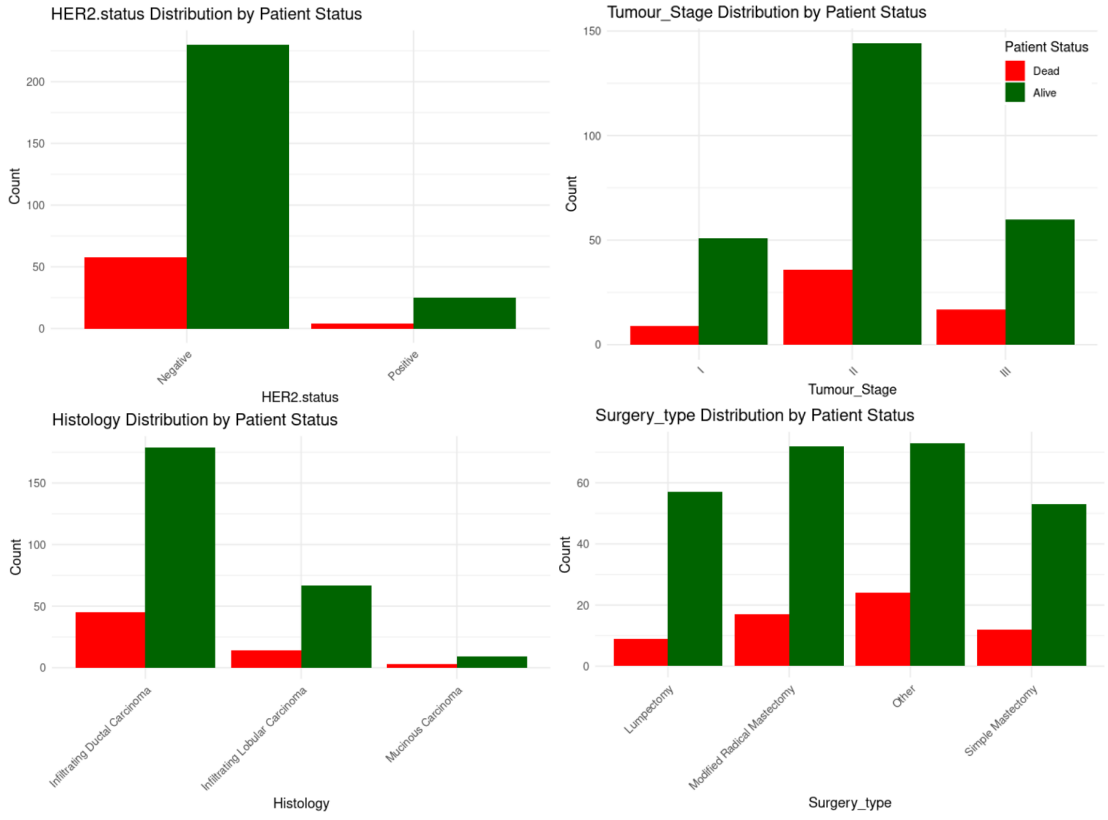
- **Histogram khoảng thời gian với tình trạng bệnh nhân**

  Đồ thị lệch phải một cách rõ rệt, tập trung quanh khoảng 300-400 ngày, từ ngày phẫu thuật đến ngày tái khám cuối cùng. Có nhiều điểm dữ liệu ngoại lai ngoài khoảng 1000 ngày, với điểm cao nhất là khoảng 3000 ngày. Nhìn chung dữ liệu này có vẻ không cho ta biết nhiều về tình trạng sống chết của bệnh nhân.

Các đồ thị hầu hết tuân theo phân phối chuẩn và hơi lệch trái hoặc phải, với số lượng người sống và người chết có tỉ lệ khá tương đồng. Từ các đồ thị này ta chưa thấy rõ được sự tương quan giữa các biến liên tục cụ thể nào đến tình trạng sống hoặc chết của bệnh nhân.

4.2.2 Biểu đồ cột cho các biến phân loại

Ta chọn các biến phân loại để vẽ biểu đồ cột. Lưu ý rằng các biến `Tumor_Stage`, `Histology`, và `Surgery_type` đã được mã hóa thành dạng one-hot encoding ở bộ dữ liệu sau tiền xử lý, do đó ta cần dùng dữ liệu các biến trước khi xử lý mã hóa. Ta cũng sử dụng `ggplot2` để vẽ biểu đồ cột.



Hình 6: Biểu đồ cột biến phân loại

```r
print("--- Plotting Bar Charts for Categorical Variables ---")

# Create a function to generate bar charts for categorical variables
create_bar_chart <- function(data, var_name) {
  # Convert Patient_Status to a factor with labels
```

```r
  data$Patient_Status_Label <- factor(data$Patient_Status,
                                      levels = c("Dead", "Alive"),
                                      labels = c("Dead", "Alive"))

  # Create the plot
  p <- ggplot(data, aes_string(x = var_name, fill = "Patient_Status_Label")) +
    geom_bar(position = "dodge") + # 'dodge' places bars side-by-side
    labs(title = paste(var_name, "Distribution by Patient Status"),
         x = var_name,
         y = "Count",
         fill = "Patient Status") +
    scale_fill_manual(values = c("Alive" = "darkgreen", "Dead" = "red")) +
    theme_minimal() +
    theme(axis.text.x = element_text(angle = 45, hjust = 1)) # Rotate labels for
        better readability

  return(p)
}

# HER2.status bar chart (using dummy_dataset instead of brc)
print("--- Plotting Bar Chart for HER2.status ---")
p_her2 <- create_bar_chart(dummy_dataset, "HER2.status")
print(p_her2)

# Tumour_Stage bar chart
print("--- Plotting Bar Chart for Tumour_Stage ---")
p_stage <- create_bar_chart(dummy_dataset, "Tumour_Stage")
print(p_stage)

# Histology bar chart
print("--- Plotting Bar Chart for Histology ---")
p_histology <- create_bar_chart(dummy_dataset, "Histology")
print(p_histology)

# Surgery_type bar chart
print("--- Plotting Bar Chart for Surgery_type ---")
p_surgery <- create_bar_chart(dummy_dataset, "Surgery_type")
print(p_surgery)
```

Mã nguồn 15: Mã R vẽ biểu đồ cột

Nhận xét:

- **HER2.status:** Từ biểu đồ, ta thấy rõ sự chênh lệch về số lượng bệnh nhân có trạng thái HER2 âm tính so với HER2 dương tính. Đa số bệnh nhân có HER2 âm tính, gấp khoảng 10 lần số người dương tính. Tuy vậy, ta không thấy được rõ ảnh hưởng của HER2 lên tình trạng của bệnh nhân. Theo biểu đồ, tỉ lệ sống và chết của bệnh nhân là tương đương nhau giữa các cột âm tính và dương tính. Nếu nhìn kỹ có thể thấy tỉ lệ chết trong nhóm positive là nhỏ hơn so với nhóm negative, tuy vậy sự chênh lệch lớn giữa số lượng mẫu 2 nhóm khiến ta chưa thể đưa ra kết luận chắc chắn.

- **Tumour_Stage:** Trong 3 giai đoạn I, II, III, thì giai đoạn 2 chiếm tỉ lệ nhiều nhất trong bộ dữ liệu. Có thể dự đoán rằng bệnh nhân thường đến khám và điều trị ung thư khi đã đến giai đoạn 2 vì giai đoạn 1 khó nhận biết ung thư, giai đoạn 2 khối u phát triển dễ nhận diện, và phần lớn bệnh nhân điều trị trước khi để ung thư phát triển đến giai đoạn 3. Tuy nhiên, ta có thể thấy tỉ lệ tử vong giai đoạn 1 thấp hơn so với giai đoạn 2 và 3 do tình trạng ung thư diễn biến chưa phức tạp và dễ kiểm soát hơn.

- **Histology:** Ta thấy rằng Infiltrating Ductal Carcinoma (IDC) (Ung thư ống xâm lấn) chiếm tỉ lệ cao nhất, giống với số liệu thực tế (70-80%)[3], tiếp đến là Infiltrating Lobular Carcinoma (ILC) (Ung thư tiểu thùy xâm lấn), và cuối cùng là Mucinous Carcinoma (Ung thư nhầy) chiếm tỉ lệ nhỏ, cũng giống thực tế (khoảng 5%).
- **Surgery_type:** Ta thấy Modified Radical Mastectomy (phẫu thuật cắt vú triệt để) chiếm tỉ trọng cao hơn Lumpectomy (phẫu thuật bảo tồn vú) và Simple Mastectomy (cắt bỏ vú đơn giản). Lý do có thể là vì Lumpectomy và Simple Mastectomy chỉ dùng khi tình trạng ung thư còn ở giai đoạn sớm (I, II), còn Modified Radical Mastectomy dùng để điều trị khi tình trạng ung thư lan đến bộ phận khác.

4.2.3 Biểu đồ boxplot cho các biến liên tục

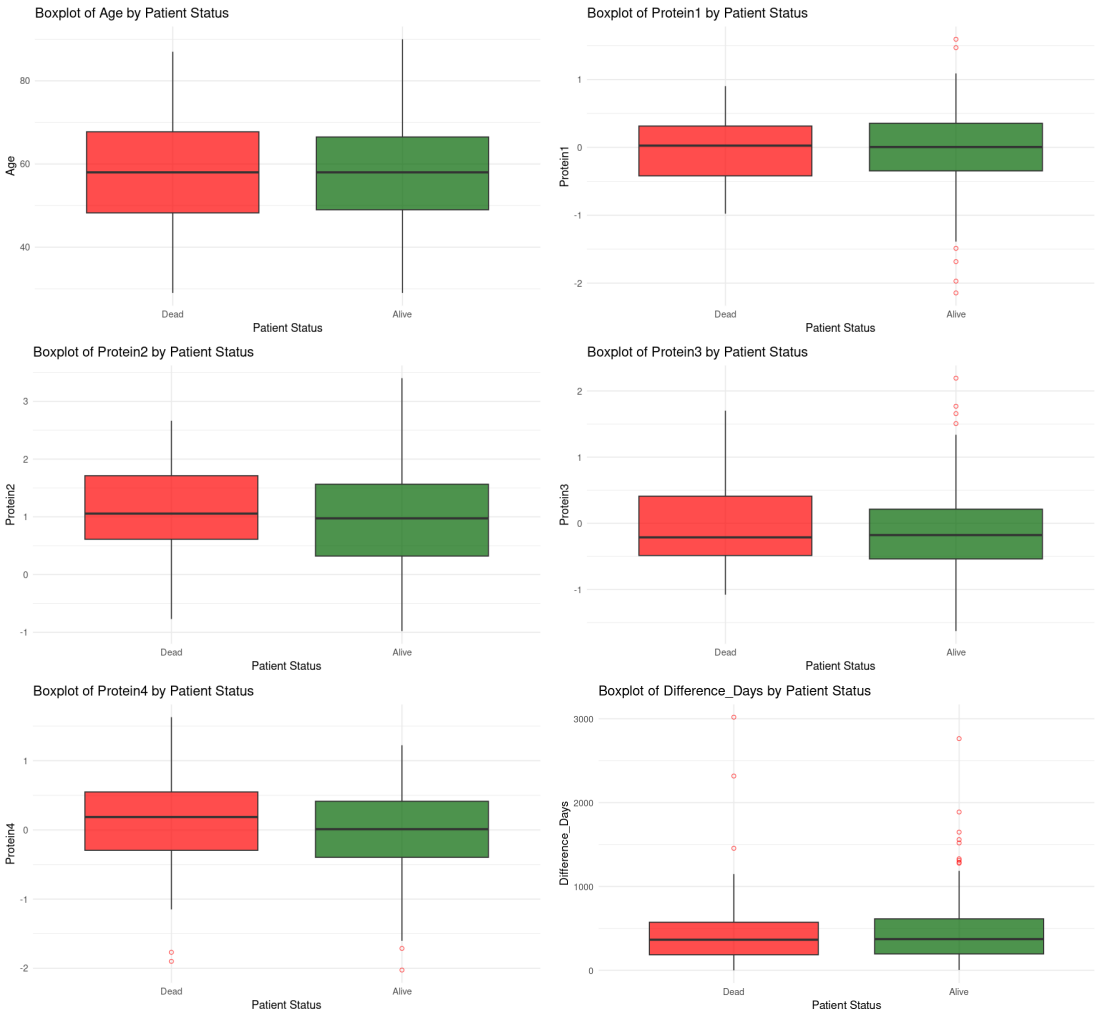
Boxplot biểu thị các đại lượng đặc trưng như kỳ vọng, khoảng tứ phân vị, điểm ngoại lai... do đó ta chỉ chọn các biến liên tục cho đồ thị này. Các đồ thị được vẽ bằng thư viện `ggplot`, thể hiện các boxplot chia ra 2 trạng thái của bệnh nhân. 6 biến liên tục bao gồm Age, Protein 1, Protein 2, Protein 3, Protein 4, Difference Days.

```
print("--- Plotting Boxplots ---")
# Loop through continuous variables and plot boxplots grouped by Patient_Status
for (v in continuous_vars) {
  plot_data <- brc
  plot_data$Patient_Status_Label <- factor(plot_data$Patient_Status, levels = c(0,
      1), labels = c("Dead", "Alive"))

  p <- ggplot(plot_data, aes_string(x = "Patient_Status_Label", y = v, fill = "
      Patient_Status_Label")) +
    geom_boxplot(alpha = 0.7, outlier.colour = "red", outlier.shape = 1) + # Show
        outliers in red
    labs(title = paste("Boxplot of", v, "by Patient Status"),
         x = "Patient Status",
         y = v) + # No fill legend needed as it's on the x-axis
    scale_fill_manual(values = c("Alive" = "darkgreen", "Dead" = "red")) +
    theme_minimal() +
    theme(legend.position = "none") # Hide the legend
  print(p)
}
```

Mã nguồn 16: Mã R vẽ boxplot

Nhận xét:

- **Boxplot độ tuổi với tình trạng bệnh nhân**
  - Nhóm Dead có trung vị tuổi gần bằng nhóm Alive (khoảng 60).
  - Độ rộng IQR của hai nhóm khá tương đồng nhưng nhóm Dead dao động rộng hơn, chịu sự dao động lớn từ khoảng 50-70 tuổi.
  - Cả hai nhóm đều có vài cá thể rất trẻ (< 40) và rất già (> 80), nhưng nhóm Alive có phần cột râu (whisker) hơi dài hơn, gợi ý độ biến thiên tuổi trong nhóm này lớn hơn.
- **Boxplot protein 1 với tình trạng bệnh nhân**
  - Nhóm Alive có trung vị Protein1 tương đương với Dead (gần 0).
  - Nhóm Alive IQR rộng tương đương Dead, nhưng khoảng IQR của Alive nằm hơi cao hơn, có thể gợi ý rằng người sống sót có chỉ số Protein 1 cao hơn.



Hình 7: Boxplot các biến liên tục và trạng thái bệnh nhân

- Nhóm Alive xuất hiện nhiều giá trị ngoại lai dương và cả vài giá trị âm sâu (< -2). Nhóm "Dead" hầu như không có ngoại lai, chủ yếu tập trung quanh –1 đến +1.

- **Boxplot protein 2 với tình trạng bệnh nhân**
  - Trung vị ở cả hai nhóm tương đương (xung quanh 1), song nhóm Dead hơi cao hơn một chút.
  - IQR nhóm Alive rộng hơn, chứng tỏ phân tán Protein2 lớn hơn ở bệnh nhân sống. IQR nhóm Alive cũng trông thấp hơn so với Dead, cho thấy người sống có xu hướng có chỉ số Protein 2 thấp hơn.
  - Nhóm Alive có nhiều giá trị cao (> 3) hơn so với nhóm Dead.

- **Boxplot protein 3 với tình trạng bệnh nhân**
  - Nhóm Dead và Alive có trung vị âm tương đương nhau (khoảng –0.2).

  - Dead có IQR rộng hơn và cao hơn so với Alive. Chứng tỏ nhóm Dead có chỉ số Protein 3 phân tán nhiều hơn và giá trị có phần cao hơn nhóm Alive.
  - Nhóm Alive có một số giá trị ngoại lai dương lớn (> 2), trong khi nhóm Dead không thấy xuất hiện ngoại lai.

- **Boxplot protein 4 với tình trạng bệnh nhân**
  - Có thể thấy trung vị Protein4 của nhóm Dead cao hơn nhóm Alive, với Dead nằm ở gần 0.2 còn Alive gần 0.
  - IQR của Dead nằm ở khoảng cao hơn, cho thấy sự khác biệt trong phân tán giữa hai nhóm. Có thể thấy trung bình người Dead có mức Protein 4 cao hơn so với Alive một chút.
  - Cả hai đều có vài outliers âm sâu (< -2), không có nhiều outliers dương. Cả giá trị lớn nhất và nhỏ nhất của Dead đều lớn hơn Alive.

- **Boxplot khoảng thời gian với tình trạng bệnh nhân**
  - Hai nhóm có trung vị gần bằng nhau, khoảng 350–400 ngày.
  - IQR của nhóm "Alive" hơi rộng hơn một chút, có thể nói bệnh nhân sống có thời gian dao động lớn hơn.
  - Nhóm Dead có một số ngoại lai lớn (trên 2.000–3.000 ngày), và nhóm Alive cũng có nhiều ngoại lai.

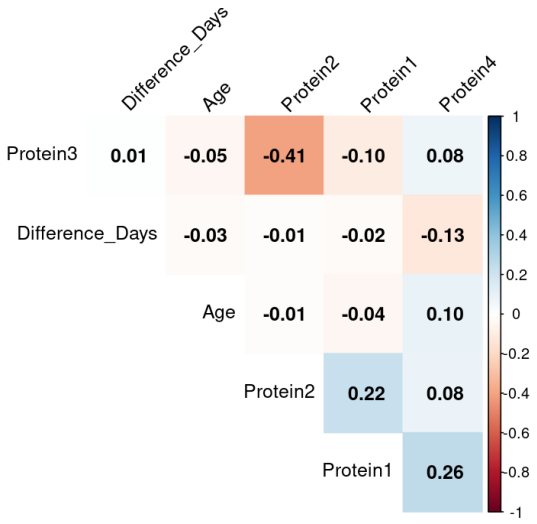
Về mặt tuổi tác, bệnh nhân tử vong có xu hướng cao tuổi hơn đôi chút nhưng phân bố khá tương đồng. Các protein cho thấy sự khác biệt nhẹ về trung vị và độ phân tán giữa hai nhóm, đặc biệt với Protein 2 và Protein 4, gợi ý những chỉ số này có thể liên quan đến tiên lượng bệnh. Difference_Days không cho thấy sự khác biệt rõ rệt ở trung vị, nhưng bệnh nhân sống có nhiều trường hợp theo dõi rất dài (outliers cao).

4.2.4 Ma trận tương quan

```r
print("--- Calculating and Plotting Correlation Matrix ---")
# Only calculate correlation between continuous variables
cor_data <- brc[, continuous_vars]
correlation_matrix <- cor(cor_data)

print("Correlation Matrix:")
print(round(correlation_matrix, 2)) # Round for better readability

# Plot the correlation matrix heatmap
print("Plotting correlation matrix heatmap...")
corrplot(correlation_matrix,
         method = "color",       # Display using colors
         addCoef.col = "black",  # Show correlation coefficients in black
         type = "upper",         # Show only the upper triangle of the matrix
         order = "hclust",       # Reorder variables based on hierarchical
             clustering
         tl.col = "black",       # Color of variable names (text labels)
         tl.srt = 45,            # Rotation angle for text labels
         diag = FALSE)           # Do not display the main diagonal
title("Correlation Matrix Heatmap of Continuous Variables", line = 5)
```

Mã nguồn 17: Mã R vẽ heatmap

Ta chỉ sử dụng các biến liên tục trong biểu đồ tương quan. Ta sử dụng `corrplot` để vẽ đồ thị tương quan giữa các biến.



Hình 8: Ma trận tương quan

Nhận xét:

- Giữa các protein có mối tương quan mạnh nhất trong bảng. Cụ thể, hệ số tương quan giữa protein 2 và protein 3 là -0.41 có thấy mức độ tương quan nghịch vừa phải giữa 2 loại protein này, cho thấy nếu protein 2 cao thì protein 3 có xu hướng thấp, và ngược lại. Các cặp protein 1 và 4 (0.26) và cặp protein 1 và 2 (0.22) cũng có độ tương quan đáng kể.
- Độ tuổi hầu như không có tác động hay mối tương quan đáng kể đến các loại protein, với các hệ số tương quan bé hơn 0.1. Tương tự, khoảng cách giữa ngày phẫu thuật đến ngày tái khám cuối cùng hầu như không có ý nghĩa đối với các chỉ số protein.

5 Thống kê suy diễn

Mục tiêu đề tài

Xác định mức độ ảnh hưởng của các yếu tố chính đến tỷ lệ tử vong sau phẫu thuật. Các yếu tố chính được xét như là tuổi tác và các giai đoạn bệnh được phát hiện tương ứng. Từ đó, khuyến khích người dân, ở đây cụ thể là khu vực Bắc Ireland, kiểm tra sức khỏe định kỳ và có biện pháp chữa trị kịp thời.

5.1 Khoảng tin cậy và kiểm định giả thuyết thống kê

5.1.1 Tỷ lệ tử vong

Xác định được tỷ lệ sống, chết sau phẫu thuật thể hiện được mức độ nghiêm trọng của vấn đề, có thể chịu nhiều ảnh hưởng bởi các yếu tố như tuổi tác, biện pháp chữa trị, giai đoạn bệnh, ... Dù nguyên nhân là gì, nếu tỷ lệ tử vong sau phẫu thuật cao, điều này cần được lưu ý nhằm kịp thời đưa ra các giải pháp can thiệp phù hợp.

Khoảng tin cậy cho tỷ lệ tổng thể

Bài toán: Ước lượng tỷ lệ tử vong sau phẫu thuật của bệnh nhân nữ tại khu vực Bắc Ireland.

Ở đây, nhóm tính toán từng đặc trưng của mẫu để lấy được khoảng tin cậy cho tỷ lệ tử vong của mẫu. Với giả định là mẫu tuân theo phân phối chuẩn, ta sử dụng phương pháp tính khoảng tin cậy cho bài toán một mẫu.

```r
num_patients    <- length(brc$Patient_Status)
dead_patients   <- sum(brc$Patient_Status == 0)
f               <- dead_patients / num_patients
z_alpha_2       <- qnorm(p= 0.05/2, lower.tail = F)
Epsilon         <- z_alpha_2 * sqrt( f*(1-f)/num_patients )
data.frame(n, dead_patients, f, z_alpha_2,
           Epsilon, Left_CI = f - Epsilon, Right_CI = f + Epsilon)
```

Mã nguồn 18: Tính khoảng tin cậy cho tỷ lệ tử vong trong R

Kết quả của đoạn chương trình:

Bảng 4: Khoảng tin cậy 95% cho tỷ lệ tử vong sau phẫu thuật

n	dead_patients	f	z_alpha_2	Epsilon	Left_CI	Right_CI
317	62	0.1955836	1.959964	0.04366416	0.1519194	0.2392478

Nhận xét: Trong tập dữ liệu, có khoảng 19.5% số bệnh nhân đã không qua khỏi sau phẫu thuật trong giai đoạn từ 2017 - 2021. Với khoảng độ tin cậy 95%, khoảng tin cậy tìm được là $(0.15, 0.24)$.

Kiểm định nhị phân một mẫu

Sau khi tìm hiểu một số thông tin liên quan đến bệnh ung thư vú ở khu vực bắc Ireland, nhóm được biết,trong khoảng năm 2018-2022, có khoảng 7,566 ca bệnh và trong số đó có 1550 ca đã không qua khỏi sau các cuộc phẫu thuật. Tỷ lệ ở quần thể này là khoảng 20.05%. [9]

Dựa vào các phân tích ở trên, nhóm sẽ kiểm định giả thuyết theo các bước sau.

Gọi $\hat{P}$ là tỷ lệ tử vong sau khi phẫu thuật của bộ dữ liệu.

p_0 là tỷ lệ tử vong sau khi phẫu thuật của bệnh nhân nữ người Ireland. Ở đây p_0 là 20.05
Đặt giả thuyết:

1. Giả thuyết H_0: $\hat{P} = p_0$
2. Giả thuyết H_1: $\hat{P} < p_0$

Ý nghĩa: Kiểm định xem là tỉ lệ tử vong của các bệnh nhân nữ mắc bệnh trong tập dữ liệu có tuân với tỉ lệ chuẩn của khu vực không.

Quy trình kiểm định ở đây là kiểm định nhị phân. Lí do nhóm chọn loại kiểm định này vì biến phụ thuộc hiện tại là biến *Patient_Status*, là biến nhị phân, kèm với đó là tập dữ liệu vừa, trên 300 bệnh nhân. Với khả năng tính toán của máy tính, ta hoàn toàn có thể sử dụng phương pháp này. quy trình này tính chính xác các chỉ số của bộ dữ liệu hơn quy trình kiểm định tỷ lệ một mẫu với giả định quy trình tuân theo phân phối chuẩn. Dưới đây là phần hiện thực của đoạn chương trình.

```
dead_patients <- sum(brc$Patient_Status == 0)
num_patients <- length(brc$Patient_Status)
binom.test(dead_patients, num_patients, alternative = "less", p = 0.205)
```

Mã nguồn 19: Kiểm định tỷ lệ tử vong sau phẫu thuật

Kết quả của đoạn chương trình như sau:

```
        Exact binomial test

data:  dead_patients and num_patients
number of successes = 62, number of trials = 317, p-value = 0.3694
alternative hypothesis: true probability of success is less than 0.205
95 percent confidence interval:
 0.0000000 0.2358775
sample estimates:
probability of success
             0.1955836
```

Mã nguồn 20: Kết quả kiểm định tỷ lệ một mẫu sử dụng kiểm định nhị phân

Nhận xét: ở đây ta chỉ cần quan tâm đến thuộc tính chủ yếu là $p\_value = 0.3694 > 0.05$. Như vậy, ta có thể kết luận là không có đủ chứng cứ để bác bỏ giả thuyết H_0, ta chấp nhận tỷ lệ tử vong sau khi phẫu thuật bằng với mức thực tế là 20.05%. Mặc dù tỷ lệ được ước tính trong bộ dữ liệu là 19.56%, lại không không có sự khác biệt ý nghĩa.

Nhận định

Tỷ lệ chết sau phẫu thuật của bệnh nhân ung thư vú ở khu vực Bắc Ireland ở mức cao, khoảng 20%.

5.1.2 Nhóm tuổi

Tuổi tác từ lâu được xem là yếu tố nguy cơ hàng đầu liên quan đến nhiều loại bệnh lý phổ biến trên thế giới. Khi con người tuổi lớn, các cơ quan và tế bào trong cơ thể dần suy giảm chức năng, làm giảm sút sức khỏe tổng thể của con người. Đây được xem là thuộc tính quan trọng để đánh giá mức độ ảnh hưởng của các yếu tố đến với bệnh tật.

Khoảng tin cậy cho trung bình tổng thể

Bài toán: Ước lượng độ tuổi trung bình của những bệnh nhân nữ tại khu vực Bắc Ireland.

Ở đây, nhóm không tính toán từng đặc trưng của mẫu để lấy được khoảng tin cậy của trung bình. Nhóm đã tận dụng thư viện có sẵn từ t-test, trong đó có thông số $confidence\_interval$ với giả định tập dữ liệu chia theo độ tuổi tuân theo phân phối chuẩn.

```
t.test(brc$Age)$conf.int
```

Mã nguồn 21: Khoảng tin cậy cho trung bình tuổi của tổng thể

Thực hiện tìm khoảng tin cậy:

```
[1] 57.30805 60.14305
attr(,"conf.level")
[1] 0.95
```

Mã nguồn 22: Kết quả tính khoảng tin cậy cho trung bình tuổi tổng thể

Nhận xét: Với độ tin cậy 95%, nhóm ước lượng được rằng tuổi trung bình của những bệnh nhân nữ nằm khoảng $(57.31, 60.14)$. Kết hợp với phần phân tích của thống kê mô tả, có thể đưa ra nhận định như sau: Khoảng tuổi này được xem là tương đối cao, tiệm cận độ tuổi nghỉ hưu, có thể tiềm ẩn nguy cơ mắc bệnh cao hơn. Đối với độ tuổi lớn hơn, dân số cũng có xu hướng giảm, dẫn đến số lượng người mắc bệnh cũng ít theo. Mặc dù vậy, dữ liệu hiện tại không đại diện cho toàn bộ quần thể bệnh nhân nữ tại Bắc Ireland, và mất đi một số thuộc tính quan trọng để đánh giá, nhóm không thực hiện kiểm định giả thuyết này cho tổng thể.

Kiểm định trung bình hai mẫu Bài toán: Kiểm định trung bình độ tuổi của hai mẫu sống và chết sau khi phẫu thuật trong khoảng thời gian 2017-2021.

Gọi μ_1 là trung bình tuổi của nhóm còn sống sau khi phẫu thuật của bộ dữ liệu.

Gọi μ_2 là trung bình tuổi của nhóm không qua khỏi sau khi phẫu thuật của bộ dữ liệu.

Đặt giả thuyết:

1. Giả thuyết H_0: $\mu_1 = \mu_2$
2. Giả thuyết H_1: $\mu_1 \neq \mu_2$

```
t.test(Age ~ Patient_Status, data = dummy_dataset)
```

Mã nguồn 23: Kiểm định trung bình tuổi của tổng thể

```
        Welch Two Sample t-test

data:  Age by Patient_Status
t = 0.22216, df = 93.905, p-value = 0.8247
alternative hypothesis: true difference in means between group Alive and group
    Dead is not equal to 0
95 percent confidence interval:
 -3.180591  3.981983
sample estimates:
mean in group Alive  mean in group Dead
           58.80392            58.40323
```

Mã nguồn 24: Kết quả kiểm định cho trung bình tuổi tổng thể

Giải thích kết quả:

1. Bậc tự do df = 93.905 làm tròn thành 94
2. Chỉ số thống kê t = 0.22216 < 1.986 (t-value của bậc tự do 94 và mức ý nghĩa $\alpha = 0.05/2$) cho thấy sự khác biệt trong trung bình hai nhóm không lớn
3. p-value = 0.8247 > 0.05 cho thấy không có đủ bằng chứng để bác bỏ giả thuyết H_0.

Kiểm định tính độc lập giữa nhóm tuổi và tỷ lệ tử vong sau phẫu thuật

Bài toán: Xác định mức độ tương quan giữa độ tuổi và tỷ lệ tử vong sau phẫu thuật của các bệnh nhân.

Nhóm lựa chọn phân chia độ tuổi thành 3 nhóm, dưới 45, từ 45 đến 75, và trên 75 tuổi. Đây là nhóm tuổi có sự chênh lệch về số lượng bệnh nhân và mức độ kỳ vọng của sự ảnh hưởng của tuổi tác lên tỷ lệ tử vong. Các bước kiểm định được tiến hành như sau.

Gọi p_1, p_2, p_3 lần lượt là tỷ lệ tử vong của các nhóm tuổi khác nhau được nêu trên.
Giả thuyết H_0: $p_1 = p_2 = p_3$.
Giả thuyết H_1: $\exists i, \exists j : p_i \neq p_j$.

Nhóm lựa chọn quy trình chi-square test để kiểm định tính phụ thuộc, vì khi phân nhóm tuổi, *Age_Group* và *Patient_Status* là các biến phân loại. Ta sẽ tạo ra một bảng tương quan để thể hiện tần suất xuất hiện của các thuộc tính, để tiện cho tính toán chi-square sử dụng thư viện có sẵn. Dưới đây là phần hiện thực của chương trình:

```r
dummy_dataset$AgeGroup <- cut(dummy_dataset$Age,
                              breaks = c(-Inf, 45, 75, Inf),
                              labels = c("<45", "45-75", "75+"),
                              right = FALSE)
tbl <- table(dummy_dataset$AgeGroup, dummy_dataset$ Patient_Status)
chisq.test(tbl)
```

Mã nguồn 25: Kiểm định tính độc lập của tuổi và tỷ lệ tử vong bằng R

```
        Pearson's Chi-squared test

data:  tbl
X-squared = 1.2148, df = 2, p-value = 0.5448
```

Mã nguồn 26: Kết quả kểm định tính độc lập của tuổi và tỷ lệ tử vong bằng R

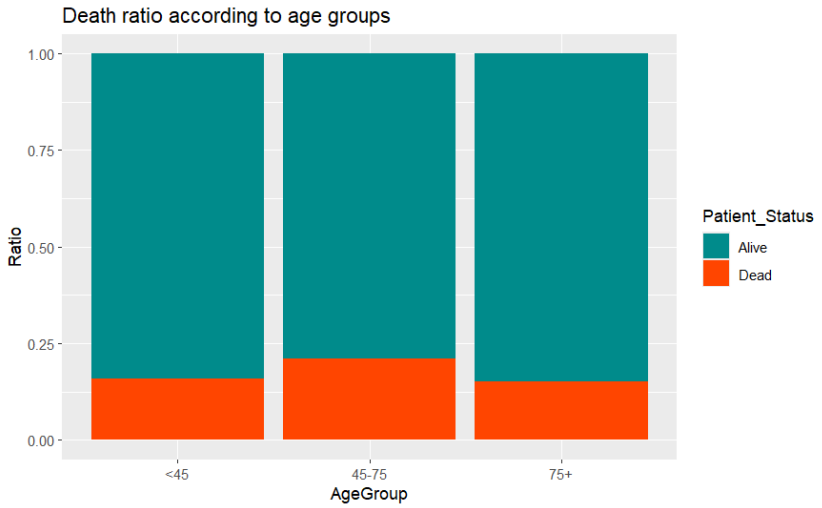
Giải thích kết quả:

1. df (bậc tự do) bằng 2, đúng với công thức tính $(col - 1) * (row - 1) = (3 - 1) * (2 - 1) = 2$.
2. X-squared (Chỉ số thống kê Chi-bình phương) = 1.21148 < 5.99 (Số tới hạn của bậc tự do 2 và mức ý nghĩa 0.05). Điều này cho thấy sự khác biệt giữa các nhóm nhỏ.
3. p-value = 0.5448 > 0.05 cho thấy không có đủ bằng chứng để bác bỏ giả thuyết H_0
4. Kết luận: Ta chấp nhận rằng không có đủ bằng chứng để chứng minh sự tương quan giữa độ tuổi và tỷ lệ tử vong trong tập dữ liệu này.

```r
ggplot(dummy_dataset, aes(x = AgeGroup, fill = Patient_Status)) +
  geom_bar(position = "fill") +
  ylab("Ratio") +
  ggtitle("Death ratio according to age groups")  +
  scale_fill_manual(values = c("Dead" = "orangered", "Alive" = "darkcyan"))
```

Mã nguồn 27: Trực quan hóa dữ liệu cho Age và Patient Status bằng R

Dưới đây là phần trực quan hóa dữ liệu cho phần tỷ lệ của các nhóm tuổi khác nhau. Có thể thấy là ở nhóm 45-75 tuổi có tỷ lệ tử vong cao hơn ít so với hai nhóm còn lại, nhưng sự khác biệt không có ý nghĩa để nhận định sự tương quan giữa nhóm tuổi và tỷ lệ tử vong.



Nhận xét tổng quan: Mặc dù theo các công bố nghiên cứu trên thế giới, tuổi tác thực sự có tác động mạnh mẽ đến khả năng sống còn của bệnh nhân sau phẫu thuật. Có thể vì bộ dữ liệu còn ít, và việc lấy mẫu chưa được hoàn chỉnh, ta không không có đủ bằng chứng để kết luận cho nhận định như các nghiên cứu trên thế giới.

Nhận định

Trong tập dữ liệu được xét, ta không có đủ bằng chứng thống kê để kết luận rằng nhóm tuổi ảnh hưởng đến tỷ lệ tử vong.

5.1.3 Giai đoạn khối u

Giai đoạn khối u, được định nghĩa trong tập dữ liệu là *Tumour_Stage*, đây là một thuộc tính quan trọng để đánh giá mức độ ảnh hưởng đến tỷ lệ tử vong sau phẫu thuật.

Phân tích ANOVA một nhân tố

Độ tuổi và giai đoạn bệnh thường có mối quan hệ mật thiết với nhau. Khi độ tuổi càng lớn, mức độ nghiêm trọng của căn bệnh càng cao, người lớn tuổi có khả năng cao bị mắc bệnh ở giai đoạn nặng hơn. Nhóm sẽ tiến hành kiểm định như sau.

Gọi μ_1, μ_2, μ_3 lần lượt là trung bình tuổi của các nhóm giai đoạn I, II, và III.

Đặt giả thuyết:

1. Giả thuyết H_0: $\mu_1 = \mu_2 = \mu_3$
2. Giả thuyết H_1: $\exists i, \exists j : \mu_i \neq \mu_j$

```
anova_result <- aov(Age ~ Tumour_Stage, data = dummy_dataset)
summary(anova_result)
```

Mã nguồn 28: Phân tích ANOVA một nhân tố độ tuổi

```
              Df Sum Sq Mean Sq F value Pr(>F)
Tumour_Stage   2   1088   544.0   3.355 0.0362 *
Residuals    314  50907   162.1
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Mã nguồn 29: Kết quả phân tích ANOVA một nhân tố bằng R

Giải thích kết quả phân tích dựa trên một số thuộc tính quan trọng như sau:

1. F-value = 3.355 > 3.0 (điểm cực hạn của phân phối fisher cho bậc tự do v1 là 2, v2 là 314, và mức ý nghĩa $\alpha = 0.05$). Điều này cho thấy có tồn tại sự khác biệt lớn giữa các nhóm trong bộ dữ liệu.
2. Pr(>F): thực ra đây là p-value. Giá trị nhận được là 0.0362 < 0.05, cho thấy có bằng để bác bỏ giả thuyết H_0.

Phân tích bội

Với phân tích ở trên, ta biết được rằng có sự khác biệt giữa các nhóm được xét. Sau đây ta sẽ tiến hành phân tích bội cho từng cặp để tìm ra sự khác biệt ý nghĩa giữa các nhóm.

```
pairwise.t.test(dummy_dataset$Age, dummy_dataset$Tumour_Stage,
                p.adjust.method = "none")
```

Mã nguồn 30: Phân tích bội cho kết quả kiểm định

```
        Pairwise comparisons using t tests with pooled SD

data:  dummy_dataset$Age and dummy_dataset$Tumour_Stage

    I    II
II  0.11 -
III 0.01 0.13

P value adjustment method: none
```

Mã nguồn 31: Kết quả Phân tích bội bằng R

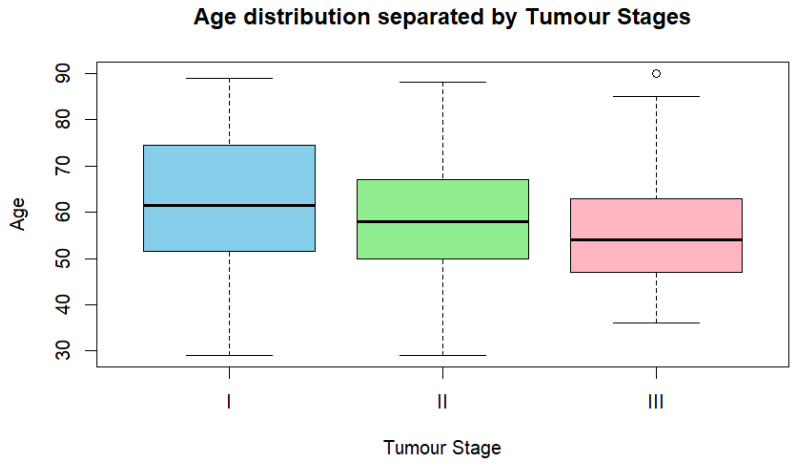
Nhận xét: Có thể thấy p-value của nhóm I-III = 0.01 < 0.05, đây được xem là có sự khác biệt ý nghĩa giữa hai nhóm. Sau đây ta sẽ thử trực quan hóa dữ liệu để có được cái nhìn tổng quan hơn.

```
boxplot(Age ~ Tumour_Stage, data = dummy_dataset,
        col = c("skyblue", "lightgreen", "lightpink"),
        main = "Age distribution separated by Tumour Stages",
        ylab = "Age",
        xlab = "Tumour Stage")
```

Mã nguồn 32: Trực quan hóa dữ liệu độ tuổi chia theo giai đoạn bằng R

Ở hình bên dưới đây, ta có thể thấy rằng là độ tuổi tủng bình của các nhóm có xu hướng giảm dần. Có thể sự khác biệt giữa hai cặp I-II và II-II là không lớn, với khoảng cách giai đoạn nhiều, ta lại thấy sự khác biệt giữa nhóm I-III.

Mặc dù theo nhận định từ các nghiên cứu trên thế giới, độ tuổi làm gia tăng mức độ nghiêm trọng của bệnh tật. Ở bộ dữ liệu này, ta lại quan sát hiện tượng ngược lại, mặc dù không quá



sâu sắc. Với tình trạng này, ta có thể đưa giả thuyết rằng là bộ dữ liệu này lấy mẫu chưa được hoàn chỉnh và có phần tập trung vào độ tuổi từ 55-65 nhiều hơn. Mặc dù vậy, ta không có đủ dữ kiện nên sẽ không thực kiểm định cho giả thuyết này.

Kiểm định tính độc lập sử dụng phương pháp Chi Bình Phương

Bài toán: Xác định sự tương quan giữa giai đoạn khối u và tỷ lệ tử vong sau phẫu thuật của bệnh nhân.

Gọi p_1, p_2, p_3 lần lượt là tỷ lệ tử vong của giai đoạn bệnh I, II, III.
Giả thuyết H_0: $p_1 = p_2 = p_3$.
Giả thuyết H_1: $\exists i, \exists j : p_i \neq p_j$.

Nhóm lựa chọn quy trình chi-square test để kiểm định tính phụ thuộc, $Tumour\_Stage$ và $Patient\_Status$ là các biến phân loại. Ta sẽ tạo ra một bảng tương quan để thể hiện tần suất xuất hiện của các thuộc tính, để tiện cho tính toán chi-square sử dụng thư viện có sẵn. Dưới đây là phần hiện thực của chương trình:

```
tbl2 <- table(dummy_dataset$Tumour_Stage, dummy_dataset$Patient_Status)
chisq.test(tbl2)
```

Mã nguồn 33: Kiểm định tính độc lập giữa giai đoạn khối u và tỷ lệ tử vong sau phẫu thuật

```
        Pearson's Chi-squared test

data:  tbl2
X-squared = 1.1254, df = 2, p-value = 0.5697
```

Mã nguồn 34: Kết quả kiểm định tính độc lập giữa giai đoạn khối u và tỷ lệ tử vong sau phẫu thuật

Phân tích kết quả:

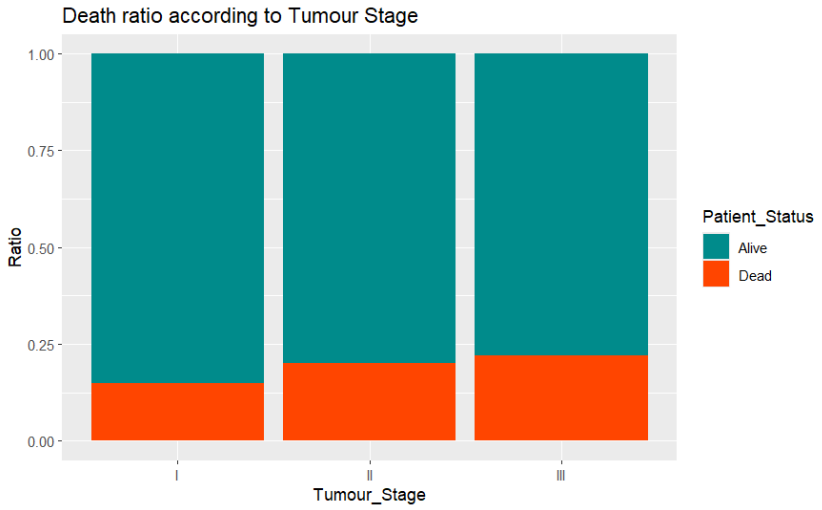
1. df (bậc tự do) bằng 2, đúng với công thức tính $(num\_Patient\_Status - 1) * (num\_Stage - 1) = (3-1) * (2-1) = 2$.
2. X-squared (Chỉ số thống kê Chi-bình phương) $= 1.1254 < 5.99$ (Số tới hạn của bậc tự do 2 và mức ý nghĩa 0.05). Điều này cho thấy sự khác biệt giữa các nhóm nhỏ.

3. p-value = 0.5697 > 0.05 cho thấy không có đủ bằng chứng để bác bỏ giả thuyết H_0
4. Kết luận: Ta chấp nhận rằng không có đủ bằng chứng để chứng minh sự tương quan giữa giai đoạn khối u và tỷ lệ tử vong trong tập dữ liệu này.

```r
tbl2 <- table(dummy_dataset$Tumour_Stage, dummy_dataset$Patient_Status)
chisq.test(tbl2)
ggplot(dummy_dataset, aes(x = Tumour_Stage, fill = Patient_Status)) +
  geom_bar(position = "fill") +
  ylab("Ratio") +
  ggtitle("Death ratio according to Tumour Stage") +
 scale_fill_manual(values = c("Dead" = "orangered", "Alive" = "darkcyan"))
```

Mã nguồn 35: Trực quan hóa dữ liệu cho Age và Tumour Stage bằng R

.



Nhận định

Trong tập dữ liệu được xét, ta không có đủ bằng chứng thống kê để kết luận rằng giai đoạn khối u ảnh hưởng đến tỷ lệ tử vong. Mặc dù, đồ thị thể hiện mức tăng nhẹ qua các giai đoạn, đây không phải là sự khác biệt có ý nghĩa.

5.1.4 Nhận xét tổng quan bộ dữ liệu

Thông qua việc kiểm định các giả thuyết thống kê vừa qua, nhóm đã có được một số nhận định về tập các bệnh nhân nữ sống ở khu vực Bắc Ireland được lấy từ bộ dữ liệu như sau:

1. Tỷ lệ tử vong sau khi phẫu thuật hơi cao, khoảng 20%.
2. Không có đủ bằng chứng thống kê để kết luận rằng nhóm tuổi ảnh hưởng đến tỷ lệ tử vong.
3. Không có đủ bằng chứng thống kê để kết luận rằng giai đoạn phát triển khối u ảnh hưởng đến tỷ lệ tử vong sau khi phẫu thuật.

5.2 Hồi quy Logistic nhị phân đa biến

5.2.1 Kiểm định giả thuyết cho các hệ số hồi quy

Trong mô hình hồi quy logistic nhị phân, mỗi hệ số hồi quy β_i được kiểm định giả thuyết như sau:

Giả thuyết không H_0: $\beta_i = 0$ – biến X_i không ảnh hưởng đến xác suất tử vong.
Giả thuyết đối H_1: $\beta_i \neq 0$ – biến X_i có ảnh hưởng đến xác suất tử vong.

Với mức ý nghĩa 5%, nếu p-value < 0.05 thì bác bỏ H_0, tức là biến đó có ảnh hưởng ý nghĩa thống kê đến xác suất tử vong. Ngược lại, nếu p-value > 0.05 thì không có đủ bằng chứng để kết luận biến đó có ảnh hưởng.

5.2.2 Quy trình phân tích và kết quả hồi quy logistic nhị phân trong bài toán

5.2.2.1 Chia tập dữ liệu Dữ liệu được chia thành tập huấn luyện (80%) và tập kiểm thử (20%) bằng hàm `sample.split` trong R để đánh giá mô hình một cách khách quan.

```
library(caTools)
set.seed(115)
split       <- sample.split(brc$Patient_Status, SplitRatio = 0.8)
train_data  <- subset(brc, split == TRUE)
test_data   <- subset(brc, split == FALSE)
```

Mã nguồn 36: Chia dữ liệu thành tập huấn luyện và kiểm thử

5.2.2.2 Xây dựng mô hình hồi quy logistic đầy đủ Mô hình ban đầu được xây dựng với tất cả các biến độc lập quan tâm bao gồm tuổi, các chỉ số protein, trạng thái HER2, giai đoạn khối u (mã hóa các mức II, III), loại mô học (mã hóa các loại L, M), loại phẫu thuật (mã hóa các loại M, O, S) và biến khoảng cách ngày giữa phẫu thuật và lần tái khám.

```
full_model <- glm(
    Patient_Status ~ Age + Protein1 + Protein2 + Protein3 + Protein4 +
    HER2.status + Tumour_Stage_II + Tumour_Stage_III + Histology_L + Histology_M +
    Surgery_type_M + Surgery_type_O + Surgery_type_S + Difference_Days,
    data = train_data, family = "binomial")
summary(full_model)
```

Mã nguồn 37: Xây dựng mô hình hồi quy logistic đầy đủ

```
    Call:
glm(formula = Patient_Status ~ Age + Protein1 + Protein2 + Protein3 +
    Protein4 + HER2.status + Tumour_Stage_II + Tumour_Stage_III +
    Histology_L + Histology_M + Surgery_type_M + Surgery_type_O +
    Surgery_type_S + Difference_Days, family = "binomial", data = train_data)

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)       1.9003095  1.0765547   1.765  0.07753 .
Age               0.0107894  0.0136526   0.790  0.42936
Protein1          0.2976489  0.3401482   0.875  0.38154
Protein2         -0.2602511  0.1986576  -1.310  0.19018
Protein3         -0.2431739  0.2968831  -0.819  0.41274
Protein4         -0.8431764  0.3216174  -2.622  0.00875 **
HER2.status       0.6435661  0.6790503   0.948  0.34326
```

```
Tumour_Stage_II   -0.3983241   0.4909462   -0.811   0.41717
Tumour_Stage_III  -0.8077462   0.5674368   -1.423   0.15459
Histology_L        0.3603756   0.4005038    0.900   0.36822
Histology_M        0.5037999   1.1287699    0.446   0.65536
Surgery_type_M    -0.1301046   0.5606498   -0.232   0.81649
Surgery_type_O    -0.7016850   0.5233850   -1.341   0.18003
Surgery_type_S    -0.5407951   0.5752906   -0.940   0.34720
Difference_Days   -0.0003042   0.0004006   -0.760   0.44755
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 251.97  on 253  degrees of freedom
Residual deviance: 234.63  on 239  degrees of freedom
AIC: 264.63

Number of Fisher Scoring iterations: 5
```

Mã nguồn 38: Kết quả mô hình hồi quy logistic đầy đủ

Phương trình log-odds sau khi ước lượng

Sau khi chạy hàm `summary(full_model)`, ta thu được các hệ số ước lượng $\hat{\beta}_i$. Khi đó, phương trình log-odds (logit) của xác suất bệnh nhân tử vong có dạng:

$$\begin{aligned}\log\left(\frac{p}{1-p}\right) = \; & 1.9003 + 0.0108 \cdot \text{Age} + 0.2976 \cdot \text{Protein1} - 0.2603 \cdot \text{Protein2} \\ & - 0.2432 \cdot \text{Protein3} - 0.8432 \cdot \text{Protein4} + 0.6436 \cdot \text{HER2.status} \\ & - 0.3983 \cdot \text{Tumour\_Stage\_II} - 0.8077 \cdot \text{Tumour\_Stage\_III} \\ & + 0.3604 \cdot \text{Histology\_L} + 0.5038 \cdot \text{Histology\_M} \\ & - 0.1301 \cdot \text{Surgery\_type\_M} - 0.7017 \cdot \text{Surgery\_type\_O} \\ & - 0.5408 \cdot \text{Surgery\_type\_S} - 0.0003 \cdot \text{Difference\_Days}\end{aligned}$$

Trong đó, chỉ có hệ số của `Protein4` là có ý nghĩa thống kê ở mức 5%, các biến còn lại đều có p-value > 0.05 nên không đủ bằng chứng để kết luận có ảnh hưởng đến log-odds tử vong.

Phân tích mô hình hồi quy logistic đầy đủ Kết quả ước lượng mô hình hồi quy logistic đầy đủ trên tập huấn luyện cho thấy:

- Hệ số hồi quy của biến `Protein4` là -0.843 với *p-value* = 0.00875, có ý nghĩa thống kê ở mức 1% (ký hiệu **). Điều này cho thấy `Protein4` có ảnh hưởng tiêu cực và có ý nghĩa đến xác suất tử vong của bệnh nhân. Cụ thể, khi giá trị `Protein4` tăng lên một đơn vị, log-odds tử vong giảm đi 0.843 đơn vị, tức là giảm xác suất tử vong.
- Các biến còn lại như tuổi (`Age`), các chỉ số protein khác, trạng thái HER2, giai đoạn khối u, loại mô học, loại phẫu thuật và khoảng cách ngày khác biệt đều không có ý nghĩa thống kê rõ ràng với *p-value* lớn hơn 0.05.
- Hệ số chặn (Intercept) có giá trị 1.900 với *p-value* khoảng 0.077, gần mức ý nghĩa 10%, cho thấy xác suất tử vong khi tất cả các biến độc lập bằng 0 là khá cao.
- Giá trị AIC của mô hình là 264.63, cho phép so sánh với các mô hình khác để chọn mô hình tối ưu.

5.2.2.3 Lựa chọn mô hình tối ưu bằng phương pháp `stepAIC` Sử dụng phương pháp `step AIC` để chọn ra mô hình hồi quy logistic có độ phù hợp tốt nhất và loại bỏ các biến không đóng góp ý nghĩa.

```r
step_model <- step(full_model)
summary(step_model)
```

```
//...
Step:  AIC=249.61
Patient_Status ~ Protein4 + Surgery_type_O

                 Df Deviance    AIC
<none>                243.61 249.61
- Surgery_type_O  1   246.32 250.32
- Protein4        1   249.75 253.75

Call:
glm(formula = Patient_Status ~ Protein4 + Surgery_type_O, family = "binomial",
    data = train_data)

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)      1.6538     0.2100   7.876 3.37e-15 ***
Protein4        -0.6992     0.2940  -2.378   0.0174 *
Surgery_type_O  -0.5534     0.3329  -1.663   0.0964 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 251.97  on 253  degrees of freedom
Residual deviance: 243.61  on 251  degrees of freedom
AIC: 249.61

Number of Fisher Scoring iterations: 4
```

Mã nguồn 39: Kết quả của phương pháp `StepAIC`

Phân tích kết quả sau khi tối ưu bằng phương pháp `stepAIC` Sau khi áp dụng phương pháp chọn mô hình stepwise dựa trên tiêu chí AIC, mô hình tối ưu chỉ còn lại hai biến:

- `Protein4` với hệ số hồi quy -0.6992 (p-*value* = 0.0174), vẫn giữ được ý nghĩa thống kê ở mức 5%.
- `Surgery_type_O` (hay là `Surgery_type_Other`) với hệ số hồi quy -0.5534 (p-*value* = 0.0964), có ý nghĩa ở mức 10%, cho thấy loại phẫu thuật khác với Lumpectomy (Cắt bỏ khối u), Simple Mastectomy (Cắt toàn bộ tuyến vú đơn giản), Modified Radical Mastectomy (Cắt toàn bộ tuyến vú có kèm nạo hạch) có xu hướng làm giảm xác suất tử vong nhưng mức ý nghĩa thấp hơn.

Mô hình tối ưu có AIC giảm xuống còn 249.61, cho thấy mô hình này phù hợp hơn và đơn giản hơn so với mô hình ban đầu.

5.2.2.4 Đánh giá mô hình Mô hình được đánh giá bằng ma trận nhầm lẫn (confusion matrix) trên tập kiểm thử để xác định độ chính xác phân loại, đồng thời vẽ đường cong ROC và tính diện tích dưới đường cong (AUC) để đánh giá khả năng phân biệt của mô hình.

```r
pred_prob                <- predict(step_model,newdata=test_data,type="response")
pred_class               <- ifelse(pred_prob > 0.5, 1, 0)
test_data$pred_class     <- factor(pred_class, levels = c(0, 1))
test_data$Patient_Status <- factor(test_data$Patient_Status, levels = c(0, 1))
library(caret)
confusionMatrix(test_data$pred_class, test_data$Patient_Status, positive = "1")
```

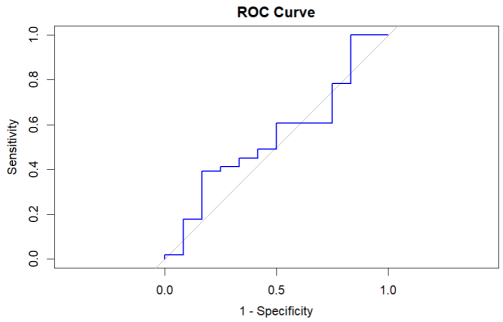
Mã nguồn 40: Đánh giá mô hình bằng ma trận nhầm lẫn

```
    Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0  0  0
         1 12 51

               Accuracy : 0.8095
                 95% CI : (0.6909, 0.8975)
    No Information Rate : 0.8095
    P-Value [Acc > NIR] : 0.576391

                  Kappa : 0

 Mcnemar's Test P-Value : 0.001496

            Sensitivity : 1.0000
            Specificity : 0.0000
         Pos Pred Value : 0.8095
         Neg Pred Value :    NaN
             Prevalence : 0.8095
         Detection Rate : 0.8095
   Detection Prevalence : 1.0000
      Balanced Accuracy : 0.5000

       'Positive' Class : 1
```

Mã nguồn 41: Ma trận nhầm lẫn

```r
library(pROC)
actual  <- as.numeric(as.character(test_data$Patient_Status))
roc_obj <- roc(actual, pred_prob)
plot(roc_obj, col = "blue", main = "ROC Curve", legacy.axes = TRUE)
auc(roc_obj)
```

Mã nguồn 42: Vẽ đường cong ROC và tính AUC



Hình 9: Đường cong ROC

```
Setting levels: control = 0, case = 1
Setting direction: controls > cases
Area under the curve: 0.5458
```

Mã nguồn 43: Diện tích dưới đường cong - Area Under the Curve

Đánh giá mô hình

- **Ma trận nhầm lẫn** cho thấy mô hình chỉ dự đoán một lớp duy nhất ("Alive"), không hề dự đoán được trường hợp "Dead" nào:

	Dự đoán sống	**Dự đoán chết**
Thực tế sống	0	0
Thực tế chết	12	51

  Điều này dẫn đến độ nhạy (Sensitivity) đạt 100% nhưng độ đặc hiệu Specificity là 0%, tức là mô hình không phân biệt được lớp "Dead" và luôn dự đoán "Alive" cho mọi trường hợp. Độ chính xác overall là 80.95% tại vì trong 63 mẫu kiểm thử có 80.95% quan sát "Alive".

- **AUC (Area Under the Curve)** chỉ đạt 0.5458, gần sát với giá trị ngẫu nhiên (0.5), cho thấy mô hình không có khả năng phân biệt thực sự giữa hai lớp "Alive" và "Dead".

- **Đường cong ROC** gần như nằm sát đường chéo, xác nhận rằng mô hình không có năng lực phân biệt thực sự.

Đây là dấu hiệu điển hình của hiện tượng **mô hình bị liệt** (degenerate model). Mặc dù accuracy khá cao (do số lượng "Alive" chiếm đa số), nhưng mô hình không có giá trị thực tiễn vì không thể dự đoán được ca "Dead" nào. Nếu ta kiểm tra tỉ lệ sống và chết ở dataset ta sẽ thấy:

```
freq(brc$Patient_Status)
```

```
        n       %       val%
0       62      19.6    19.6
1       255     80.4    80.4
```

Thì 80.4% quan sát trong mẫu là có trạng thái "Alive".

5.2.3 Kết luận

Liên hệ với các kiểm định ở mục 5.1

- **Nhóm tuổi không ảnh hưởng tới tỷ lệ tử vong.** Kết quả kiểm định cho thấy biến tuổi không có ý nghĩa thống kê trong mô hình hồi quy logistic ($p-value > 0.05$), phù hợp với kiểm định ở phần 5.1.2.
- **Giai đoạn khối u ảnh hưởng tới tỷ lệ tử vong.** Kết quả kiểm định cho thấy biến giai đoạn khối u không ảnh hưởng ý nghĩa đến tỷ lệ tử vong ($p-value > 0.05$), phù hợp với kiểm định ở phần 5.1.3

Nhưng dựa vào các tiêu chí đánh giá, mô hình hồi quy logistic hiện tại *không* đáp ứng được yêu cầu phân biệt hai trạng thái sống/chết của bệnh nhân.

5.3 Cây quyết định

Một cách tiếp cận khác cho bài toán dự đoán tình trạng bệnh nhân là mô hình cây quyết định. Cũng giống như mô hình hồi quy logistic, để thực hiện mô hình cây quyết định, ta tiến hành chia tập dữ liệu

5.3.1 Chia tập dữ liệu

Dữ liệu được chia thành 2 phần gọi là tập huấn luyện và tập kiểm tra theo tỉ lệ 8:2

```
library(caTools)
set.seed(115)
split       <- sample.split(brc$Patient_Status, SplitRatio = 0.8)
train_data  <- subset(brc, split == TRUE)
test_data   <- subset(brc, split == FALSE)
```

Mã nguồn 44: Chia dữ liệu thành tập huấn luyện và kiểm thử

5.3.2 Xây dựng mô hình

Mô hình ban đầu được xây dựng với tất cả các biến độc lập quan tâm như ở phần xây dựng mô hình hồi quy logistic đầy đủ. Nếu ở giai đoạn chia tập giữ liệu, ta set.seed(123) thì hình ảnh cây sau khi huấn luyện sẽ như sau:

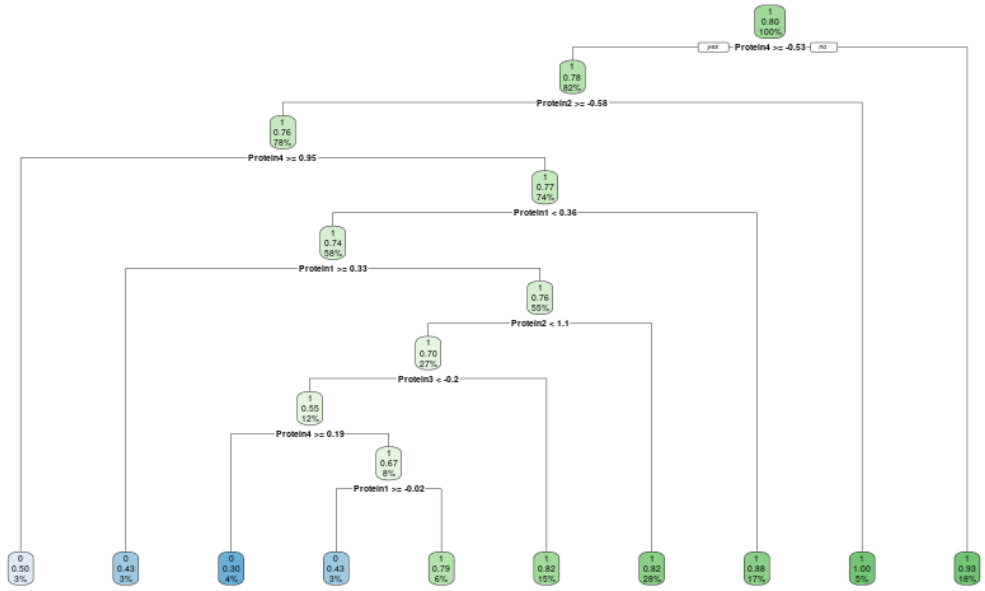
Hình 10: Cây quyết định khi set.seed(123)

Điều này cho thấy các biến đầu vào của tập huấn luyện với set.seed(123) không có đủ "sức mạnh dự đoán" để phân chia biến mục tiêu hay nói cách khác, cây quyết định bị ảnh hưởng nhiều bởi tập huấn luyện. Vì vậy, nhóm đã chọn giá trị khác, cụ thể là set.seed(115) để xây dựng cây.

```
library(rpart)
library(rpart.plot)
tree_model <- rpart(Patient_Status ~  Age + Protein1 + Protein2 + Protein3 +
    Protein4 + HER2.status + Tumour_Stage_II + Tumour_Stage_III + Histology_L +
    Histology_M + Surgery_type_M + Surgery_type_O + Surgery_type_S + Difference_
    Days,
                  data = train_data,
                  method = "class")
summary(tree_model)
rpart.plot(tree_model, extra = 106, fallen.leaves = TRUE)
```

Mã nguồn 45: Xây dựng mô hình cây quyết định chưa tinh chỉnh

Kết quả cây sau khi huấn luyện:



Hình 11: Cây quyết định chưa tinh chỉnh

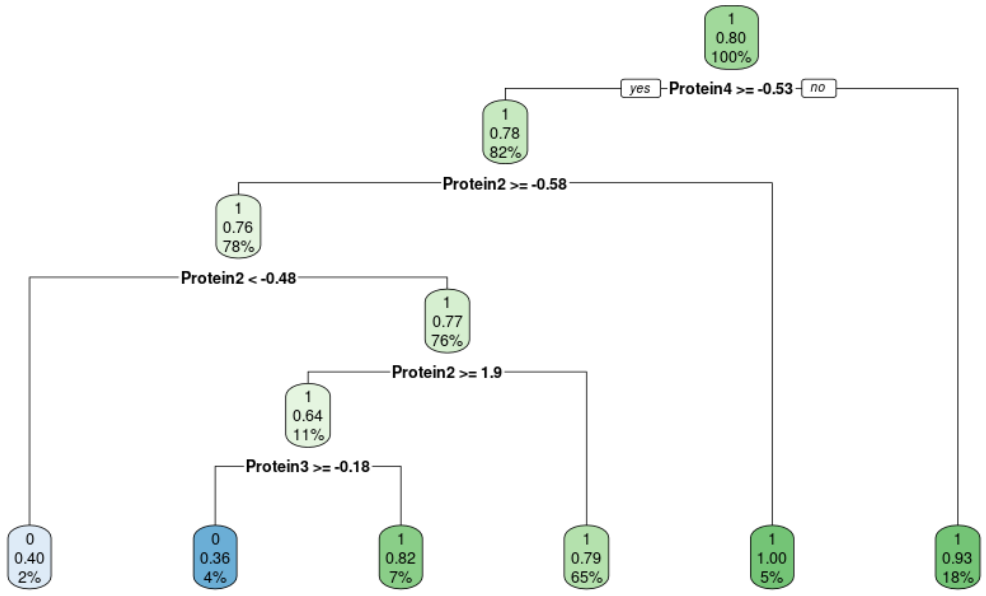
Vì cây quá sâu nên nhóm sử dụng dùng tham số control để tinh chỉnh cây với các tham số tinh chỉnh đã được chú thích trong phần code. Đây cũng là cây mà nhóm sử dụng làm mô hình.

```
custom_tree_model <- rpart(
  Patient_Status ~ Age + Protein1 + Protein2 + Protein3 + Protein4 +
                   HER2.status + Tumour_Stage_II + Tumour_Stage_III + Histology_L +
                   Histology_M + Surgery_type_M + Surgery_type_O +
                   Surgery_type_S + Difference_Days,
  data = train_data,
```

```
  method = "class",
  control = rpart.control(
    minsplit = 10,      # So luong mau toi thieu de chia nhanh
    minbucket = 5,      # So luong mau toi thieu trong nut la
    cp = 0.01,          # Nguong cai thien do phuc tap
    maxdepth = 5        # Do sau toi da cua cay
  )
)
summary(custom_tree_model)
rpart.plot(custom_tree_model, extra = 106, fallen.leaves = TRUE)
```

Mã nguồn 46: Xây dựng mô hình cây quyết định đã tinh chỉnh

Kết quả cây sau khi tinh chỉnh:



Hình 12: Cây quyết định sau khi tinh chỉnh

Phân tích sau khi xây dựng mô hình Kết quả cây sau khi xây dựng chỉ lựa chọn các thuộc tính liên quan đến Protein để xây dựng cây. Đây là các chỉ số liên quan yếu tố sinh học mà mô hình dự đoán có mối quan hệ với tình trạng bệnh nhân trong số các thuộc tính của tập dữ liệu.

5.3.3 Đánh giá mô hình

Tương tự mô hình hồi quy logistic, mô hình cũng dùng ma trận nhầm lẫn, đường cong ROC và diện tích dưới đường cong AUC để đánh giá khả năng của mô hình.

```
library(caret)
tree_pred_class <- predict(custom_tree_model, newdata = test_data, type = "class")
test_data$tree_pred_class <- factor(tree_pred_class, levels = c(0, 1))
```

```r
test_data$Patient_Status <- factor(test_data$Patient_Status, levels = c(0, 1))
confusionMatrix(test_data$tree_pred_class, test_data$Patient_Status, positive = "1")
```

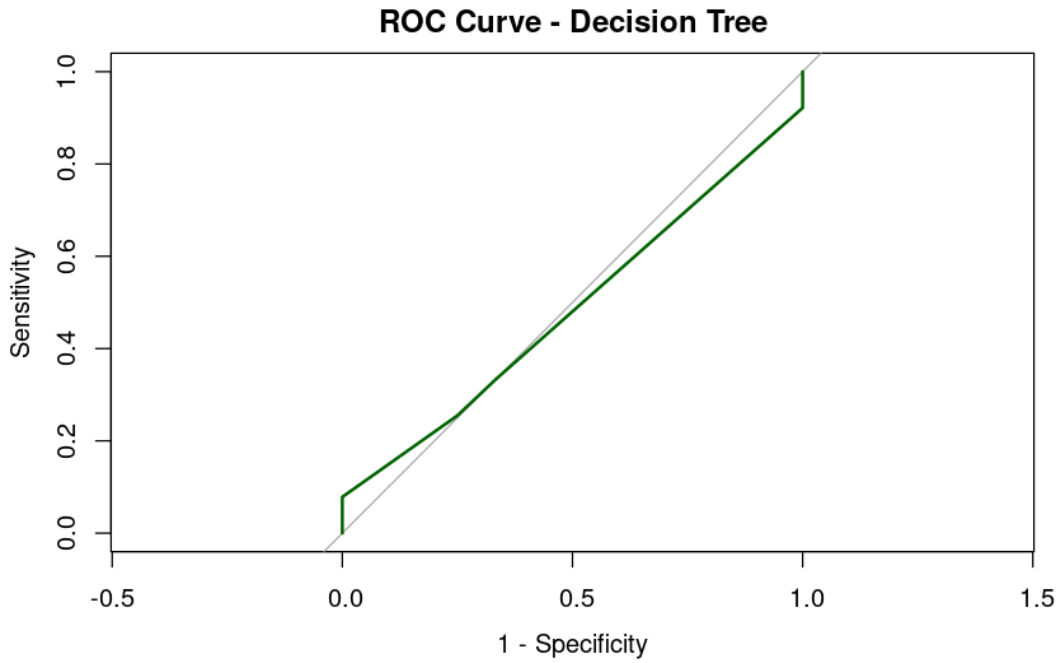
Mã nguồn 47: Phần code về ma trận nhầm lẫn

```
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0  0  4
         1 12 47

               Accuracy : 0.746
                 95% CI : (0.6206, 0.8473)
    No Information Rate : 0.8095
    P-Value [Acc > NIR] : 0.92150

                  Kappa : -0.1053

 Mcnemar's Test P-Value : 0.08012

            Sensitivity : 0.9216
            Specificity : 0.0000
         Pos Pred Value : 0.7966
         Neg Pred Value : 0.0000
             Prevalence : 0.8095
         Detection Rate : 0.7460
   Detection Prevalence : 0.9365
      Balanced Accuracy : 0.4608

       'Positive' Class : 1
```

Mã nguồn 48: Kết quả ma trận nhầm lẫn

```r
library(pROC)
actual <- as.numeric(as.character(test_data$Patient_Status))
tree_pred_prob <- predict(custom_tree_model, newdata = test_data, type = "prob")[,2]
roc_tree <- roc(actual, tree_pred_prob)
plot(roc_tree, col = "darkgreen", lwd = 2, main = "ROC Curve - Decision Tree", legacy.axes = TRUE)
auc(roc_tree)
```

Mã nguồn 49: Phần code về đường cong ROC và diện tích AUC



Hình 13: Đường cong ROC - cây quyết định

```
Setting levels: control = 0, case = 1
Setting direction: controls < cases
Area under the curve: 0.4845
```

Mã nguồn 50: Diện tích AUC

Đánh giá

- **Ma trận nhầm lẫn**: Lớp 1 chiếm 80.95% (Prevalence = 0.8095), lớp 0 chỉ 19.05% cho thấy tình trạng mất cân bằng lớp nghiêm trọng. Trong khi đó, Accuracy chỉ đạt 74.6 % dự đoán đúng. Mô hình không phân biệt được lớp 0 (Specificity = 0.0000) cho thấy mô hình vô dụng với lớp thiểu số.
- **Đường cong ROC** gần như sát đường chéo. Cho thấy mô hình không đủ tin cậy để dự đoán.
- **Diện tích AUC** Kết quả AUC = 0.4845 từ mô hình cho thấy hiệu suất rất kém, mô hình tệ hơn mô hình dự đoán ngẫu nhiên (AUC = 0.5). Lí do có thể là mô hình thiên vị dự đoán lớp 1 (chiếm đa số), dẫn đến AUC thấp. Từ đó cho thấy mô hình không có khả năng phân biệt thực sự giữa hai lớp.

5.3.4 Kết luận

Mô hình cây quyết định chưa đáp ứng được yêu cầu phân biệt trạng thái sống/chết của bệnh nhân.

6 Kết luận

Trong bài báo cáo này, nhóm đã sử dụng bộ dữ liệu "Real Breast Cancer Data" để tiến hành xử lý, phân tích dữ liệu với mục tiêu là phân biệt và dự đoán trạng thái sống chết của bệnh nhân. Thông qua các bước xử lý dữ liệu, phân tích ý nghĩa các biến dữ liệu, kiểm định giả thuyết thống kê, lựa chọn và xây dựng mô hình dự đoán, nhóm đã xây dựng được 2 mô hình gồm: Hồi quy Logistic nhị phân đa biến, Cây quyết định.

Với những công việc đã được trình bày ở bên trên, nhóm xin kết luận về kết quả báo cáo như sau: Cả 2 mô hình chưa đáp ứng được yêu cầu phân biệt trạng thái sống chết của bệnh nhân. Vì vậy, báo cáo này chưa đạt được mục tiêu đề ra. Những nguyên nhân cho việc này có thể kể đến như sau:

- Mô hình chưa thực sự phù hợp: Vì lượng kiến thức còn hạn chế, nhóm chỉ sử dụng những mô hình thống kê đơn giản, vì vậy hiệu suất của mô hình chưa cao, chưa đạt được mục tiêu đề ra.
- Kích thước bộ dữ liệu nhỏ: Số lượng các bệnh nhân trong bộ dữ liệu là khá ít (khoảng 300 người), vì vậy mô hình có thể không học được các đặc trưng quan trọng.
- Dữ liệu không đại diện: Một số biến trong bộ dữ liệu chưa phản ánh đúng với thực tế. Chẳng hạn như phân phối về tuổi trong bộ dữ liệu là khá cân đối, trái ngược với phân phối thực tế là người lớn tuổi hơn sẽ có tỷ lệ mắc ung thư vú cao hơn.
- Lựa chọn mục tiêu chưa phù hợp: Mục tiêu phân biệt và dự đoán trạng thái sống chết của bệnh nhân có thể không phù hợp với dữ liệu và kiến thức hiện tại.

Trong tương lai, nhóm sẽ nghiên cứu thêm nhiều kiến thức khác về thống kê để có thể xây dựng được mô hình dự đoán phù hợp hơn với bộ dữ liệu này. Đồng thời, nhóm cũng sẽ tiến hành phân tích thêm các bộ dữ liệu uy tín khác để có thể nhìn nhận rõ hơn về các mô hình đã được xây dựng ở trên.

Tài liệu tham khảo

[1] Hirotugu Akaike. A new look at the statistical model identification. *IEEE transactions on automatic control*, 19(6):716–723, 1974.

[2] Guy Blanc, Jane Lange, Ali Malik, and Li-Yang Tan. Popular decision tree algorithms are provably noise tolerant. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 2091–2106. PMLR, 17–23 Jul 2022.

[3] Breastcancer.org. Invasive ductal carcinoma (idc), 2025. Accessed: May 03, 2025.

[4] Tom Fawcett. An introduction to roc analysis. *Pattern recognition letters*, 27(8):861–874, 2006.

[5] David W Hosmer, Stanley Lemeshow, and Rodney X Sturdivant. *Applied Logistic Regression*. John Wiley & Sons, 2013.

[6] Geoffrey J McLachlan and Thriyambakam Krishnan. Maximum likelihood estimation. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 1–38, 1987.

[7] David M Powers. *Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation*, volume 2. 2011.

[8] J. R. Quinlan. Induction of decision trees. *Machine Learning*, 1(1):81–106, 1986.

[9] Northern Ireland Cancer Registry. Female breast cancer: 1993–2022. Technical report, Northern Ireland Cancer Registry, Queen's University Belfast, 2024. An official statistics publication.