ĐẠI HỌC QUỐC GIA HÀ NỘI
TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN
KHOA TOÁN - CƠ - TIN HỌC

BÁO CÁO MÔN HỌC

ĐỀ TÀI NGHIÊN CỨU

ỨNG DỤNG AI CHẨN ĐOÁN BỆNH UNG THƯ VÚ

GV hướng dẫn: NGUYỄN HẢI VINH

Nhóm sinh viên thực hiện:

Họ và tên	*MSSV*	*Mã lớp*
Cao Duy Ninh	22001627	K67A4
Trương Đan Vi	22001655	K67A4
Nguyễn Quang Việt	22001659	K67A4

Hà Nội, 2024

Mục lục

LỜI CẢM ƠN

Trong quá trình thực hiện đề tài "Ứng dụng trí tuệ nhân tạo phát hiện tiền ung thư và ung thư" thuộc môn Nhập môn Trí tuệ nhân tạo, nhóm chúng em đã nhận được sự hỗ trợ và giúp đỡ quý báu từ nhiều phía. Những kiến thức lý thuyết trên lớp, cùng với sự hướng dẫn tận tình và những giờ học thực hành, đã trở thành nguồn động lực to lớn giúp chúng em hoàn thành đề tài này. Chúng em xin gửi lời cảm ơn chân thành đến: Trước hết, chúng em muốn bày tỏ lòng biết ơn sâu sắc đến thầy Nguyễn Hải Vinh, giảng viên tại Trường Đại học Khoa học Tự Nhiên, người đã trực tiếp hướng dẫn và luôn sẵn sàng hỗ trợ nhóm chúng em. Thầy đã không ngần ngại dành thời gian và công sức để hướng dẫn chi tiết từng bước, từ việc lựa chọn phương pháp nghiên cứu cho đến việc phân tích kết quả. Những buổi gặp gỡ và trao đổi với thầy không chỉ giúp chúng em hiểu rõ hơn về các khái niệm lý thuyết mà còn truyền cảm hứng cho chúng em trong quá trình thực hiện đề tài. Thầy đã tạo điều kiện thuận lợi nhất để nhóm có thể hoàn thành đề tài một cách tốt nhất. Chúng em cũng xin gửi lời cảm ơn đến các thầy cô giáo khác trong khoa, những người đã truyền đạt cho chúng em những kiến thức quý giá và tạo ra một môi trường học tập tích cực. Sự nhiệt huyết và tâm huyết của các thầy cô đã góp phần không nhỏ vào sự hình thành và phát triển của chúng em. Cuối cùng, chúng em muốn cảm ơn những người bạn cùng lớp, những người đã cùng nhau chia sẻ tài liệu, kiến thức và cùng nhau học tập. Sự hỗ trợ và động viên từ các bạn đã giúp chúng em vượt qua những khó khăn trong quá trình thực hiện đề tài này. Những buổi thảo luận sôi nổi và những giờ phút làm việc nhóm đầy hứng khởi đã tạo ra không chỉ những kết quả học tập tốt mà còn những kỷ niệm đáng nhớ trong quãng đời sinh viên của chúng em. Chúng em xin chân thành cảm ơn tất cả mọi người đã đồng hành cùng chúng em trong hành trình này. Sự giúp đỡ và ủng hộ của mọi người chính là động lực lớn lao để chúng em tiếp tục phấn đấu và phát triển trong tương lai

Phân công nhiệm vụ

Dự án được thực hiện với sự tham gia của 4 thành viên, mỗi người đảm nhận các nhiệm vụ cụ thể như sau:

Cao Duy Ninh:

- Tham gia làm báo cáo chương I , II .
- Tìm hiểu về thuật toán Random Forest.
- Tham gia code và xây dựng mô hình.

Trương Đan Vi

- Làm báo cáo chương II , III, IV.
- Tìm hiểu về thuật toán Logistic Regresion.
- Tham gia code và xây dựng mô hình.

Nguyễn Quang Việt

- Làm báo cáo chương II , III.
- Tìm hiểu về thuật toán Navie Bayes.
- Tham gia code và xây dựng mô hình.

Mỗi thành viên đều hỗ trợ lẫn nhau trong các giai đoạn quan trọng của dự án nhằm đảm bảo hoàn thành công việc đúng tiến độ và đạt chất lượng tốt nhất.

1 Mở đầu

1.1 Lý do chọn đề tài

- Thống kê của Tổ chức Ung thư toàn cầu (GLOBOCAN) cho thấy, mỗi năm nước ta ghi nhận khoảng 21.555 ca mắc mới ung thư vú, chiếm 25% tổng số các bệnh ung thư ở nữ giới. Ở Việt Nam, ung thư vú cũng là loại ung thư đứng hàng đầu, chiếm tới 25,8% các trường hợp ung thư ở nữ giới với 21.555 người mới được phát hiện bệnh và 9.345 bệnh nhân tử vong.

- Hiện nay, ở Việt Nam, đội ngũ bác sĩ lành nghề chưa đáp ứng đủ nhu cầu xã hội, nhất là ở tuyến Huyện và Tỉnh. Việc thiếu đội ngũ Y bác sĩ chất lượng cao dẫn tới việc chẩn đoán các bệnh về ung thư còn chưa có độ chính xác cao (ở các tuyến huyện và tuyến dưới hơn). Việc chẩn đoán sai bệnh có thể gây ra nhiều hậu quả đáng tiếc

- Việc chẩn đoán đúng nguy ung thư lành tính và ác tính có ý nghĩa vô cùng quan trọng trong việc hỗ trợ các bác sĩ, chuyên gia y tế và bệnh nhân trong việc phát hiện và can thiệp kịp thời, giảm thiểu nguy cơ tử vong cũng như giảm bớt hậu quả lâu dài do bệnh ung thư gây ra.

1.2 Mục tiêu dự án

Nhận thấy sự cấp thiết đó, nhóm em đã tiến hành xây dựng mô hình chẩn đoán ung thư vú(lành tính và ác tính) sử dụng một số phương pháp học máy cơ bản. Mô hình này nhằm hỗ trợ phát hiện sớm các nguy cơ ung thư vú, giúp các bác sĩ và bệnh nhân có thể can thiệp kịp thời, giảm thiểu rủi ro tử vong cũng như hạn chế các di chứng nghiêm trọng có thể xảy ra. Với các phương pháp học máy: Naive Bayes, Random Forest, và Logistic Regression, mô hình được kỳ vọng sẽ cung cấp công cụ dự đoán chính xác, dễ triển khai và có thể ứng dụng thực tiễn trong môi trường y tế.

2 Chương I: Cơ sở lý thuyết

Trong lĩnh vực y tế, việc thiết lập các mô hình thông qua phương pháp học máy có thể hỗ trợ các bác sĩ cải thiện tỷ lệ chẩn đoán ung thư (lành tính và ác tính), nhằm đạt được mục đích phát hiện sớm và điều trị sớm. Phương pháp học máy đã mang lại kết quả tốt trong chẩn đoán ung thư. Mặc dù hiện có nhiều phương pháp học máy được áp dụng để phân loại tế bào ung thư vú nhưng không có thuật toán đơn lẻ nào có thể áp dụng cho tất cả các vấn đề. Mỗi loại thuật toán học máy đều có lĩnh vực chuyên môn riêng, do đó việc lựa chọn thuật toán sẽ khác nhau trong các tình huống khác nhau.

2.1 Vấn Đề Cần Giải Quyết

Ung thư vú là một trong những bệnh ung thư phổ biến và là nguyên nhân gây tử vong hàng đầu ở phụ nữ. Việc phát hiện sớm và dự đoán nguy cơ ung thư vú có thể cứu sống hàng triệu người. Các vấn đề chính trong việc dự đoán ung thư vú bao gồm:

- Xử lý và phân tích các đặc trưng của dữ liệu bệnh nhân, như sinh thiết hoặc các chỉ số xét nghiệm.
- Xây dựng mô hình dự đoán chính xác để phân loại các trường hợp ung thư vú thành các nhóm ác tính và lành tính
- Đảm bảo tính chính xác và độ tin cậy của mô hình để hỗ trợ bác sĩ đưa ra quyết định điều trị.

2.2 Tổng Quan Về Dự Đoán Ung Thư Vú

Dự đoán ung thư vú có thể được thực hiện qua nhiều phương pháp và công cụ khác nhau. Các mô hình máy học, bao gồm các thuật toán học có giám sát như Navie Bayes, Random Forest, và Logistic Regresion, đã được sử dụng để dự đoán sự xuất hiện và giai đoạn của ung thư vú. Dữ liệu đầu vào của mô hình dự đoán ung thư vú bao gồm các đặc trưng (features) được tính toán từ các tế bào nhân vú, cùng với nhãn lớp (class) thể hiện kết quả chẩn đoán là **benign** (khối u lành tính) hoặc **malignant** (khối u ác tính). Dưới đây là mô tả chi tiết về các thuộc tính trong bộ dữ liệu:

ID number

- **Mô tả**: Số nhận dạng duy nhất của mỗi bệnh nhân hoặc mẫu.

Diagnosis (M, B)

- **Mô tả**: Nhãn phân loại kết quả chẩn đoán.
  - **M**: Malignant (Ác tính)
  - **B**: Benign (Lành tính)

Các đặc trưng hình học và thống kê của tế bào (tính toán từ các tế bào nhân vú)

Các đặc trưng này được tính toán từ các hình ảnh chụp tế bào và bao gồm các chỉ số mô tả hình học, kết cấu và độ mịn của các khối u trong vú. Mỗi đặc trưng có ba giá trị được tính: giá trị trung bình, độ lệch chuẩn và giá trị "tồi tệ nhất" (largest).
*Các đặc trưng trung bình

- **Radius**: Giá trị trung bình của các khoảng cách từ tâm đến các điểm trên đường viền của tế bào.
- **Texture**: Độ lệch chuẩn của các giá trị độ xám trong hình ảnh.
- **Perimeter**: Chu vi của khối u.
- **Area**: Diện tích của khối u.
- **Smoothness**: Sự thay đổi trong độ dài của bán kính tại các điểm khác nhau trên tế bào.
- **Compactness**: Tính chặt chẽ của khối u, tính toán bằng công thức $\left(\frac{\text{Perimeter}^2}{\text{Area}}\right) - 1.0$.
- **Concavity**: Mức độ nghiêm trọng của các phần lõm của khối u.
- **Concave Points**: Số lượng các phần lõm của đường viền khối u.
- **Symmetry**: Độ đối xứng của khối u.
- **Fractal Dimension**: Đo lường độ phức tạp của hình dạng khối u, thường được sử dụng để mô phỏng dạng của các "đường bờ biển".

Các đặc trưng độ lệch chuẩn

Các giá trị này mô tả sự biến thiên của các đặc trưng trong các hình ảnh.

- Ví dụ: **Radius SE** (Độ lệch chuẩn của bán kính).

Các đặc trưng "tồi tệ nhất"

Các giá trị tồi tệ nhất được tính là trung bình của ba giá trị lớn nhất của mỗi đặc trưng.

- Ví dụ: **Worst Radius** (Bán kính lớn nhất trong ba giá trị lớn nhất).

Class Distribution

- **357 benign** (lành tính)
- **212 malignant** (ác tính)

Missing Values

- **Không có giá trị thiếu**: Dữ liệu không có giá trị thiếu cho bất kỳ thuộc tính nào.

Tổng quan về đặc trưng

Bộ dữ liệu có tổng cộng **30 đặc trưng**, bao gồm các đặc trưng tính trung bình, độ lệch chuẩn và giá trị tồi tệ nhất của các yếu tố như bán kính, chu vi, diện tích, tính mịn, độ chặt chẽ, độ lõm, v.v. Nhãn lớp sẽ giúp phân loại mẫu là **benign** hoặc **malignant**, đây là mục tiêu cuối cùng của mô hình học máy trong dự đoán ung thư vú.

2.3 Tiêu chí đánh giá các đặc trưng

2.3.1 Điểm Lõm (Concave Points)

U ác tính:

- Giá trị trung bình cao hơn.
- Độ biến thiên lớn hơn.
- Số lượng điểm lõm nhiều hơn.

2.3.2 Chu Vi (Perimeter)

Đặc điểm ở u ác tính

- Chu vi lớn hơn đáng kể.
- Tăng dần ở các nhóm đặc trưng (mean, SE, worst).
- Biểu thị sự phát triển không kiểm soát của khối u.

2.3.3 Diện Tích (Area)

Tiêu chí phân biệt

- U ác tính có diện tích lớn hơn.
- Tăng dần ở các nhóm đặc trưng.
- Phản ánh sự phát triển khối u.

2.3.4 Bán Kính (Radius)

Đặc điểm phân loại

- U ác tính có bán kính lớn hơn.
- Tăng dần từ nhóm trung bình đến nhóm giá trị cực đại.
- Cho thấy sự lan rộng của khối u.

2.3.5 Độ Chặt (Compactness)

Tiêu chí quan trọng

- U ác tính có độ chặt thấp hơn.
- Biểu thị sự không đều đặn của khối u.
- Phản ánh tính xâm lấn.

2.4 Nguyên tắc phân loại tổng thể

2.4.1 Nguyên tắc cơ bản

- Không dựa vào một đặc trưng duy nhất.
- Xem xét tổng thể các đặc trưng.
- Đánh giá mối tương quan giữa các đặc trưng.

2.5 Các Phương Pháp Dự Đoán Ung Thư Vú

2.5.1 Phương Pháp Học Máy (Machine Learning)

Học máy là một trong những công cụ chính trong dự đoán ung thư vú. Các thuật toán học máy có thể học và phân loại các mẫu dữ liệu để xác định sự hiện diện của ung thư. Các phương pháp học máy phổ biến bao gồm:

- Navie Bayes: là một phương pháp phân loại dựa trên định lý Bayes với giả định độc lập giữa các đặc trưng. Phương pháp này đặc biệt hữu ích khi dữ liệu có nhiều đặc trưng và có thể được mô hình hóa tốt với phân phối xác suất.

- Logistic Regression: là một phương pháp phân loại tuyến tính sử dụng hàm logistic (hay còn gọi là hàm sigmoid) để dự đoán xác suất của các lớp phân loại. Mặc dù tên gọi là hồi quy, logistic regression thực sự là một phương pháp phân loại.

- Random Forest: là một phương pháp học máy ensemble, trong đó sử dụng nhiều cây quyết định (decision trees) để đưa ra dự đoán. Mỗi cây quyết định trong rừng được huấn luyện trên một mẫu ngẫu nhiên của dữ liệu và các đặc trưng ngẫu nhiên.

2.5.2 Chuẩn hóa dữ liệu

Trước khi áp dụng các thuật toán học máy, dữ liệu cần được chuẩn hóa để đảm bảo tính đồng nhất và dễ dàng xử lý. Một trong các phương pháp chuẩn hóa phổ biến là StandardScaler, giúp chuẩn hóa các đặc trưng về cùng một phạm vi (giữa 0 và 1), điều này rất quan trọng trong các mô hình học máy.

2.5.3 Đánh Giá Mô Hình

Để đảm bảo mô hình dự đoán ung thư vú chính xác, cần phải đánh giá hiệu suất của mô hình qua các chỉ số như độ chính xác (accuracy), độ nhạy (sensitivity), độ đặc hiệu (specificity).

2.6 Tầm Quan Trọng Của Dự Đoán Sớm

Phát hiện ung thư vú ở giai đoạn sớm có thể giúp tăng khả năng điều trị thành công. Các phương pháp dự đoán ung thư vú giúp giảm thiểu số lượng ca bệnh tiến triển và giảm tỷ lệ tử vong do ung thư vú, mang lại lợi ích lớn cho sức khỏe cộng đồng.

3 Chương II: Thiết kế hệ thống, thuật toán và dữ liệu

3.1 Tiền xử lý dữ liệu

3.1.1 Tổng quan dữ liệu

Bộ dữ liệu Breast Cancer Wisconsin (Diagnostic Dataset) là một bộ dữ liệu phổ biến được sử dụng trong các bài toán phân loại ung thư. Nó được công bố lần đầu tiên bởi Dr. William H. Wolberg tại University of Wisconsin, và hiện tại thường xuyên được sử dụng để nghiên cứu và thực hành trong lĩnh vực học máy (machine learning) và phân tích dữ liệu y tế.
Các đặc điểm được tính toán từ hình ảnh số hóa của một mẫu chọc hút bằng kim nhỏ (FNA) của khối u vú. Chúng mô tả các đặc điểm của nhân tế bào có trong hình ảnh.
Bộ dữ liệu chứa 569 mẫu thử nghiệm, bao gồm 357 mẫu lành tính và 212 mẫu ung thư vú ác tính.
Các thuộc tính trong bộ dữ liệu:

- 1. Số ID
- 2. Chẩn đoán (M = ác tính, B = lành tính)
- 3-32) Mười đặc điểm có giá trị thực được tính toán cho mỗi nhân tế bào:
  - a) bán kính (trung bình khoảng cách từ tâm đến các điểm trên chu vi)
  - b) kết cấu (độ lệch chuẩn của các giá trị thang độ xám)
  - c) chu vi
  - d) diện tích
  - e) độ mịn (biến thiên cục bộ trong độ dài bán kính)
  - f) độ chặt ($chuvi^2/dintch - 1,0$)
  - g) độ lõm (mức độ nghiêm trọng của các phần lõm của đường đồng mức)
  - h) các điểm lõm (số phần lõm của đường đồng mức)
  - i) tính đối xứng
  - j) chiều fractal ("xấp xỉ đường bờ biển" - 1)
  - Giá trị trung bình, lỗi chuẩn và "tệ nhất" hoặc lớn nhất (giá trị trung bình của ba giá trị lớn nhất) của các đặc điểm này được tính toán cho mỗi hình ảnh, tạo ra 30 đặc điểm. Ví dụ, trường 3 là Bán kính trung bình, trường 13 là Bán kính SE, trường 23 là Bán kính tệ nhất.

Tất cả các thuộc tính được mã hóa lại bằng bốn chữ số có nghĩa. Giá trị thuộc tính bị thiếu: không có
Nhãn phân loại đại diện cho loại ung thư vú. Do đó, tập dữ liệu mẫu chứa tổng cộng 30 đặc điểm và một đặc điểm nhãn mẫu (ác tính và lành tính).

3.1.2 Chuẩn hóa dữ liệu

Dữ liệu chẩn đoán ung thư vú chứa nhiều đặc trưng (như radius-mean, texture-mean, area-mean,...) với đơn vị đo lường và phạm vi giá trị khác nhau. Điều này có thể gây ảnh hưởng xấu đến hiệu suất của các thuật toán học máy, đặc biệt là các thuật toán dựa trên khoảng cách hoặc độ lớn giá trị (ví dụ: KNN, Logistic Regression). Trong báo cáo này, theo đặc điểm của bộ dữ liệu ung thư vú WDBC, tiêu chuẩn hóa điểm Z đã được chọn để xử lý dữ liệu. Dữ liệu được xử lý theo tiêu chuẩn hóa điểm Z tuân theo phân phối chuẩn chuẩn, nghĩa là giá trị trung bình là 0 và phương sai là 1.
Công thức chuẩn hóa điểm Z như sau: $x^* = \frac{x - \text{mean}}{\text{std}}$ trong đó giá trị trung bình là giá trị trung bình của dữ liệu đặc tính mẫu và std là độ lệch chuẩn của dữ liệu đặc tính mẫu.
Việc chuẩn hóa dữ liệu bằng `StandardScaler` mang lại các lợi ích sau:

- Cân bằng giá trị của các đặc trưng: Sau khi chuẩn hóa, các đặc trưng có trung bình 0 và độ lệch chuẩn 1, giúp mô hình hoạt động ổn định hơn.
- Cải thiện hiệu suất của các thuật toán học máy: Các thuật toán dựa trên khoảng cách (ví dụ: KNN, SVM) hoặc tối ưu hóa gradient (như Logistic Regression) trở nên hiệu quả hơn khi dữ liệu đã được chuẩn hóa.
- Đảm bảo vai trò công bằng giữa các đặc trưng: Đặc trưng có phạm vi giá trị lớn sẽ không áp đảo đặc trưng có giá trị nhỏ hơn.
- Tăng tốc độ hội tụ: Trong các mô hình dựa trên tối ưu hóa gradient, việc chuẩn hóa giúp gradient được tính toán chính xác hơn, từ đó tăng tốc độ hội tụ của thuật toán.

3.2 Mô hình dự đoán dựa trên thuật toán Naïve Bayes

3.2.1 Ý tưởng thuật toán

Naïve Bayes là một thuật toán học máy phân loại dựa trên Định lý Bayes.

Định lý Bayes được biểu diễn dưới dạng công thức như sau:

$$P(C|X) = \frac{P(X|C) \cdot P(C)}{P(X)}$$

Trong đó:

- $P(C|X)$: Xác suất hậu nghiệm (*posterior probability*) – xác suất để X thuộc lớp C.
- $P(X|C)$: Xác suất có điều kiện (*likelihood*) – xác suất để X xảy ra khi biết lớp C.
- $P(C)$: Xác suất tiên nghiệm (*prior probability*) – xác suất xảy ra của lớp C.
- $P(X)$: Xác suất xảy ra của X – đóng vai trò là hằng số trong bài toán phân loại.

3.2.2 Giả định "Naïve"

Ta gọi định lí này là "Naïve" (ngây thơ) vì nó giả định rằng các yếu tố dự báo trong mô hình Naïve Bayes là độc lập có điều kiện hoặc không liên quan đến bất kỳ đặc trưng nào khác trong mô hình. Nó cũng giả định rằng tất cả các đặc trưng đều đóng góp như nhau vào kết quả. Mặc dù các giả định này thường bị vi phạm trong các tình huống thực tế (ví dụ: một từ tiếp theo trong email phụ thuộc vào từ trước đó), nhưng nó đơn giản hóa vấn đề phân loại bằng cách làm cho nó dễ tính toán hơn. Nghĩa là, bây giờ chỉ cần một xác suất duy nhất cho mỗi biến, điều này giúp cho việc tính toán mô hình dễ dàng hơn. Bất chấp giả định độc lập không thực tế này, thuật toán phân loại hoạt động tốt, đặc biệt là với kích thước mẫu nhỏ.
Nhờ giả định này, xác suất $P(X|C)$ có thể được tính như sau:

$$P(X|C) = P(X_1|C) \cdot P(X_2|C) \cdot \ldots \cdot P(X_n|C)$$

3.2.3 Thuật toán Naïve Bayes

- Xét các bài toán phân lớp với C class khác nhau. Thay vì tìm ra chính xác label của mỗi điểm dữ liệu $x \in R^d$, ta có thể đi tìm xác suất để đầu ra đó rơi vào mỗi class: $p(y = c|x)$, hoặc viết gọn thành $p(c|x)$.

- Cách thuật toán hoạt động :

  1. Tính toán xác suất: Sử dụng công thức định lý Bayes, xác suất của một nhãn C được tính như sau:

$$P(c|x) = \frac{P(c).P(x|c)}{P(x)}$$

2. Ta cần tính :

$$c = \arg \max_{c \in \{1,\dots,C\}} p(c|x) \tag{1}$$

Biểu thức trong dấu argmax ở (1) nhìn chung khó có cách tính trực tiếp. Thay vào đó, quy tắc Bayes thường được sử dụng:

$$c = \arg\max_c p(c|x) = \arg\max_c \frac{p(x|c)p(c)}{p(x)} \propto \arg\max_c p(x|c)p(c)$$

3. Ước lượng xác suất có điều kiện $P(x|c)$: Với giả định độc lập của Naive Bayes:

$$p(x|c) = p(x_1, x_2, \dots, x_d|c) = \prod_{i=1}^{d} p(x_i|c)$$

4. Lựa chọn nhãn: Nhãn được dự đoán là nhãn C có xác suất $P(c|x)$ lớn nhất.

- Ở bước huấn luyện, các phân phối $p(c)$ và $p(x_i|c), i = 1, \dots, d$ sẽ được xác định dựa vào dữ liệu huấn luyện. Việc xác định các giá trị này có thể dựa vào MLE hoặc MAP.

- Việc tính toán $p(x_i|c)$ phụ thuộc vào loại dữ liệu. Có ba loại phân bố xác suất thường được dùng: *Gaussian Naive Bayes*, *Multinomial Naive Bayes*, và *Bernoulli Naive Bayes*.

3.2.4 Ước lượng tham số MLE

Trong bài toán phát hiện ung thư lành tính và ác tính , ta sử dụng mô hình *Gaussian Naive Bayes* . Nên MLE (ước lượng tham số cực đại) được tính như sau :

1. hàm phân phối chuẩn $N(0, 1)$

$$f(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

2. Likelihood Function :

$$L(\mu, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right)$$

3. Log-Likelihood Function :

$$\ell(\mu, \sigma^2) = \log L(\mu, \sigma^2) = \sum_{i=1}^{n} \log \left(\frac{1}{\sqrt{2\pi\sigma^2}} \exp \left(-\frac{(x_i - \mu)^2}{2\sigma^2} \right) \right)$$

$$\ell(\mu, \sigma^2) = \sum_{i=1}^{n} \left[\log \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right) - \frac{(x_i - \mu)^2}{2\sigma^2} \right]$$

$$\ell(\mu, \sigma^2) = \sum_{i=1}^{n} \left[-\frac{1}{2} \log(2\pi) - \frac{1}{2} \log(\sigma^2) - \frac{(x_i - \mu)^2}{2\sigma^2} \right]$$

$$\ell(\mu, \sigma^2) = -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log(\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2$$

4. Maximizing the Log-Likelihood :
   Đạo hàm của $\ell(\mu, \sigma^2)$ với μ là:

$$\frac{\partial \ell(\mu, \sigma^2)}{\partial \mu} = -\frac{1}{\sigma^2} \sum_{i=1}^{n} (x_i - \mu)$$

   Cho $\frac{\partial \ell(\mu, \sigma^2)}{\partial \mu} = 0$, ta có:

$$\sum_{i=1}^{n} (x_i - \mu) = 0$$

   Từ đó ta có :

$$\mu = \frac{1}{n} \sum_{i=1}^{n} x_i$$

   Đạo hàm của $\ell(\mu, \sigma^2)$ với σ^2 là:

$$\frac{\partial \ell(\mu, \sigma^2)}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2} \sum_{i=1}^{n} (x_i - \mu)^2$$

   Cho $\frac{\partial \ell(\mu, \sigma^2)}{\partial \sigma^2} = 0$, ta có:

$$-\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2} \sum_{i=1}^{n} (x_i - \mu)^2 = 0$$

   Nhân với $2(\sigma^2)^2$:

$$-n\sigma^2 + \sum_{i=1}^{n} (x_i - \mu)^2 = 0$$

Từ đó , ta có σ^2:

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$$

5. Kết luận
   Ước lượng tham số MLE cho các tham số của phân phối chuẩn Gaussion là

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$$

3.2.5 Điểm mạnh và hạn chế

Điểm mạnh :

- Ít phức tạp hơn : So với các bộ phân loại khác, Naïve Bayes được coi là bộ phân loại đơn giản hơn vì các tham số dễ ước tính hơn. Do đó, đây là một trong những thuật toán đầu tiên được học trong các khóa học về khoa học dữ liệu và học máy.
- Có khả năng mở rộng tốt : So với hồi quy logistic, Naïve Bayes được coi là một bộ phân loại nhanh và hiệu quả, khá chính xác khi giả định độc lập có điều kiện được giữ nguyên. Nó cũng có yêu cầu lưu trữ thấp
- Có thể xử lý dữ liệu có nhiều chiều : Các trường hợp sử dụng, chẳng hạn như phân loại tài liệu, có thể có số lượng chiều cao, gây khó khăn cho các bộ phân loại khác trong việc quản lý.

Hạn chế :

- Giả định cốt lõi không thực tế : Mặc dù giả định độc lập có điều kiện nhìn chung có hiệu quả, nhưng giả định này không phải lúc nào cũng đúng, dẫn đến phân loại không chính xác.

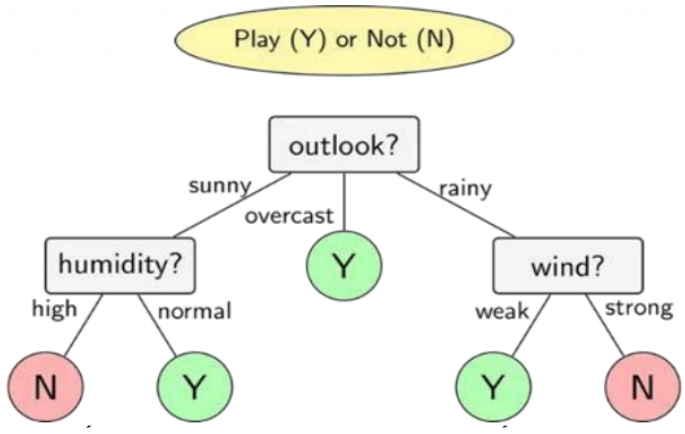
3.3 Mô hình dự đoán dựa trên thuật toán Random Forest

3.3.1 Khái niệm Decision trees

Vì mô hình rừng ngẫu nhiên được tạo thành từ nhiều decision trees, nên sẽ hữu ích khi bắt đầu bằng cách mô tả ngắn gọn thuật toán decision trees. Cây quyết định (Decision Tree) là một cây phân cấp có cấu trúc được dùng để phân lớp các đối tượng dựa vào dãy các luật. Nó cho dữ liệu về các đối tượng gồm các thuộc tính cùng với lớp (classes) của nó, cây quyết định sẽ sinh ra các luật để dự đoán lớp của các dữ liệu chưa biết.

3.3.2 Vai trò của Decision trees

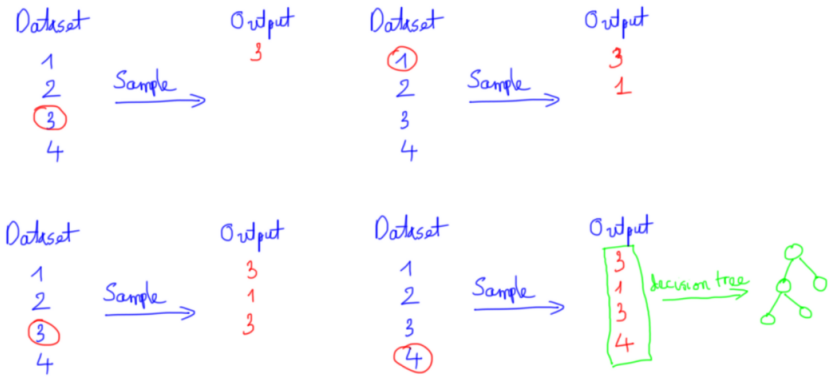
Mô hình cây quyết định là một phương pháp đơn giản có thể được sử dụng để phân loại các đối tượng theo các tính năng của chúng.
Ví dụ: Bạn có thể có một cây quyết định bạn có ra ngoài đi chơi hay không dựa trên các thuộc tính sau: thời tiết, độ ẩm và gió hoặc bạn có thể có một cây quyết định cho bạn biết đối tượng của bạn có phải là một quả táo hay không dựa trên các thuộc tính sau: màu sắc, kích thước và trọng lượng.
Một cây quyết định hoạt động bằng cách đi xuống từ nút gốc cho đến khi nó đạt đến nút quyết định. Các nút quyết định có các nhánh đưa chúng ta đến các nút lá hoặc nhiều nút quyết định hơn. Các nút lá là các nút đầu cuối trình bày quyết định cuối cùng giống như tên của chúng cho thấy.



3.3.3 Phương pháp Ensemble

Random Forest là một phần mở rộng của phương pháp bagging nên chúng ta sẽ tìm hiểu thêm về phương pháp Ensemble và bagging.

Các phương pháp Ensemble được tạo thành từ một tập hợp các bộ phân loại(ví dụ: decision trees)và các dự đoán của chúng được tổng hợp để xác định kết quả phổ biến nhất. Các phương pháp Ensemble nổi tiếng nhất là bagging, còn được gọi là tổng hợp bootstrap và boosting.Trong phương pháp này, một mẫu dữ liệu ngẫu nhiên trong tập huấn luyện được chọn có thể được chọn nhiều lần. Sau khi một số mẫu dữ liệu được tạo, các mô hình này sẽ được train độc lập và tùy thuộc vào loại chức năng(tức là hồi quy hoặc phân loại) phần lớn các dự đoán đó mang lại ước tính chính xác hơn. Cách tiếp cận này thường được sử dụng để giảm phương sai trong tập dữ liệu nhiễu.

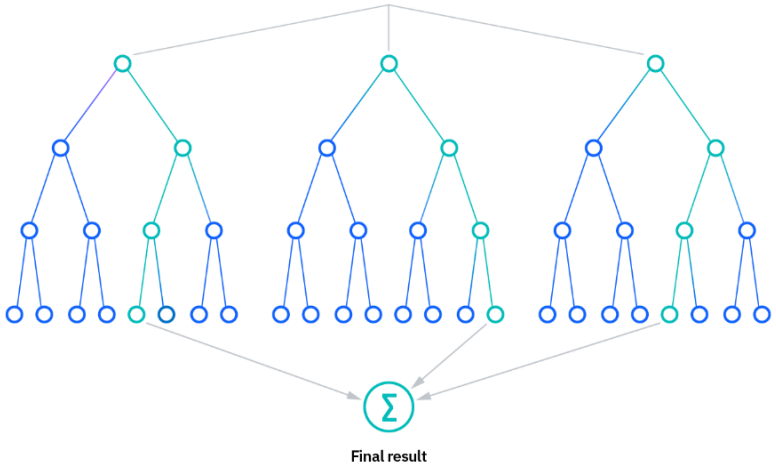


3.3.4 Thuật toán Random Forest

Thuật toán Random Forest sử dụng cả tính chất bagging và tính ngẫu nhiên đặc trưng để tạo ra một rừng decision trees không tương quan. Tính ngẫu nhiên của tính năng, còn được gọi là tính năng bagging hoặc "phương pháp không gian con ngẫu nhiên" , tạo ra một tập hợp con các tính năng ngẫu nhiên, đảm bảo mối tương quan thấp giữa các decision trees. Đây là điểm khác biệt chính giữa decisoin trees và random forest. Trong khi decisoin trees xem xét tất cả các phân chia đặc điểm có thể có thì các random forest chỉ chọn một tập hợp con của các đặc điểm đó.
Nếu chúng ta quay lại câu hỏi "tôi có nên lướt web không?" Ví dụ: những câu hỏi mà tôi có thể hỏi để xác định dự đoán có thể không toàn diện như bộ câu hỏi của người khác. Bằng cách tính đến tất cả những biến đổi tiềm ẩn trong dữ liệu, chúng tôi có thể giảm nguy cơ khớp quá mức, sai lệch và phương sai tổng thể, dẫn đến dự đoán chính xác hơn.

3.3.5 Random Forest hoạt động như thế nào?

Random Forest có ba tham số chính cần được đặt trước khi đào tạo. Chúng bao gồm kích thước nút, số lượng cây và số lượng đối tượng được lấy mẫu. Từ đó, random forest classifier có thể được sử dụng để giải quyết các vấn đề hồi quy hoặc phân loại. Random forest được tạo thành từ một tập hợp các decison trees và mỗi cây trong quần thể bao gồm một mẫu dữ liệu được lấy từ tập huấn luyện có thay thế, được gọi là mẫu bootstrap. Trong mẫu đào tạo đó, một phần ba trong số đó được dùng làm test data, được gọi là out-of-bag (oob). Sau đó, một trường hợp ngẫu nhiên khác được đưa vào thông qua tính năng bagging, tăng thêm tính đa dạng cho tập dữ liệu và giảm mối tương quan giữa các decison trees. Tùy thuộc vào loại vấn đề, việc xác định dự đoán sẽ khác nhau. Đối với hồi quy, các decision trees riêng lẻ sẽ được tính trung bình và đối với phân loại, phiếu bầu đa số sẽ đưa ra kết quả được dự đoán. Cuối cùng, mẫu oob sau đó được sử dụng để xác thực chéo, hoàn thiện dự đoán đó.



3.3.6 Điểm mạnh và Hạn chế

Điểm mạnh :

- Giảm nguy cơ over fitting: Cây quyết định có nguy cơ quá khớp vì chúng có xu hướng khớp chặt tất cả các mẫu trong dữ liệu đào tạo. Tuy nhiên, khi có một số lượng lớn cây quyết định trong một khu rừng ngẫu nhiên, bộ phân loại sẽ không khớp quá mức mô hình vì việc tính trung bình các cây không tương quan làm giảm phương sai tổng thể và lỗi dự đoán.
- Tăng tính linh hoạt: Vì random forest có thể xử lý cả nhiệm vụ hồi quy và phân loại với độ chính xác cao nên đây là phương pháp phổ biến

trong số các nhà khoa học dữ liệu. Feature bagging cũng làm cho trình phân loại random forest trở thành một công cụ hiệu quả để ước tính các giá trị bị thiếu vì nó duy trì độ chính xác khi một phần dữ liệu bị thiếu.

- Dễ dàng phân loại độ quan trọng của đặc trưng: Rừng ngẫu nhiên giúp dễ dàng đánh giá tầm quan trọng của biến hoặc đóng góp của biến vào mô hình. Có một số cách để đánh giá tầm quan trọng của tính năng. Tầm quan trọng của Gini và độ giảm tạp chất trung bình (MDI) thường được sử dụng để đo mức độ chính xác của mô hình giảm khi một biến nhất định bị loại trừ. Tuy nhiên, tầm quan trọng của hoán vị, còn được gọi là độ chính xác giảm trung bình (MDA), là một biện pháp quan trọng khác. MDA xác định mức giảm độ chính xác trung bình bằng cách hoán vị ngẫu nhiên các giá trị tính năng trong các mẫu oob.

Hạn chế

- Quy trình tốn thời gian: Vì thuật toán rừng ngẫu nhiên có thể xử lý các tập dữ liệu lớn nên chúng có thể cung cấp các dự đoán chính xác hơn, nhưng có thể xử lý dữ liệu chậm vì chúng phải tính toán dữ liệu cho từng cây quyết định riêng lẻ.
- Yêu cầu nhiều tài nguyên hơn: Vì rừng ngẫu nhiên xử lý các tập dữ liệu lớn hơn nên chúng sẽ yêu cầu nhiều tài nguyên hơn để lưu trữ dữ liệu đó.
- Phức tạp hơn: Việc dự đoán một cây quyết định đơn lẻ dễ diễn giải hơn so với một rừng cây quyết định.

3.4 Mô hình dự đoán dựa trên thuật toán Logistic Regression

3.4.1 Khái niệm Logistic Regression

Logistic Regression là một mô hình học máy được sử dụng để dự đoán xác suất của một sự kiện, thường là một biến phân loại (binary classification). Mặc dù tên gọi có "regression" (hồi quy) trong đó, nhưng Logistic Regression chủ yếu được dùng cho bài toán phân loại chứ không phải hồi quy liên tục

3.4.2 Thuật toán Logistic Regression

- Phương trình Logistic Regression :

$$P(Y = 1 \mid X) = \frac{1}{1 + \exp\left(w_0 + \sum_i w_i X_i\right)} = \frac{1}{1 + e^{-z}}$$

Trong đó

- $P(Y = 1 \mid X)$: Xác suất sự kiện $Y = 1$ xảy ra với điều kiện đầu vào $X = (X_1, X_2, \ldots, X_n)$.
- w_0: Hệ số chặn (intercept).
- w_i: Trọng số (weight) của biến X_i.
- X_i: Giá trị của biến đầu vào thứ i.
- exp: Hàm mũ (exponential function).
- Nếu $P(Y = 1 \mid X) > 0.5$, ta phân loại $Y = 1$ (sự kiện xảy ra).
- Nếu $P(Y = 1 \mid X) \leq 0.5$, ta phân loại $Y = 0$ (sự kiện không xảy ra).

3.4.3 Xây dựng và tối đa hóa hàm mất mát

Với các mô hình LR, ta có thể giả sử rằng xác suất để một điểm dữ liệu x rơi vào lớp thứ nhất là $f(\mathbf{w}^\top \mathbf{x})$ và rơi vào lớp còn lại là $1 - f(\mathbf{w}^\top \mathbf{x})$:

$$p(y_i = 1 \mid \mathbf{x}_i; \mathbf{w}) = f(\mathbf{w}^\top \mathbf{x}_i) \quad (14.4)$$

$$p(y_i = 0 \mid \mathbf{x}_i; \mathbf{w}) = 1 - f(\mathbf{w}^\top \mathbf{x}_i) \quad (14.5)$$

Trong đó $p(y_i = 1 \mid \mathbf{x}_i; \mathbf{w})$ được hiểu là xác suất xảy ra sự kiện đầu ra $y_i = 1$ khi biết tham số mô hình $\mathbf{w}$ và dữ liệu đầu vào $\mathbf{x}_i$.
Ta cần giải bài toán tối ưu:

$$\mathbf{w} = \arg\max_{\mathbf{w}} p(\mathbf{y} \mid \mathbf{X}; \mathbf{w}) \quad (14.7)$$

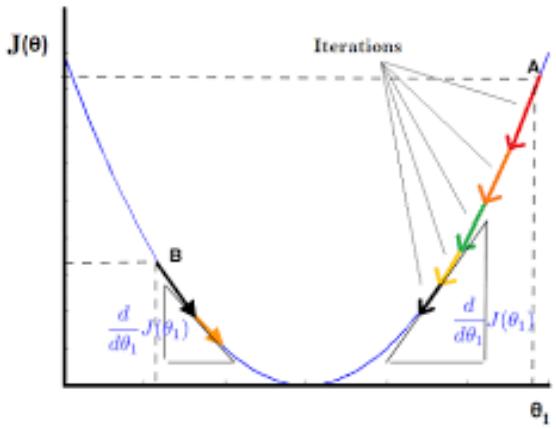
Nguyên lý MLE, tính sự hợp lý (likelihood)

$$L(w, w_0) = P(D) = \prod_{i=1}^{n} P(y_i|x_i) = \prod_{i=1}^{n} \mu_i^{y_i}(1 - \mu_i)^{1-y_i}$$

Lấy negative-loglikelihood (NLL)

$$\ell(w, w_0) = -\log L(w, w_0) = \sum_{i=1}^{n} -y_i \log \mu_i - (1 - y_i) \log(1 - \mu_i)$$

Xuống đồi bằng đạo hàm (gradient descent)



Hình 1: Xuống đồi bằng đạo hàm

Tính đạo hàm, đi ngược hướng đạo hàm, với $\lambda > 0$

$$w \leftarrow w - \lambda \nabla_w \ell(w, w_0)$$
$$w_0 \leftarrow w_0 - \lambda \nabla_{w_0} \ell(w, w_0)$$

Đạo hàm

$$\mu_i = \sigma(w^T x_i + w_0 \cdot 1)$$
$$\sigma'(z) = \sigma(z)(1 - \sigma(z))$$
$$\begin{aligned}
\nabla_w \ell(w, w_0) &= \sum_{i=1}^{n} \frac{\partial \ell}{\partial \mu_i} \frac{\partial \mu_i}{\partial w} \\
&= \sum_{i=1}^{n} \left(-\frac{y_i}{\mu_i} + \frac{1 - y_i}{1 - \mu_i} \right) \mu_i (1 - \mu_i) x_i \\
&= \sum_{i=1}^{n} (-y_i(1 - \mu_i) + (1 - y_i)\mu_i) x_i \\
&= \sum_{i=1}^{n} (\mu_i - y_i) x_i \\
\nabla_{w_0} \ell(w, w_0) &= \sum_{i=1}^{n} (\mu_i - y_i)
\end{aligned}$$

Nếu $y_i = 0$ thì mình muốn $\mu_i = 0$

Nếu $y_i = 1$ thì mình muốn $\mu_i = 1$

3.4.4 Điểm mạnh và hạn chế

Điểm mạnh :

- Độ chính xác tốt cho nhiều tập dữ liệu đơn giản và hoạt động tốt khi dữ liệu có thể phân tách tuyến tính.
- Nó không chỉ cung cấp thước đo độ phù hợp của một yếu tố dự báo (kích thước hệ số) mà còn cung cấp hướng kết nối của yếu tố đó (tích cực hay tiêu cực)

Hạn chế :

- Nếu số lượng sát thương ít hơn số lượng tính năng thì không nên sử dụng Hồi quy logistic, nếu không có thể dẫn đến tình trạng quá trùng khớp
- Nó chỉ có thể được sử dụng để mong đợi các chức năng rời rạc. Do đó, các biến phụ thuộc của Hồi quy Logistic bị buộc phải bỏ đi.

4 Chương III: Thực thi mô hình và đánh giá kết quả

4.1 Thực thi mô hình

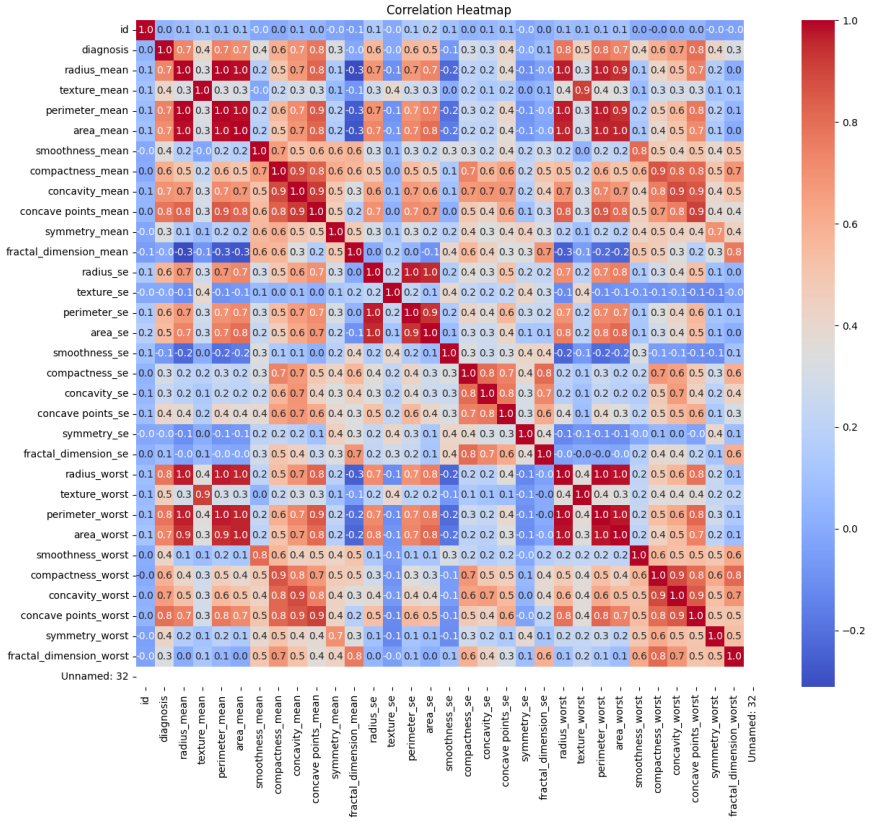
https://colab.research.google.com/drive/1nWnl4tWVkGxb-6rVIaley-sA58jj-h5b

4.1.1 Mô tả đặc điểm của các loại khối u dựa trên các đặc trưng

Trong dữ liệu, các đặc trưng như `radius_mean`, `texture_mean`, `area_mean`, và nhiều thông số khác được sử dụng để phân tích sự khác biệt giữa hai loại u:

- **Benign (B):** U lành tính.
- **Malignant (M):** U ác tính.

Các chỉ số thống kê quan trọng được hiển thị qua heatmap dưới đây, cho thấy mối tương quan giữa các đặc trưng và chẩn đoán:



Biểu đồ tương quan cho thấy các đặc trưng như `radius_mean`, `area_mean`, và `concave_points_mean` có mối quan hệ mạnh mẽ với chẩn đoán (`diagnosis`).

Điều này gợi ý rằng các đặc trưng này có thể đóng vai trò quan trọng trong việc phân loại khối u.

4.1.2 Mô hình Naive Bayes

Sau khi thực thi mô hình, ta có báo cáo phân tích của mô hình Naive Bayes như sau:

```
Naive Bayes Classifier Model Performance:
Accuracy: 0.9240
Classification Report:
              precision    recall  f1-score   support

           0       0.94      0.94      0.94       108
           1       0.89      0.90      0.90        63

    accuracy                           0.92       171
   macro avg       0.92      0.92      0.92       171
weighted avg       0.92      0.92      0.92       171
```

Hình 2: Classification report của mô hình hồi quy Naive Bayes

4.1.3 Mô hình Random Forest

Sau khi thực thi mô hình, ta có báo cáo phân tích của mô hình Random Forest như sau:

```
Random Forest Classifier Model Performance:
Accuracy: 0.9649
Classification Report:
              precision    recall  f1-score   support

           0       0.97      0.97      0.97       108
           1       0.95      0.95      0.95        63

    accuracy                           0.96       171
   macro avg       0.96      0.96      0.96       171
weighted avg       0.96      0.96      0.96       171
```

Hình 3: Classification report của mô hình Random Forest

4.1.4 Mô hình Logistic Regression

Sau khi thực thi mô hình, ta có báo cáo phân tích của mô hình Logistic Regression như sau:

```
Logistic Regression Model Performance:
Accuracy: 0.9766
Classification Report:
              precision    recall  f1-score   support

           0       0.97      0.99      0.98       108
           1       0.98      0.95      0.97        63

    accuracy                           0.98       171
   macro avg       0.98      0.97      0.97       171
weighted avg       0.98      0.98      0.98       171
```

Hình 4: Classification report của mô hình hồi quy logistic

4.2 Đánh giá kết quả

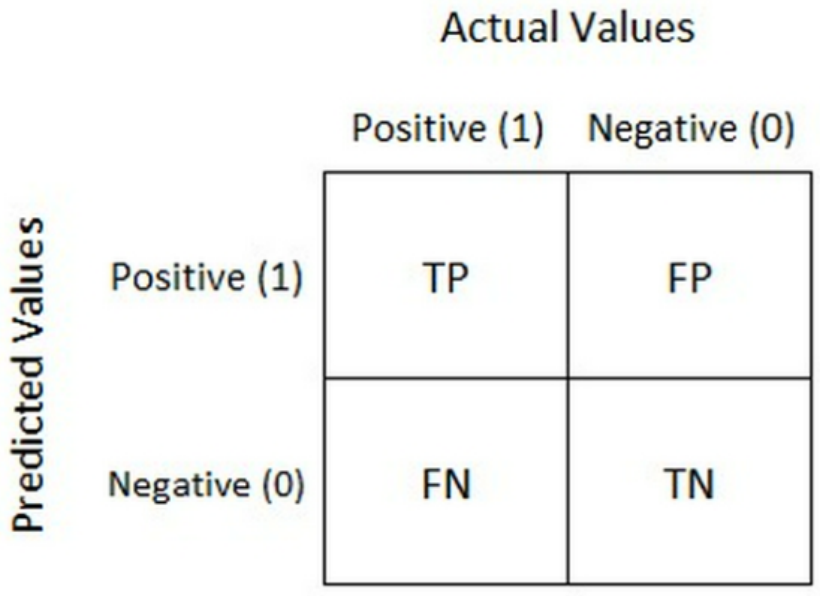
4.2.1 Một số khái niệm cần biết

- Ma trận nhầm lân (Confusion Matrix) là một phương pháp đánh giá kết quả của những bài toán phân loại với việc xem xét cả những chỉ số về độ chính xác và độ bao quát của các dự đoán cho từng lớp. Một confusion matrix gồm 4 chỉ số sau đối với mỗi lớp phân loại: trong đó:

- True Positive (TP): số lượng điểm của lớp positive được phân loại đúng là positive.
- True Negative (TN): số lượng điểm của lớp negative được phân loại đúng là negative.
- False Positive (FP): số lượng điểm của lớp negative bị phân loại nhầm thành positive.
- False Negative (FN): số lượng điểm của lớp positiv bị phân loại nhầm thành negative

- Độ chính xác(accuracy) là tỷ lệ phần trăm tất cả các phân loại chính xác, cho dù là phân loại positive hay negative. Giá trị này được định nghĩa theo toán học là:

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



Hình 5: Caption

- Precision là tỷ lệ phần trăm tất cả các kết quả phân loại dương tính của mô hình thực sự là dương tính. Nó được định nghĩa về mặt toán học là:

$$Precision = \frac{TP}{TP + FP}$$

- Recall là tỷ lệ phần trăm tất cả các kết quả dương tính thực tế được phân loại chính xác là dương tính. Nó được định nghĩa về mặt toán học là:

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

Từ định nghĩa về precision va recall, có thể precision cao đồng nghĩa với việc độ chính xác của các điểm tìm được là cao. Recall cao đồng nghĩa với việc True Positive Rate cao, tức tỉ lệ bỏ sót các điểm thực sự positive là thấp.
- F_1 score, hay F1-score, là harmonic mean của precision và recall (giả sử rằng hai đại lượng này khác không):

$$F_1 score = \frac{2.precision.recall}{precision.recall}$$

4.2.2 So sánh tương quan giữa các mô hình

Ta có bảng so sánh về độ chính xác giữa các mô hình

Mô hình	Accuracy
Naive Bayes	0.924
Random Forest	0.9649
Logistic Regression	0.9766

Bảng 1: Bảng so sánh về độ chính xác (accuracy) giữa các mô hình

Từ bảng trên, ta thấy chỉ số accuracy của mô hình logistic regression là cao nhất, nhưng câu hỏi đặt ra là chỉ số accuracy cao nhất có phải là đủ để quyết định lựa chọn mô hình nào hay chưa?
Trên thực tế, chỉ số accuracy chưa thể chứng minh hoàn toàn được mô hình đó có đủ tốt hay không? Trong bài toán chẩn đoán ung thư đang được thực hiện, giả sử có hai mô hình như sau:

- Model A đạt accuracy là 99%. Tuy nhiên mô hình đã để lọt một trường hợp bị ung thư ác tính nhưng mô hình dự đoán là ung thư lành và người bệnh không điều trị kịp thời dẫn đến hậu quả xấu nhất là tử vong.
- Model B đạt accuracy thấp hơn là 92%. Nhưng sau khi thực thi thì không có trường hợp nào bị ung thư bị bỏ sót. Model chỉ đạt 92% do có vài trường hợp không bị ung thư và bị chỉ định nhầm thành có ung thư ác tính (nhưng sau khi xét nghiệm kỹ lại thì đã khẳng định không bị ung thư và xuất viện rồi).

Từ ví dụ trên, có thể thấy mô hình B dù có độ chính xác thấp hơn nhưng sẽ phù hợp hơn cho bài toán cần thực hiện.
Tương tự đối với bài toán chẩn đoán ung thư đang được thực hiện, yêu cầu đặt ra là giảm thiểu tối đa khả năng dự đoán False Negative (tức là bệnh nhân ung thư ác tính bị dự đoán sai thành ung thư lành tính) nên chỉ số recall của ung thư ác tính cần được lưu ý.
Theo kết quả thực thi của các mô hình đã cho ở trên, ta thấy chỉ số Recall của 3 mô hình Naive Bayes, Random Forest và Logistic Regression lần lượt là: 0,9; 0,95 và 0,95 nên mô hình Random Forest và mô hình Logistic Regression phù hợp hơn so với mô hình Naive Bayes cho bài toán chẩn đoán ung thư đang được thực hiện.

5 Chương IV: Kết luận

Mô hình đã được xây dựng nhằm phân loại ung thư vú thành hai loại: ác tính (malignant) và lành tính (benign) dựa trên các đặc trưng trong bộ dữ liệu Breast Cancer Wisconsin. Bộ dữ liệu này bao gồm các thông tin về đặc tính tế bào học của khối u, như là kích thước hạt nhân, hình dạng và kết cấu.
Bằng cách sử dụng Google Colab, tổng thời gian chạy của mỗi thuật toán là khoảng dưới 1 phút. Độ chính xác (accuracy) của các mô hình được sử dụng trong cuộc điều tra này đều ở mức khá cao trên 92%, cho thấy những mô hình này có độ tin cậy nhất định.
Ngoài độ chính xác (accuracy), chỉ số recall cũng là nhân tố quan trọng trọng việc lựa chọn mô hình phù hợp. Từ kết quả thực thi đã có, ta thấy mô hình Logistic Regression có độ chính xác (accuracy) và chỉ số recall cao nhất nên mô hình này phù hợp nhất với yêu cầu bài toán và bộ dữ liệu cho trước.
Một số đề xuất để cải thiện mô hình dự đoán:

- Thử nghiệm thêm các phương pháp tiên tiến hơn như Deep Learning hoặc tăng cường dữ liệu (Data Augmentation) để cải thiện kết quả.
- Tăng cường việc chọn lọc và xử lý dữ liệu để giảm nhiễu, nâng cao chất lượng dự đoán.
- Nghiên cứu thêm các nguồn dữ liệu thực tế khác để đánh giá khả năng tổng quát hóa của mô hình.

Mô hình phân loại dựa trên bộ dữ liệu Breast Cancer Wisconsin đã cho thấy tiềm năng ứng dụng cao trong việc hỗ trợ chẩn đoán ung thư vú, giảm thiểu sai sót và hỗ trợ các bác sĩ trong việc ra quyết định. Tuy nhiên, cần tiếp tục cải tiến và kiểm định trên các dữ liệu mới để đảm bảo tính chính xác và khả năng áp dụng rộng rãi.

6 TÀI LIỆU THAM KHẢO

References

[1] Dr. Michael Bowles, Machine Learning in Python® : Essential Techniques for Predictive Analysis, May 2015

[2] W. Wolberg, O. Mangasarian, N. Street, and W. Street. "Breast Cancer Wisconsin (Diagnostic)," UCI Machine Learning Repository, 1993. [Online]. Available: https://doi.org/10.24432/C5DW2B.

[3] Chen, Hua, Wang, Nan, Du, Xueping, Mei, Kehui, Zhou, Yuan, Cai, Guangxing, Classification Prediction of Breast Cancer Based on Machine Learning, Computational Intelligence and Neuroscience, 2023, 6530719, 9 pages, 2023. https://doi.org/10.1155/2023/6530719

[4] Monirujjaman Khan, Mohammad, Islam, Somayea, Sarkar, Srobani, Ayaz, Fozayel Ibn, Kabir, Md. Mursalin, Tazin, Tahia, Albraikan, Amani Abdulrahman, Almalki, Faris A., [Retracted] Machine Learning Based Comparative Analysis for Breast Cancer Prediction, Journal of Healthcare Engineering, 2022, 4365855, 15 pages, 2022. https://doi.org/10.1155/2022/4365855