TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN &

TRUYỀN THÔNG VIỆT HÀN

**Khoa Khoa Học Máy Tính**



ĐỒ ÁN MÔN HỌC

**ĐỀ TÀI : DỰ ĐOÁN UNG THƯ VÚ BẰNG HỒI QUY LOGISTIC**

Sinh viên thực hiện: **PHAN MINH ĐỨC – 20SE2**

Giảng viên hướng dẫn: **TS. NGUYỄN SỸ THÌN**

Đà Nẵng, tháng 6 năm 2023

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN &

TRUYỀN THÔNG VIỆT HÀN

**Khoa Khoa Học Máy Tính**



ĐỒ ÁN MÔN HỌC

**ĐỀ TÀI : DỰ ĐOÁN UNG THƯ VÚ BẰNG HỒI QUY LOGISTIC**

Sinh viên thực hiện: **PHAN MINH ĐỨC – 20SE2**

Giảng viên hướng dẫn: **TS. NGUYỄN SỸ THÌN**

Đà Nẵng, tháng 6 năm 2023

NHẬN XÉT CỦA GIẢNG VIÊN HƯỚNG DẪN

# 

# 

# 

# 

# 

LỜI CẢM ƠN

Đầu tiên, nhóm em xin gửi lời cảm ơn chân thành đến Trường đại học Công nghệ Thông tin và Truyền thông Việt - Hàn đã đưa môn học Học máy vào trương trình giảng dạy. Đặc biệt, chúng em xin gửi lời cảm ơn sâu sắc đến giảng viên bộ môn – ThS. TS. Nguyễn Sĩ Thìn đã dạy, truyền đạt những kiến thức quý báu cho chúng em trong suốt thời gian học tập vừa qua. Trong thời gian tham gia lớp đồ án của thầy, em đã có thêm cho mình nhiều kiến thức bổ ích, tinh thần học tập hiệu quả, nghiêm túc. Đây chắc chắn sẽ là những kiến thức quý báu, là hành trang để chúng em có thể vững bước sau này.

Bộ môn Học máy là môn học thú vị, vô cùng bổ ích và có tính thực tế cao. Đảm bảo cung cấp đủ kiến thức, gắn liền với nhu cầu thực tiễn của sinh viên. Tuy nhiên, do vốn kiến thức còn nhiều hạn chế và khả năng tiếp thu thực tế còn nhiều bỡ ngỡ khi thực hiện đề tài.

Mặc dù nhóm chúng em đã có sự cố gắng tìm hiểu trong khi thực hiện đề tài nhưng do điều kiện và thời gian có hạn, cũng như khả năng quản lý dự án chưa có nên trong bài báo cáo này chúng em không tránh khỏi thiếu sót, chúng em rất mong nhận được sự giúp đỡ từ phía các thầy cô và các bạn.

Chúng em xin chân thành cảm ơn!

*Sinh viên,*

MỤC LỤC

[CHƯƠNG 1. TỔNG QUAN VỀ BÀI TOÁN 1](#_Toc137026152)

[1. Giới thiệu về bài toán 1](#_Toc137026153)

[2. Lý do chọn đề tài 1](#_Toc137026154)

[CHƯƠNG 2. GIỚI THIỆU THUẬT TOÁN HỒI QUY LOGISTIC 2](#_Toc137026155)

[1. Giới thiệu hồi quy logistic 2](#_Toc137026156)

[1.1. Sự quan trong của hồi quy logistic 2](#_Toc137026157)

[1.2. Sự quan trong của hồi quy logistic 2](#_Toc137026158)

[2. Hoạt động của phân tích hồi quy 3](#_Toc137026159)

[3. Hoạt động của Mô hình hồi quy logistic 3](#_Toc137026160)

[4. Phân tích hồi quy logistic có những loại nào? 6](#_Toc137026161)

[CHƯƠNG 3. SỬ DỤNG HỒI QUY LOGISTIC ĐỂ DỰ ĐOÁN 7](#_Toc137026162)

[1. Mô tả dữ liệu : 7](#_Toc137026163)

[2. Các bước thực hiện 7](#_Toc137026164)

[2.1. Mô tả dữ liệu 8](#_Toc137026165)

[2.2. Phân tích định lượng/định tính dữ liệu, mối quan hệ giữa các đặc tính dữ liệu 8](#_Toc137026166)

[2.3. Tiền xử lý dữ liệu 9](#_Toc137026167)

[2.4. Sử dụng thuật toán Logistic Regression để dự đoán dữ liệu 10](#_Toc137026168)

[2.5. Đánh giá mô hình 10](#_Toc137026169)

[KẾT LUẬN 12](#_Toc137026170)

[1. Kết quả đạt được 12](#_Toc137026171)

[2. Hướng nghiên cứu 12](#_Toc137026172)

[PHỤ LỤC ii](#_Toc137026173)

# 

# **CHƯƠNG 1. TỔNG QUAN VỀ BÀI TOÁN**

1. Giới thiệu về bài toán

Bài toán "Breast cancer prediction using Logistic Regression" (Dự đoán ung thư vú bằng Hồi quy Logistic) nhằm dự đoán khả năng mắc bệnh ung thư vú dựa trên các đặc tính và thông tin có sẵn về bệnh nhân. Ung thư vú là một trong những căn bệnh phổ biến và nguy hiểm đối với phụ nữ trên toàn thế giới. Việc phát hiện sớm ung thư vú có thể cải thiện khả năng điều trị và cứu sống nhiều người.

Quá trình dự đoán ung thư vú thường sử dụng phương pháp học máy để xây dựng mô hình dự đoán. Trong trường hợp này, em chọn sử dụng Logistic Regression, một thuật toán học máy phân loại phổ biến và phù hợp cho các bài toán dự đoán nhị phân..

## 2. Lý do chọn đề tài

Lựa chọn đề tài "Breast cancer prediction using Logistic Regression" có một số lý do quan trọng:

* Tầm quan trọng của bài toán: Ung thư vú là căn bệnh phổ biến và có tác động lớn đến sức khỏe và cuộc sống của nhiều phụ nữ. Dự đoán khả năng mắc bệnh ung thư vú có thể giúp trong việc sàng lọc và phát hiện sớm, tăng cơ hội điều trị hiệu quả và cứu sống nhiều người.
* Ứng dụng trong lĩnh vực y học: Bài toán này áp dụng phương pháp học máy vào lĩnh vực y học, mang lại sự tiện lợi và hiệu quả trong việc dự đoán và phân loại bệnh. Sử dụng Logistic Regression, một thuật toán phân loại phổ biến, cung cấp một phương pháp đơn giản nhưng mạnh mẽ để dự đoán ung thư vú.
* Tiềm năng ứng dụng thực tế: Kết quả của bài toán có thể được áp dụng trong thực tế y tế để hỗ trợ các chuyên gia y tế trong quá trình đánh giá nguy cơ ung thư vú và đưa ra quyết định điều trị. Điều này có thể giúp tăng khả năng phát hiện sớm và cải thiện kết quả điều trị cho bệnh nhân.

Việc chọn đề tài "Breast cancer prediction using Logistic Regression" không chỉ đáp ứng nhu cầu quan trọng trong lĩnh vực y học mà còn mang lại tiềm năng ứng dụng thực tế để cải thiện chẩn đoán và điều trị ung thư vú.

# **CHƯƠNG 2.** **GIỚI THIỆU THUẬT TOÁN HỒI QUY LOGISTIC**

## 1. Giới thiệu hồi quy logistic

Hồi quy logistic là một kỹ thuật phân tích dữ liệu sử dụng toán học để tìm ra mối quan hệ giữa hai yếu tố dữ liệu. Sau đó, kỹ thuật này sử dụng mối quan hệ đã tìm được để dự đoán giá trị của những yếu tố đó dựa trên yếu tố còn lại. Dự đoán thường cho ra một số kết quả hữu hạn, như có hoặc không.

### 1.1. Sự quan trong của hồi quy logistic

Hồi quy logistic là một kỹ thuật quan trọng trong lĩnh vực trí tuệ nhân tạo và máy học (AI/ML). Mô hình ML là các chương trình phần mềm có thể được đào tạo để thực hiện các tác vụ xử lý dữ liệu phức tạp mà không cần sự can thiệp của con người. Mô hình ML được xây dựng bằng hồi quy logistic có thể giúp các tổ chức thu được thông tin chuyên sâu hữu ích từ dữ liệu kinh doanh của mình. Họ có thể sử dụng những thông tin chuyên sâu này để phân tích dự đoán nhằm giảm chi phí hoạt động, tăng độ hiệu quả và đổi chỉnh quy mô nhanh hơn. Ví dụ: doanh nghiệp có thể khám phá các mẫu hình cải thiện khả năng giữ chân nhân viên hoặc tạo ra thiết kế sản phẩm mang về nhiều lợi nhuận hơn.

Dưới đây là một số lợi ích của việc sử dụng hồi quy logistic so với các kỹ thuật ML khác.

**Tính đơn giản**

Các mô hình hồi quy logistic ít phức tạp về mặt toán học hơn các phương pháp ML khác. Do đó, bạn có thể triển khai chúng ngay cả khi đội ngũ của bạn không ai có chuyên môn sâu về ML.

**Tốc độ**

Các mô hình hồi quy logistic có thể xử lý khối lượng lớn dữ liệu ở tốc độ cao bởi chúng cần ít khả năng điện toán hơn, chẳng hạn như bộ nhớ và sức mạnh xử lý. Điều này khiến các mô hình hồi quy logistic trở nên lý tưởng đối với những tổ chức đang bắt đầu với các dự án ML để đạt được một số thành tựu nhanh chóng.

**Sự linh hoạt**

Bạn có thể sử dụng hồi quy logistic để tìm đáp án cho các câu hỏi có hai hoặc nhiều kết quả hữu hạn. Bạn cũng có thể sử dụng phương pháp này để xử lý trước dữ liệu. Ví dụ: bạn có thể sắp xếp dữ liệu với một phạm vi giá trị lớn, chẳng hạn như giao dịch ngân hàng, thành một phạm vi giá trị hữu hạn, nhỏ hơn nhờ hồi quy logistic. Sau đó, bạn có thể xử lý tập dữ liệu nhỏ hơn này với các kỹ thuật ML khác để phân tích chính xác hơn.

**Khả năng hiển thị**

Phân tích hồi quy logistic cung cấp cho nhà phát triển khả năng nhìn nhận các quy trình phần mềm nội bộ rõ hơn so với các kỹ thuật phân tích dữ liệu khác. Khắc phục sự cố và sửa lỗi cũng trở nên dễ dàng hơn do các phép toán ít phức tạp hơn.

### 1.2. Sự quan trong của hồi quy logistic

Hồi quy logistic có một số ứng dụng thực tế trong nhiều ngành công nghiệp khác nhau.

**Sản xuất**

Các công ty sản xuất áp dụng phân tích hồi quy logistic để ước tính xác suất xảy ra sự cố ở bộ phận trong máy móc. Sau đó, họ sẽ lên lịch bảo trì dựa trên xác suất đã ước tính này để giảm thiểu sự cố trong tương lai.

**Chăm sóc sức khỏe**

Các nhà nghiên cứu y khoa lên kế hoạch điều trị và chăm sóc dự phòng bằng cách dự đoán khả năng mắc bệnh ở bệnh nhân. Họ sử dụng các mô hình hồi quy logistic để so sánh tác động của tiền sử gia đình hoặc của bộ gen lên bệnh tật.

**Tài chính**

Các công ty tài chính phải phân tích các giao dịch tài chính để đề phòng gian lận, xem xét các đơn xin vay và đơn bảo hiểm để đề phòng rủi ro. Những vấn đề này phù hợp với mô hình hồi quy logistic bởi chúng có kết quả cụ thể, chẳng hạn như rủi ro cao hoặc rủi ro thấp và gian lận hoặc không gian lận.

**Bộ phận Tiếp thị**

Các công cụ quảng cáo trực tuyến sử dụng mô hình hồi quy logistic để dự đoán xem người dùng sẽ nhấp vào một quảng cáo hay không. Kết quả là, các nhà tiếp thị có thể phân tích phản ứng của người dùng đối với những từ ngữ và hình ảnh khác nhau, tạo ra các quảng cáo hiệu suất cao có khả năng thu hút khách hàng.

## 2. Hoạt động của phân tích hồi quy

Hồi quy logistic là một trong số ít các kỹ thuật phân tích hồi quy khác nhau thường được các nhà khoa học dữ liệu sử dụng trong máy học (ML). Để hiểu rõ về hồi quy logistic, trước tiên chúng ta phải hiểu kỹ thuật phân tích hồi quy cơ bản. Dưới đây là một ví dụ về phân tích hồi quy tuyến tính để cho thấy cách thức hoạt động của phân tích hồi quy.

**Xác định câu hỏi**

Bất kỳ quá trình phân tích dữ liệu nào cũng bắt đầu bằng một câu hỏi kinh doanh.

**Thu thập dữ liệu lịch sử**

Sau khi xác định câu hỏi, bạn cần xác định các yếu tố dữ liệu có liên quan. Sau đó, bạn sẽ thu thập dữ liệu trước đây cho tất cả các yếu tố.

**Đào tạo mô hình phân tích hồi quy**

Bạn sẽ xử lý dữ liệu lịch sử bằng phần mềm hồi quy. Phần mềm sẽ xử lý các điểm dữ liệu khác nhau và kết nối chúng theo phương thức toán học bằng cách sử dụng phương trình.

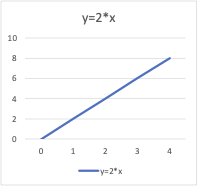
**Dự đoán các giá trị không xác định**

Đối với các giá trị không xác định, phần mềm sẽ áp dụng phương trình để dự đoán.

## 3. Hoạt động của Mô hình hồi quy logistic

**Phương trình**

Trong toán học, phương trình cho ta mối quan hệ giữa hai biến: x và y. Bạn có thể sử dụng các phương trình hoặc hàm này để vẽ đồ thị theo trục x và trục y bằng cách nhập các giá trị khác nhau của x và y. Ví dụ: nếu bạn vẽ đồ thị cho hàm y = 2\*x, bạn sẽ có một đường thẳng như hình dưới đây. Do đó hàm này còn được gọi là hàm tuyến tính.



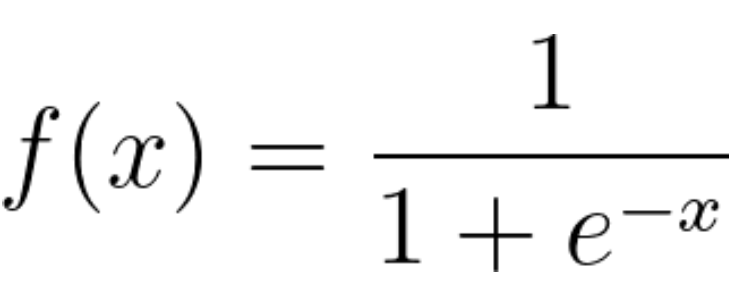
**Biến**

Trong thống kê, biến là các yếu tố dữ liệu hoặc thuộc tính có giá trị khác nhau. Bất kỳ phân tích nào cũng có một số biến nhất định là biến độc lập hoặc biến giải thích. Những thuộc tính này là nguyên nhân của một kết quả. Các biến khác là biến phụ thuộc hoặc biến đáp ứng; giá trị của chúng phụ thuộc vào các biến độc lập. Nhìn chung, hồi quy logistic khám phá cách các biến độc lập ảnh hưởng đến một biến phụ thuộc bằng cách xem xét các giá trị dữ liệu lịch sử của cả hai biến.

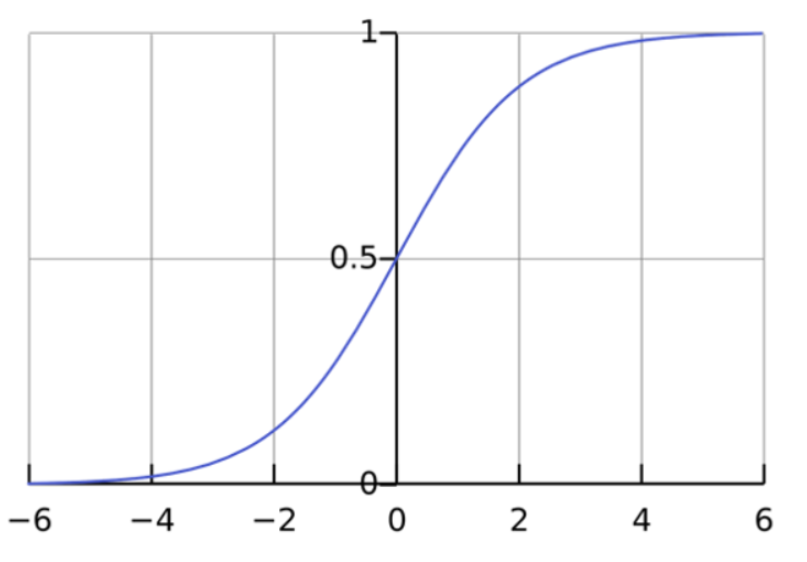
Trong ví dụ ở trên của chúng tôi, x được gọi là biến độc lập, biến dự đoán hoặc biến giải thích vì nó có một giá trị đã xác định. Y được gọi là biến phụ thuộc, biến kết quả hoặc biến đáp ứng vì giá trị của nó không xác định.

**Hàm hồi quy logistic**

Hồi quy logistic là một mô hình thống kê sử dụng hàm logistic, hay hàm logit trong toán học làm phương trình giữa x và y. Hàm logit ánh xạ y làm hàm sigmoid của x.



Nếu vẽ phương trình hồi quy logistic này, bạn sẽ có một đường cong hình chữ S như hình dưới đây.



Như bạn có thể thấy, hàm logit chỉ trả về các giá trị giữa 0 và 1 cho biến phụ thuộc, dù giá trị của biến độc lập là gì. Đây là cách hồi quy logistic ước tính giá trị của biến phụ thuộc. Phương pháp hồi quy logistic cũng lập mô hình phương trình giữa nhiều biến độc lập và một biến phụ thuộc.

**Phân tích hồi quy logistic với nhiều biến độc lập**

Trong nhiều trường hợp, nhiều biến giải thích ảnh hưởng đến giá trị của biến phụ thuộc. Để lập mô hình các tập dữ liệu đầu vào như vậy, công thức hồi quy logistic phải giả định mối quan hệ tuyến tính giữa các biến độc lập khác nhau. Bạn có thể sửa đổi hàm sigmoid và tính toán biến đầu ra cuối cùng như sau

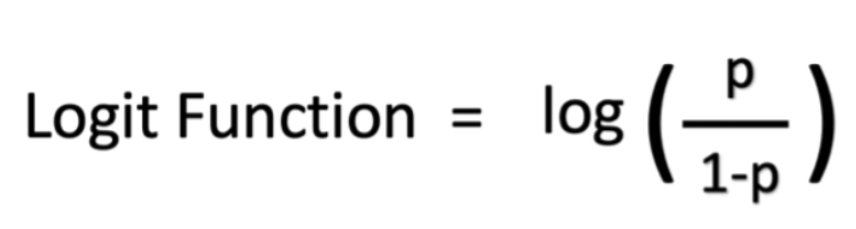
y = f(β0 + β1x1 + β2x2+… βnxn)

Ký hiệu β đại diện cho hệ số hồi quy. Mô hình logit có thể đảo ngược tính toán các giá trị hệ số này khi bạn cho nó một tập dữ liệu thực nghiệm đủ lớn có các giá trị đã xác định của cả hai biến phụ thuộc và biến độc lập.

**Log của tỷ số odds**

Mô hình logit cũng có thể xác định tỷ số thành công trên thất bại hay log của tỷ số odds. Ví dụ: nếu bạn đang chơi poker với bạn bè và thắng bốn ván trên mười ván, tỷ số chiến thắng của bạn là bốn phần sáu, hoặc 4/6, và đó là tỷ số thành công trên thất bại của bạn. Mặt khác, xác suất thắng là 4/10.

Về mặt toán học, tỷ số odds về mặt xác suất của bạn là p/(1 - p) và log của tỷ số odds là log (p/(1 - p)). Bạn có thể biểu diễn hàm logistic bằng log của tỷ số odds như hình dưới đây:



## ****4. Phân tích hồi quy logistic có những loại nào?****

Có ba cách tiếp cận phân tích hồi quy logistic dựa trên kết quả của biến phụ thuộc.

**Hồi quy logistic nhị phân**

Hồi quy logistic nhị phân phù hợp với các vấn đề phân lớp nhị phân chỉ có hai kết quả có thể xảy ra. Biến phụ thuộc chỉ có thể có hai giá trị, chẳng hạn như có và không hoặc 0 và 1.

Dù hàm logistic tính toán một phạm vi giá trị giữa 0 và 1, mô hình hồi quy nhị phân vẫn sẽ làm tròn kết quả đến các giá trị gần nhất. Nói chung, kết quả dưới 0,5 sẽ được làm tròn thành 0 và kết quả trên 0,5 sẽ được làm tròn thành 1, do đó hàm logistic trả về một kết quả nhị phân.

**Hồi quy logistic đa thức**

Hồi quy đa thức có thể phân tích các vấn đề có một số kết quả có thể xảy ra, miễn là số kết quả hữu hạn. Ví dụ: kỹ thuật này có thể dự đoán xem giá nhà sẽ tăng 25%, 50%, 75% hay 100% dựa trên dữ liệu dân số, nhưng sẽ không thể dự đoán được giá trị chính xác của một ngôi nhà.

Hồi quy logistic đa thức hoạt động bằng cách ánh xạ các giá trị kết quả cho các giá trị khác nhau giữa 0 và 1. Hàm logistic có thể trả về một khoảng dữ liệu liên tục như 0,1, 0,11, 0,12, v.v., do đó hồi quy đa thức cũng nhóm đầu ra đến các giá trị gần nhất có thể có.

**Hồi quy logistic thứ tự**

Hồi quy logistic thứ tự, hay mô hình logit có thứ tự, là một loại hồi quy đa thức đặc biệt cho các vấn đề trong đó các số đại diện cho các bậc chứ không phải là giá trị thực tế. Ví dụ: bạn sẽ sử dụng hồi quy thứ tự để dự đoán đáp án cho câu hỏi khảo sát yêu cầu khách hàng đánh giá dịch vụ của bạn ở mức kém, ổn, tốt hoặc xuất sắc dựa trên một giá trị số, chẳng hạn như số lượng mặt hàng họ mua từ bạn trong năm.

# **CHƯƠNG 3. SỬ DỤNG HỒI QUY LOGISTIC ĐỂ DỰ ĐOÁN**

## 1. Mô tả dữ liệu :

Dữ liệu được mô tả trong đề tài là kết quả của việc phân tích hình ảnh số hóa từ một loạt các tế bào nhân có trong hình ảnh khúc xạ kim tiêm mỏng (FNA) của khối u vú. Các đặc trưng này mô tả các đặc điểm của các nhân tế bào.

Thông tin về các thuộc tính:

Số ID

Chẩn đoán (M = ác tính, B = lành tính)

Các đặc trưng có giá trị thực được tính cho mỗi nhân tế bào :

a) Bán kính (giá trị trung bình của khoảng cách từ tâm đến các điểm trên đường viền)

b) Độ nhám (độ lệch chuẩn của giá trị mức xám)

c) Chu vi

d) Diện tích

e) Mịn màng (biến thiên cục bộ trong độ dài bán kính)

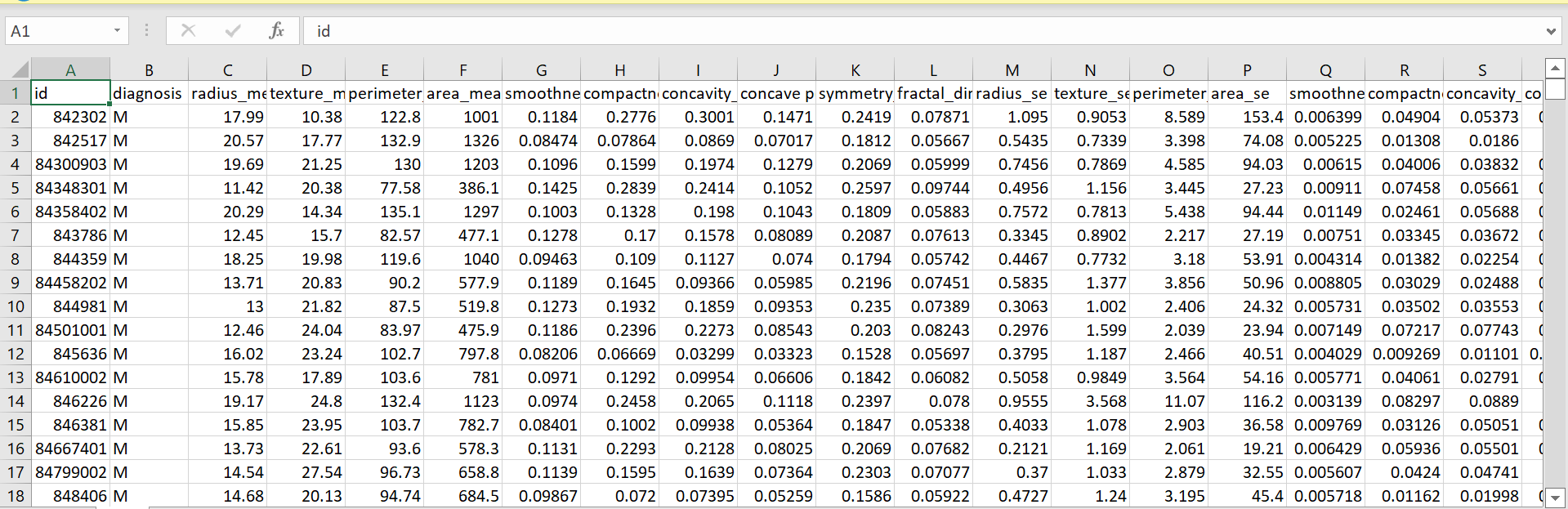
f) Độ nhồi (chu vi^2 / diện tích - 1.0)

g) Độ lõm (mức độ lõm của các phần lõm trên đường viền)

h) Điểm lõm (số lượng phần lõm trên đường viền)

i) Đối xứng

j) Chiều không gian fractal ("xấp xỉ bờ biển" - 1)



## 2. Các bước thực hiện

Bài toán được thực hiện theo các bước sau :

- Mô tả dữ liệu: các thông tin dữ liệu, các đặc tính và mối liên quan giữa các đặc tính

- Phân tích định lượng/ định tính dữ liệu, mối quan hệ giữa các đặc tính dữ liệu

- Tiền xử lý dữ liệu: các dữ liệu khuyết (missing data), dữ liệu không đồng nhất (ví dụ: low fat, lf, LowFat,...), tìm dữ liệu outlier...

- Sử dụng bất kỳ thuật toán học máy nào (knn, kmean, hồi quy tuyến tính,...) để dự đoán dữ liệu

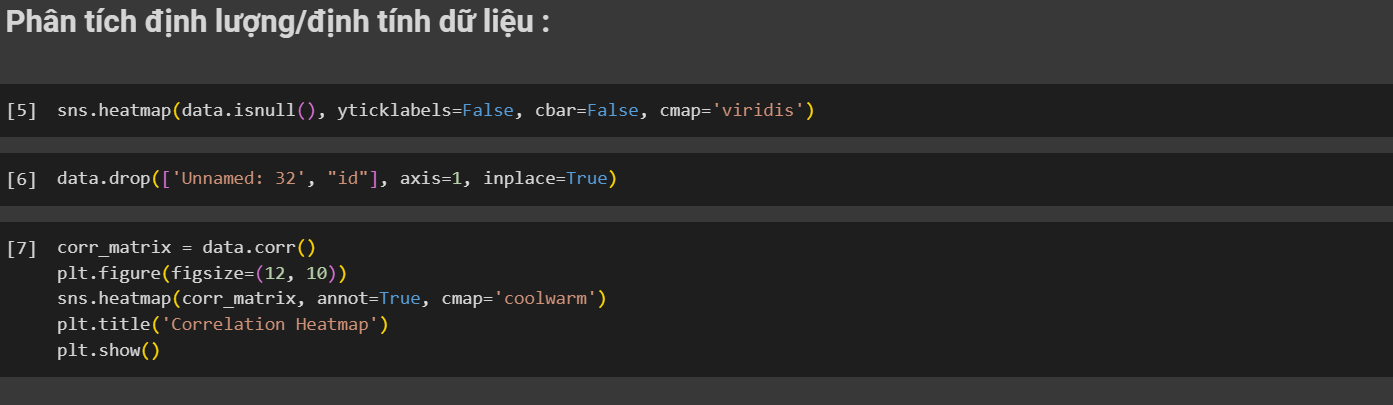
- So sánh các thuật toán và đưa ra kết luận cho dữ liệu

### 2.1. Mô tả dữ liệu

**

Sau khi đọc dữ liệu, ta đã sử dụng các hàm head(), info() và describe() để có cái nhìn tổng quan về dữ liệu. Hàm head() hiển thị một số dòng đầu tiên của dữ liệu, giúp ta nhanh chóng xem qua cấu trúc và giá trị ban đầu của dữ liệu. Hàm info() cung cấp thông tin về kiểu dữ liệu và số lượng mẫu dữ liệu. Hàm describe() cho ta một báo cáo thống kê về các đặc trưng, bao gồm số lượng, giá trị trung bình, độ lệch chuẩn, giá trị tối thiểu và tối đa của từng đặc trưng.

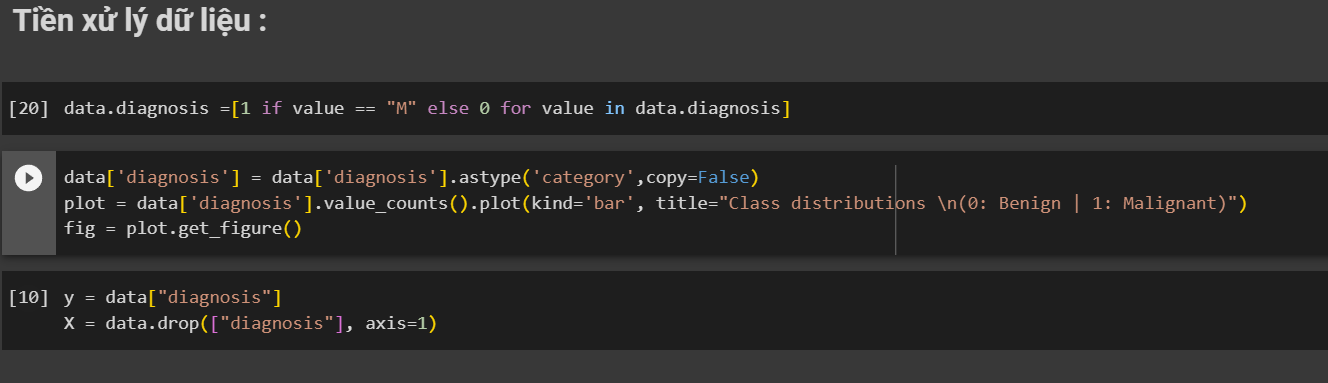
### 2.2. Phân tích định lượng/định tính dữ liệu, mối quan hệ giữa các đặc tính dữ liệu



Phân tích định tính dữ liệu liên quan đến việc xác định các thuộc tính rời rạc hoặc hạng mục của các đặc trưng. Trong trường hợp này, cột chẩn đoán (diagnosis) được xem như một đặc tính định tính, với hai giá trị "M" và "B" đại diện cho ung thư ác tính và ung thư lành tính, tương ứng. Ta đã sử dụng phương pháp value\_counts() để đếm số lượng mẫu thuộc mỗi giá trị và tạo biểu đồ cột để trực quan hóa phân phối của các giá trị chẩn đoán.

Mối quan hệ giữa các đặc tính dữ liệu có thể được xác định thông qua ma trận tương quan. Bằng cách tính toán ma trận tương quan và sử dụng biểu đồ heatmap, ta có thể hiển thị mối quan hệ tuyến tính giữa các đặc trưng. Trong code của bạn, ta đã sử dụng corr\_matrix = data.corr() để tính toán ma trận tương quan và sns.heatmap(corr\_matrix, annot=True, cmap='coolwarm') để tạo biểu đồ heatmap. Biểu đồ này cho phép ta nhìn thấy mức độ tương quan giữa các đặc trưng, trong đó màu sắc và giá trị số được sử dụng để biểu thị mức độ tương quan.

### 2.3. Tiền xử lý dữ liệu



Bước tiền xử lý dữ liệu trong trường hợp này đã được thực hiện trước khi ta đọc dữ liệu từ tệp CSV và sau đó xem thông tin về dữ liệu. Dưới đây là mô tả các bước tiền xử lý đã được thực hiện:

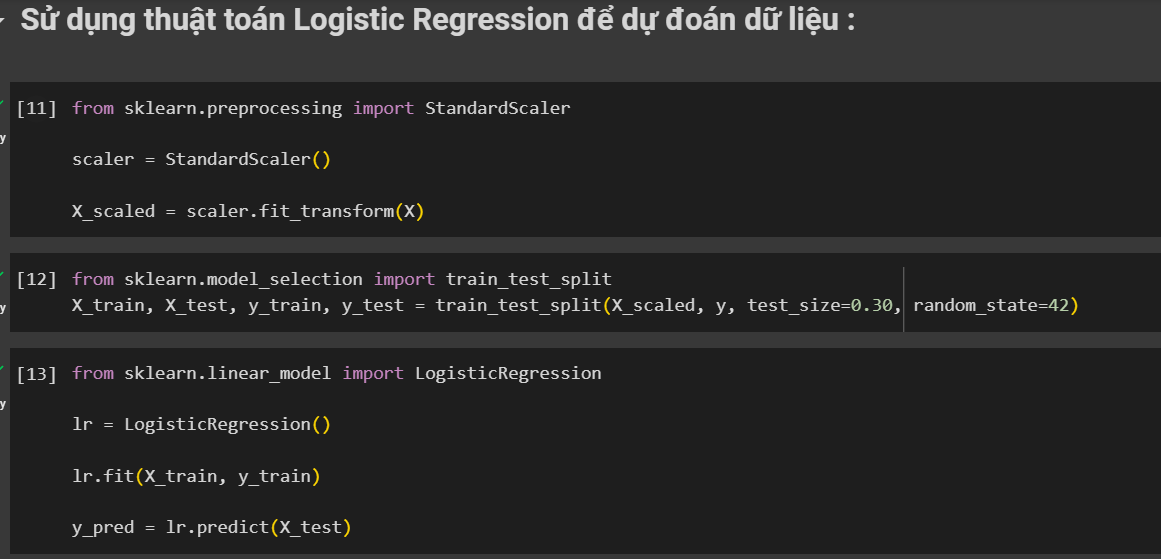
Đọc dữ liệu từ tệp CSV: Sử dụng thư viện pandas, ta đã sử dụng hàm pd.read\_csv() để đọc dữ liệu từ tệp CSV có tên là "data.csv".

Xem một số dòng đầu tiên của dữ liệu: Sử dụng hàm data.head() để hiển thị một số dòng đầu tiên của dữ liệu. Điều này giúp ta có cái nhìn tổng quan về cấu trúc và giá trị ban đầu của dữ liệu.

Xem thông tin về dữ liệu: Sử dụng hàm data.info() để hiển thị thông tin về dữ liệu. Hàm này cung cấp thông tin về kiểu dữ liệu của từng cột và số lượng mẫu dữ liệu. Điều này giúp ta kiểm tra xem có dữ liệu bị thiếu hay không và có cần thực hiện các bước xử lý dữ liệu khác để điều chỉnh kiểu dữ liệu.

Xem mô tả thống kê của dữ liệu: Sử dụng hàm data.describe() để hiển thị mô tả thống kê của dữ liệu. Hàm này tính toán các giá trị thống kê như giá trị trung bình, độ lệch chuẩn, giá trị tối thiểu và tối đa của từng đặc trưng. Điều này giúp ta có cái nhìn tổng quan về phân bố và đặc tính của các đặc trưng trong dữ liệu.

### **2.4. Sử dụng thuật toán Logistic Regression để dự đoán dữ liệu**



**Chuẩn bị dữ liệu:** Trước khi áp dụng thuật toán Logistic Regression, ta cần chuẩn bị dữ liệu cho quá trình huấn luyện và kiểm tra. Trong code của bạn, ta đã chia dữ liệu thành hai phần: dữ liệu huấn luyện (X\_train, y\_train) và dữ liệu kiểm tra (X\_test, y\_test) bằng cách sử dụng hàm train\_test\_split từ thư viện sklearn. Dữ liệu đã được chia theo tỷ lệ 70% cho huấn luyện và 30% cho kiểm tra.

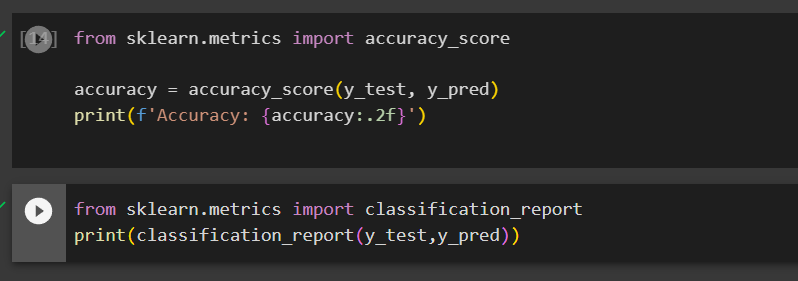
Tiền xử lý dữ liệu: Trước khi áp dụng thuật toán Logistic Regression, ta đã tiền xử lý dữ liệu bằng cách sử dụng StandardScaler từ thư viện sklearn để chuẩn hóa dữ liệu thành các giá trị có mean bằng 0 và độ lệch chuẩn bằng 1. Việc chuẩn hóa dữ liệu giúp cân bằng và tối ưu hiệu suất của mô hình.

Xây dựng mô hình Logistic Regression: Ta đã sử dụng lớp LogisticRegression từ thư viện sklearn để xây dựng mô hình Logistic Regression. Mô hình được khởi tạo bằng cách sử dụng lr = LogisticRegression().

Huấn luyện mô hình: Sau khi xây dựng mô hình, ta đã sử dụng phương thức fit trên dữ liệu huấn luyện để huấn luyện mô hình Logistic Regression. Điều này có nghĩa là mô hình sẽ học từ dữ liệu huấn luyện để tìm ra các tham số tốt nhất cho việc dự đoán chẩn đoán.

**Dự đoán dữ liệu:** Sau khi mô hình được huấn luyện, ta đã sử dụng phương thức predict để dự đoán nhãn cho dữ liệu kiểm tra (X\_test). Dự đoán này dựa trên mô hình đã học được từ dữ liệu huấn luyện

2.5. Đánh giá mô hình



Sử dụng các phương thức accuracy\_score và classification\_report từ thư viện sklearn để đánh giá hiệu suất của mô hình Logistic Regression trên dữ liệu kiểm tra. Accuracy là tỷ lệ phân loại chính xác của mô hình, và classification report cung cấp thông tin về precision, recall và f1-score cho từng lớp chẩn đoán (benign và malignant).

# **KẾT LUẬN**

## 1. Kết quả đạt được

Hoàn thành và thực thi được các chức năng và tác vụ cơ bản: Các chức năng cơ bản như đọc dữ liệu từ tệp CSV, mô tả dữ liệu, phân tích định lượng/định tính dữ liệu, tiền xử lý dữ liệu, sử dụng thuật toán Logistic Regression để dự đoán dữ liệu đã được thực hiện thành công.

## 2. Hướng nghiên cứu

Sử dụng các thuật toán học máy khác: Ngoài Logistic Regression, bạn có thể thử nghiệm với các thuật toán khác như Support Vector Machines (SVM), Random Forest, Gradient Boosting, Neural Networks và các phương pháp Ensemble để so sánh hiệu suất và tìm ra thuật toán tốt nhất cho bài toán dự đoán ung thư vú.

**TÀI LIỆU THAM KHẢO**

[1] “ https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data ”

[2] “ <https://www.youtube.com/watch?v=My4JgIeFdWk> ”

x

x

x

**PHỤ LỤC**