**HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG**

**----- 🙠🕮🙢 -----**



**BÁO CÁO TIỂU LUẬN**

**LƯU TRỮ VÀ PHÂN TÍCH DỮ LIỆU**

**ĐỀ TÀI**

|  |
| --- |
| **PHÂN TÍCH VÀ DỰ ĐOÁN KHỐI U ÁC TÍNH CỦA BỆNH UNG THƯ VÚ** |

|  |  |
| --- | --- |
| Giảng viên hướng dẫn | Nguyễn Minh Tuấn |
| Nhóm thực hiện | 02 |
| Nhóm môn học | 01 |
| Thành viên | Hoàng Đông Tuấn – B20DCVT334 |
|  |  |

**Hà Nội, 5/2024**

Mục lục

[LỜI NÓI ĐẦU 3](#_Toc168726745)

[CHƯƠNG 1. TỔNG QUAN VỀ PHÂN TÍCH DỮ LIỆU VÀ MÔ HÌNH HỌC MÁY 4](#_Toc168726746)

[**1.1. Tổng quan về phân tích dữ liệu** 4](#_Toc168726747)

[**1.1.1. Khái niệm dữ liệu** 4](#_Toc168726748)

[**1.1.2. Khái niệm phân tích dữ liệu** 5](#_Toc168726749)

[**1.2. Mô hình học máy** 5](#_Toc168726750)

[**1.2.1. Học máy có giám sát** 5](#_Toc168726751)

[**1.2.2. Học máy không giám sát** 6](#_Toc168726752)

[**1.3. Tổng kết chương** 7](#_Toc168726753)

[CHƯƠNG 2. GIỚI THIỆU BÀI TOÁN VÀ CÁC KIẾN THỨC LIÊN QUAN 8](#_Toc168726754)

[**2.1. Giới thiệu bài toán** 8](#_Toc168726755)

[**2.2. Các kiến thức liên quan** 9](#_Toc168726756)

[**2.2.1 Thuật toán Random Forest** 9](#_Toc168726757)

[**2.2.2 Thuật toán Logistic Regression** 11](#_Toc168726758)

[**2.2.3. XGBoost** 14](#_Toc168726759)

[**2.2.4. Tham số đánh giá** 16](#_Toc168726760)

[**2.3. Tổng kết chương 2** 17](#_Toc168726761)

[CHƯƠNG 3. ỨNG DỤNG TRONG PHÂN TÍCH VÀ DỰ ĐOÁN NGUY CƠ MẮC UNG THƯ VÚ 18](#_Toc168726762)

[**3.1. Giới thiệu tệp dữ liệu** 18](#_Toc168726763)

[**3.2. Khai phá dữ liệu (EDA)** 19](#_Toc168726764)

[**3.2.1. Thống kê tóm tắt** 19](#_Toc168726765)

[**3.3. Tiền xử lí dữ liệu** 20](#_Toc168726766)

[**3.3.1 Làm sạch dữ liệu** 20](#_Toc168726767)

[**3.3.2. Trực quan hóa dữ liệu** 26](#_Toc168726768)

[**3.3.3 Kiểm tra giả thuyết** 27](#_Toc168726769)

[**3.3.4 Lấy mẫu và mã hóa nhãn** 28](#_Toc168726770)

[**3.3.5 Kiểm tra mối tương quan** 29](#_Toc168726771)

[**3.4. Huấn luyện mô hình** 31](#_Toc168726772)

[**3.5. Kết luận** 37](#_Toc168726773)

# LỜI NÓI ĐẦU

Theo Globocan 2020, có 2,261,419 ca ung thu vú được phát hiện mỗi năm ở cả 2 giới, chiếm đến 11.7 % tổng số các loại bệnh ung thư, điều đáng lo ngại là con số này đang có xu hướng gia tăng và độ tuổi mắc bệnh ngày càng trẻ hóa. Nhận biết các dấu hiệu cảnh báo ung thư vú như thế nào? Làm sao để có thể kiểm soát phát hiện sớm và điều trị hiệu quả căn bệnh này?

Theo Tổ chức Y tế Thế giới (WHO-2023), ung thư vú có đặc điểm là sự phát triển không kiểm soát của các tế bào bất thường ở vú, dẫn đến hình thành các khối u. Nếu không được điều trị, những khối u này có khả năng di căn, lan sang các bộ phận khác của cơ thể và có nguy cơ gây tử vong. Năm 2020, có gần 2,3 triệu phụ nữ được chẩn đoán mắc bệnh ung thư vú và 685.000 ca tử vong trên toàn cầu. Tính đến cuối năm 2020, có 7,8 triệu phụ nữ còn sống được chẩn đoán mắc bệnh ung thư vú trong 5 năm qua, khiến đây trở thành căn bệnh ung thư phổ biến nhất trên thế giới. Ung thư vú xảy ra ở mọi quốc gia trên thế giới ở phụ nữ ở mọi lứa tuổi sau tuổi dậy thì nhưng với tỷ lệ ngày càng tăng ở tuổi sau.

Vì muốn dự đoán được khối u lành tính hay ác tính ở giai đoạn sớm để giúp tìm ra pháp đồ điều trị và giảm nguy cơ tử vong nên chúng em có dự án sử dụng thuật toán để phân tích và dự đoán khối u lành tích hay ác tính qua dữ liệu có sẵn được thu thập từ các nguồn khác nhau.

# CHƯƠNG 1. TỔNG QUAN VỀ PHÂN TÍCH DỮ LIỆU VÀ MÔ HÌNH HỌC MÁY

## **1.1. Tổng quan về phân tích dữ liệu**

### **1.1.1. Khái niệm dữ liệu**

Trong thời đại công nghệ và khoa học có những sự phát triển vượt bậc và mạnh mẽ, dữ liệu đã dần trở thành yếu tố then chốt, quan trọng trong nhiều lĩnh vực. Việc sử dụng ngày càng nhiều phương tiện và mạng vật lý thông qua mạng cảm biến cho các mục đích kinh doanh và cá nhân tạo ra một lượng lớn dữ liệu. Điều này nhằm thay đổi quy trình kinh doanh và mở ra cơ hội mới trên toàn thế giới.

Trong phân tích dữ liệu, "dữ liệu" là tập hợp các thông tin, sự kiện hoặc dấu hiệu được thu thập hoặc sản sinh để phân tích và rút ra các kết luận. Dữ liệu có thể là các con số, văn bản, hình ảnh, âm thanh, hoặc bất kỳ loại thông tin nào khác có thể được biểu diễn và lưu trữ trong các hệ thống máy tính

Dữ liệu có thể được phân loại thành hai loại chính:

* **Dữ liệu có cấu trúc:** Đây là loại dữ liệu được tổ chức trong các bảng, cơ sở dữ liệu hoặc các cấu trúc dữ liệu khác, trong đó mỗi mẫu dữ liệu được lưu trữ theo một định dạng cố định và có các trường dữ liệu riêng biệt. Ví dụ bao gồm bảng Excel, cơ sở dữ liệu SQL, hoặc các tệp tin CSV.
* **Dữ liệu không có cấu trúc:** Loại dữ liệu này không được tổ chức theo cấu trúc cố định, và thường chứa các dạng thông tin không theo quy luật như văn bản tự do, email, tweet, hình ảnh, video, và âm thanh. Dữ liệu không có cấu trúc thường đòi hỏi các kỹ thuật xử lý đặc biệt để trích xuất thông tin hữu ích từ chúng, như phân loại, tổng hợp, hoặc trích xuất đặc trưng.

Bên cạnh đó, dữ liệu cũng có thể được phân loại theo đặc điểm khác nhau như:

* **Dữ liệu thời gian:** Các dữ liệu liên quan đến thời gian và thường được sắp xếp theo các dấu hiệu thời gian như ngày, giờ, tháng, năm, v.v.
* **Dữ liệu địa lý:** Dữ liệu liên quan đến vị trí địa lý và có thể được biểu diễn bằng tọa độ địa lý (latitude, longitude) hoặc các tên địa lý.
* **Dữ liệu đa phương tiện:** Bao gồm hình ảnh, video, âm thanh, văn bản, v.v.

Việc hiểu rõ về loại dữ liệu mà bạn đang làm việc là quan trọng để áp dụng các phương pháp phân tích phù hợp và rút ra những kết luận có ý nghĩa từ dữ liệu đó.

### **Top of Form**

### **1.1.2. Khái niệm phân tích dữ liệu**

Phân tích dữ liệu là quá trình nghiên cứu và tìm hiểu các dữ liệu để nhận biết các mẫu, quan hệ và thông tin hữu ích từ chúng. Mục tiêu của phân tích dữ liệu là chuyển đổi dữ liệu thành thông tin hữu ích để hỗ trợ quyết định và hiểu biết sâu hơn về một vấn đề cụ thể.

Dưới đây là một số khái niệm quan trọng liên quan đến phân tích dữ liệu:

* **Thu thập dữ liệu:** Quá trình thu thập thông tin từ các nguồn khác nhau như cơ sở dữ liệu, tệp tin, hoặc các nguồn dữ liệu trực tuyến.
* **Tiền xử lý dữ liệu:** Bao gồm các bước làm sạch dữ liệu, xử lý dữ liệu bị thiếu, loại bỏ nhiễu, và chuẩn hóa dữ liệu để chuẩn bị cho quá trình phân tích.
* **Khám phá dữ liệu:** Quá trình này liên quan đến việc khám phá và hiểu cấu trúc, đặc điểm, và các mẫu tồn tại trong dữ liệu thông qua việc sử dụng các công cụ và kỹ thuật như phân tích thống kê, trực quan hóa dữ liệu, và khai phá dữ liệu.
* **Phân tích dữ liệu:** Bước này bao gồm việc áp dụng các phương pháp phân tích để phát hiện ra mối quan hệ, xu hướng, và thông tin quan trọng từ dữ liệu. Các phương pháp phổ biến bao gồm phân tích đa biến, phân tích hồi quy, phân tích chuỗi thời gian, và phân tích phân cụm.
* **Tạo mô hình dữ liệu:** Quá trình xây dựng các mô hình dữ liệu để dự đoán hoặc phân loại các sự kiện dựa trên dữ liệu đã được phân tích.
* **Đánh giá và diễn giải kết quả:** Cuối cùng, kết quả của quá trình phân tích được đánh giá và diễn giải để đưa ra những hiểu biết và quyết định có ích.

Phân tích dữ liệu là một phần quan trọng của khoa học dữ liệu và có ứng dụng rộng rãi trong nhiều lĩnh vực như kinh doanh, y tế, khoa học xã hội, và nghiên cứu khoa học.

## **1.2. Mô hình học máy**

### **1.2.1. Học máy có giám sát**

Học máy có giám sát là một phương pháp trong lĩnh vực học máy, trong đó một mô hình máy tính được đào tạo bằng cách sử dụng dữ liệu đã được gán nhãn trước để học cách dự đoán hoặc phân loại dữ liệu mới. Trong quá trình huấn luyện, mô hình máy tính sẽ được cung cấp các ví dụ dữ liệu, mỗi ví dụ đi kèm với một nhãn hay một kết quả mong muốn. Mục tiêu là tìm ra mối quan hệ giữa các đặc trưng của dữ liệu và nhãn tương ứng, từ đó mô hình có thể học cách dự đoán nhãn cho dữ liệu mới chưa được nhãn.

Dưới đây là một số ví dụ về các ứng dụng của học máy có giám sát:

* Phân loại ảnh: Một mô hình có thể được huấn luyện để phân loại ảnh vào các nhóm khác nhau như chó, mèo, hoặc xe hơi dựa trên ảnh mà nó đã nhìn thấy trước đó và nhãn được gán cho từng ảnh.
* Dự đoán giá nhà: Dựa trên các đặc trưng của căn nhà như diện tích, vị trí, và số phòng ngủ, một mô hình có thể được đào tạo để dự đoán giá nhà tương ứng.
* Phát hiện email rác: Một mô hình có thể được huấn luyện để phân loại email là rác hoặc không rác dựa trên nội dung và các đặc trưng khác của email.
* Dự đoán tổn thất trong bảo hiểm: Một công ty bảo hiểm có thể sử dụng học máy để dự đoán khả năng tổn thất cho mỗi chính sách bảo hiểm dựa trên các thông tin như tuổi, giới tính, và lịch sử lái xe của khách hàng.

Trong học máy có giám sát, một phần quan trọng của quá trình là đánh giá hiệu suất của mô hình trên dữ liệu chưa nhìn thấy trước đó, để đảm bảo rằng nó có thể tổng quát hóa tốt và có thể dự đoán chính xác trên dữ liệu mới. Điều này thường được thực hiện bằng cách chia tập dữ liệu thành tập huấn luyện và tập kiểm tra, và sau đó đánh giá mô hình trên tập kiểm tra.**Top of Form**

### **1.2.2. Học máy không giám sát**

Học máy không giám sát là một phương pháp trong lĩnh vực học máy mà mô hình máy tính được huấn luyện từ dữ liệu không có nhãn hoặc chỉ có một phần dữ liệu được gán nhãn. Trong học máy không giám sát, mục tiêu chính là tìm ra cấu trúc ẩn trong dữ liệu hoặc phát hiện các mẫu không được biết trước.

Dưới đây là một số ví dụ về các ứng dụng của học máy không giám sát:

* Phân cụm dữ liệu: Một mô hình có thể được sử dụng để phân nhóm các điểm dữ liệu vào các cụm dựa trên sự tương đồng giữa chúng, mà không cần biết trước về các nhãn.
* Phát hiện bất thường: Học máy không giám sát có thể được sử dụng để phát hiện các điểm dữ liệu bất thường hoặc không bình thường trong một tập dữ liệu, như là các giao dịch tài chính gian lận hoặc các sự kiện hỏng hóc trong một hệ thống.
* Reduction chiều dữ liệu: Các phương pháp giảm chiều dữ liệu như PCA (Principal Component Analysis) hoặc t-SNE (t-distributed Stochastic Neighbor Embedding) có thể được sử dụng để giảm số chiều của dữ liệu mà vẫn giữ lại các đặc điểm quan trọng nhất.
* Học biểu diễn dữ liệu: Một mô hình có thể được huấn luyện để học biểu diễn dữ liệu dưới dạng một không gian mới, mà dữ liệu có thể được biểu diễn một cách tốt hơn cho các nhiệm vụ sau này như phân loại hoặc dự đoán.

Trong học máy không giám sát, việc đánh giá hiệu suất của mô hình có thể trở nên khó khăn hơn so với học máy có giám sát do thiếu các nhãn đúng cho so sánh. Do đó, các phương pháp đánh giá chất lượng của mô hình thường dựa trên các tiêu chí như sự đồng nhất trong cụm, khả năng phát hiện bất thường, hoặc khả năng khôi phục lại cấu trúc ban đầu của dữ liệu.

## **Top of Form**

## **1.3. Tổng kết chương**

Trong chương này, chúng ta đã bắt đầu với một cái nhìn tổng quan về hai lĩnh vực quan trọng trong khoa học dữ liệu: Phân tích Dữ liệu và Mô hình Học máy.

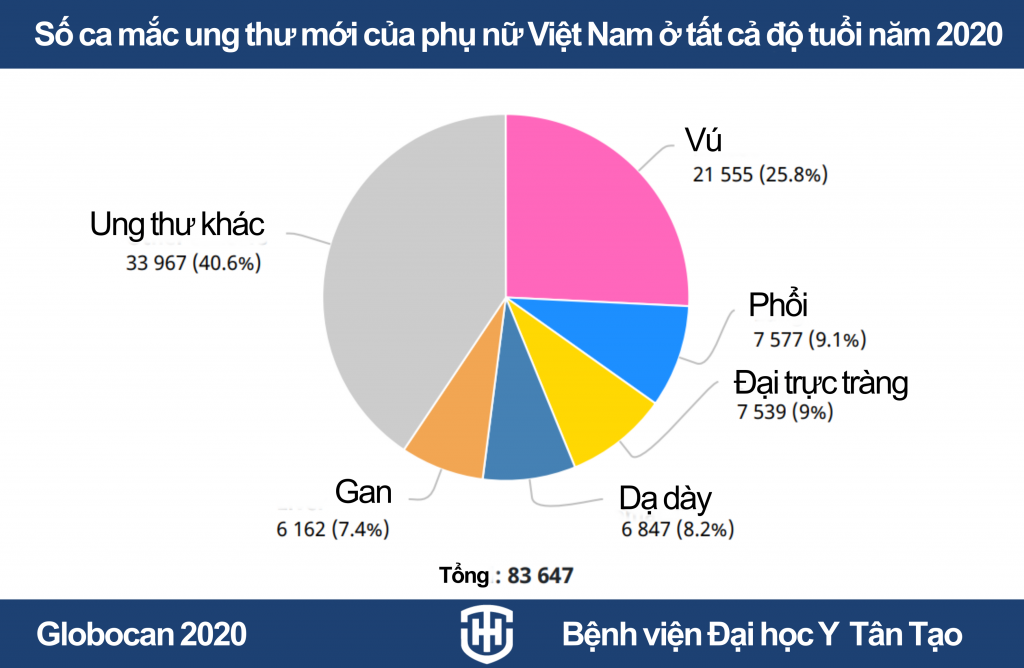
Phân tích Dữ liệu: Chương bắt đầu bằng việc giới thiệu về khái niệm dữ liệu và vai trò quan trọng của nó trong thế giới hiện đại, nơi mà dữ liệu ngày càng trở nên phong phú và quyết định. Chúng ta đã thảo luận về hai loại chính của dữ liệu: dữ liệu có cấu trúc và dữ liệu không có cấu trúc, cũng như các loại dữ liệu khác như dữ liệu thời gian, dữ liệu địa lý, và dữ liệu đa phương tiện. Tiếp theo, chúng ta đã tiếp tục với một cái nhìn tổng quan về Mô hình Học máy, với một sự tập trung đặc biệt vào hai phương pháp chính: Học máy có giám sát và Học máy không giám sát.

Tóm lại, chương này đã cung cấp cho chúng ta một cái nhìn tổng quan về hai lĩnh vực quan trọng trong khoa học dữ liệu và đã giúp chúng ta hiểu rõ hơn về vai trò và quy trình của Phân tích Dữ liệu và Mô hình Học máy trong việc nắm bắt và sử dụng thông tin từ dữ liệu.

# CHƯƠNG 2. GIỚI THIỆU BÀI TOÁN VÀ CÁC KIẾN THỨC LIÊN QUAN

## **2.1. Giới thiệu bài toán**

* **Bối cảnh**

Ung thư vú là một trong những loại ung thư phổ biến nhất ở phụ nữ trên toàn thế giới. Theo Tổ chức Y tế Thế giới (WHO), ung thư vú là nguyên nhân hàng đầu gây tử vong do ung thư ở phụ nữ. Tại Việt Nam, ung thư vú cũng là một vấn đề sức khỏe đáng quan tâm, với số ca mắc mới và tử vong ngày càng gia tăng. 

***Hình 2.1. Thống kê thông tin về bệnh ung thư vú tại Việt Nam***

Sự gia tăng về mức độ phổ biến và tác động nghiêm trọng của ung thư vú đặt ra một thách thức lớn cho các nước và tổ chức y tế trên toàn thế giới trong việc phòng ngừa, chẩn đoán và điều trị bệnh này. Việc phát hiện sớm ung thư vú có thể cải thiện đáng kể tỷ lệ sống sót của bệnh nhân.

* **Mục đích**

Nắm bắt được bối cảnh và tình trạng của ung thư vú hiện nay, bài toán phân tích tập dữ liệu để dự đoán nguy cơ mắc bệnh ung thư vú được coi như một giải pháp góp phần mang lại nhiều lợi ích trong y tế. Bài toán này sử dụng các phương pháp phân tích dữ liệu và mô hình học máy để dự đoán và phân loại nguy cơ mắc bệnh của mỗi người thành hai nhóm: có nguy cơ bị ung thư vú và không có nguy cơ mắc ung thư vú.

Để giải quyết bài toán này, ta sẽ sử dụng một tập dữ liệu có sẵn về thông tin sức khỏe của mỗi người, bao gồm các đặc trưng như chu vi, bán kính, diện tích khối u, v.v... Tập dữ liệu này sẽ được chuẩn bị và tiền xử lý để có thể sử dụng cho việc huấn luyện mô hình học máy. Sau khi chuẩn bị dữ liệu, ta có thể sử dụng các kỹ thuật phân tích dữ liệu để khám phá mối quan hệ giữa các đặc trưng và nguy cơ mắc bệnh ung thư vú. Sau đó, ta sẽ sử dụng mô hình học máy để huấn luyện và đánh giá hiệu quả của mô hình trên tập dữ liệu huấn luyện và tập dữ liệu kiểm tra.

Kết quả của bài toán này là một mô hình học máy có thể dự đoán nguy cơ mắc bệnh ung thư vú của mỗi người dựa trên các đặc trưng sức khỏe của họ. Kết quả này có thể được sử dụng để giúp các bác sĩ chẩn đoán bệnh cho các bệnh nhân và đưa ra các biện pháp phòng ngừa, điều trị phù hợp và kịp thời, từ đó nâng cao hiệu quả chẩn đoán và điều trị ung thư vú, góp phần giảm tỷ lệ tử vong do căn bệnh này.

## **2.2. Các kiến thức liên quan**

### **2.2.1 Thuật toán Random Forest**

* **Định nghĩa**: Random forest là một phương pháp thống kê mô hình hóa bằng máy (machine learning statis-tic) dùng để phục vụ các mục đích phân loại, tính hồi quy và các nhiệm vụ khác bằng cách xây dựng nhiều cây quyết định (Decision tree). Một cây quyết định là một cách đơn giản để biểu diễn một giao thức (Protocol).

Nói cách khác, cây quyết định biểu diễn một kế hoạch, trả lời câu hỏi phải làm gì trong một hoàn cảnh nhất định. Mỗi Node của cây sẽ là các thuộc tính, và các nhánh là giá trị lựa chọn của thuộc tính đó. Bằng cách đi theo các giá trị thuộc tính trên cây, cây quyết định sẽ cho ta biết giá trị dự đoán. Nhóm thuật toán cây quyết định có một điểm mạnh đó là có thể sử dụng cho cả bài toán. Phân loại (Classification) và Hồi quy (Regression). Random Forest có khả năng tìm ra thuộc tính nào quan trọng hơn so với những thuộc tính khác. Trên thực tế, nó còn có thể chỉ ra rằng một số thuộc tính là không có tác dụng trong cây quyết định.

**Mô tả phương pháp random forest**

* **Lựa chọn cây quyết định (decision tree learning)**

Cây quyết định là một phương pháp phổ biến cho các nhiệm vụ mô hình hóa bằng máy (machine learning). Các cây quyết định được lựa chọn với các tiêu chí phù hợp để đáp ứng các yêu cầu nhiệm vụ khai thác dữ liệu. Các cây quyết định dược thiết kế với xu hướng nhận biết được cả những yếu tó bất thường : phù hợp với các mẫu có độ lệch nhỏ nhưng phương sai lớn.

A diagram of a tree

Description automatically generated

Hình 2.2 Mô tả lựa chọn cây quyết định

Nó hoạt động theo bốn bước:

1. Chọn các mẫu ngẫu nhiên từ tập dữ liệu đã cho.
2. Thiết lập cây quyết định cho từng mẫu và nhận kết quả dự đoán từ mỗi quyết định cây.
3. Hãy bỏ phiếu cho mỗi kết quả dự đoán.
4. Chọn kết quả được dự đoán nhiều nhất là dự đoán cuối cùng.

A diagram of a training set

Description automatically generated

*Hình 2.3: Thuật toán Random Forest*

Ưu nhược điểm của thuật toán

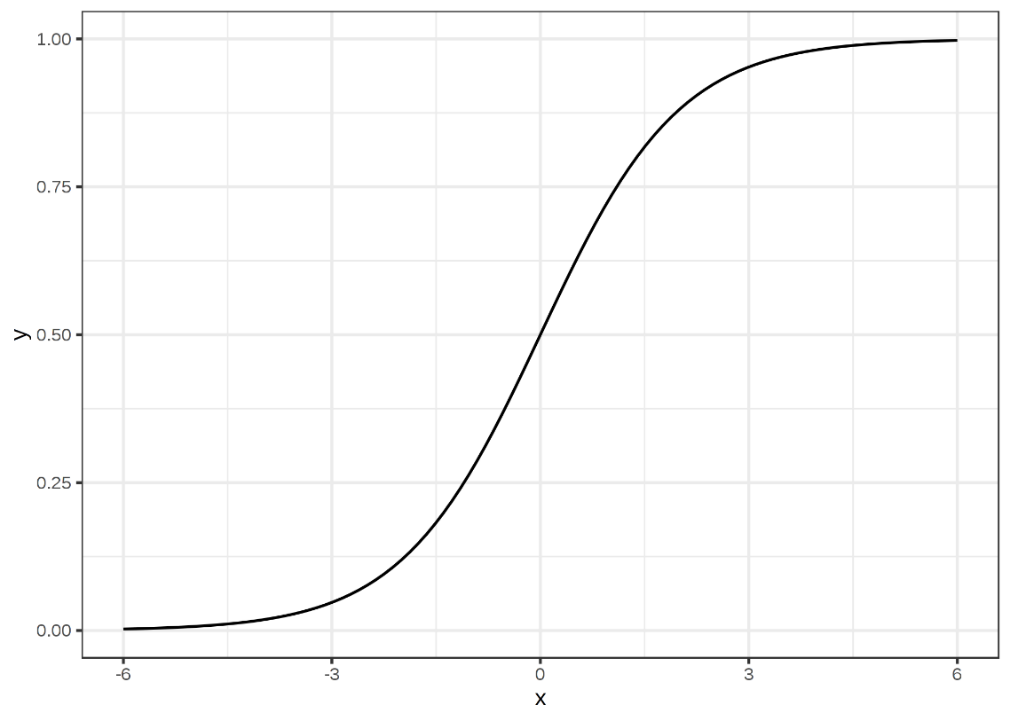
- Ưu điểm: Random forests được coi là một phương pháp chính xác và mạnh mẽ vì số cây quyết định tham gia vào quá trình này.

- Nhược điểm: Random forests chậm tạo dự đoán bởi vì nó có nhiều cây quyết định.

### **2.2.2 Thuật toán Logistic Regression**

* **Định nghĩa**: Logistic Regression là 1 thuật toán phân loại (Classification) được dùng để gán các đối tượng cho 1 tập hợp giá trị rời rạc (như 0, 1, 2, ...). Đây là phương thức tốt nhất cho các vấn đề phân loại nhị phân (các bài toán với các giá trị 2 lớp). [3]
* **Hàm Sigmoid**:
* Logistic Regression sử dụng hàm sigmoid (hoặc hàm logistic) để dự đoán chuyển đổi đầu ra từ một biến độc lập tuyến tính thành một giá trị xác suất nằm trong khoảng từ 0 đến 1 biểu diễn xác suất của từng từng mẫu. Hàm sigmoid có công thức: [3]

với x là tổ hợp tuyến tính của các biến độc lập



***Hình 2.4. Hàm Sigmoid***

* Đặc điểm:
* Hàm liên tục và luôn đưa ra giá trị trong khoảng (0, 1)
* Có đạo hàm tại mọi điểm nên có thể dùng Gradient Descent
* **Tính toán các công thức liên quan**
* **Hàm dự đoán (Prediction Function)**
* Với dòng thức i trong dữ liệu, gọi , , …, là các vector đặc trưng của data sample thứ i
* Xác suất mà thuật toán dự đoán cho label có phân loại là 1 là:
* Xác suất mà thuật toán dự đoán cho label có phân loại là 0 là:
* Hàm Sigmoid
* Tương tự như hàm dự đoán trong Linear Regression là

Trong đó: là hệ số chặn, hay giá trị dự đoán của **y** khi tất cả các biến độc lập bằng 0, i= 1,2,…k

, là Hệ số hồi quy (coefficients) cho các biến độc lập

* Áp dụng trên trong Logistic Regression ta có hàm dự đoán như sau:
* **Hàm mất mát (Loss Function)**
* Nếu nhãn của vector đặc trưng đó bằng 1, thì ta mong muốn càng gần 1 càng tốt hay thuật toán dự đoán xác suất nhãn của data sample thứ i càng gần 1 càng tốt. [3]
* Nếu nhãn của vector đặc trưng đó bằng 0, thì ta mong muốn càng gần 0 càng tốt hay thuật toán dự đoán xác suất nhãn của data sample thứ i càng gần 0 càng tốt.
* Với mỗi , ta có hàm mất mát



* Ta có hàm mất mát trên tất cả bộ dữ liệu như sau:

A close up of symbols

Description automatically generated

* Nhận xét với trường hợp :

A blue line graph with numbers

Description automatically generated

***Hình 2.5. Hàm mất mát với label =0***

* Hàm L tăng trong khoảng [0,1]
* Khi thuật toán dự đoán gần 0, tức là giá trị dự đoán gần với giá trị thật thì L nhỏ, xấp xỉ bằng 0
* Khi thuật toán dự đoán gần 1, tức là giá trị dự đoán ngược lại giá trị thật thì L lớn.
* Hàm L nhỏ khi giá trị model gần với giá trị thật và rất lớn khi model dự đoán sai, hay nói cách khác L càng nhỏ thì model dự đoán càng gần với giá trị thật. => Bài toán toán quy về tìm giá trị nhỏ nhất của L

### **2.2.3. XGBoost**

XGBoost (extreme Gradient Boosting ) là phiên bản cải tiến của Gradient Boosting. Ưu điểm vượt trội của nó được chứng minh ở các khía cạnh :

Gradient Boosting là một dạng tổng quát hóa của AdaBoost. Cụ thể như sau, vẫn vấn đề tối ưu ban đầu

min*cn*​,*wn*​​*L*(*y*,*Wn*−1​+*cn*​*wn*​))

Trước tiên mình xin nhắc lại một chút lí thuyết mà các bạn đã khá quen trong neural network: [**Gradient Descent**](https://machinelearningcoban.com/2017/01/12/gradientdescent/)

*θn*​=*θn*−1​−*η*∂*θ*∂​*L*(*θn*−1​)

Phía trên là công thức cập nhật tham số mô hình theo hướng giảm của đạo hàm (Gradient Descent). Công thức này được sử dụng không gian tham số, tuy nhiên, để liên hệ với bài toán chúng ta đang xét, mình chuyển công thức sang góc nhìn của không gian hàm số.

Khá đơn giản thôi, nếu chúng ta coi chuỗi các model boosting là một hàm số 𝑊*W*, thì mỗi hàm learner có thể coi là một tham số 𝑤*w*. Đến đây, để cực tiểu hóa hàm loss 𝐿(𝑦,𝑊)*L*(*y*,*W*), chúng ta áp dụng Gradient Descent

*Wn*​=*Wn*−1​−*η*∂*w*∂​*L*(*Wn*−1​)

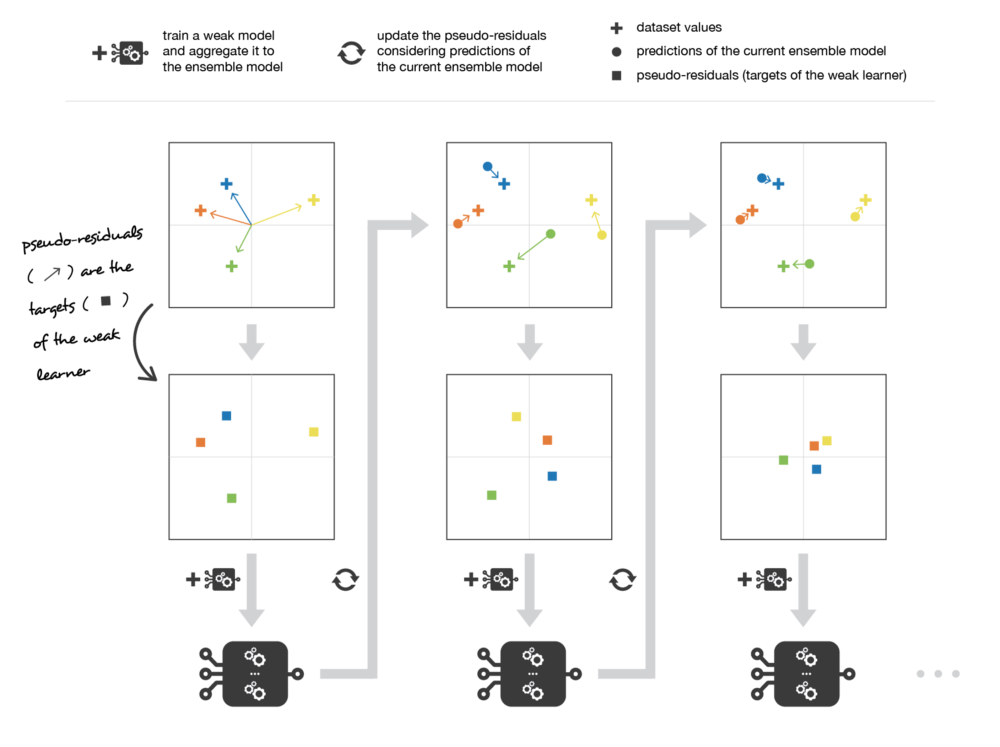
Đến đây, ta có thể thấy mối quan hệ liên quan sau

*cn*​*wn*​≈−*η*∂*w*∂​*L*(*Wn*−1​)

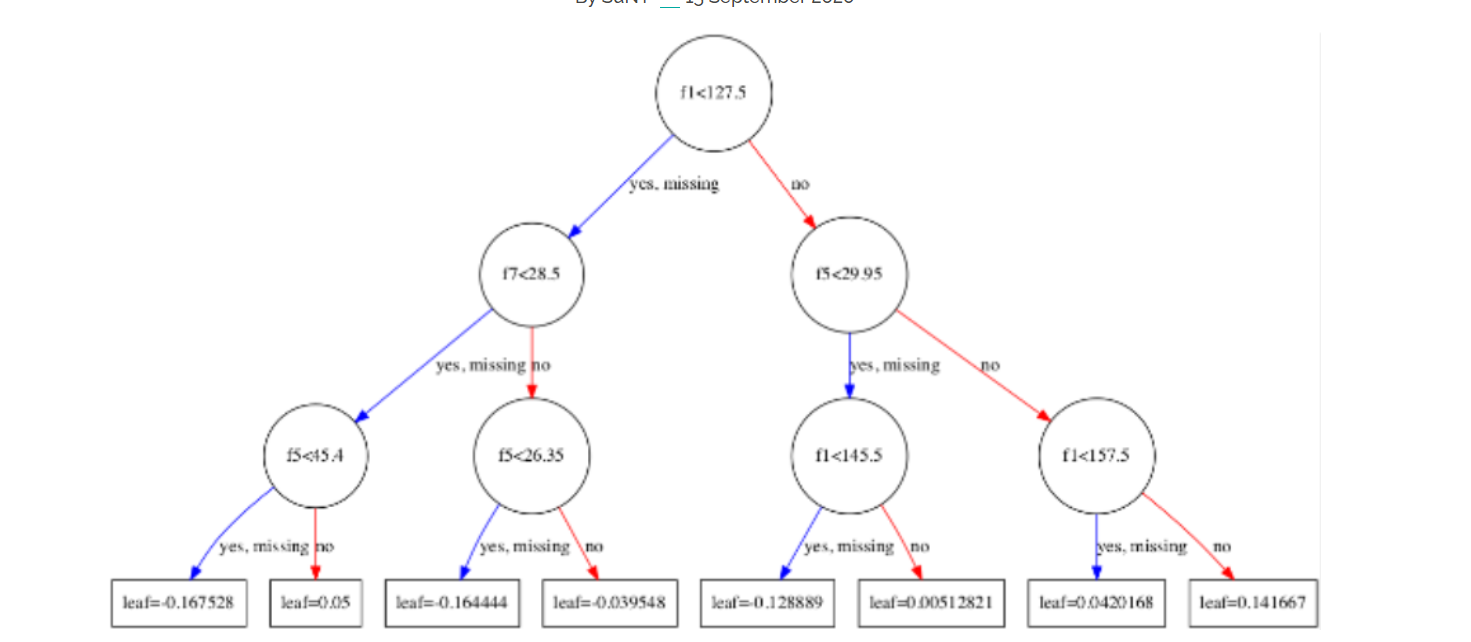
với 𝑤𝑛*wn*​ là model được thêm vào tiếp theo. Khi đó, model mới cần học để fit để vào giá trị −𝜂∂∂𝑤𝐿(𝑊𝑛−1)−*η*∂*w*∂​*L*(*Wn*−1​). (Giá trị −𝜂∂∂𝑤𝐿(𝑊𝑛−1)−*η*∂*w*∂​*L*(*Wn*−1​) còn có 1 tên gọi khác là **pseudo-residuals**)

Tóm lại, chúng ta có thể tóm tắt quá trình triển khai thuật toán như sau:

* Khởi tạo giá trị pseudo-residuals là bằng nhau cho từng điểm dữ liệu
* Tại vòng lặp thứ i
  + Train model mới được thêm vào để fit vào giá trị của pseudo-residuals đã có
  + Tính toán giá trị confidence score 𝑐𝑖*ci*​ của model vừa train
  + Cập nhật model chính 𝑊=𝑊+𝑐𝑖∗𝑤𝑖*W*=*W*+*ci*​∗*wi*​
  + Cuối cùng, tính toán giá trị pseudo-residuals −𝜂∂∂𝑤𝐿(𝑊𝑛−1)−*η*∂*w*∂​*L*(*Wn*−1​) để làm label cho model tiếp theo
* Sau đó lặp lại với vòng lặp i + 1.

Nếu bạn để ý thì phương pháp cập nhật lại trọng số của điểm dữ liệu của AdaBoost cũng là 1 trong các case của Gradient Boosting. Do đó, Gradient Boosting bao quát được nhiều trường hợp hơn.

Hình 2.6 Hình mô tả thuật toán XGBoost



Hình 2.7 Mô tả cây quyết định

### **2.2.4. Tham số đánh giá**

* **Ma trận nhầm lẫn :** So sánh giá trị thực tế với giá trị dự đoán

**A screenshot of a cell phone

Description automatically generated**

***Hình 2.8. Ma trận nhầm lẫn***

* Âm tính thực (TN): Nếu 1 mẫu được dự đoán là sai và thực tế nó sai thì đó gọi là âm tính thực
* Dương tính thực (TP): Nếu 1 mẫu được dự đoán là đúng và thực tế nó đúng thì đó gọi là dương tính thực

Vậy âm tính thực (TN) và dương tính thực (TP) là những dự đoán chính xác của mô hình.

* Âm tính giả (FN): Nếu 1 mẫu được dự đoán là sai nhưng thực tế nó lại đúng thì đó gọi là âm tính giả
* Dương tính giả (FP): Nếu 1 mẫu được dự đoán là đúng nhưng thực tế nó là sai thì được gọi là dương tính giả.

Vậy âm tính giả và dương tính giả là những dự đoán sai lệch của mô hình

* **Độ chính xác toàn thể (Accuracy):** Phần trăm khả năng dự đoán chính xác của mô hình trên tổng mẫu quan sát

Gọi tổng số mẫu quan sát là N = TP + TN + FP + FN

Độ chính xác được xác định bởi công thức:

Accuracy = =

* Sai số toàn thể = 1 – Accuracy = 1 -
* **Độ nhạy mô hình (Recall):** Đặc trưng cho trong tổng số mẫu thực tế đúng thì có bao nhiêu phần trăm là của mô hình dự đoán đúng

Recall =

* **Độ chính xác phân loại (Precision):** Đặc trưng cho số mẫu được dự đoán là đúng với thực tế là đúng trong tổng số mẫu được dự đoán là đúng của mô hình

Precision =

* **F1-score:** Trung bình giữa Precision và Recall

F1-score =2 \*

## **2.3. Tổng kết chương 2**

Chương 2 đã cung cấp một cái nhìn tổng quan về bài toán được nghiên cứu cùng với các kiến thức liên quan. Trước tiên, chương này đã giới thiệu về bối cảnh của bài toán và muc đích của bài toán cũng như ý nghĩa của việc giải quyết nó trong ngữ cảnh hiện tại. Chương này đã đưa ra hướng nghiên cứu và động lực của đề tài.

Tiếp theo, chương đã đề cập đến các kiến thức cơ bản liên quan đến đề tài. Điều này bao gồm các khái niệm, định nghĩa và phương pháp liên quan. Các khái niệm này không chỉ giúp người đọc hiểu rõ hơn về vấn đề được nghiên cứu mà còn giúp xác định cách tiếp cận và phương pháp phân tích đề tài.

# CHƯƠNG 3. ỨNG DỤNG TRONG PHÂN TÍCH VÀ DỰ ĐOÁN NGUY CƠ MẮC UNG THƯ VÚ

## **3.1. Giới thiệu tệp dữ liệu**

Tập dữ liệu được tạo ra vào tháng 11 năm 1995 bởi Tiến sĩ William H. Wolberg, W. Nick Street và Olvi L. Mangasarian từ Khoa Khoa học Máy tính tại Đại học Wisconsin.Sử dụng các đặc điểm của khối u để dự đoán tình trạng bệnh nhân và hướng điều trị kịp thời.

Các đặc điểm được chia ra thành 3 nhóm chính với đuôi \_mean là giá trị trung bình và 2 nhóm có đuôi :

\_se là viết tắt của Sai số chuẩn và \_worst là giá trị trung bình của ba giá trị lớn nhất.

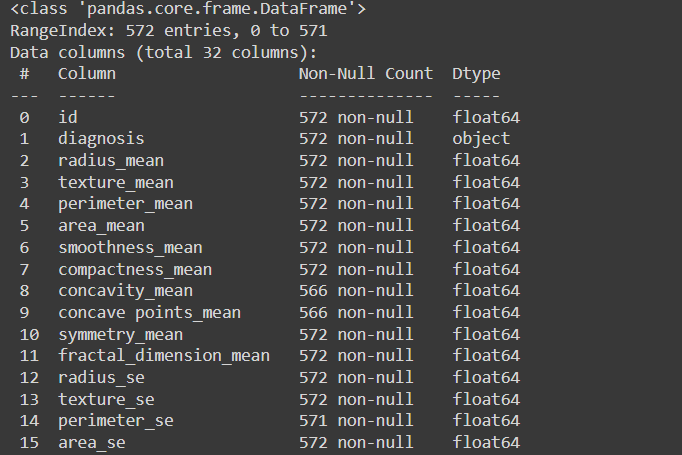
Mô tả các thông tin về tình trạng của mỗi người khảo sát được lưu ở tệp dữ liệu:

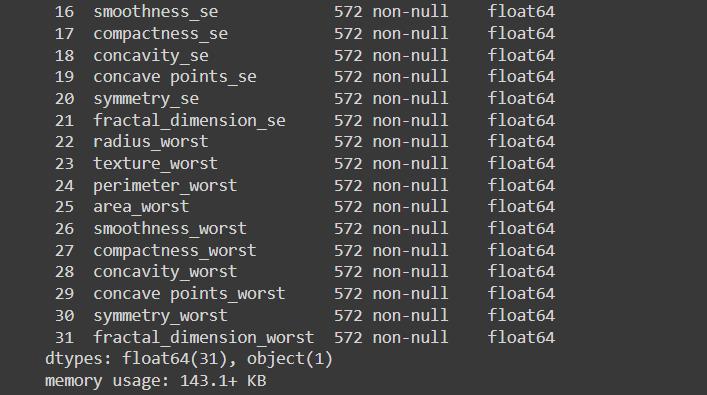
|  |  |  |
| --- | --- | --- |
| **Biến** | **Mô tả** | **Định nghĩa** |
| Id | Chỉ số |  |
| Diagnosis | Chuẩn đoán | M( malignant)= ác tính  B( bening)= lành tính |
| Radius | Bán kính | Trung bình khoảng cách từ tâm đến các điểm trên chu vi |
| Texture | Kết cấu | Độ lệch chuẩn của các giá trị thang màu xám |
| Perimeter | Chu vi |  |
| Area | Diện tích |  |
| Smoothness | Độ mịn | Sự thay đổi cục bộ về chiều dài bán kính |
| Compactness | Độ nén | Chu vi^2/ diện tích – 1.0 |
| Concavity | Độ lõm | Mức độ nghiêm trọng của các phần lõm của đường viền |
| Concave points | Điểm lõm | Số phần lõm của đường viền |
| Symmetry | Tính  đối xứng |  |
| Dimension fractal | Kích thước fractal |  |

## **3.2. Khai phá dữ liệu (EDA)**

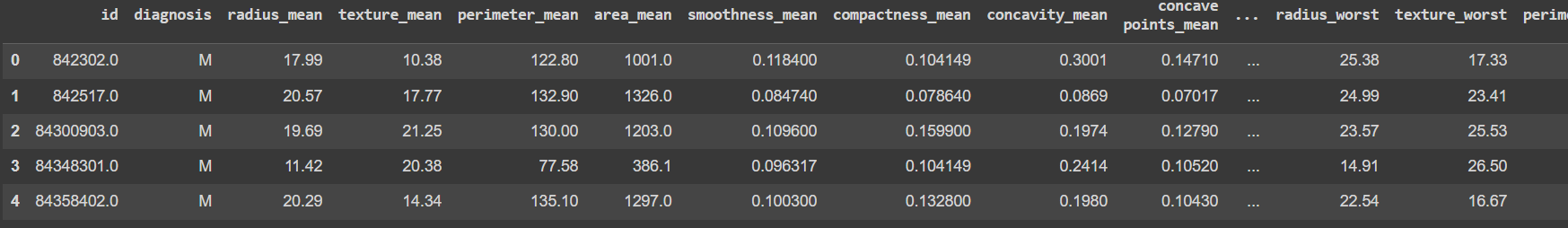
### **3.2.1. Thống kê tóm tắt**

* **Thông tin về tệp dữ liệu:**

****

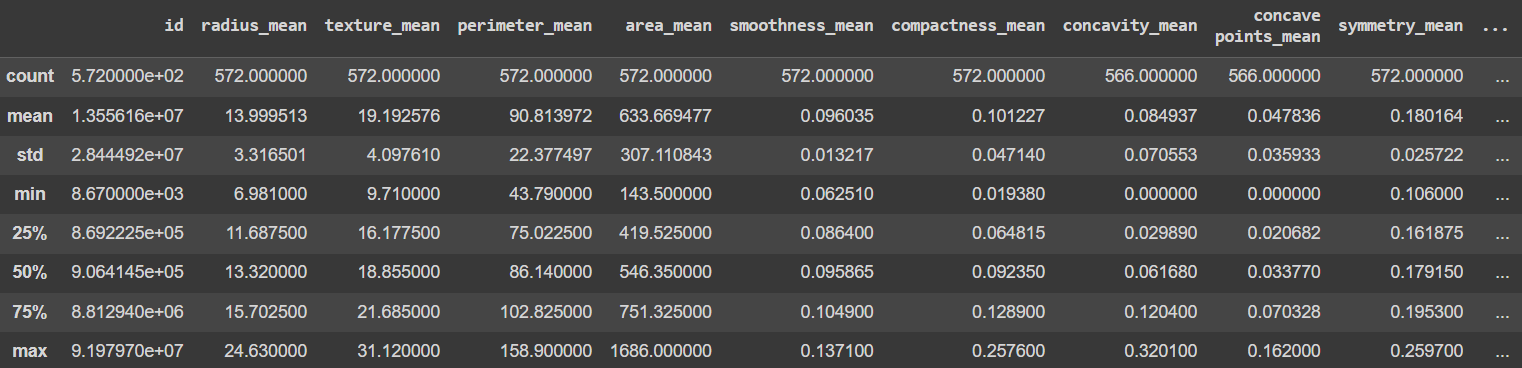
****

Tệp dữ liệu sử dụng bao gồm 572 mẫu dữ liệu, 31 đặc trưng và 1 biến mục tiêu



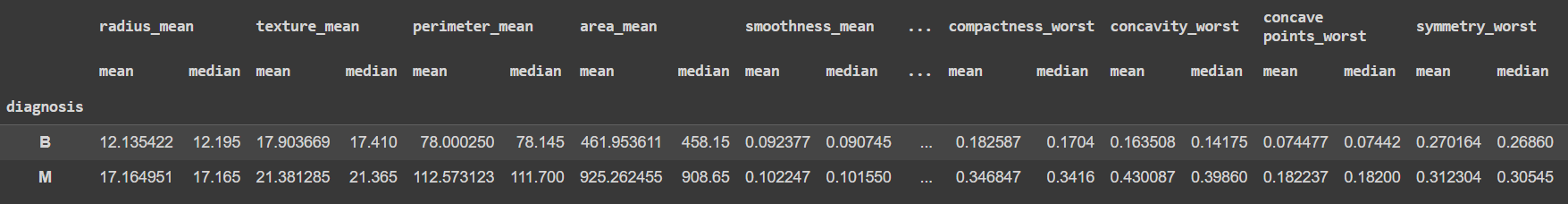
Biểu diễn 5 dòng đầu tiên của dữ liệu

* **Đặc điểm các kiểu đặc trưng kiểu dạng số**

****

**Từ bảng trên ta thấy :**

* Cột “id” chứa giá trị số và có một số dữ liệu bị thiếu là concavity\_mean, concave points\_mean
* Các cột "radius\_mean", "texture\_mean", "perimeter\_mean", "area\_mean" có giá trị trung bình tương đối và độ lệch chuẩn khá lớn, cho thấy sự biến động đáng kể trong dữ liệu.
* Các cột "smoothness\_mean", "compactness\_mean", "concavity\_mean", "concave points\_mean", "symmetry\_mean" có giá trị trung bình và độ lệch chuẩn khá nhỏ, cho thấy sự ổn định hơn trong dữ liệu.

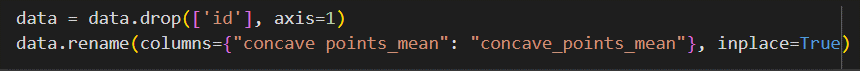


Groupby theo biến mục tiêu diagnosis ta thấy sự phân bố không đều của giá trị trung bình (mean) và giá trị trung vị (median) ở một vài cột đặc trưng

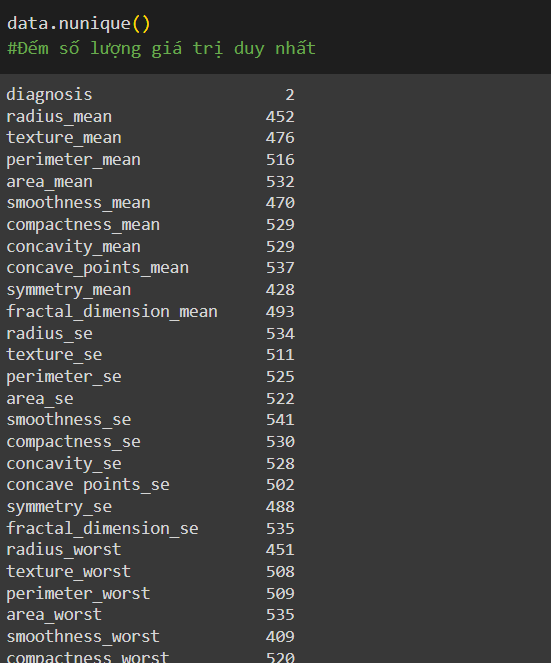
## **3.3. Tiền xử lí dữ liệu**

### **3.3.1 Làm sạch dữ liệu**

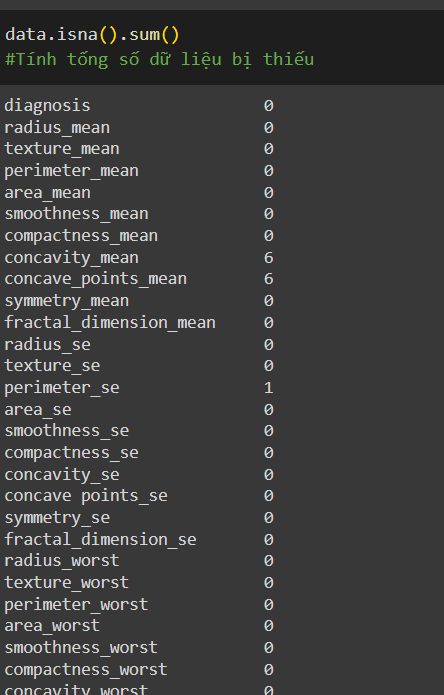
* Loại bỏ cột “id” và đổi tên cột theo định dạng



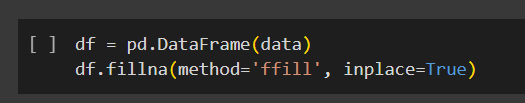
* Đếm số lượng giá trị duy nhất



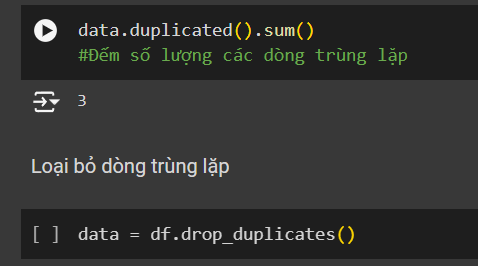
* Tính tổng số giá trị bị thiếu



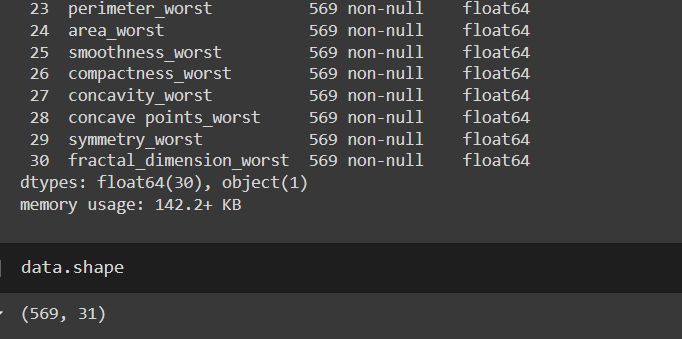
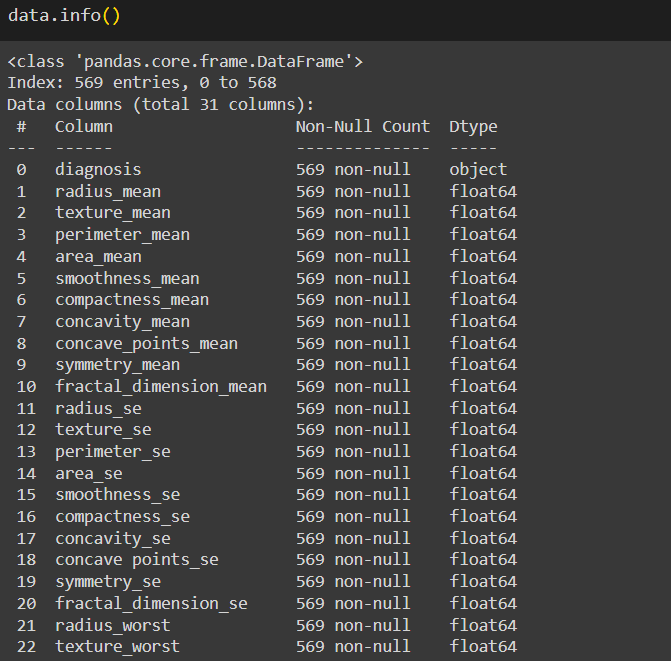
* Xử lí giá trị bị thiếu bằng phương pháp điền giá trị liền trước

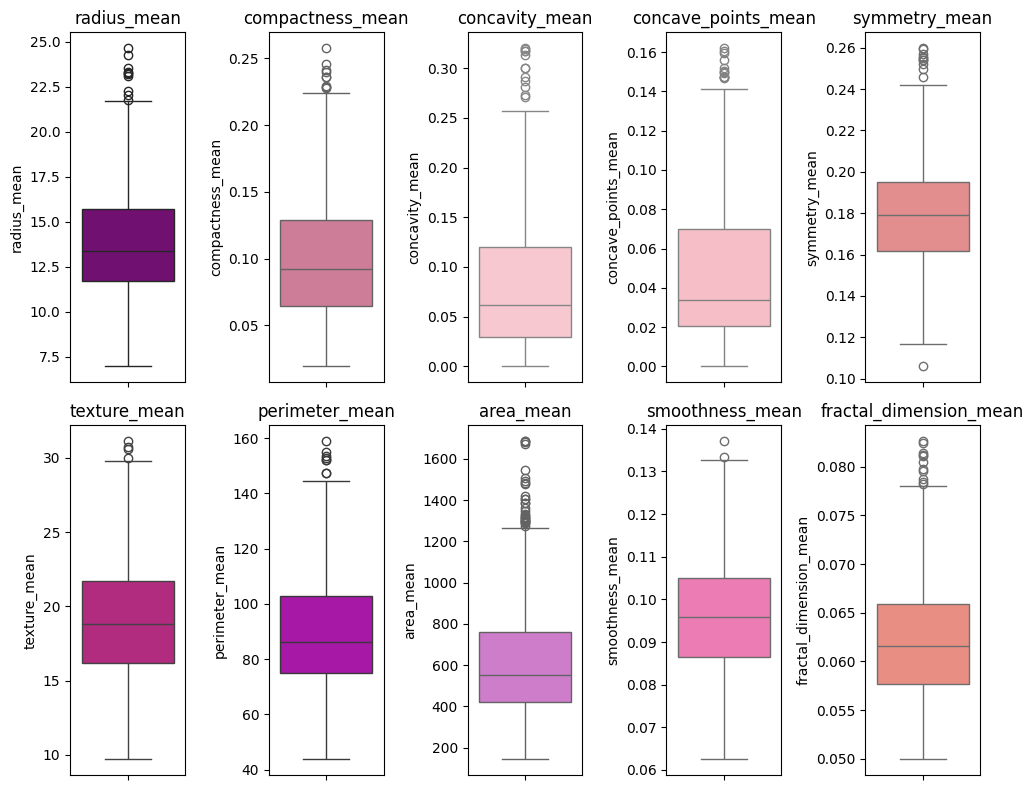


* Tiếp theo là đếm và loại bỏ các giá trị trùng lặp



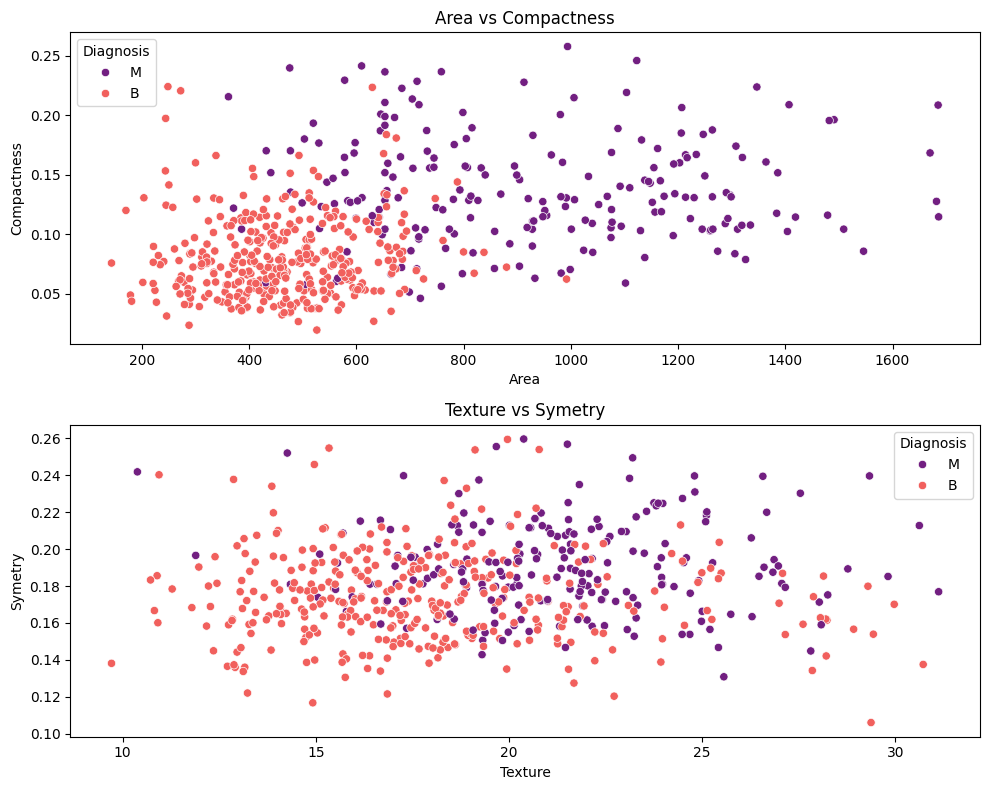
Sau khi làm sạch dữ liệu ta thu được dữ liệu với 569 hàng và 31 cột với tỉ lệ loại bỏ là 3/572 xấp xỉ 0.5% là 1 giá trị chấp nhận được cho việc đưa vào học máy





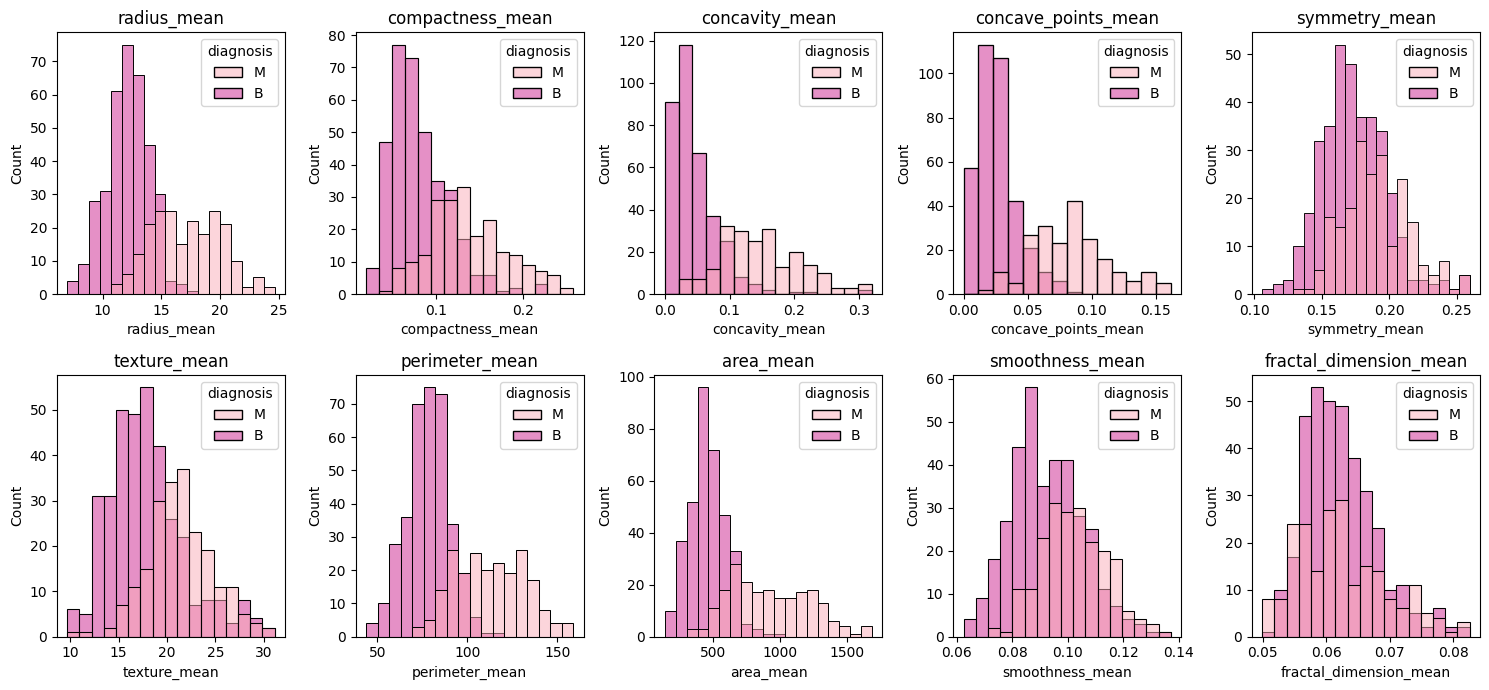
Hình 3.1 Hình ảnh biển diễn dữ liệu sau khi làm sạch dữ liệu

### **3.3.2. Trực quan hóa dữ liệu**



Hình 3.2 Biểu đồ so sánh giữa “Area vs Compactness” và “ Texture vs Symetry” dựa trên tham số mục tiêu

* Biểu đồ so sánh giữa “ Area vs Compactness” cho thấy mối quan hệ giữa diện tích và độ đặc của khối u. Nếu các điểm dữ liệu của khối u ác tính tập trung ở vùng có diện tích lớn và độ đặc cao hơn so với khối u lành tính cho thấy khối u ác tính thường có xu hướng đặc hơn và lớn hơn.
* Có sự phân cụm rõ ràng giữa hai nhóm ( lành tính và ác tính ), cho thấy rằng diện tích và độ đặc có thể là đặc trưng quan trọng để phân loại khối u.
* Tương tự như vậy đối với biểu đồ “ Texture vs Symetry”

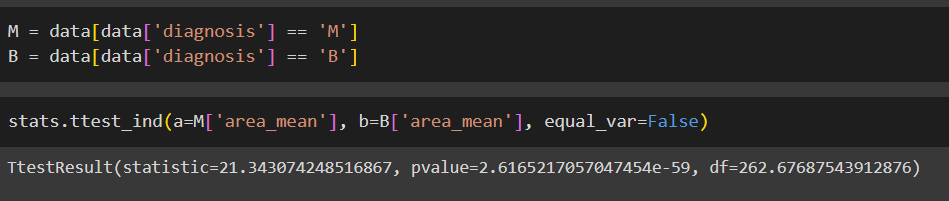


Hình 3.3 Biểu đồ so sánh phân phối các đặc trưng của khối u theo biến mục tiêu “diagnosis”

* Bộ biểu đồ này giúp chúng ta dễ dàng nhận biết những đặc trưng nào có sự khác biệt đáng kể giữa khối u lành tính và ác tính. Ví dụ, nếu chúng ta thấy biểu đồ radius\_mean cho thấy khối u ác tính thường có bán kính lớn hơn khối u lành tính, điều đó có thể cung cấp thông tin hữu ích cho việc chẩn đoán.

### **3.3.3 Kiểm tra giả thuyết**

* Giả thuyết không (H0): Không có sự khác biệt có ý nghĩa thống kê về diện tích trung bình giữa khối u ác tính và khối u lành tính.
* Giả thuyết thay thế (H1): Có sự khác biệt có ý nghĩa thống kê về diện tích trung bình giữa khối u ác tính và khối u lành tính.



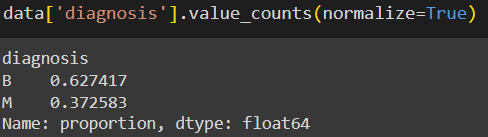
Kết quả đầu ra cho thấy :

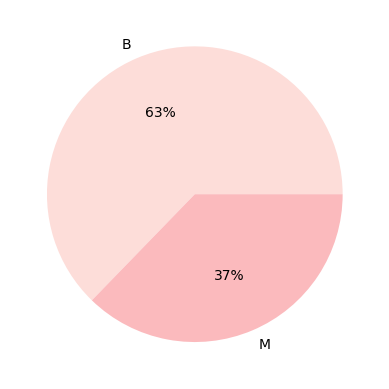
* statistic=21.343074248516867: Đây là giá trị thống kê t được tính toán. Giá trị này rất lớn, cho thấy có sự khác biệt lớn giữa hai nhóm mẫu.
* pvalue=2.6165217057047454e-59: Đây là giá trị p, xác suất để quan sát thấy một giá trị thống kê t cực đoan như vậy hoặc cực đoan hơn, nếu giả thuyết null là đúng (tức là không có sự khác biệt giữa hai nhóm). Giá trị p này cực kỳ nhỏ, gần như bằng 0.
* df=262.67687543912876: Đây là bậc tự do của kiểm định t

Với giá trị p cực kỳ nhỏ (gần như bằng 0), chúng ta có thể bác bỏ giả thuyết null ở bất kỳ mức ý nghĩa thông thường nào (ví dụ: 0.05, 0.01). Điều này có nghĩa là có bằng chứng rất mạnh mẽ để kết luận rằng có sự khác biệt có ý nghĩa thống kê giữa diện tích trung bình của khối u ác tính (M['area\_mean']) và diện tích trung bình của khối u lành tính (B['area\_mean']). Cụ thể hơn, khối u ác tính có xu hướng có diện tích lớn hơn đáng kể so với khối u lành tính.

### **3.3.4 Lấy mẫu và mã hóa nhãn**

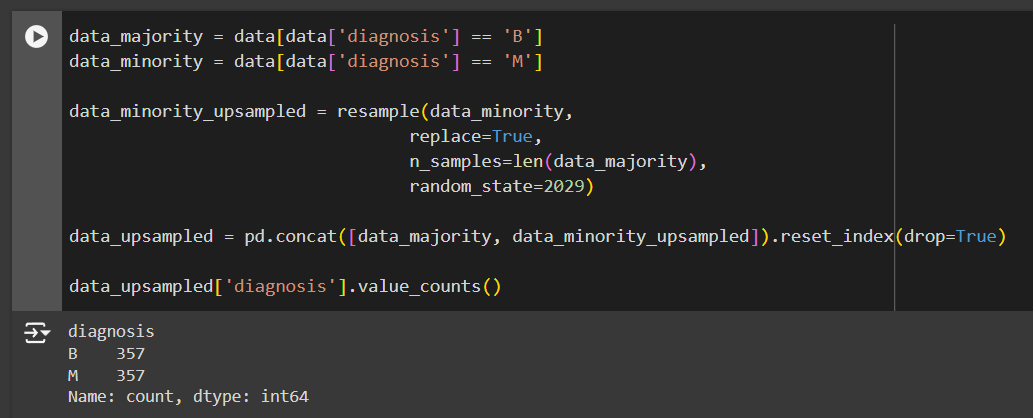
Tính phần trăm theo biến mục tiêu ta được





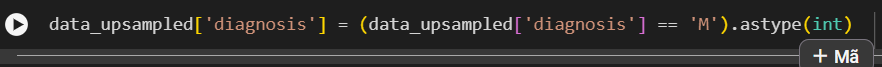
Hình 3.4 Biểu đồ mẫu theo thông số mục tiêu “diagnosis”

Từ biểu đồ trên ta thấy sự chênh lệch về mẫu dữ liệu, với số lượng mẫu B( lành tính ) nhiều hơn đáng kể mẫu M( ác tính) do vậy ta cần cân bằng lại dữ liệu

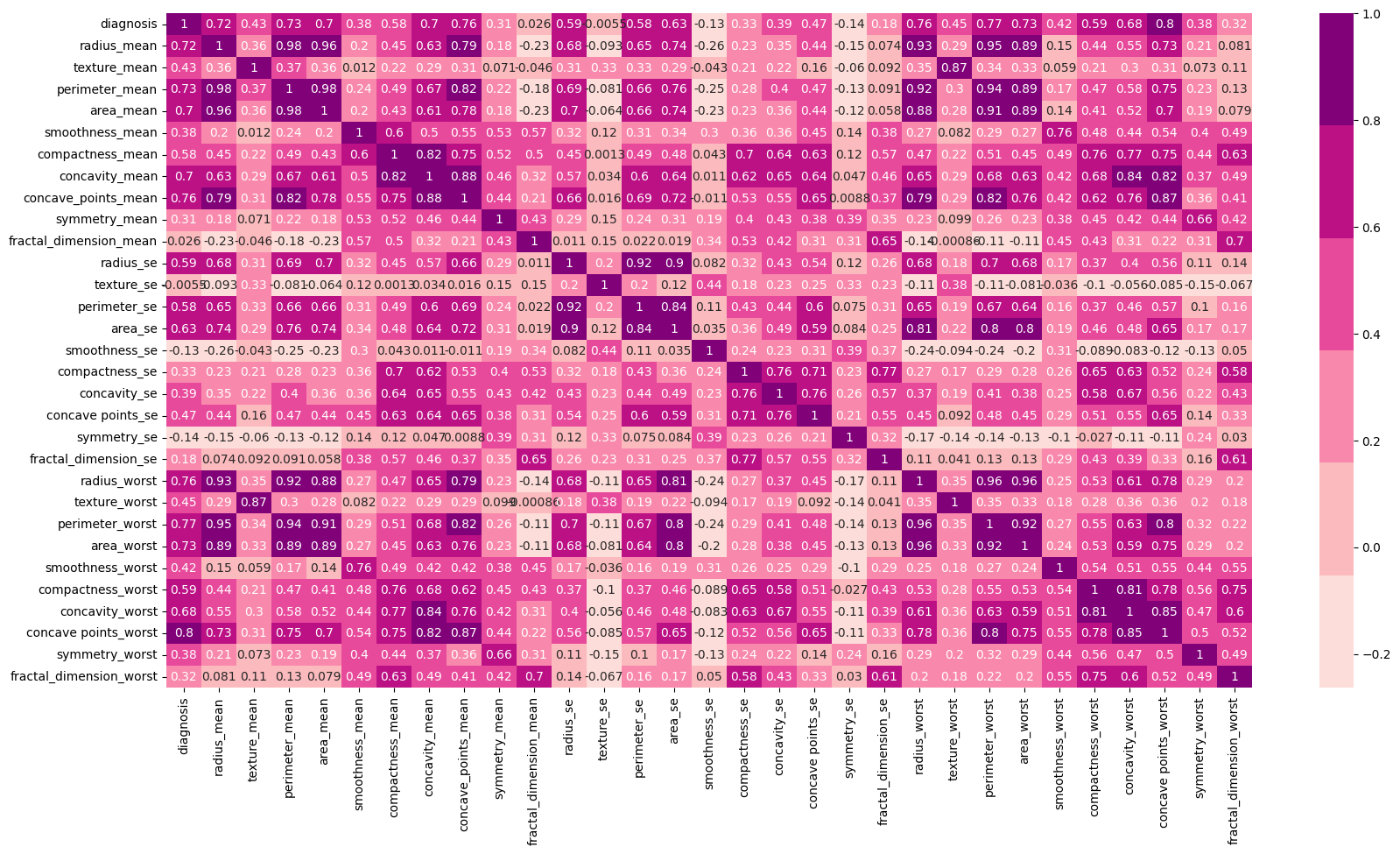


Ta dùng kĩ thuật Upsampling: Kỹ thuật upsampling giúp giải quyết vấn đề này bằng cách tăng số lượng mẫu của lớp thiểu số lên bằng cách sao chép ngẫu nhiên các mẫu hiện có.

Tiếp theo là mã hóa nhãn theo thông số mục tiêu và chuyển biến mục tiêu về dạng 0 và 1:

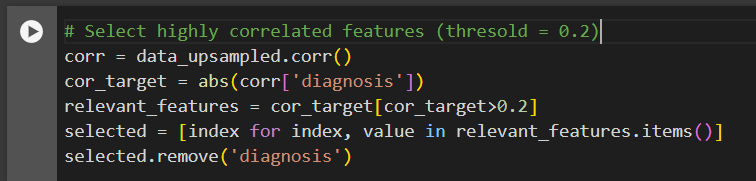


### **3.3.5 Kiểm tra mối tương quan**

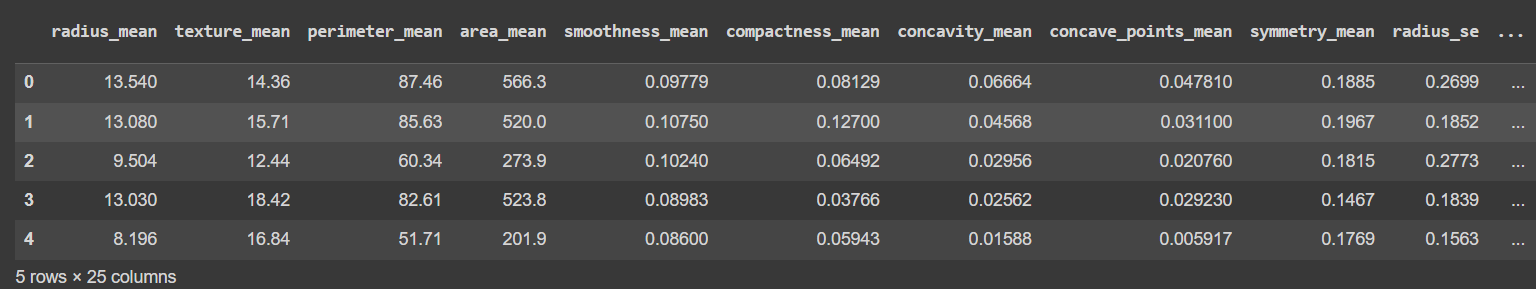


Hình 3.5 Ma trận tương quan giữa các đặc trương

Từ ma trận tương quan ta thấy có rất nhiều đặc trưng ảnh hưởng đến việc chuẩn đoán nhưng để tránh hiện tượng **đa cộng tuyến tính** thì chỉ những đặc điểm có giá trị tương quan cao được chọn và thêm vào biến độc lập



Mức chọn =0.2 dựa trên tỉ lệ tương quan giữa các biến và có nghiên cứu từ các thông số do bác sĩ đưa ra



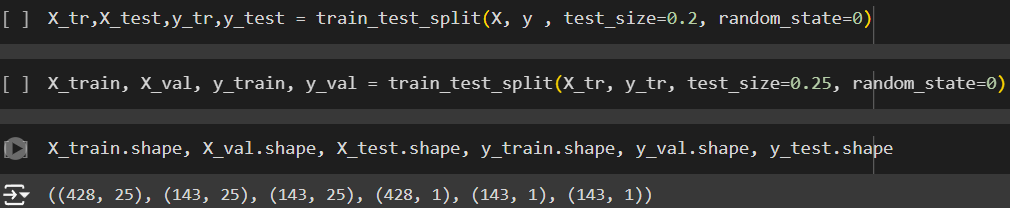
Hình ảnh dữ liệu sau khi chọn tương quan với target > 0.2 thì chỉ còn lại 25 đặc trưng và chuẩn bị đưa vào học máy

## **3.4. Huấn luyện mô hình**

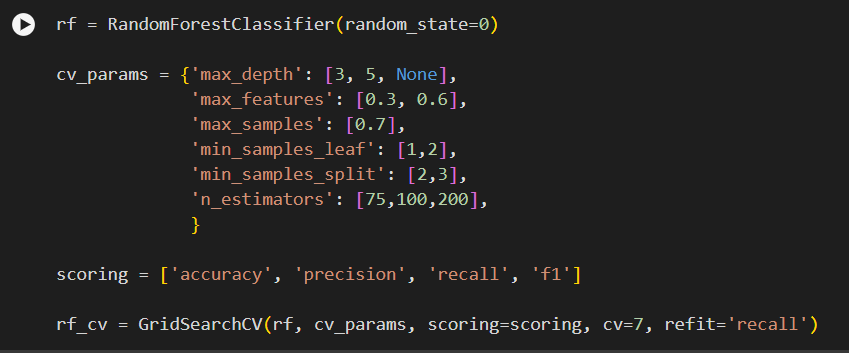
Tiến hành chia tập dữ liệu thành các tập train, test, value

Chia dữ liệu được sử dụng làm tập kiểm tra 20% và 80% làm tập huấn luyện

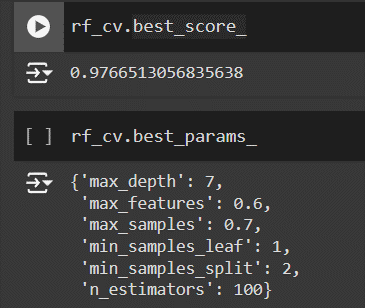
Tỉ lệ dữ liệu từ tập huấn luyện được sử dụng 25% làm tập xác thực



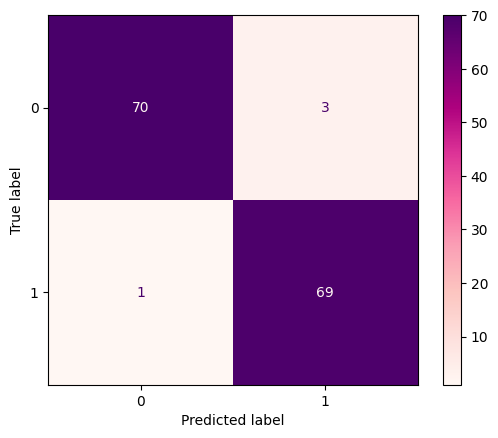
**Khởi tạo mô hình Random forest** và tìm kiếm siêu tham số với GridSearchCV và sau khi tìm được siêu tham số tốt nhất, mô hình cuối cùng sẽ được huấn luyện lại trên toàn bộ tập huấn luyện bằng cách sử dụng siêu tham số tối đa hóa độ nhạy ( recall)



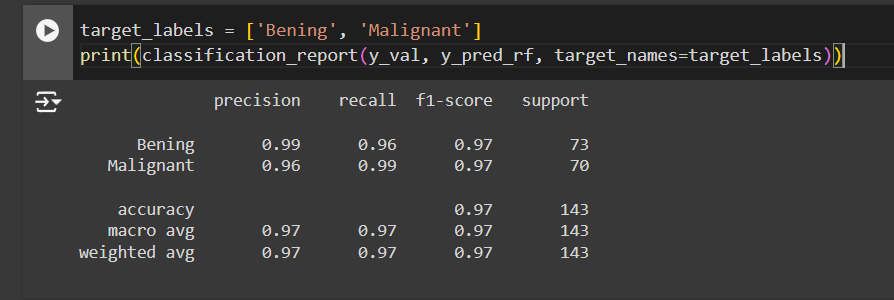
Giá trị siêu tham số sau quá trì tìm kiếm thông số tốt nhất



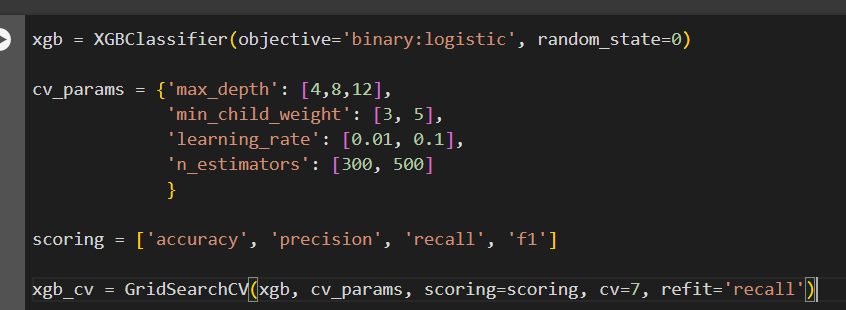
Ma trận nhầm lẫn của mô hình Random forest

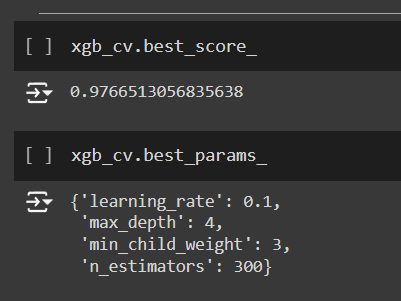


Báo cáo phân loại của mô hình Random forest

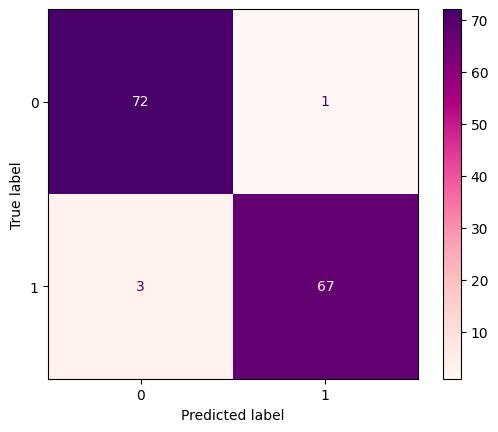


**Khởi tạo mô hình XGBoost** và tìm siêu thông số tối ưu

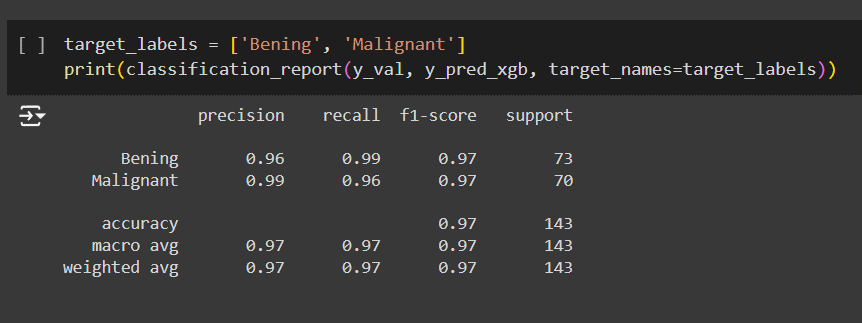




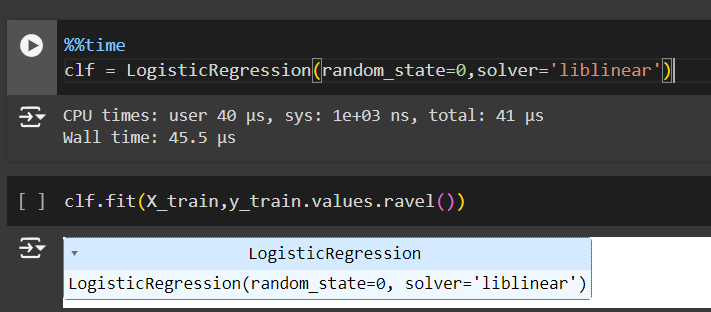
Ma trận nhầm lẫn



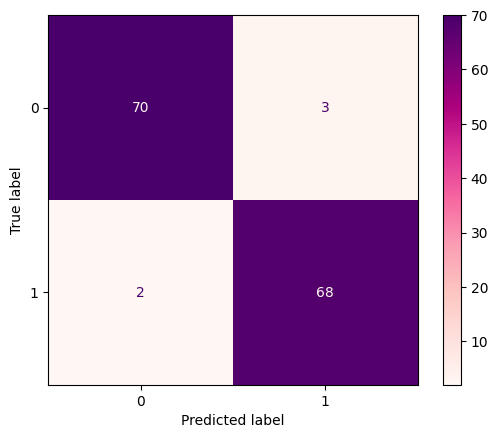
Báo cáo phân loại của XGBoost



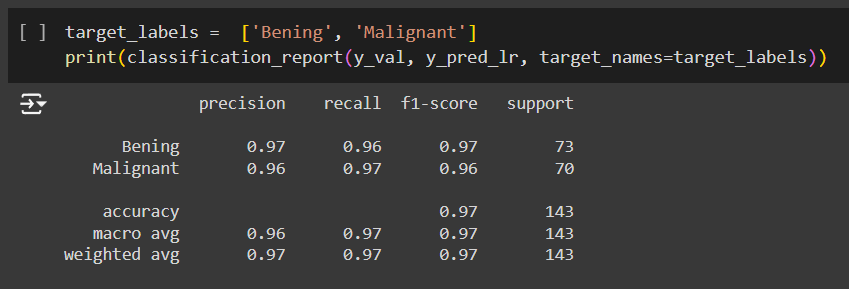
**Khởi tạo mô hình hồi quy logistic**

****

**Ma trận nhầm lẫn**



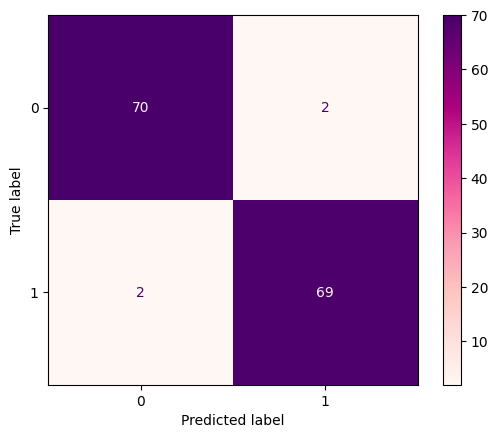
Báo cáo phân loại mô hình hồi quy logistic



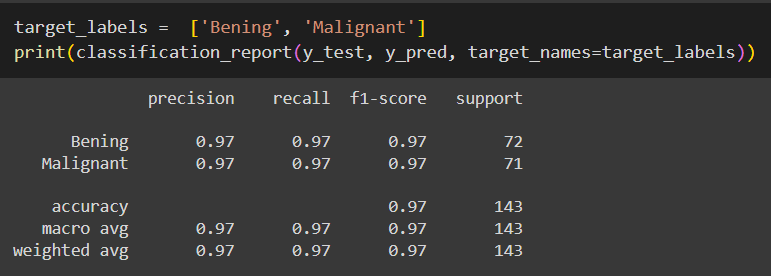
Từ 3 mô hình huấn luyện ta có thể thấy mô hình sử dụng thuật toán Random forest đạt hiệu quả tốt nhất trong 3 mô hình

Ta tiến hành dự đoán với mô hình Random forest với thông số tốt nhất trên tập kiểm tra (X\_test)

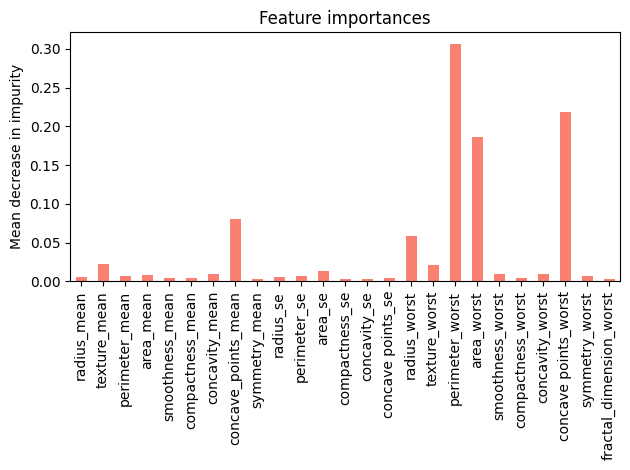
Ma trận nhầm lẫn của tệp kiểm tra



Báo cáo chi tiết của tập dữ liệu sau huấn luyện



Biểu đồ biểu hiện đặc trưng ảnh hưởng đến khả năng dự đoán



## **3.5. Kết luận**

Ung thư vú là một trong những mối đe dọa sức khỏe nghiêm trọng nhất đối với phụ nữ, gây ra gánh nặng lớn cho xã hội và hệ thống y tế. Nghiên cứu về phân tích và dự đoán nguy cơ ung thư vú đã và đang đóng vai trò quan trọng trong việc phát hiện sớm, điều trị hiệu quả và giảm tỷ lệ tử vong do căn bệnh này.

Thông qua việc phân tích dữ liệu, chúng ta đã xác định được các đặc trưng quan trọng đóng góp vào việc dự đoán ung thư vú, nổi bật là các đặc trưng thuộc nhóm "\_worst" (giá trị trung bình của ba giá trị lớn nhất). Đặc biệt, khi tập trung vào nhóm '\_mean', các đặc trưng như độ lõm (concavity), chu vi (perimeter) và diện tích (area) của khối u có vai trò quan trọng trong việc phát hiện chính xác ung thư vú. Mô hình dự đoán đã đạt được độ chính xác 97%, cho thấy tiềm năng to lớn của việc ứng dụng các phương pháp học máy trong lĩnh vực y tế.

Nghiên cứu này không chỉ cung cấp cái nhìn sâu sắc về các yếu tố nguy cơ ung thư vú mà còn mở ra hướng đi mới trong việc phát triển các công cụ hỗ trợ chẩn đoán và dự đoán tiên tiến. Việc ứng dụng rộng rãi các mô hình dự đoán như vậy có thể giúp phát hiện sớm ung thư vú, từ đó nâng cao hiệu quả điều trị và giảm thiểu nguy cơ tử vong cho bệnh nhân.

Trong tương lai, việc tiếp tục nghiên cứu và phát triển các phương pháp học máy mới, kết hợp với dữ liệu đa dạng và phong phú hơn, sẽ tiếp tục cải thiện khả năng dự đoán và hỗ trợ quyết định lâm sàng trong việc phòng ngừa, chẩn đoán và điều trị ung thư vú. Đồng thời, việc chia sẻ và hợp tác nghiên cứu giữa các chuyên gia y tế, nhà khoa học dữ liệu và các bên liên quan khác cũng đóng vai trò quan trọng trong việc thúc đẩy sự phát triển của lĩnh vực này, hướng tới mục tiêu chung là bảo vệ sức khỏe và nâng cao chất lượng cuộc sống cho phụ nữ.