**BỘ CÔNG THƯƠNG**

**TRƯỜNG ĐẠI HỌC CÔNG THƯƠNG TPHCM**

**KHOA CÔNG NGHỆ THÔNG TIN**

------------------

A logo with a book and a hat

Description automatically generated

**PHÂN TÍCH DỮ LIỆU VÀ DỰ BÁO**

**BREAST-CANCER**

**Giảng viên hướng dẫn:** Đinh Nguyễn Trọng Nghĩa

**Sinh viên thực hiện:**

Huỳnh Thị Kim Quý\_2001210787

Nguyễn Lê Hoàng Sơn\_2001216109

Đặng Anh Tú\_2001216269

Lê Thùy Linh\_2001211752

*TP HỒ CHÍ MINH, tháng 11 năm 2024*

**BỘ CÔNG THƯƠNG**

**TRƯỜNG ĐẠI HỌC CÔNG THƯƠNG TPHCM**

**KHOA CÔNG NGHỆ THÔNG TIN**

------------------

A logo with a book and a hat

Description automatically generated

**PHÂN TÍCH DỮ LIỆU VÀ DỰ BÁO**

**BREAST-CANCER**

**Giảng viên hướng dẫn:** Đinh Nguyễn Trọng Nghĩa

**Sinh viên thực hiện:**

Huỳnh Thị Kim Quý\_2001210787

Nguyễn Lê Hoàng Sơn\_2001216109

Đặng Anh Tú\_2001216269

Lê Thùy Linh\_2001211752

*TP HỒ CHÍ MINH, tháng 11 năm 2024*

**NHẬN XÉT CỦA GIẢNG VIÊN HƯỚNG DẪN**

**Nội dung nhận xét:**

....................................................................................................................................... ....................................................................................................................................... ....................................................................................................................................... ....................................................................................................................................... ....................................................................................................................................... ....................................................................................................................................... ....................................................................................................................................... ....................................................................................................................................... ....................................................................................................................................... ....................................................................................................................................... ....................................................................................................................................... ....................................................................................................................................... .......................................................................................................................................

.......................................................................................................................................

TP Hồ Chí Minh, ngày …. tháng .... năm

**Giảng viên hướng dẫn**

*(Ký và ghi rõ họ tên)*

LỜI CAM ĐOAN

Nhóm xin cam đoan đây là công trình nghiên cứu của nhóm. Các số liệu, kết quả nêu trong đồ án là trung thực và chưa từng được ai công bố trong bất kỳ công trình nào khác.

Nhóm xin cam đoan rằng mọi sự giúp đỡ cho việc thực hiện đồ án này   
đã được cảm ơn và các thông tin trích dẫn trong đồ án đã được chỉ rõ nguồn gốc.

**Sinh viên thực hiện đồ án**

*(Ký và ghi rõ họ tên)*

LỜI CẢM ƠN

Đồ án này đạt kết quả tốt đẹp, nhóm chúng em đã nhận được sự hỗ trợ, giúp đỡ của nhiều cơ quan, tổ chức, cá nhân. Với tình cảm sâu sắc, chân thành, cho phép nhóm chúng em được bày tỏ lòng biết ơn sâu sắc đến tất cả các cá nhân và cơ quan đã tạo điều kiện giúp đỡ em trong quá trình học tập và nghiên cứu đề tài.

Trước hết, nhóm chúng em xin gửi lời chào trân trọng và lòng biết ơn sâu sắc nhất đến quý thầy cô trong Khoa Công Nghệ Thông Tin tại Trường Đại Học Công Thương Thành Phố Hồ Chí Minh. Sự quan tâm và lòng nhiệt huyết của thầy cô đã là nguồn động viên không ngừng, giúp chúng tôi vượt qua khó khăn và hoàn thành khóa luận này.

Đặc biệt, chúng em muốn bày tỏ lòng biết ơn chân thành đặc biệt thầy Đinh Nguyễn Trọng Nghĩa**.** Sự hướng dẫn tận tâm và kiến thức sâu rộng của thầy không chỉ giúp nhóm chúng em hoàn thành dự án một cách thành công mà còn giúp chúng tôi phát triển kiến thức và kỹ năng nghiên cứu của mình.

Mặc dù nhóm chúng em đã cố gắng hết sức trong quá trình này, nhưng chắc chắn rằng nó không hoàn hảo. Nhóm chúng em đánh giá cao mọi ý kiến đóng góp và sự chỉ bảo của quý thầy để chúng em có thể hoàn thiện và nâng cao chất lượng của sản phẩm này cũng như kiến thức của nhóm chúng em trong tương lai.

Nhóm chúng em xin chân thành cảm ơn!

**Mở đầu**

Bệnh ung thư vú đã trở thành một trong những vấn đề sức khỏe nghiêm trọng nhất toàn cầu, đe dọa tính mạng của hàng triệu phụ nữ mỗi năm. Theo thống kê của Tổ chức Y tế Thế giới (WHO), ung thư vú không chỉ là loại ung thư phổ biến nhất mà còn đứng đầu trong các nguyên nhân gây tử vong ở nữ giới. Mặc dù đã có những tiến bộ trong việc điều trị, tỷ lệ tử vong do ung thư vú vẫn ở mức cao nếu bệnh không được phát hiện kịp thời. Chính sự gia tăng không ngừng của các ca mắc mới và mức độ nghiêm trọng của bệnh đã đặt ra một thách thức lớn cho cả cộng đồng và hệ thống y tế toàn cầu.

1. **Tầm Quan Trọng của Nghiên Cứu**

Trong bối cảnh này, việc phát hiện ung thư vú ở giai đoạn sớm trở thành yếu tố quyết định giúp nâng cao tỷ lệ sống sót của bệnh nhân. Tuy nhiên, các yếu tố nguy cơ như độ tuổi, di truyền, thói quen sinh hoạt và yếu tố môi trường vẫn đang là những yếu tố cần được nghiên cứu kỹ lưỡng. Nghiên cứu ung thư vú không chỉ là một nhiệm vụ khoa học mà còn là hành động thiết thực để cứu sống hàng triệu người. Chính vì thế, việc phân tích các yếu tố tác động đến bệnh lý và tìm ra những mối liên hệ giữa chúng trở thành một phần không thể thiếu trong việc xây dựng các chiến lược điều trị và phòng ngừa hiệu quả.

1. **Mục Tiêu của Đồ Án**

Đồ án này nhằm mục đích tạo ra một công cụ mạnh mẽ và dễ sử dụng để trực quan hóa dữ liệu bệnh ung thư vú, đồng thời cung cấp những phân tích sâu sắc về các yếu tố ảnh hưởng đến nguy cơ mắc bệnh. Mục tiêu cụ thể của đồ án bao gồm:

* **Khám Phá Dữ Liệu:** Phân tích những yếu tố chính ảnh hưởng đến nguy cơ ung thư vú, từ đó phát hiện các mối quan hệ quan trọng, giúp mở ra những hướng nghiên cứu mới.
* **Trực Quan Hóa Dữ Liệu:** Xây dựng một dashboard trực quan, dễ hiểu và dễ sử dụng, giúp người dùng có thể nhanh chóng truy cập và hiểu rõ thông tin bệnh lý từ dữ liệu.
* **Dự Đoán Nguy Cơ:** Áp dụng các mô hình học máy tiên tiến như hồi quy logistic và cây quyết định để dự đoán nguy cơ mắc ung thư vú, hỗ trợ các bác sĩ trong việc đưa ra quyết định điều trị chính xác và kịp thời.

1. **Phương Pháp Nghiên Cứu**

Đồ án sử dụng tập dữ liệu phong phú từ các nghiên cứu trước đây để phân tích và xây dựng mô hình dự báo. Các phương pháp phân tích dữ liệu chủ yếu bao gồm:

* **ANOVA:** Phân tích sự khác biệt giữa các nhóm bệnh nhân (lành tính và ác tính), giúp xác định các yếu tố có tác động mạnh mẽ đến sự phát triển của bệnh.
* **Hồi Quy Logistic:** Áp dụng hồi quy logistic để phân tích và dự đoán khả năng mắc ung thư vú, nhận diện các yếu tố nguy cơ chủ yếu.
* **Cây Quyết Định:** Sử dụng cây quyết định để xây dựng mô hình phân loại, giúp giải thích rõ ràng cách thức các yếu tố quyết định nguy cơ mắc bệnh.

1. **Cấu Trúc Báo Cáo**

Báo cáo này được tổ chức thành các phần chính, bao gồm:

* **Giới thiệu:** Tóm tắt về bệnh ung thư vú và lý do tại sao nghiên cứu này lại quan trọng.
* **Tiền xử lý dữ liệu:** Cung cấp thông tin chi tiết về tập dữ liệu sử dụng, quy trình kiểm tra chất lượng dữ liệu và các phương pháp phân tích đã áp dụng trong nghiên cứu.
* **Phân tích dữ liệu:** Trình bày các bước tiền xử lý dữ liệu và lý do chọn các

kỹ thuật phân tích.

* **Kết Quả Phân Tích:** Cung cấp kết quả phân tích từ các mô hình, kèm theo các biểu đồ, bảng số liệu minh họa để giúp người đọc hiểu rõ hơn.
* **Thiết Kế Dashboard:** Trình bày về cấu trúc và tính năng của dashboard, tập trung vào việc mang đến trải nghiệm trực quan và dễ dàng tương tác.
* **Kết Luận:** Phân tích những phát hiện quan trọng, giải thích các mối liên hệ giữa các yếu tố và bệnh ung thư vú, đồng thời đưa ra các đề xuất cho nghiên cứu tiếp theo.

# PHÂN CÔNG THÀNH VIÊN TRONG NHÓM

|  |  |  |  |
| --- | --- | --- | --- |
| **HỌ TÊN** | **MSSV** | **CÔNG VIỆC** | **%** |
| Huỳnh Thị Kim Quý | 2001210787 | Phân tích dữ liệu | 100% |
| Nguyễn Lê Hoàng Sơn | 2001216109 | Thuật toán | 100% |
| Đặng Anh Tú | 2001216269 | Phân tích dữ liệu | 100% |
| Lê Thùy Linh | 2001211752 | Tiền xử lý dữ liệu, powerpoint, word. | 100% |

**Mục lục**

[PHÂN CÔNG THÀNH VIÊN TRONG NHÓM 5](#_Toc184593816)

[CHƯƠNG 1 : CƠ SỞ LÝ THUYẾT 8](#_Toc184593817)

[1. Giới Thiệu Nền Tảng Kaggle 8](#_Toc184593818)

[1.1. Nguồn Dữ Liệu Đa Dạng và Phong Phú 8](#_Toc184593819)

[1.2. Cuộc Thi Dữ Liệu 9](#_Toc184593820)

[1.3. Kaggle Notebooks 9](#_Toc184593821)

[1.4. Các Khóa Học Miễn Phí 10](#_Toc184593822)

[1.5. Ưu Điểm của Kaggle 10](#_Toc184593823)

[1.6. Nhược Điểm của Kaggle 11](#_Toc184593824)

[1.7. Ứng Dụng Trong Đồ Án 11](#_Toc184593825)

[2. Dashboard 12](#_Toc184593826)

[2.1. Biểu Đồ và Đồ Thị 12](#_Toc184593827)

[2.2. Tùy Chọn Lọc Dữ Liệu 13](#_Toc184593828)

[3. Các Kỹ Thuật Phân Tích Dữ Liệu và Thống Kê 14](#_Toc184593829)

[3.1. Phân Tích Mô Tả 14](#_Toc184593830)

[3.2. Phân Tích So Sánh 15](#_Toc184593831)

[3.3. Phân Tích Quan Hệ 15](#_Toc184593832)

[3.4. Phân Tích Dự Đoán 15](#_Toc184593833)

[3.5. Khám Phá Dữ Liệu (Data Mining) 16](#_Toc184593834)

[CHƯƠNG 2 : GIỚI THIỆU ĐỀ TÀI 17](#_Toc184593835)

[1. Mô Tả Tập Dữ Liệu 17](#_Toc184593836)

[2. Mục Tiêu Phân Tích 18](#_Toc184593837)

[2.1. Phân Tích và Trực Quan Hóa Dữ Liệu: 18](#_Toc184593838)

[2.2. Dự Đoán Nguy Cơ Mắc Bệnh: 18](#_Toc184593839)

[2.3. Các bước thực hiện 19](#_Toc184593840)

[CHƯƠNG 3 : TIỀN XỬ LÝ DỮ LIỆU 19](#_Toc184593841)

[1. Khám phá dữ liệu 19](#_Toc184593842)

[1.1. Kiểm tra dữ liệu thiếu 19](#_Toc184593843)

[1.2. Phương Pháp Kiểm Tra Dữ Liệu Thiếu 20](#_Toc184593844)

[2. Tiền xử lý dữ liệu 21](#_Toc184593845)

[2.1. Xóa Dữ Liệu Thiếu 22](#_Toc184593846)

[2.2. Chuyển Đổi Biến 22](#_Toc184593847)

[2.3. Chuẩn hóa dữ liệu 22](#_Toc184593848)

[2.4. Kiểm tra dữ liệu sau khi tiền xử lý 23](#_Toc184593849)

[3. Trực quan hóa dữ liệu 23](#_Toc184593850)

[4. Kết Luận 30](#_Toc184593851)

[CHƯƠNG 4 : PHÂN TÍCH DỮ LIỆU 30](#_Toc184593852)

[1. Phân tích dữ liệu 30](#_Toc184593853)

[1.1. Hồi quy tuyến tính 30](#_Toc184593854)

[Phân Tích Mối Quan Hệ Tuyến Tính Giữa Độ Nhẵn và Độ Lõm trong Khối U Vú 30](#_Toc184593855)

[1.2. Hồi quy Logistic 36](#_Toc184593856)

[1.3. Cây quyết định 42](#_Toc184593857)

[4. Kết Luận 46](#_Toc184593858)

# CHƯƠNG 1 : CƠ SỞ LÝ THUYẾT

## Giới Thiệu Nền Tảng Kaggle

Kaggle là một nền tảng trực tuyến hàng đầu trong lĩnh vực khoa học dữ liệu và học máy, được thành lập vào năm 2010. Ban đầu, Kaggle được biết đến như một nơi tổ chức các cuộc thi về phân tích dữ liệu, nhưng với sự phát triển nhanh chóng, nó đã trở thành một cộng đồng lớn mạnh với hàng triệu người dùng từ khắp nơi trên thế giới. Nền tảng này không chỉ phục vụ như một kho dữ liệu phong phú mà còn là một không gian học tập và phát triển kỹ năng trong lĩnh vực dữ liệu.

### Nguồn Dữ Liệu Đa Dạng và Phong Phú

Kaggle cung cấp hàng triệu tập dữ liệu từ nhiều lĩnh vực khác nhau, bao gồm y tế, tài chính, giáo dục, thể thao và nhiều lĩnh vực khác. Điều này tạo ra cơ hội tuyệt vời cho những người làm nghiên cứu, nhà phân tích và lập trình viên trong việc tìm kiếm và sử dụng dữ liệu cho các dự án và nghiên cứu của họ.

* **Tập Dữ Liệu Y Tế**: Trong bối cảnh nghiên cứu về bệnh ung thư vú, Kaggle có nhiều tập dữ liệu quan trọng như:
  + **Breast Cancer Dataset**: Đây là một trong những tập dữ liệu nổi tiếng nhất về ung thư vú, bao gồm thông tin chi tiết về các đặc điểm của khối u như kích thước, hình dạng, kết cấu và các chỉ số lâm sàng khác. Tập dữ liệu này không chỉ cung cấp thông tin về khối u mà còn đi kèm với nhãn phân loại (ác tính hoặc lành tính), rất hữu ích cho việc xây dựng các mô hình học máy.
* **Chất Lượng Dữ Liệu**: Kaggle đảm bảo rằng các tập dữ liệu được cung cấp đều có chất lượng cao, có mô tả rõ ràng và hướng dẫn sử dụng, giúp người dùng dễ dàng khai thác và áp dụng vào thực tế.
* **Khả Năng Tìm Kiếm Nâng Cao**: Với các bộ lọc thông minh, người dùng có thể dễ dàng tìm kiếm các tập dữ liệu phù hợp với nhu cầu của mình, tiết kiệm thời gian và nâng cao hiệu quả làm việc.

### Cuộc Thi Dữ Liệu

Một trong những điểm nổi bật của Kaggle là tổ chức nhiều cuộc thi dữ liệu, nơi mà người dùng có thể tham gia giải quyết các bài toán thực tế từ các tổ chức và công ty lớn. Những cuộc thi này không chỉ mang lại cơ hội học hỏi mà còn thúc đẩy sự sáng tạo và khả năng giải quyết vấn đề.

* **Thách Thức Thực Tế**: Các cuộc thi thường liên quan đến các vấn đề trong thế giới thực, từ dự đoán doanh thu đến phân tích rủi ro y tế, giúp người tham gia áp dụng lý thuyết vào thực tiễn. Chẳng hạn, trong các cuộc thi về phân tích bệnh ung thư vú, người dùng có thể phát triển các mô hình dự đoán dựa trên các đặc điểm của bệnh nhân.
* **Cộng Đồng Hỗ Trợ**: Tham gia vào các cuộc thi trên Kaggle giúp người dùng kết nối với những người có cùng sở thích và đam mê. Họ có thể chia sẻ ý tưởng, chiến lược và nhận phản hồi từ cộng đồng, điều này không chỉ làm phong phú thêm trải nghiệm học tập mà còn thúc đẩy sự hợp tác.
* **Giải Thưởng và Công Nhận**: Người chiến thắng trong các cuộc thi thường nhận được giải thưởng hấp dẫn, cùng với sự công nhận từ cộng đồng. Điều này không chỉ giúp nâng cao hồ sơ chuyên môn mà còn tạo động lực cho người tham gia.

### Kaggle Notebooks

Kaggle Notebooks là một trong những tính năng nổi bật nhất của nền tảng, cho phép người dùng thực hiện lập trình và phân tích dữ liệu trực tuyến mà không cần cài đặt phần mềm.

* **Môi Trường Lập Trình Tích Hợp**: Hỗ trợ nhiều ngôn ngữ lập trình như Python và R, người dùng có thể dễ dàng triển khai các mô hình học máy, thực hiện phân tích thống kê và trực quan hóa dữ liệu trong cùng một không gian làm việc. Điều này rất thuận tiện cho việc tạo ra các báo cáo và chia sẻ kết quả.
* **Chia Sẻ và Hợp Tác**: Người dùng có thể chia sẻ notebook của mình với cộng đồng, cho phép người khác xem và bình luận. Đây là một cách tuyệt vời để học hỏi và cải thiện kỹ năng lập trình, cũng như để nhận được phản hồi từ những người có kinh nghiệm.
* **Tài Nguyên Hỗ Trợ**: Kaggle cung cấp nhiều tài liệu hướng dẫn và mẫu notebook cho các loại phân tích khác nhau, giúp người dùng có thể bắt đầu nhanh chóng với dự án của mình và giảm thiểu thời gian tìm hiểu.

### Các Khóa Học Miễn Phí

Kaggle không chỉ là nơi tìm kiếm dữ liệu mà còn là một nền tảng học tập phong phú với nhiều khóa học miễn phí về phân tích dữ liệu và học máy.

* **Khóa Học Chất Lượng Cao**: Các khóa học được thiết kế bởi các chuyên gia trong ngành, giúp người dùng nắm rõ kiến thức cơ bản và nâng cao về phân tích dữ liệu, bao gồm các chủ đề như học máy, xử lý dữ liệu và trực quan hóa.
* **Thực Hành Thực Tế**: Mỗi khóa học thường đi kèm với các bài tập thực hành, giúp người học áp dụng kiến thức vào thực tế và củng cố kỹ năng của mình. Việc thực hành không chỉ giúp người học ghi nhớ kiến thức mà còn tăng khả năng ứng dụng trong các tình huống thực tế.
* **Chứng Nhận**: Sau khi hoàn thành các khóa học, người dùng có thể nhận chứng nhận, điều này có thể giúp họ cải thiện hồ sơ cá nhân và thu hút nhà tuyển dụng.

### Ưu Điểm của Kaggle

* **Kho Dữ Liệu Phong Phú**: Kaggle cung cấp một kho dữ liệu đa dạng từ nhiều lĩnh vực khác nhau, cho phép người dùng dễ dàng tìm kiếm và tải về thông tin cần thiết.
* **Cộng Đồng Hỗ Trợ**: Với hàng triệu người dùng, Kaggle tạo ra một cộng đồng mạnh mẽ, nơi mọi người có thể học hỏi lẫn nhau và chia sẻ kinh nghiệm.
* **Cuộc Thi Dữ Liệu**: Các cuộc thi giúp người dùng thử thách bản thân, cải thiện kỹ năng và nhận được sự công nhận từ cộng đồng.
* **Kaggle Notebooks**: Môi trường lập trình trực tuyến tiện lợi cho việc phân tích dữ liệu, giúp người dùng dễ dàng chia sẻ và hợp tác.
* **Khóa Học Miễn Phí**: Nhiều tài nguyên học tập giúp người dùng nâng cao kiến thức mà không phải trả phí.

### Nhược Điểm của Kaggle

* **Chất Lượng Dữ Liệu Không Đều**: Một số tập dữ liệu có thể thiếu thông tin hoặc không đầy đủ, điều này có thể làm giảm độ chính xác của các phân tích.
* **Cạnh Tranh Cao**: Trong các cuộc thi, sự cạnh tranh lớn có thể khiến một số người cảm thấy áp lực. Người mới bắt đầu có thể gặp khó khăn trong việc so sánh với những người có kinh nghiệm hơn.
* **Thiếu Tài Nguyên Học Tập Nâng Cao**: Mặc dù có nhiều khóa học miễn phí, nhưng một số người dùng có thể thấy rằng các tài nguyên học tập không đủ sâu sắc hoặc không đáp ứng nhu cầu của họ khi họ muốn tìm hiểu các chủ đề nâng cao.
* **Phụ Thuộc Vào Kết Nối Internet**: Kaggle là một nền tảng trực tuyến, do đó người dùng cần có kết nối Internet ổn định để truy cập và làm việc với dữ liệu. Điều này có thể gây khó khăn cho những người sống ở khu vực có kết nối kém.
* **Khó Khăn Trong Việc Ứng Dụng Thực Tế**: Một số người dùng có thể gặp khó khăn trong việc áp dụng các kỹ thuật học máy đã học từ Kaggle vào các tình huống thực tế, do sự khác biệt giữa môi trường học tập và môi trường làm việc.
* **Thiếu Tính Riêng Tư**: Khi chia sẻ dữ liệu và notebook, người dùng cần phải cân nhắc về vấn đề bảo mật và quyền riêng tư, đặc biệt khi làm việc với dữ liệu nhạy cảm.

### Ứng Dụng Trong Đồ Án

Trong đồ án "Nghiên Cứu và Phân Tích Bệnh Ung Thư Vú", Kaggle sẽ được áp dụng một cách hiệu quả như sau:

* **Khai Thác Tập Dữ Liệu**: Sử dụng Breast Cancer Wisconsin Dataset để thực hiện các bước phân tích thống kê và xây dựng các mô hình học máy nhằm dự đoán và phân loại trạng thái của khối u. Điều này không chỉ giúp hiểu rõ hơn về bệnh lý mà còn có thể hỗ trợ trong việc đưa ra quyết định điều trị.
* **Tiền Xử Lý Dữ Liệu**: Các bước này bao gồm làm sạch dữ liệu, xử lý giá trị thiếu, mã hóa biến phân loại và chuẩn hóa dữ liệu để đảm bảo tính chính xác của các mô hình phân tích. Việc này rất quan trọng, vì dữ liệu chất lượng cao sẽ dẫn đến kết quả phân tích đáng tin cậy hơn.
* **Áp Dụng Kỹ Thuật Phân Tích**:
  + **ANOVA**: Phân tích phương sai sẽ được sử dụng để xác định sự khác biệt giữa các nhóm bệnh nhân (ác tính và lành tính) dựa trên các chỉ số lâm sàng.
  + **Chi-square**: Kiểm định này sẽ đánh giá mối liên hệ giữa biến phân loại diagnosis và các yếu tố khác như tuổi tác, kích thước khối u.
  + **Hồi Quy Tuyến Tính**: Sẽ được sử dụng để dự đoán các biến liên tục như area\_mean từ các biến độc lập, giúp cung cấp cái nhìn sâu sắc về các yếu tố ảnh hưởng đến kích thước khối u.
  + **Hồi Quy Logistic**: Sẽ giúp phân loại bệnh nhân thành ác tính hoặc lành tính, tạo điều kiện cho việc ra quyết định điều trị.
  + **Cây Quyết Định**: Sẽ mô hình hóa quá trình ra quyết định dựa trên các biến độc lập, giúp hiểu rõ hơn về các yếu tố ảnh hưởng đến trạng thái bệnh.

## Dashboard

Dashboard là một công cụ trực quan hóa dữ liệu quan trọng, giúp người dùng theo dõi, phân tích và trực quan hóa thông tin một cách dễ dàng và hiệu quả. Một dashboard hiệu quả không chỉ giúp người xem nhanh chóng hiểu được dữ liệu mà còn hỗ trợ việc ra quyết định dựa trên các thông tin quan trọng. Dưới đây là các yếu tố chính để xây dựng một dashboard hiệu quả, cùng với những chi tiết cụ thể hơn về từng yếu tố.

### Biểu Đồ và Đồ Thị

Biểu đồ và đồ thị là công cụ trực quan hóa dữ liệu chính trong dashboard, giúp người dùng dễ dàng nhận diện các xu hướng, mẫu và điểm bất thường trong dữ liệu. Dưới đây là một số loại biểu đồ phổ biến và ứng dụng của chúng:

* **Biểu Đồ Cột**:
  + **Mô Tả**: Biểu đồ cột sử dụng các cột thẳng đứng để biểu thị giá trị của các danh mục khác nhau.
  + **Ứng Dụng**: Thích hợp cho việc so sánh dữ liệu giữa các nhóm, chẳng hạn như doanh thu theo từng khu vực hoặc số lượng bệnh nhân theo từng loại bệnh.
* **Biểu Đồ Đường**:
  + **Mô Tả**: Biểu đồ đường kết nối các điểm dữ liệu bằng các đường thẳng, thể hiện sự thay đổi của dữ liệu theo thời gian.
  + **Ứng Dụng**: Thích hợp để theo dõi các chỉ số theo thời gian, chẳng hạn như số ca mắc bệnh theo tháng hoặc doanh số bán hàng theo quý.
* **Biểu Đồ Tròn**:
  + **Mô Tả**: Biểu đồ tròn chia một hình tròn thành các phần tương ứng với tỷ lệ phần trăm của các thành phần trong tổng thể.
  + **Ứng Dụng**: Dùng để thể hiện tỷ lệ phần trăm, ví dụ như tỷ lệ phần trăm các loại bệnh trong tổng số bệnh nhân.
* **Biểu Đồ Tán Xạ**:
  + **Mô Tả**: Biểu đồ tán xạ sử dụng các điểm để biểu thị mối quan hệ giữa hai biến, với một biến trên trục x và một biến trên trục y.
  + **Ứng Dụng**: Giúp phân tích mối quan hệ, chẳng hạn như mối liên hệ giữa tuổi tác và kích thước khối u trong nghiên cứu ung thư.
* **Biểu Đồ Heatmap**:
  + **Mô Tả**: Biểu đồ heatmap sử dụng các ô màu để thể hiện mật độ hoặc giá trị của các biến.
  + **Ứng Dụng**: Thích hợp cho việc phân tích và trực quan hóa dữ liệu đa chiều, như sự phân bố của các bệnh nhân theo địa lý.

### Tùy Chọn Lọc Dữ Liệu

Tùy chọn lọc dữ liệu là một phần quan trọng của dashboard, cho phép người dùng tương tác với dữ liệu và tùy chỉnh thông tin hiển thị theo nhu cầu cá nhân. Các tính năng lọc dữ liệu nên bao gồm:

* **Bộ Lọc Thời Gian**:
  + **Chức Năng**: Cho phép người dùng chọn khoảng thời gian cụ thể để phân tích dữ liệu, chẳng hạn như ngày, tuần, tháng hoặc năm.
  + **Lợi Ích**: Giúp người dùng dễ dàng theo dõi xu hướng theo thời gian và nhận diện các biến động quan trọng.
* **Bộ Lọc Danh Mục**:
  + **Chức Năng**: Giúp người dùng chỉ hiển thị dữ liệu cho một hoặc nhiều danh mục cụ thể, như các nhóm bệnh nhân hoặc các loại bệnh.
  + **Lợi Ích**: Tăng cường khả năng phân tích chi tiết và giúp người dùng tập trung vào những thông tin quan trọng nhất cho nghiên cứu.
* **Bộ Lọc Đa Lựa Chọn**:
  + **Chức Năng**: Cho phép người dùng chọn nhiều giá trị từ một danh sách, ví dụ như chọn nhiều khu vực địa lý để phân tích.
  + **Lợi Ích**: Tạo sự linh hoạt trong việc phân tích dữ liệu và cho phép người dùng dễ dàng so sánh các nhóm khác nhau.
* **Tìm Kiếm Nâng Cao**:
  + **Chức Năng**: Cung cấp chức năng tìm kiếm để người dùng nhanh chóng tìm thấy dữ liệu theo từ khóa hoặc thuộc tính nhất định.
  + **Lợi Ích**: Giúp tiết kiệm thời gian và nâng cao hiệu quả trong việc tìm kiếm thông tin cụ thể trong tập dữ liệu lớn.

## Các Kỹ Thuật Phân Tích Dữ Liệu và Thống Kê

Phân tích dữ liệu và thống kê là những công cụ thiết yếu trong việc hiểu và rút ra thông tin từ các tập dữ liệu lớn. Khi áp dụng các kỹ thuật này, người dùng có thể tìm kiếm các mẫu, xu hướng và mối quan hệ trong dữ liệu, từ đó đưa ra những quyết định thông minh và hiệu quả hơn. Dưới đây là một cái nhìn sâu hơn về các kỹ thuật phân tích dữ liệu và thống kê phổ biến, cùng với các ứng dụng cụ thể của chúng trong nghiên cứu bệnh ung thư vú.

### Phân Tích Mô Tả

**Mô Tả**: Phân tích mô tả cung cấp cái nhìn tổng quan về dữ liệu thông qua các chỉ số thống kê cơ bản, giúp người dùng hiểu rõ hơn về cấu trúc và đặc điểm của tập dữ liệu.

* **Kỹ Thuật Cụ Thể**:
  + **Độ Trung Bình**: Tính toán giá trị trung bình của một biến giúp hiểu rõ hơn về xu hướng chung. Ví dụ, trong nghiên cứu bệnh nhân ung thư vú, tính trung bình tuổi của bệnh nhân có thể giúp xác định nhóm tuổi chính trong nghiên cứu.
  + **Độ Lệch Chuẩn**: Đo lường sự phân tán của dữ liệu xung quanh giá trị trung bình, giúp xác định mức độ biến động. Độ lệch chuẩn cao có thể chỉ ra rằng dữ liệu có sự đa dạng lớn giữa các mẫu.
  + **Phân Tích Tần Suất**: Xác định số lượng và tỷ lệ phần trăm của các giá trị khác nhau trong tập dữ liệu, giúp người dùng nhận diện những đặc điểm nổi bật.
* **Ứng Dụng trong Nghiên Cứu Ung Thư Vú**:
  + Tạo ra các báo cáo mô tả về độ tuổi, giai đoạn bệnh, và tỷ lệ sống sót của bệnh nhân, từ đó giúp các nhà nghiên cứu và bác sĩ có cái nhìn tổng quan về tình hình bệnh lý.

### Phân Tích So Sánh

**Mô Tả**: Phân tích so sánh giúp người dùng đánh giá sự khác biệt giữa các nhóm dữ liệu, cho phép xác định các yếu tố ảnh hưởng đến kết quả.

* **Kỹ Thuật Cụ Thể**:
  + **Kiểm Định T-Test**: Dùng để so sánh trung bình của hai nhóm độc lập. Ví dụ, so sánh tỷ lệ sống sót giữa bệnh nhân được điều trị bằng thuốc A và thuốc B.
  + **ANOVA (Phân Tích Phương Phân)**: Dùng để so sánh trung bình giữa ba nhóm trở lên, như so sánh tỷ lệ hồi phục ở ba nhóm bệnh nhân khác nhau theo độ tuổi.
* **Ứng Dụng**:
  + Phân tích hiệu quả của các phương pháp điều trị khác nhau, từ đó xác định phương pháp điều trị tốt nhất cho từng nhóm bệnh nhân.

### Phân Tích Quan Hệ

**Mô Tả**: Phân tích quan hệ giúp xác định mối liên hệ giữa hai hoặc nhiều biến, từ đó đưa ra các kết luận về các yếu tố có thể ảnh hưởng đến kết quả.

* **Kỹ Thuật Cụ Thể**:
  + **Hệ Số Tương Quan (Correlation Coefficient)**: Đo lường mức độ liên hệ giữa hai biến. Hệ số này nằm trong khoảng từ -1 đến 1, với 0 cho thấy không có mối liên hệ.
  + **Phân Tích Hồi Quy (Regression Analysis)**: Dùng để dự đoán giá trị của một biến dựa trên một hoặc nhiều biến độc lập. Ví dụ, hồi quy tuyến tính có thể dự đoán tỷ lệ sống sót của bệnh nhân dựa trên các yếu tố như tuổi tác và kích thước khối u.
* **Ứng Dụng**:
  + Xác định các yếu tố nguy cơ như mối quan hệ giữa chỉ số khối cơ thể (BMI) và tỷ lệ mắc ung thư vú, từ đó giúp xây dựng các chương trình phòng ngừa.

### Phân Tích Dự Đoán

**Mô Tả**: Phân tích dự đoán sử dụng các kỹ thuật học máy để xây dựng mô hình dự đoán cho các kết quả trong tương lai, giúp đưa ra các quyết định dựa trên dự đoán.

* **Kỹ Thuật Cụ Thể**:
  + **Cây Quyết Định**: Một phương pháp trực quan cho phép phân loại hoặc dự đoán kết quả dựa trên các đặc điểm đầu vào. Cây quyết định có thể được sử dụng để phân loại bệnh nhân thành các nhóm có nguy cơ cao và thấp.
  + **Mô Hình Hồi Quy Logistic**: Dùng để dự đoán xác suất của một sự kiện xảy ra (như khả năng sống sót sau khi điều trị).
* **Ứng Dụng**:
  + Dự đoán khả năng tái phát bệnh ung thư vú dựa trên các đặc điểm lâm sàng và hình ảnh y tế, từ đó giúp bác sĩ lập kế hoạch điều trị hiệu quả hơn.

### Khám Phá Dữ Liệu (Data Mining)

**Mô Tả**: Khám phá dữ liệu là quá trình tìm kiếm các mẫu và xu hướng trong dữ liệu lớn, giúp người dùng phát hiện ra những thông tin mới và quý giá.

* **Kỹ Thuật Cụ Thể**:
  + **Phân Nhóm (Clustering)**: Nhóm các đối tượng tương tự lại với nhau, giúp phát hiện các mẫu trong dữ liệu. Sử dụng thuật toán như K-means hoặc DBSCAN để phân nhóm bệnh nhân theo đặc điểm lâm sàng.
  + **Quy Tắc Kết Hợp (Association Rules)**: Tìm kiếm các mối quan hệ giữa các biến trong dữ liệu, như xác định các yếu tố nguy cơ của ung thư vú thông qua phân tích mối quan hệ giữa các biến.
* **Ứng Dụng**:
  + Phát hiện các nhóm bệnh nhân có nguy cơ cao dựa trên các yếu tố như di truyền, lối sống và tiền sử bệnh tật.

# CHƯƠNG 2 : GIỚI THIỆU ĐỀ TÀI

## Mô Tả Tập Dữ Liệu

Tập dữ liệu được sử dụng trong đồ án này là một tập hợp thông tin phong phú về bệnh nhân ung thư vú, được thu thập từ các nghiên cứu y tế và cơ sở dữ liệu bệnh viện. Tập dữ liệu này bao gồm 569 bệnh nhân, với các biến quan trọng như sau:

* **id**: Mã định danh duy nhất cho từng bệnh nhân, giúp theo dõi và quản lý dữ liệu một cách chính xác.
* **diagnosis**: Kết quả chẩn đoán bệnh, được phân loại thành hai nhóm: "M" (malignant - ác tính) và "B" (benign - lành tính). Biến này là yếu tố quyết định trong việc phân loại và phân tích dữ liệu bệnh.
* **radius\_mean**: Trung bình kích thước bán kính của khối u, được đo bằng cm. Đây là yếu tố quan trọng trong việc đánh giá mức độ nghiêm trọng của bệnh.
* **texture\_mean**: Trung bình độ nhám của khối u, phản ánh tính chất bề mặt của khối u.
* **perimeter\_mean**: Trung bình chu vi của khối u, cũng được đo bằng cm, giúp đánh giá hình dạng của khối u.
* **area\_mean**: Diện tích trung bình của khối u, cung cấp thông tin về kích thước tổng thể.
* **smoothness\_mean**: Độ mịn trung bình của khối u, đo lường sự đồng đều của bề mặt khối u.
* **compactness\_mean**: Độ chặt chẽ trung bình, cho biết mức độ hình dạng của khối u.
* **concavity\_mean**: Độ lõm trung bình, thể hiện sự hiện diện của các vùng lõm trong khối u.
* **concave points\_mean**: Số lượng điểm lõm trung bình của khối u.
* **symmetry\_mean**: Độ đối xứng trung bình của khối u, cho thấy sự đồng đều trong hình dạng.
* **fractal\_dimension\_mean**: Đánh giá độ phức tạp hình học của khối u.

Ngoài ra, tập dữ liệu còn bao gồm các biến đo độ sai số và các đặc trưng xấu nhất của khối u, như:

* **radius\_worst, texture\_worst, perimeter\_worst, area\_worst**: Các giá trị tồi tệ nhất trong các đặc trưng này, giúp đưa ra cái nhìn sâu sắc hơn về tình trạng bệnh.

## Mục Tiêu Phân Tích

Mục tiêu chính của đồ án này được chia thành hai phần rõ ràng:

### Phân Tích và Trực Quan Hóa Dữ Liệu:

* **Mục đích**: Sử dụng các kỹ thuật phân tích thống kê và trực quan hóa dữ liệu để hiểu rõ hơn về các yếu tố ảnh hưởng đến nguy cơ mắc ung thư vú.
* **Phương pháp**:
* **Phân tích mô tả**: Tóm tắt các đặc điểm của dữ liệu thông qua các chỉ số thống kê như trung bình, độ lệch chuẩn, và phân phối.
* **Trực quan hóa**: Sử dụng biểu đồ (như biểu đồ hộp, biểu đồ phân tán) để minh họa mối quan hệ giữa các biến. Ví dụ, phân tích mối liên hệ giữa kích thước khối u và giai đoạn bệnh hoặc giữa các yếu tố nguy cơ và kết quả chẩn đoán.
* **Phân tích tương quan**: Xác định các mối liên hệ giữa các yếu tố nguy cơ và kết quả bệnh, từ đó tìm ra các yếu tố có ảnh hưởng lớn nhất đến nguy cơ mắc bệnh.

### Dự Đoán Nguy Cơ Mắc Bệnh:

* **Mục đích**: Phát triển các mô hình học máy để dự đoán nguy cơ mắc ung thư vú dựa trên các yếu tố đầu vào từ tập dữ liệu.
* **Phương pháp**:
  + **Xây dựng mô hình**: Áp dụng các thuật toán học máy như hồi quy logistic, cây quyết định, và hồi quy hồi quy để xây dựng mô hình dự đoán.
  + **Đánh giá mô hình**: Sử dụng các chỉ số như độ chính xác, độ nhạy và độ đặc hiệu để đánh giá hiệu suất của mô hình.
  + **Khuyến nghị lâm sàng**: Dựa trên kết quả phân tích và dự đoán, cung cấp các khuyến nghị cho bác sĩ và nhà nghiên cứu trong việc phát hiện sớm và điều trị bệnh ung thư vú.

### Các bước thực hiện

* **Tiền xử lý dữ liệu**: Làm sạch và chuẩn hóa dữ liệu để đảm bảo tính chính xác cho các phân tích tiếp theo.
* **Phân tích các biến liên quan**: Khám phá và đánh giá mối quan hệ giữa các đặc điểm của khối u và biến mục tiêu.
* **Đề xuất và đánh giá mô hình dự báo**: Xây dựng các mô hình học máy để dự đoán tình trạng khối u, bao gồm hồi quy logistic và cây quyết định.
* **Xây dựng dashboard**: Tạo giao diện trực quan giúp người dùng dễ dàng tương tác với dữ liệu và hiểu rõ hơn về kết quả phân tích.
* **Viết báo cáo**: Tổng hợp và trình bày tất cả các kết quả nghiên cứu một cách rõ ràng, làm nổi bật các phát hiện chính và các khuyến nghị cho nghiên cứu trong tương lai.

# CHƯƠNG 3 : TIỀN XỬ LÝ DỮ LIỆU

## Khám phá dữ liệu

Trong hành trình tìm hiểu và phân tích dữ liệu bước đầu tiên là thực hiện chính là khám phá dữ liệu. Đây không chỉ là một thao tác đơn thuần mà là một cuộc hành trình khám phá những câu chuyện tiềm ẩn và những bí mật mà dữ liệu mang trong mình. Khám phá dữ liệu giúp chúng ta có cái nhìn tổng quát rõ nét hơn về cấu trúc, chất lượng và các yếu tố ảnh hưởng đến dữ liệu từ đó định hướng cho các bước phân tích sâu hơn.

### Kiểm tra dữ liệu thiếu

Kiểm tra dữ liệu thiếu là một công đoạn cần thiết trong quy trình tiền xử lý dữ liệu, đóng vai trò quan trọng trong việc đảm bảo rằng tập dữ liệu không chỉ đầy đủ mà còn đáng tin cậy trước khi tiến hành phân tích và xây dựng mô hình. Dữ liệu thiếu không chỉ đơn giản là những khoảng trống trong tập dữ liệu mà chúng còn có thể dẫn đến những hiểu lầm nghiêm trọng trong các phân tích và dự đoán, ảnh hưởng đến các quyết định quan trọng, đặc biệt là trong lĩnh vực y tế.

* **Tại sao phải kiểm tra dữ liệu thiếu?**
* Độ tin cậy của mô hình: để xây dựng một mô hình học máy hiệu quả dữ liệu đầu vào cần phải đầy đủ và chính xác. Sự thiếu hụt trong bất kỳ biến nào có thể dẫn đến việc mô hình không học được các mối quan hệ quan trọng trong dữ liệu từ đó làm giảm độ chính xác của các dự đoán.
* Kết quả phân tích đáng tin cậy: nhiều phương pháp thống kê yêu cầu dữ liệu không có giá trị thiếu để đảm bảo rằng các phân tích đưa ra các kết luận chính xác. Giá trị thiếu có thể làm sai lệch các phép thử và dẫn đến những quyết định không hợp lý.
* Hệ lụy trong quyết định y tế: trong lĩnh vực y tế nơi mà tính chính xác và độ tin cậy là tối quan trọng dữ liệu thiếu có thể dẫn đến những quyết định sai lầm, ảnh hưởng nghiêm trọng đến chăm sóc bệnh nhân và kết quả điều trị. Mỗi chuẩn đoán đều có thể trở thành yếu tố quyết định trong việc đưa ra phác đồ điều trị hợp lý.
* Chi phí kinh tế: dữ liệu thiếu không chỉ ảnh hưởng đến chất lượng phân tích mà còn có thể gây tốn kém về mặt tài chính. Việc khắc phục những thiếu sót trong dữ liệu có thể đòi hỏi thời gian và nguồn lực, dẫn đến việc chậm trễ trong nghiên cứu và phát triển.

### Phương Pháp Kiểm Tra Dữ Liệu Thiếu

Kiểm tra dữ liệu thiếu là một quy trình quan trọng và không thể thiếu trong hành trình phân tích dữ liệu. Nó không chỉ giúp chúng ta phát hiện những khoảng trống trong thông tin mà còn đảm bảo rằng chúng ta đang làm việc với một tập dữ liệu đầy đủ và đáng tin cậy. Dưới đây là những phương pháp hiệu quả để thực hiện kiểm tra dữ liệu thiếu.

* **Đếm Giá Trị Thiếu**

Sử dụng hàm isnull(): Một trong những cách đơn giản và trực quan nhất để xác định số lượng giá trị thiếu trong từng biến là sử dụng hàm isnull() trong Pandas. Phương pháp này cho phép chúng ta nhanh chóng nắm bắt mức độ thiếu hụt của dữ liệu, từ đó đưa ra những hành động cần thiết.

* **Tạo Bảng Tóm Tắt Dữ Liệu Thiếu**

Tổng hợp thông tin: Sau khi xác định số lượng giá trị thiếu, chúng ta có thể tổng hợp thông tin này vào một bảng tóm tắt. Bảng này không chỉ trình bày số lượng và tỷ lệ phần trăm của dữ liệu thiếu cho từng biến mà còn giúp chúng ta dễ dàng nhận diện những biến cần được chú trọng. Một cái nhìn tổng quát sẽ giúp định hướng cho các quyết định tiếp theo.

* **Trực Quan Hóa Dữ Liệu Thiếu**

Biểu đồ trực quan: Việc trực quan hóa dữ liệu thiếu thông qua các biểu đồ như biểu đồ cột hoặc biểu đồ tròn sẽ giúp người phân tích dễ dàng nhận ra các mẫu và xu hướng. Sử dụng các thư viện như Matplotlib hoặc Seaborn trong Python sẽ biến những con số khô khan thành những hình ảnh sống động và đáng chú ý.

* **Phân Tích Theo Nhóm**

Kiểm tra dữ liệu thiếu theo nhóm: Phân tích dữ liệu thiếu theo từng nhóm có thể cung cấp cái nhìn sâu sắc về cách thức mà dữ liệu thiếu phân bổ. Điều này không chỉ giúp xác định các yếu tố có thể liên quan đến việc thiếu dữ liệu mà còn mở ra những hướng đi mới trong việc xử lý thông tin.

* **Ứng Dụng Các Kỹ Thuật Thống Kê**

Phân tích hệ số hồi quy: Trong quá trình xây dựng mô hình hồi quy việc kiểm tra các hệ số hồi quy có thể chỉ ra rằng một số biến không đủ dữ liệu, điều này có thể ảnh hưởng đến độ chính xác của mô hình. Việc này giúp chúng ta điều chỉnh và cải thiện mô hình một cách hiệu quả và chính xác hơn.

## Tiền xử lý dữ liệu

Sau khi đã hoàn thành giai đoạn khám phá dữ liệu chuyển sang bước tiền xử lý, một giai đoạn quan trọng nhằm làm sạch và chuẩn hóa dữ liệu sẽ tạo ra nền tảng vững chắc cho các phân tích và mô hình học máy tiếp theo. Đây là thời điểm chuẩn bị cho những khám phá sâu hơn, đảm bảo rằng dữ liệu hoàn toàn sẵn sàng các bước phân tích tiếp theo.

### Xóa Dữ Liệu Thiếu

Trong quá trình khám phá dữ liệu đã sử lý các giá trị thiếu. Tuy nhiên, việc chuẩn bị cho các tình huống có thể xảy ra trong tương lai vẫn là một yếu tố vô cùng quan trọng. Trong những nghiên cứu khác, việc xử lý dữ liệu thiếu có thể trở thành một thách thức đáng kể, đòi hỏi phải có những chiến lược ứng phó hiệu quả

* **Loại bỏ mẫu**: Đối với những mẫu có giá trị thiếu, nếu số lượng này nhỏ và không ảnh hưởng đến kết quả tổng thể, chúng tôi sẽ cân nhắc việc loại bỏ chúng. Việc làm này không chỉ giúp bảo toàn tính chính xác của dữ liệu mà còn giữ cho quy trình phân tích không bị gián đoạn.
* **Điền giá trị**: Trong trường hợp cần thiết, nếu việc loại bỏ mẫu không khả thi hoặc có thể gây ra sự thiên lệch trong kết quả, chúng tôi sẽ sử dụng các giá trị thống kê như giá trị trung bình hoặc trung vị để thay thế cho các giá trị thiếu. Phương pháp này giúp duy trì tính toàn vẹn và độ chính xác của tập dữ liệu, đảm bảo rằng các phân tích tiếp theo vẫn có thể diễn ra một cách hợp lý.

### Chuyển Đổi Biến

Trong hành trình tiền xử lý dữ liệu một trong những bước thiết yếu và quan trọng nhất là mã hóa biến mục tiêu. Việc thực hiện mã hóa không chỉ tạo điều kiện thuận lợi cho quá trình phân tích mà còn đảm bảo rằng các thuật toán học máy có thể tiếp cận và xử lý dữ liệu một cách hiệu quả và chính xác.

Đối với nghiên cứu bệnh ung thư vú việc chuyển đổi biến diagnosis thành các giá trị số theo quy tắc sau:

* **B (Lành tính)** → 0
* **M (Ác tính)** → 1

### Chuẩn hóa dữ liệu

* Để tạo ra một nền tảng vững chắc cho các mô hình học máy, nhóm thực hiện chuẩn hóa dữ liệu bằng phương pháp MinMaxScaler. Phương pháp này sẽ đưa tất cả các giá trị về khoảng [0, 1], đảm bảo rằng mọi biến liên tục đều nằm trong một khoảng giá trị đồng nhất.
* Chuẩn hóa không chỉ đơn thuần là một bước kỹ thuật nó là chìa khóa giúp cải thiện hiệu suất của các mô hình học máy. Khi các giá trị được điều chỉnh về cùng một quy chuẩn để giảm thiểu độ lệch do sự khác biệt về quy mô giữa các biến. Điều này cho phép các thuật toán học máy hoạt động một cách tối ưu học hỏi và đưa ra dự đoán chính xác hơn.
* Hơn nữa, việc tạo ra sự đồng nhất giữa các biến sẽ giúp giảm thiểu nguy cơ mắc phải các vấn đề trong quá trình tối ưu hóa từ đó giúp mô hình hội tụ nhanh chóng và ổn định hơn. Với sự chuẩn bị kỹ lưỡng này, dữ liệu sẽ trở thành một nguồn tài nguyên quý giá mở ra những cơ hội khám phá sâu sắc hơn về bệnh ung thư vú và nâng cao chất lượng chăm sóc sức khỏe cho bệnh nhân.

### Kiểm tra dữ liệu sau khi tiền xử lý

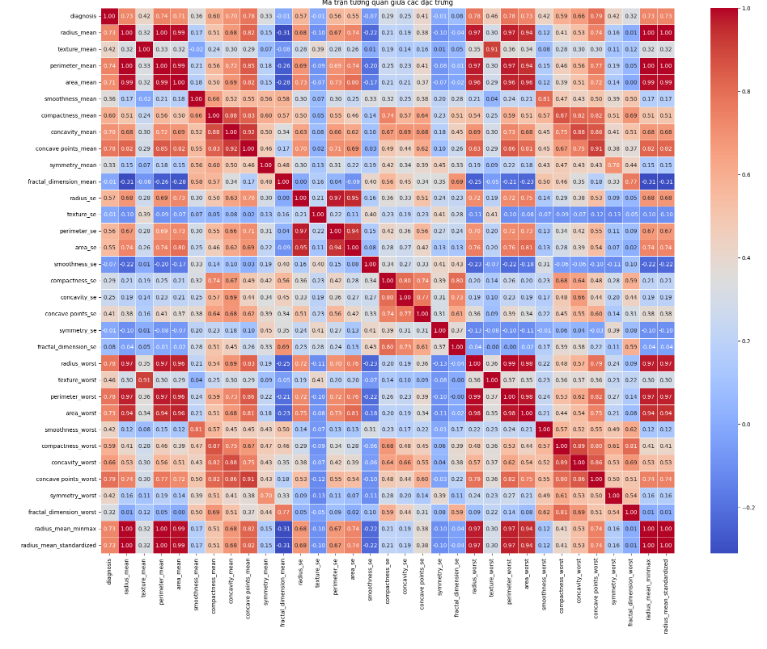
Sau khi hoàn tất các bước tiền xử lý, sẽ tiến hành kiểm tra lại dữ liệu để đảm bảo rằng mọi biến đã được mã hóa và chuẩn hóa một cách chính xác. Bước kiểm tra này không chỉ giúp củng cố sự tự tin trong quá trình phân tích mà còn bảo đảm rằng dữ liệu sử dụng là chính xác và đáng tin cậy.

Để thực hiện kiểm tra này, sẽ áp dụng một số phương pháp thống kê cơ bản. Cụ thể, các biến đã được chuẩn hóa sẽ được mô tả để nắm bắt các chỉ số quan trọng như giá trị trung bình, độ lệch chuẩn, giá trị tối thiểu và tối đa. Đồng thời, cũng sẽ xác minh số lượng các giá trị trong biến mục tiêu nhằm đảm bảo rằng việc mã hóa đã diễn ra đúng cách.

## Trực quan hóa dữ liệu

Trực quan hóa dữ liệu phân loại là một bước quan trọng trong quá trình phân tích dữ liệu để hiểu rõ hơn về các đặc điểm phân phối của các lớp dữ liệu. Dưới đây là một số phương pháp trực quan hóa phổ biến khi làm việc với dữ liệu phân loại, đặc biệt đối với tập dữ liệu breastcancer đang xử lý.

1. **Ma trận tương quan**

****

Mạnh: |r| > 0.7 (hoặc 0.8 tùy theo lĩnh vực)

Trung bình: 0.5 < |r| ≤ 0.7 (hoặc 0.5 < |r| ≤ 0.8)

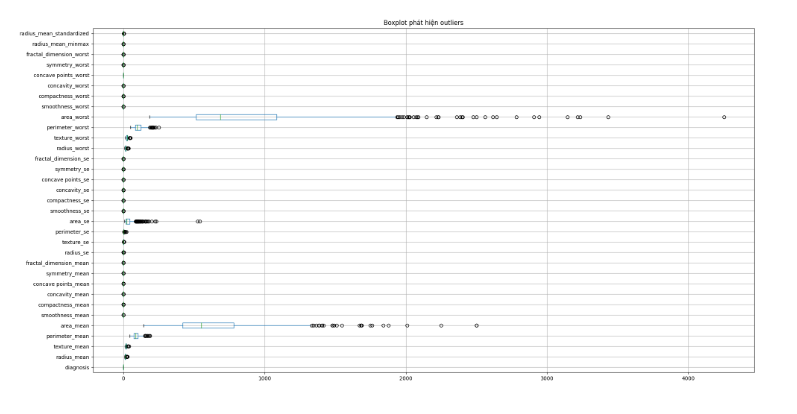
Yếu: |r| ≤ 0.5

**Nhận xét:**

* **Các nhóm biến có tương quan mạnh:**
* **Kích thước (Mean):** radius\_mean, perimeter\_mean, và area\_mean có tương quan mạnh mẽ với nhau (gần 1). Điều này là hiển nhiên do mối quan hệ hình học giữa bán kính, chu vi và diện tích. Chỉ cần một trong ba biến này là đủ để đại diện cho kích thước khối u trong mô hình.
* **Kích thước (Worst):** radius\_worst, perimeter\_worst, và area\_worst cũng có tương quan cực kỳ mạnh mẽ với nhau (gần 1), tương tự như nhóm mean.
* **Hình dạng (Mean và Worst):** Các biến liên quan đến hình dạng (compactness, concavity, concave points) của cả nhóm mean và worst có tương quan mạnh mẽ với nhau. Ví dụ, concavity\_mean có tương quan mạnh với concave points\_mean và compactness\_mean. Tương tự, tương quan tương tự được quan sát ở nhóm worst. Điều này cho thấy các đặc điểm hình dạng của khối u có liên hệ chặt chẽ với nhau.
* **Mean vs. Worst:** Nhiều biến mean có tương quan mạnh mẽ với các biến tương ứng của worst. Ví dụ, radius\_mean tương quan mạnh với radius\_worst. Điều này cho thấy mối liên hệ giữa kích thước ban đầu và kích thước cuối cùng của khối u.
* **Tương quan với biến Diagnosis:**
* **Các biến "Worst" có tương quan mạnh nhất:** Các biến "worst" (radius\_worst, perimeter\_worst, area\_worst, compactness\_worst, concavity\_worst, concave points\_worst) có tương quan mạnh mẽ nhất với diagnosis. Tương quan dương mạnh cho thấy các biến này là các chỉ số quan trọng để phân biệt giữa ung thư ác tính và lành tính.
* **Tương quan mạnh của các biến "Mean":** Một số biến "mean" (như radius\_mean, perimeter\_mean, area\_mean, concavity\_mean, concave points\_mean) cũng có tương quan đáng kể với diagnosis, nhưng yếu hơn so với nhóm "worst".
* **Biến SE có tương quan yếu:** Các biến "SE" nói chung có tương quan yếu hơn với diagnosis so với "mean" và "worst".
* **Đa cộng tuyến:**

Như đã đề cập trước đó, sự tương quan mạnh mẽ giữa các biến (đặc biệt là trong các nhóm "mean" và "worst") cho thấy sự hiện diện của đa cộng tuyến. Điều này có thể làm ảnh hưởng đến độ ổn định và khả năng giải thích của mô hình. Vì vậy, cần áp dụng các kỹ thuật giảm chiều như PCA hoặc chọn lọc biến để giải quyết vấn đề này. Ví dụ: có thể giữ lại chỉ một trong ba biến radius\_mean, perimeter\_mean, và area\_mean.

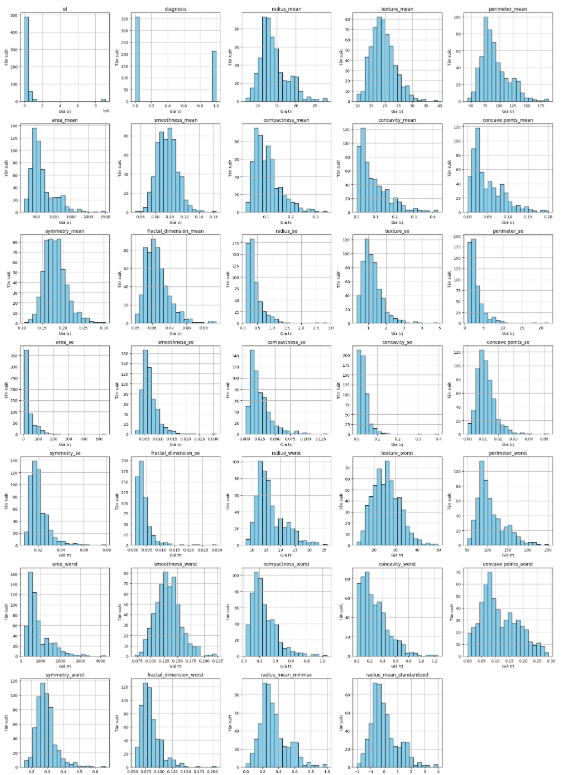
1. **Biểu đồ Boxplot**



**Nhận xét:**

* **Biến Mean:** (radius\_mean, texture\_mean, perimeter\_mean, are
* **Area\_mean và Perimeter\_mean có outliers rõ rệt:** Những biến này có những điểm outlier nằm xa so với phần còn lại của dữ liệu, chỉ ra sự khác biệt đáng kể về kích thước và chu vi của khối u so với phần lớn các trường hợp khác. Đây là những biến quan trọng cần được xem xét kỹ trong việc xây dựng mô hình.
* **Các biến khác:** C
* **Biến Standard Error (SE):** (radius\_se, texture\_se, perimeter\_se, area\_se)
* **Nhiều outliers:** Nhóm biến này có rất nhiều outliers, đặc biệt là ở các biến liên quan đến concavity, compactness, và perimeter. Điều này cho thấy sự thay đổi đáng kể trong các đặc điểm tế bào giữa các lần đo lường khác nhau trong cùng một mẫu. Sự hiện diện nhiều outliers ở đây có thể do tính chất phức tạp của quá trình đo lường hoặc sự biến đổi lớn về hình thái tế bào.
* **Xử lý outliers:** Việc xử lý outliers trong nhóm này cần thận trọng. Loại bỏ chúng có thể dẫn đến mất thông tin quan trọng về sự biến đổi trong dữ liệu. Các phương pháp thay thế như winsorizing hoặc robust statistical methods có thể phù hợp hơn.
* **Biến Worst:** (radius\_worst, texture\_worst)
* **Outliers tập trung ở đuôi phải:** Tương tự như biến mean, nhóm này cũng hiển thị rõ ràng các outliers ở đuôi phải. Điều này đặc biệt đáng chú ý đối với area\_worst và perimeter\_worst, cho thấy sự khác biệt đáng kể trong kích thước và chu vi của khối u ở giai đoạn cuối.
* **Ảnh hưởng đến phân tích:** Những outliers này có thể ảnh hưởng lớn đến các mô hình phân tích thống kê, đặc biệt là các mô hình nhạy cảm với outliers như hồi quy tuyến tính.
* **Biến Diagnosis:**
* **Không phải là dữ liệu liên tục:** Biến này là biến phân loại, vì vậy boxplot không thể cung cấp thông tin phân bố chi tiết. Tuy nhiên, ta có thể ước lượng tỷ lệ giữa malignant và benign từ chiều dài của boxplot.

1. **Biểu đồ histogram**



**Nhận xét chi tiết từng nhóm biến:**

* **Biến Mean:** Các biến này thường có phân bố lệch phải. area\_mean và perimeter\_mean có phân bố lệch phải mạnh nhất, phù hợp với nhận xét từ boxplot trước đó.
* **Biến Standard Error (SE):** Các biến SE thường có phân bố tập trung hơn, nhưng vẫn có một số biến có đuôi phải dài. Điều này cho thấy sự biến động của các đặc điểm đo lường không phải lúc nào cũng đối xứng.
* **Biến Worst:** Như đã đề cập, các biến "worst" thường cho thấy phân bố lệch phải mạnh mẽ nhất. Điều này có thể phản ánh sự gia tăng đáng kể của các đặc điểm tế bào trong các trường hợp ung thư ác tính.
* **Biến diagnosis:** Biểu đồ cho thấy sự mất cân bằng giữa hai lớp (malignant và benign), với số lượng mẫu lành tính (B) nhiều hơn mẫu ác tính (M). Điều này cần được cân nhắc trong quá trình xây dựng và đánh giá mô hình phân loại. Các kỹ thuật xử lý mất cân bằng lớp như oversampling, undersampling, hoặc cost-sensitive learning có thể được áp dụng.
* **radius\_mean\_standardized:** Biến này đã được chuẩn hóa, có phân phối gần với phân phối chuẩn hơn các biến khác.

**Nhận xét chung:**

* **Phân bố không đồng nhất:** Hầu hết các biến đều không tuân theo phân phối chuẩn (normal distribution). Nhiều biến có vẻ lệch phải (right-skewed), có nghĩa là có nhiều giá trị nhỏ hơn và một số ít giá trị lớn hơn. Điều này có thể ảnh hưởng đến các giả định của một số phương pháp phân tích thống kê.
* **Cần biến đổi dữ liệu:** Do sự không đối xứng của nhiều biến, việc biến đổi dữ liệu (data transformation) có thể cần thiết trước khi tiến hành phân tích. Các biến đổi phổ biến bao gồm logarit (log transformation), bình phương căn bậc hai (square root transformation), hoặc Box-Cox transformation. Việc này sẽ giúp dữ liệu gần giống với phân phối chuẩn hơn, cải thiện độ chính xác của mô hình.
* **Sự khác biệt giữa các nhóm biến:** Có thể thấy sự khác biệt rõ rệt về hình dạng histogram giữa các nhóm biến (mean, SE, worst). Nhóm biến "worst" thường có phân bố lệch phải rõ rệt hơn so với nhóm "mean" và "SE".

## Kết Luận

Chương 3 đã khép lại một giai đoạn quan trọng trong quy trình nghiên cứu, đó là tiền xử lý dữ liệu. Qua từng bước thực hiện, từ việc kiểm tra dữ liệu thiếu đến mã hóa và chuẩn hóa, một nền tảng vững chắc đã được xây dựng, sẵn sàng cho các phân tích sâu sắc và có ý nghĩa hơn trong những chương tiếp theo.

Việc xử lý dữ liệu thiếu không chỉ giúp đảm bảo rằng tập dữ liệu đầy đủ và chính xác, mà còn phản ánh sự cẩn trọng và trách nhiệm trong nghiên cứu. Mã hóa biến diagnosis thành các giá trị số đã tạo điều kiện thuận lợi cho việc áp dụng các kỹ thuật học máy, cho phép khai thác thông tin một cách hiệu quả hơn. Đồng thời, chuẩn hóa các biến liên tục đã giúp loại bỏ sự thiên lệch do sự khác biệt về quy mô, từ đó tối ưu hóa hiệu suất của các mô hình phân tích.

Những bước đi này không chỉ là những thao tác kỹ thuật mà còn là những quyết định chiến lược, thể hiện cam kết đối với chất lượng và độ tin cậy của dữ liệu. Với sự chuẩn bị kỹ lưỡng này, nghiên cứu không chỉ sẵn sàng đối mặt với các thách thức mà còn mở ra cơ hội khám phá những hiểu biết mới về bệnh ung thư vú.

Khi tiến vào chương tiếp theo, những nền tảng vững chắc đã được thiết lập sẽ tạo điều kiện cho những phân tích sâu sắc hơn, khám phá các mối quan hệ phức tạp và từ đó, góp phần nâng cao chất lượng chăm sóc sức khỏe cho cộng đồng. Sự tỉ mỉ trong mỗi bước chuẩn bị sẽ là chìa khóa để đạt được những kết quả có giá trị và ứng dụng thực tiễn trong lĩnh vực y tế.

# CHƯƠNG 4 : PHÂN TÍCH DỮ LIỆU

## Phân tích dữ liệu

### Hồi quy tuyến tính

## Phân Tích Mối Quan Hệ Tuyến Tính Giữa Độ Nhẵn và Độ Lõm trong Khối U Vú

**Câu Hỏi Nghiên Cứu:**

Liệu Độ nhẵn của khối u có liên quan đến sự gia tăng Độ lõm bên trong khối u không?

* **Độ nhẵn (smoothness\_mean)**: Mức độ mượt mà của bề mặt khối u, đo lường sự đồng đều của các điểm trên bề mặt khối u. Là một biến định lượng.
* **Độ lõm bên trong khối u (concave points\_mean)**: Đo lường độ lõm vào của các điểm biên của khối u. Là một biến định lượng.

**Phương Pháp Phân Tích:**

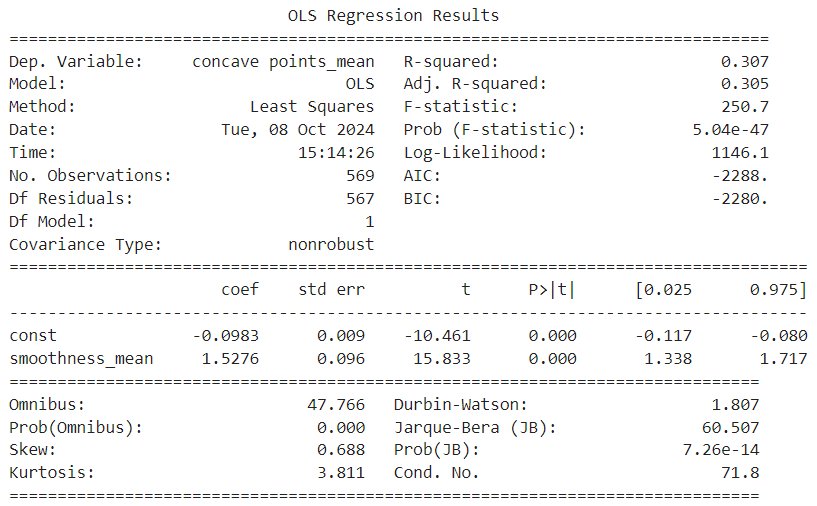
Để trả lời câu hỏi trên, tôi đã thực hiện phân tích **hồi quy tuyến tính** giữa hai biến định lượng **Độ nhẵn** và **Độ lõm**.

* **Phương trình hồi quy tuyến tính**:

concave points\_mean=−0.0983+1.5276×smoothness\_mean

Phương trình này cho thấy rằng, mỗi khi **Độ nhẵn** tăng thêm 1 đơn vị, **Độ lõm** dự kiến sẽ tăng thêm khoảng **1.5276 đơn vị**.

**Kết Quả Phân Tích Hồi Quy Tuyến Tính:**

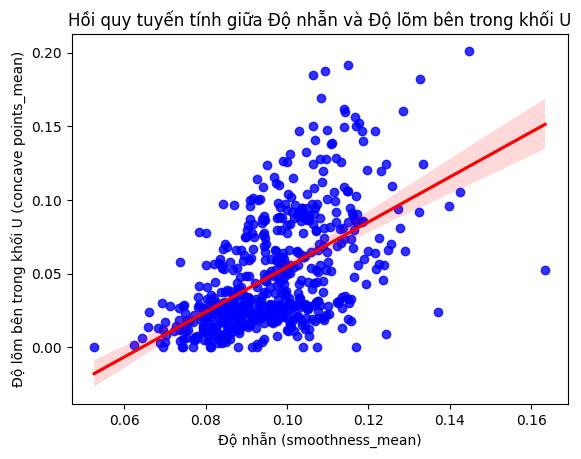


* **F-statistic** = 250.7 và **p-value** = 5.04e-47:  
  Giá trị p-value rất nhỏ cho thấy mô hình hồi quy này có ý nghĩa thống kê cao và mối quan hệ giữa **Độ nhẵn** và **Độ lõm** là có ý nghĩa.
* **R-squared** = 0.307:  
  Mô hình giải thích được khoảng **30.7%** sự biến động trong **Độ lõm** dựa trên **Độ nhẵn**. Mặc dù mức độ giải thích không quá cao, nhưng vẫn có mối quan hệ rõ ràng giữa hai biến này.

**Kết Luận:**

Từ kết quả hồi quy, có thể kết luận rằng có một **mối quan hệ tuyến tính dương** giữa **Độ nhẵn** và **Độ lõm** bên trong khối u. Khi **Độ nhẵn** của khối u tăng, **Độ lõm** cũng có xu hướng tăng theo. Điều này có thể chỉ ra rằng các khối u có bề mặt nhẵn hơn thường có xu hướng có các điểm lõm sâu hơn.

**Biểu Đồ Trực Quan:**



Biểu đồ hồi quy dưới đây cho thấy **đường hồi quy** (màu đỏ) và các điểm dữ liệu (màu xanh). Đường hồi quy thể hiện rõ mối quan hệ dương giữa hai biến, với các điểm dữ liệu có xu hướng bám sát đường hồi quy, mặc dù vẫn có một số sự phân tán.

Code:  
import pandas as pd

import statsmodels.api as sm

import statsmodels.formula.api as smf

import seaborn as sns

import matplotlib.pyplot as plt

# Đọc dữ liệu từ file CSV

data = pd.read\_csv('/kaggle/input/breast-cancer-2001216269/17-breast-cancer.csv')

# Chọn 2 biến định lượng cần phân tích

X = data['smoothness\_mean']

Y = data['concave points\_mean']

# Thêm hằng số vào X để thực hiện hồi quy tuyến tính

X = sm.add\_constant(X)

# Thực hiện hồi quy tuyến tính

model = sm.OLS(Y, X).fit()

# In kết quả của hồi quy

print(model.summary())

# Vẽ biểu đồ hồi quy tuyến tính

sns.regplot(x='smoothness\_mean', y='concave points\_mean', data=data, scatter\_kws={"color": "blue"}, line\_kws={"color": "red"})

plt.xlabel('Độ nhẵn (smoothness\_mean)')

plt.ylabel('Độ lõm bên trong khối U (concave points\_mean)')

plt.title('Hồi quy tuyến tính giữa Độ nhẵn và Độ lõm bên trong khối U')

plt.show()

**Phân Tích Mối Quan Hệ Tuyến Tính Giữa Độ Lõm và Khả Năng Ác Tính của Khối U Vú**

**Câu Hỏi Nghiên Cứu:**

Liệu **Độ lõm** có thể là một yếu tố dự đoán cho **khả năng ác tính** của khối u vú không?

* **Độ lõm (concave points\_mean)**: Đo lường độ sâu của các điểm lõm trên bề mặt khối u. Là một biến định lượng.
* **Khả năng ác tính (diagnosis)**: Biến phân loại cho biết khối u là lành tính (0) hay ác tính (1). Là một biến phân loại.

**Phương Pháp Phân Tích:**

Để trả lời câu hỏi trên, tôi đã thực hiện phân tích **hồi quy tuyến tính** giữa **Độ lõm** và **Khả năng ác tính**. Phương trình hồi quy tuyến tính được mô tả như sau:

concave points\_mean=0.0257+0.0623×diagnosis

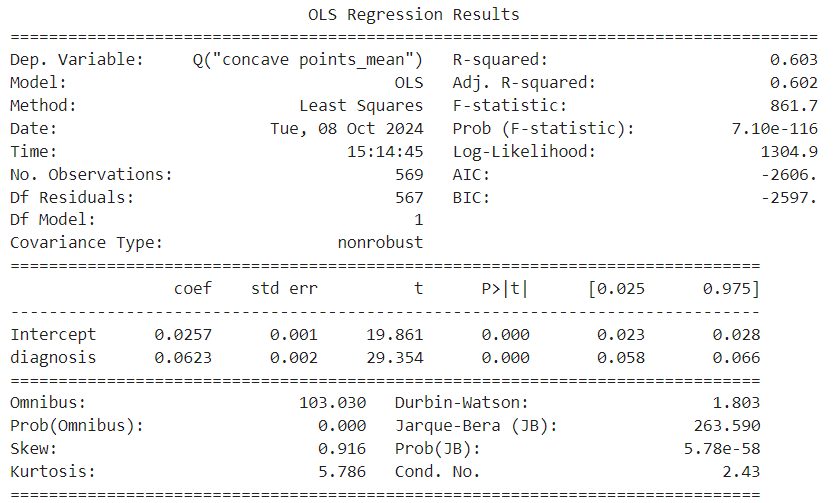
* **Đối với khối u lành tính (diagnosis = 0)**:

concave points\_mean=0.0257+0.0623×0=0.0257

* **Đối với khối u ác tính (diagnosis = 1)**:

concave points\_mean=0.0257+0.0623×1=0.0257+0.0623=0.0880

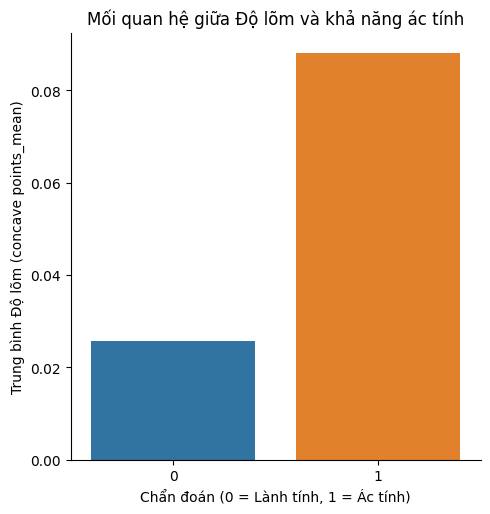
**Kết Quả Phân Tích Hồi Quy Tuyến Tính:**



* **F-statistic** = 861.7 và **p-value** = 7.10e-116:  
  Giá trị p-value rất nhỏ cho thấy mô hình hồi quy có ý nghĩa thống kê cao, và **Độ lõm** có mối quan hệ rõ rệt với **Khả năng ác tính** của khối u.
* **R-squared** = 0.603:  
  Mô hình giải thích được **60.3%** sự biến động trong **Độ lõm** dựa trên **Khả năng ác tính**.

**Kết Luận:**

Từ kết quả hồi quy, có thể kết luận rằng **Độ lõm** của khối u vú có mối quan hệ tuyến tính với **Khả năng ác tính**. Các khối u ác tính có **Độ lõm** cao hơn so với các khối u lành tính, cho thấy rằng **Độ lõm** có thể là một yếu tố dự đoán khả năng ác tính của khối u.

**Biểu Đồ Trực Quan:**

Biểu đồ cột dưới đây thể hiện **giá trị trung bình** của **Độ lõm** cho cả hai nhóm:

* Nhóm lành tính (diagnosis = 0) có giá trị trung bình khoảng **0.0257**.
* Nhóm ác tính (diagnosis = 1) có giá trị trung bình khoảng **0.088**.

Sự khác biệt rõ rệt giữa hai nhóm cho thấy **Độ lõm** của các khối u ác tính có xu hướng cao hơn so với các khối u lành tính.

Code:  
# Chuyển đổi cột "diagnosis" và "concave points\_mean" sang dạng số nếu chưa

data['diagnosis'] = data['diagnosis'].apply(lambda x: 1 if x == 'M' else 0) # 1 là ác tính, 0 là lành tính

data['concave points\_mean'] = pd.to\_numeric(data['concave points\_mean'], errors='coerce')

# Lọc dữ liệu cho các giá trị không NaN trong cột 'concave points\_mean'

data\_clean = data[['concave points\_mean', 'diagnosis']].dropna()

# Thực hiện hồi quy tuyến tính

reg = smf.ols('Q("concave points\_mean") ~ diagnosis', data=data\_clean).fit()

# In ra kết quả hồi quy

print(reg.summary())

# Vẽ biểu đồ cột để trực quan hóa

sns.catplot(x='diagnosis', y='concave points\_mean', data=data\_clean, kind='bar', errorbar=None)

plt.xlabel('Chẩn đoán (0 = Lành tính, 1 = Ác tính)')

plt.ylabel('Trung bình Độ lõm (concave points\_mean)')

plt.title('Mối quan hệ giữa Độ lõm và khả năng ác tính')

plt.show()

### Hồi quy Logistic

**Phân Tích Hồi Quy Logistic Giữa Kích Thước Khối U và Khả Năng Ác Tính**

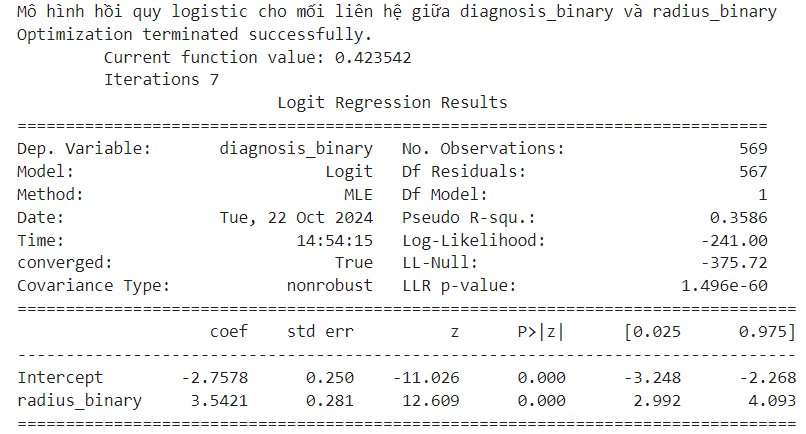
**Câu Hỏi Nghiên Cứu:**

Liệu kích thước khối u có ảnh hưởng đến khả năng ác tính của khối u vú? Cụ thể, chúng tôi muốn phân tích mối liên hệ giữa **radius\_binary** (kích thước khối u: to hay nhỏ) và **diagnosis\_binary** (khả năng ác tính: có hay không).

**Dữ Liệu:**

* **radius\_binary**: Biến nhị phân được tạo ra từ **radius\_mean**. Nếu **radius\_mean** lớn hơn giá trị trung vị, biến này nhận giá trị 1 (khối u lớn), ngược lại nhận giá trị 0 (khối u nhỏ).
* **diagnosis\_binary**: Biến nhị phân thể hiện khả năng ác tính của khối u, với giá trị 1 nếu khối u ác tính và 0 nếu khối u lành tính.

**Mô Hình Hồi Quy Logistic:**



Trong đó:

* **Intercept (Giao điểm)** = -2.76: Khi khối u nhỏ (radius\_binary = 0), khả năng khối u ác tính rất thấp.
* **radius\_binary = 3.54**: Nếu khối u lớn (radius\_binary = 1), khả năng khối u ác tính sẽ tăng 3.54 lần so với khối u nhỏ.

**Kết Quả Hồi Quy Logistic:**

* **Hệ số Intercept** = -2.76: Giá trị này cho thấy khi kích thước khối u nhỏ, khả năng khối u ác tính rất thấp.
* **Hệ số radius\_binary** = 3.54: Nếu khối u lớn, khả năng khối u ác tính tăng **3.54 lần** so với khối u nhỏ.
* **Pseudo R-squared** = 0.3586: Mô hình giải thích khoảng 35.86% sự biến động trong khả năng ác tính dựa trên kích thước khối u.
* **p-value (cho biến radius\_binary)** = 1.496e-60: Giá trị p rất nhỏ chứng tỏ rằng kích thước khối u có mối quan hệ đáng kể với khả năng ác tính.

**Tỷ Số Chênh (Odds Ratio):**

* **ODD = 3.54**: Điều này có nghĩa là khối u lớn có khả năng ác tính cao hơn **3.5 lần** so với khối u nhỏ.

**Hãy suy ngẫm:**

Mặc dù mô hình hồi quy logistic cho thấy mối quan hệ mạnh mẽ giữa kích thước khối u và khả năng ác tính, nhưng cần lưu ý rằng **kích thước khối u lớn không phải là nguyên nhân trực tiếp** gây ra tính ác tính. Đây chỉ là một mối liên hệ được phát hiện trong dữ liệu và có thể có những yếu tố khác cũng góp phần vào khả năng ác tính của khối u.

**Biểu Đồ Trực Quan:**

Ảnh có chứa văn bản, ảnh chụp màn hình, biểu đồ, Sơ đồ

Description automatically generated

* Khi kích thước khối u **lớn** (biến **radius\_binary** = 1), xác suất khối u là ác tính **tăng cao**. Đây là dấu hiệu cho thấy khối u lớn có xu hướng ác tính hơn so với khối u nhỏ.
* Ngược lại, khi kích thước khối u **nhỏ** (biến **radius\_binary** = 0), xác suất khối u là ác tính thấp hơn rất nhiều.
* Biểu đồ này làm nổi bật sự liên quan mạnh mẽ giữa kích thước của khối u và khả năng ác tính. Mặc dù mô hình không thể chắc chắn xác định tất cả các trường hợp, nhưng nó cho thấy rằng kích thước lớn là một yếu tố có thể giúp xác định khối u ác tính.

Code:  
import numpy as np

import pandas as pd

import statsmodels.api as sm

import statsmodels.formula.api as smf

import seaborn as sns

import matplotlib.pyplot as plt

# Đọc dữ liệu

df = pd.read\_csv('/kaggle/input/breast-cancer-2001216269/17-breast-cancer.csv')

# Tạo biến nhị phân dựa trên giá trị trung vị của 'radius\_mean'

median\_radius = df['radius\_mean'].median()

df['radius\_binary'] = (df['radius\_mean'] > median\_radius).astype(int)

# Chuyển đổi cột 'diagnosis' thành nhị phân

df['diagnosis\_binary'] = df['diagnosis'].apply(lambda x: 1 if x == 'M' else 0)

# Xây dựng và huấn luyện mô hình hồi quy logistic

reg\_logit = smf.logit('diagnosis\_binary ~ radius\_binary', data=df).fit()

# In bảng tóm tắt kết quả hồi quy logistic

print(reg\_logit.summary())

import seaborn as sns

import matplotlib.pyplot as plt

# Vẽ biểu đồ Bar Plot

sns.catplot(x='radius\_binary', hue='diagnosis\_binary', kind='count', data=df, height=6, aspect=1.5)

plt.title('Sự phân bố của Khả Năng Ác Tính Theo Kích Thước Khối U')

plt.xlabel('Kích Thước Khối U (0: Nhỏ, 1: Lớn)')

plt.ylabel('Số lượng Khối U')

plt.xticks([0, 1], ['Khối u Nhỏ', 'Khối u Lớn'])

# Hiển thị biểu đồ

plt.show()

**Phân Tích Hồi Quy Logistic Giữa Ba Biến Giải Thích và Khả Năng Ác Tính**

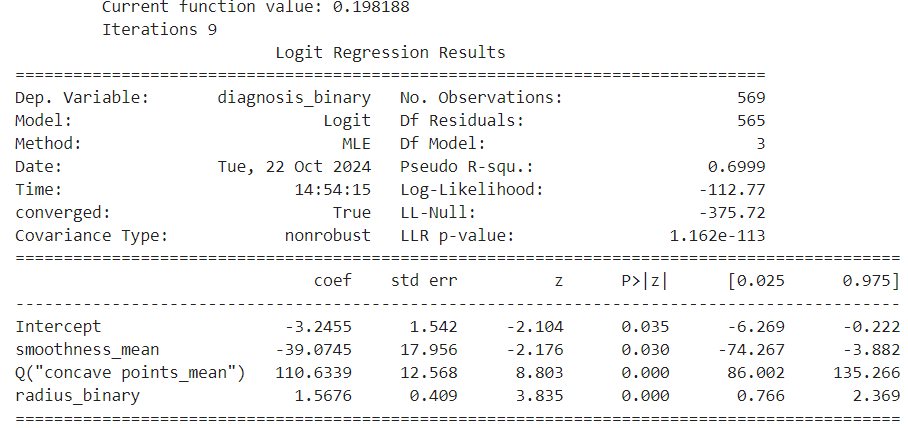
**Câu Hỏi Nghiên Cứu:**

Liệu độ nhẵn, độ lõm và kích thước khối u có thể dự đoán khả năng ác tính của khối u không?

**Phương Pháp:**

Áp dụng mô hình hồi quy logistic với ba biến giải thích: **smoothness\_mean**, **concave points\_mean**, và **radius\_binary** để phân tích khả năng ác tính (**diagnosis\_binary**).

**Mô hình hồi quy logistic cho kết quả:**



**Kết Quả Mô Hình Hồi Quy:**

* Intercept (Giao điểm): -3.25  
  Khi tất cả các biến giải thích bằng 0 (khối u nhỏ, nhẵn và không lõm), khả năng khối u ác tính rất thấp.
* smoothness\_mean: -39.07  
  Với mỗi đơn vị tăng của độ nhẵn, khả năng khối u ác tính giảm mạnh (ODD giảm gấp 39.07 lần). Điều này cho thấy khối u có độ nhẵn cao có xu hướng ít ác tính hơn.
* concave points\_mean: 110.63  
  Với mỗi đơn vị tăng của độ lõm, khả năng khối u ác tính tăng mạnh (ODD tăng gấp 110.63 lần). Điều này cho thấy độ lõm là yếu tố quan trọng nhất trong việc dự đoán khả năng ác tính của khối u.
* radius\_binary: 1.57  
  Nếu khối u lớn, khả năng khối u ác tính tăng gấp 1.57 lần so với khối u nhỏ. Mặc dù không mạnh bằng độ lõm, kích thước khối u vẫn là yếu tố quan trọng trong việc phân biệt giữa các khối u lành tính và ác tính.

**Giải Thích Kết Quả:**

* Độ nhẵn (smoothness\_mean): Độ nhẵn của khối u có tác động ngược chiều với khả năng ác tính. Cụ thể, khối u nhẵn hơn có xu hướng ít ác tính hơn, với ODD giảm gấp 39.07 lần khi độ nhẵn tăng thêm 1 đơn vị.
* Độ lõm (concave points\_mean): Độ lõm là yếu tố có ảnh hưởng mạnh nhất đến khả năng ác tính. Mỗi đơn vị tăng của độ lõm làm tăng gấp 110.63 lần khả năng ác tính. Điều này cho thấy các khối u có độ lõm sâu có xu hướng ác tính hơn rất nhiều so với các khối u có bề mặt nhẵn.
* Kích thước khối u (radius\_binary): Mặc dù không mạnh như độ lõm, kích thước khối u vẫn có tác động đáng kể đến khả năng ác tính. Các khối u lớn có khả năng ác tính cao hơn so với các khối u nhỏ.

**Kết Luận:**

* Độ lõm (concave points\_mean) là yếu tố mạnh mẽ nhất trong việc dự đoán khả năng ác tính của khối u. Khối u có độ lõm sâu hơn có khả năng ác tính cao hơn nhiều so với các khối u nhẵn.
* Độ nhẵn (smoothness\_mean) có tác động tiêu cực đến khả năng ác tính. Các khối u nhẵn có xu hướng ít ác tính hơn.
* Kích thước khối u (radius\_binary) cũng ảnh hưởng đến khả năng ác tính, nhưng yếu tố độ lõm có tác động mạnh mẽ nhất.

Code:  
import statsmodels.formula.api as smf

# In thông báo về mô hình hồi quy logistic

print("Mô hình hồi quy logistic với các biến giải thích: smoothness\_mean, concave points\_mean, radius\_binary")

# Xây dựng và huấn luyện mô hình hồi quy logistic

reg\_logit = smf.logit('diagnosis\_binary ~ smoothness\_mean + Q("concave points\_mean") + radius\_binary', data=df).fit()

# In bảng tóm tắt kết quả hồi quy logistic

print(reg\_logit.summary())

### Cây quyết định

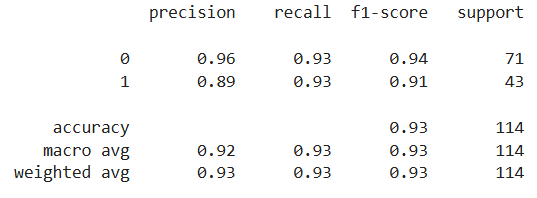
**Phân Tích Mô Hình Cây Quyết Định cho Dự Báo Khả Năng Ác Tính của Khối U Ung Thư Vú**

**Giới thiệu Mô Hình**

Trong bài toán phân loại này, chúng ta đã sử dụng mô hình **Cây Quyết Định (Decision Tree)** để dự đoán khả năng ác tính của các khối u ung thư vú. Các đặc trưng đầu vào là các thuộc tính của bộ dữ liệu ung thư vú, bao gồm các chỉ số như độ nhẵn (smoothness\_mean), độ lõm (concave points\_mean), và các đặc trưng khác liên quan đến kích thước và hình dáng khối u.

Mô hình cây quyết định được huấn luyện trên một tập dữ liệu huấn luyện và kiểm tra hiệu suất của mô hình trên tập dữ liệu kiểm tra.

**Kết Quả Mô Hình:**

****

Dưới đây là các chỉ số chính của mô hình phân loại cây quyết định:

* **Độ chính xác (Accuracy):** Mô hình đạt được độ chính xác là **93%** trên tập kiểm tra, cho thấy mô hình có khả năng phân loại tốt giữa hai lớp: **Benign** (khối u lành tính) và **Malignant** (khối u ác tính).
* **Precision, Recall, F1-Score:**
  + **Benign (0)**:
    - **Precision:** 0.96, có nghĩa là khi mô hình dự đoán một khối u lành tính, 96% trong số đó là đúng.
    - **Recall:** 0.93, có nghĩa là 93% trong số các khối u lành tính thực tế đã được mô hình nhận diện đúng.
    - **F1-Score:** 0.94, là chỉ số tổng hợp giữa Precision và Recall, cho thấy mô hình đạt hiệu suất rất tốt trong việc phân loại khối u lành tính.
  + **Malignant (1)**:
    - **Precision:** 0.89, có nghĩa là khi mô hình dự đoán một khối u ác tính, 89% trong số đó là đúng.
    - **Recall:** 0.93, có nghĩa là 93% trong số các khối u ác tính thực tế đã được mô hình nhận diện đúng.
    - **F1-Score:** 0.91, cho thấy mô hình cũng có hiệu suất cao trong việc phân loại các khối u ác tính.
* **Macro Average**: Tính toán trung bình của các chỉ số Precision, Recall và F1-Score cho tất cả các lớp (Benign và Malignant). Các chỉ số này đều đạt khoảng **0.92 - 0.93**, cho thấy mô hình có khả năng phân loại khá đồng đều giữa các lớp.
* **Weighted Average**: Tính toán trung bình có trọng số theo tỉ lệ mẫu của các lớp. Kết quả cũng dao động từ **0.93** cho cả Precision, Recall và F1-Score, cho thấy mô hình thực hiện rất ổn định trên cả hai lớp.

**Nhận Xét về Mô Hình**

* **Hiệu suất tốt:** Với độ chính xác **93%** và các chỉ số F1-Score cao cho cả hai lớp, mô hình cây quyết định này hoạt động rất hiệu quả trong việc phân loại khối u ung thư vú thành hai nhóm lành tính và ác tính.
* **Điều chỉnh mô hình:** Mặc dù mô hình hoạt động tốt, nhưng có thể cần điều chỉnh một số siêu tham số (như độ sâu cây, số lượng lá, hoặc các thuật toán cắt tỉa cây) để cải thiện hiệu suất, đặc biệt trong việc phân loại khối u ác tính. Precision của lớp ác tính hơi thấp (0.89), điều này có thể cải thiện với việc điều chỉnh các tham số mô hình.
* **Ý nghĩa trong ứng dụng y tế:** Mô hình cây quyết định có thể được áp dụng trong thực tế để hỗ trợ bác sĩ trong việc dự đoán khả năng ác tính của khối u. Với các đặc trưng như độ nhẵn, độ lõm và kích thước khối u, mô hình có thể đưa ra cảnh báo kịp thời về các khối u có nguy cơ cao.

**Cấu trúc của Cây Quyết Định:**

Ảnh có chứa biểu đồ, hàng, Kế hoạch, thiết kế

Description automatically generated

Cây quyết định là một mô hình phân loại trong đó các quyết định được đưa ra dựa trên các thuộc tính của dữ liệu. Cây chia dữ liệu thành các nhánh từ gốc cho đến các lá, trong mỗi nhánh, mô hình dựa vào giá trị của một thuộc tính cụ thể để phân loại. Dưới đây là một số điểm nổi bật từ cây này:

* **Nút Gốc** (Root Node): Đây là nơi bắt đầu của cây và là bước phân chia đầu tiên trong quá trình ra quyết định. Cây sẽ phân chia dựa trên một đặc trưng cụ thể, và ở đây ta thấy cây quyết định bắt đầu từ thuộc tính **concave points\_mean**. Nếu giá trị của thuộc tính này nhỏ hơn hoặc lớn hơn một ngưỡng cụ thể, mô hình sẽ tiếp tục phân chia xuống các nhánh tiếp theo.
* **Các Nút Nội Bộ**: Mỗi nút trong cây đại diện cho một phân chia tiếp theo dựa trên giá trị của một thuộc tính khác. Trong trường hợp này, cây tiếp tục phân chia dữ liệu theo các đặc trưng như **radius\_mean**, **smoothness\_mean**, và các thuộc tính khác. Mỗi nhánh quyết định xem liệu khối u có phải là ác tính hay lành tính, với các phân chia theo các ngưỡng cụ thể của từng thuộc tính.
* **Các Lá**: Các lá của cây quyết định là nơi mà mô hình đưa ra dự đoán cuối cùng. Trong trường hợp này, các lá sẽ cho kết quả là 0 (Lành tính) hoặc 1 (Ác tính), tùy thuộc vào giá trị của các thuộc tính ở các nút phân chia.
* **Các Quyết Định Quan Trọng:**
* **Concave Points Mean**: Thuộc tính **concave points\_mean** có vẻ đóng vai trò quan trọng nhất trong việc phân loại các khối u lành tính và ác tính. Cây quyết định sử dụng nó làm phân chia chính tại nút gốc. Điều này cho thấy rằng độ lõm của khối u có mối liên hệ chặt chẽ với khả năng ác tính của khối u.
* **Radius Mean và Smoothness Mean**: Các đặc trưng **radius\_mean** và **smoothness\_mean** xuất hiện trong các phân chia tiếp theo. Mặc dù không phải là yếu tố phân chia đầu tiên, nhưng chúng vẫn góp phần quan trọng trong việc phân loại. **Radius\_mean** có thể liên quan đến kích thước của khối u, trong khi **smoothness\_mean** có thể đo lường độ đồng đều của bề mặt khối u.
* **Phân chia thành hai nhánh**: Như biểu đồ cho thấy, mô hình chia khối u thành hai nhóm: khối u lành tính (0) và khối u ác tính (1). Khi một khối u có các đặc điểm đặc biệt như độ lõm lớn hoặc kích thước lớn, khả năng ác tính sẽ cao hơn.
* **Quyết Định ở Các Lá:**

Các lá của cây cho thấy kết quả cuối cùng của việc phân loại, với khả năng khối u lành tính hoặc ác tính. Ví dụ, khi **concave points\_mean** nhỏ và **radius\_mean** nhỏ, mô hình phân loại khối u lành tính (0). Ngược lại, khi **concave points\_mean** cao, mô hình có xu hướng phân loại khối u là ác tính (1).

* **Đặc điểm của Cây Quyết Định:**
* **Cây đơn giản và dễ hiểu:** Cây quyết định giúp chúng ta dễ dàng theo dõi quá trình ra quyết định của mô hình. Chúng ta có thể dễ dàng nhận ra các yếu tố nào đóng vai trò quan trọng trong việc phân loại khối u lành tính và ác tính.
* **Không cần xử lý dữ liệu phức tạp:** Mô hình cây quyết định thường không yêu cầu phải chuẩn hóa hoặc làm sạch dữ liệu phức tạp, vì nó có thể hoạt động tốt với dữ liệu thô và phân loại tốt ngay cả khi dữ liệu có sự bất cân xứng.
* **Khả năng Giải Thích:**

Một trong những ưu điểm lớn của cây quyết định là khả năng giải thích được quá trình ra quyết định. Đây là một yếu tố quan trọng trong lĩnh vực y tế, nơi mà bác sĩ có thể dễ dàng hiểu lý do tại sao mô hình phân loại một khối u là lành tính hay ác tính dựa trên các yếu tố như độ lõm và kích thước khối u.

* **Những Hạn Chế Của Cây Quyết Định:**
* **Overfitting:** Mặc dù mô hình cây quyết định có thể rất mạnh mẽ, nhưng nó cũng dễ bị overfitting nếu cây quá phức tạp, nghĩa là nó quá phụ thuộc vào các đặc điểm cụ thể của tập huấn luyện và không thể tổng quát tốt trên dữ liệu mới.
* **Không có khả năng nắm bắt quan hệ phi tuyến tính phức tạp:** Các cây quyết định cơ bản có thể gặp khó khăn trong việc mô hình hóa các quan hệ phức tạp giữa các đặc trưng, mặc dù có thể cải thiện vấn đề này bằng cách sử dụng các kỹ thuật như cây quyết định ngẫu nhiên (Random Forest) hoặc cây quyết định sâu (XGBoost).
* **Kết Luận:**
* Cây quyết định giúp mô hình hóa mối quan hệ giữa các đặc trưng của khối u và khả năng ác tính một cách rõ ràng và dễ hiểu. Cây này cho thấy **concave points\_mean** là đặc trưng quan trọng nhất trong việc phân loại khối u lành tính và ác tính, với các đặc trưng khác như **radius\_mean** và **smoothness\_mean** cũng đóng góp vào quá trình phân loại.
* Mặc dù cây quyết định hoạt động tốt trong việc phân loại các khối u, việc điều chỉnh các tham số mô hình để tránh overfitting và cải thiện khả năng tổng quát của mô hình có thể là một bước quan trọng tiếp theo.

## Kết Luận

Chương 4 đã tiến hành phân tích sâu về mối quan hệ giữa các yếu tố liên quan đến bệnh ung thư vú thông qua các kiểm định giả thuyết và mô hình hồi quy. Cụ thể, các kết quả nghiên cứu cho thấy:

Giai đoạn phát hiện bệnh và kết quả điều trị: Kiểm định giả thuyết cho thấy giai đoạn phát hiện bệnh có mối liên hệ chặt chẽ với kết quả điều trị ung thư vú. Điều này khẳng định rằng việc phát hiện sớm bệnh có ảnh hưởng lớn đến khả năng điều trị và cải thiện tỷ lệ sống sót của bệnh nhân. Các biện pháp sàng lọc và phát hiện sớm cần được chú trọng để nâng cao hiệu quả điều trị.

Đặc tính khối u và tỷ lệ sống sót: Các đặc tính của khối u, như kích thước và độ ác tính, có ảnh hưởng mạnh mẽ đến tỷ lệ sống sót của bệnh nhân. Những bệnh nhân có khối u lớn hơn và có độ ác tính cao có xu hướng có tiên lượng xấu hơn. Điều này chỉ ra rằng việc theo dõi các đặc tính này trong quá trình điều trị là rất quan trọng.

Mô hình hồi quy: Các mô hình hồi quy được xây dựng để dự đoán khả năng ác tính của khối u cho thấy kích thước và độ ác tính có thể được sử dụng như những yếu tố dự đoán chính xác tình trạng bệnh. Mô hình hồi quy có thể hỗ trợ các bác sĩ trong việc đưa ra quyết định điều trị chính xác và cá nhân hóa cho từng bệnh nhân.

Nhìn chung, các kết quả từ chương này đã chỉ ra tầm quan trọng của việc phát hiện sớm và theo dõi đặc tính khối u trong việc đưa ra quyết định điều trị cho bệnh nhân ung thư vú. Các kết quả này có thể góp phần vào việc cải thiện tỷ lệ sống sót và nâng cao chất lượng cuộc sống cho bệnh nhân.