BỘ GIÁO DỤC VÀ ĐÀO TẠO

TRƯỜNG ĐẠI HỌC MỎ ĐỊA CHẤT HÀ NỘI

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

A blue circle with white text and a globe

Description automatically generated

**BÁO CÁO CUỐI KỲ**

**ĐỀ TÀI:** **XÂY DỰNG MÔ HÌNH DỰ ĐOÁN KHẢ NĂNG MẮC BỆNH GAN Ở MỸ**

|  |  |  |
| --- | --- | --- |
| **Nhóm sinh viên thực hiện:** | **Đỗ Phương Mai Anh** | **2121051257** |
| **Nguyễn Hồng Khanh** | **2121050122** |
| **Phạm Anh Tuấn** | **2121050107** |
| **Giảng viên hướng dẫn:** | **TS. Nguyễn Hoàng Long** | |

**Hà Nội, 1/2025**

# MỤC LỤC

[MỤC LỤC 2](#_Toc187266380)

[DANH MỤC CÁC HÌNH VẼ 4](#_Toc187266381)

[MỞ ĐẦU 6](#_Toc187266382)

[CHƯƠNG 1 TỔNG QUAN VỀ BỆNH GAN 8](#_Toc187266383)

[1.1. Bệnh gan là gì? 8](#_Toc187266384)

[1.2. Các bệnh lý nguy hiểm về gan và triệu chứng sớm nhất 8](#_Toc187266385)

[1.2.1. Gan nhiễm mỡ 8](#_Toc187266386)

[1.2.2. Viêm gan 8](#_Toc187266387)

[1.2.3. Xơ gan 9](#_Toc187266388)

[1.2.4. Ung thư gan 10](#_Toc187266389)

[1.2.5. Áp xe gan 10](#_Toc187266390)

[1.3. Đối tượng có nguy cơ mắc bệnh gan cao 10](#_Toc187266391)

[1.4. Biến chứng nguy hiểm 10](#_Toc187266392)

[1.5. Chẩn đoán và điều trị bệnh hiệu quả 11](#_Toc187266393)

[1.5.1. Chẩn đoán 11](#_Toc187266394)

[1.5.2. Điều trị 12](#_Toc187266395)

[1.6. Cách phòng ngừa bệnh gan 13](#_Toc187266396)

[1.7. Người bệnh gan nên ăn gì và kiêng gì? 13](#_Toc187266397)

[CHƯƠNG 2 TÌM HIỂU VÀ XỬ LÝ DỮ LIỆU 15](#_Toc187266398)

[2.1. Tìm hiểu dữ liệu 15](#_Toc187266399)

[2.1.1. Nguồn gốc 15](#_Toc187266400)

[2.1.2. Ý nghĩa và thực tế của các biến 15](#_Toc187266401)

[2.2. Xử lý dữ liệu 16](#_Toc187266402)

[2.2.1. Kiểm tra giá trị trùng lặp 17](#_Toc187266403)

[2.2.2. Kiểm tra dữ liệu thiếu 19](#_Toc187266404)

[2.2.3. Kiểm tra lỗi cấu trúc 20](#_Toc187266405)

[2.2.4. Kiểm tra giá trị ngoại lai 22](#_Toc187266406)

[CHƯƠNG 3 QUY TRÌNH XÂY DỰNG MÔ HÌNH DỰ ĐOÁN BỆNH GAN 29](#_Toc187266407)

[3.1. Các kĩ thuật và mô hình sử dụng 29](#_Toc187266408)

[3.1.1. Kĩ thuật phân loại 29](#_Toc187266409)

[3.1.2. Mô hình RandomForest 29](#_Toc187266410)

[3.1.3. Mô hình Autogluon 30](#_Toc187266411)

[3.1.4. Các chỉ số để đánh giá mô hình 30](#_Toc187266412)

[3.2. Xây dựng mô hình 36](#_Toc187266413)

[3.2.1. Quy trình xây dựng mô hình 36](#_Toc187266414)

[3.2.2. Xây dựng và huấn luyện mô hình với các lần kiểm thử khác 37](#_Toc187266415)

[KẾT LUẬN 44](#_Toc187266416)

[TÀI LIỆU THAM KHẢO 45](#_Toc187266417)

[PHỤ LỤC 46](#_Toc187266418)

# DANH MỤC CÁC HÌNH VẼ

[Hình 1.1 Hình ảnh gan khỏe mạnh (bên trái) và gan bị xơ (bên phải) 9](#_Toc187266351)

[Hình 1.2 Xét nghiệm máu và men gan 12](#_Toc187266352)

[Hình 1.3. Các loại rau củ quả là thực phẩm rất tốt cho người bệnh gan 14](#_Toc187266353)

[Hình 2.1. Thông tin dữ liệu 15](#_Toc187266354)

[Hình 2.2. Các bước xử lý dữ liệu 17](#_Toc187266355)

[Hình 2.3. Các giá trị trùng lặp 18](#_Toc187266356)

[Hình 2.4 Bản ghi còn lại sau khi xoá trùng lặp 19](#_Toc187266357)

[Hình 2.5. Kiểm tra các thuộc tính bị thiếu 19](#_Toc187266358)

[Hình 2.6. Các bản ghi sau khi xử lý dữ liệu thiếu 20](#_Toc187266359)

[Hình 2.7. Kiểm tra lỗi cấu trúc 21](#_Toc187266360)

[Hình 2.8. Dữ liệu sau khi sửa lỗi cấu trúc 22](#_Toc187266361)

[Hình 2.9. Đặc trưng thống kê 22](#_Toc187266362)

[Hình 2.10. Hiển thị chỉ số Bilirubin toàn phần của các bệnh nhân trên 45mg/dl 24](#_Toc187266363)

[Hình 2.11. Bệnh nhân có chỉ số Alkphos cao hơn 1200 24](#_Toc187266364)

[Hình 2.12. Bệnh nhân có chỉ số Sgpt cap hơn 56 25](#_Toc187266365)

[Hình 2.13. Bênh nhân có chỉ số Sgot cao 25](#_Toc187266366)

[Hình 2.14. Bệnh nhân có chỉ số ALB cao 25](#_Toc187266367)

[Hình 2.15. Bệnh nhân có A/G Ratio cao 26](#_Toc187266368)

[Hình 3.1. ROC và AUC của một mô hình hoàn hảo theo giả định. 32](#_Toc187266369)

[Hình 3.2. ROC và AUC của những phỏng đoán hoàn toàn ngẫu nhiên. 34](#_Toc187266370)

[Hình 3.3. ROC và AUC của hai mô hình giả định. Đường cong trên phải, có AUC lớn hơn, đại diện cho hai mô hình tốt hơn 34](#_Toc187266371)

[Hình 3.4. Điểm được gắn nhãn đại diện cho các ngưỡng. 35](#_Toc187266372)

[Hình 3.5.Kết quả các chỉ số 36](#_Toc187266373)

[Hình 3.6. Hình ảnh quy trình các thử nghiệm 38](#_Toc187266374)

[Hình 3.7. Các chỉ số dự đoán kiểm thử lần 2 39](#_Toc187266375)

[Hình 3.8. Trước cân bằng và sau cân bằng 40](#_Toc187266376)

[Hình 3.9. Các chỉ số đánh giá cho mô hình thử nghiệm 3 41](#_Toc187266377)

[Hình 3.10.Mô hình và chỉ số từ mô hình Autogluon 42](#_Toc187266378)

[Hình 3.11. Bảng chỉ số của các thử nghiệm 42](#_Toc187266379)

Phân công công việc

# 

# MỞ ĐẦU

1. Lý do chọn đề tài

Bệnh gan là một trong những nguyên nhân gây tử vong hàng đầu trên toàn cầu, với hàng triệu người bị ảnh hưởng mỗi năm. Theo các tổ chức y tế, tỷ lệ mắc bệnh gan, bao gồm viêm gan, xơ gan và ung thư gan, đang gia tăng đáng kể, đặc biệt ở các khu vực có chế độ ăn uống không lành mạnh, lối sống ít vận động, và mức độ tiếp xúc với các yếu tố độc hại cao. Xuất phát từ thực tế trên, đề tài Xây dựng mô hình dự đoán bệnh gan ra đời nhằm giúp các bệnh viện, tổ chức y tế dự đoán nhanh xem một bệnh nhân có nguy cơ mắc bệnh gan không để sớm có phương pháp điều trị kịp thời.

1. Tầm quan trọng của đề tài

Phát hiện bệnh gan sớm có ý nghĩa quyết định trong việc cải thiện chất lượng điều trị và giảm thiểu nguy cơ tử vong. Tuy nhiên, việc chẩn đoán sớm thường gặp thách thức do các triệu chứng không rõ ràng ở giai đoạn đầu và sự phức tạp của các chỉ số y sinh học. Việc sử dụng các mô hình dự đoán dựa trên dữ liệu thực tế từ các bệnh nhân cung cấp một công cụ hỗ trợ hiệu quả cho các bác sĩ trong việc ra quyết định lâm sàng. Nghiên cứu này không chỉ tối ưu hóa khả năng dự đoán bệnh gan mà còn giúp xây dựng các hệ thống y tế thông minh, góp phần cải thiện chất lượng chăm sóc sức khỏe cộng đồng.

1. Mục tiêu nghiên cứu

Nghiên cứu tập trung vào việc xây dựng mô hình dự đoán chính xác bệnh gan với trên 80% khả năng dự đoán chính xác để có thể dự đoán chính xác khả năng mắc bệnh của bệnh nhân.

1. Bố cục của đồ án

* Mở đầu
* Chương 1: Chương này, giới thiệu vệ bệnh gan là gì?, các bệnh lý nguy hiểm về gan và triệu chứng sớm nhất. Giới thiệu về đối tượng có nguy cơ mắc bệnh gan cao, các biến chứng nguy hiểm, chuẩn doán và điều trị hiẹu quả. Cuối cùng là phòng ngửa và người bị bệnh nên ăn gì hay kiêng ăn gì.
* Chương 2: Đến với chương này nhóm tìm hiểu dữ liệu nguồn gốc ,ý nghĩa và thực tế của các biến. Xử lý dữ liệu theo 4 bước trùng lặp, kiểm tra dữ liệu thiếu, kiểm tra lỗi cấu trúc, kiểm tra ngoại lai.
* Chương 3: Chương này viết về quy trình xây dựng mô hình dự đoán bệnh gan. Giới thiệu về các kĩ thuật mô hình xử dụng, cuối cùng là xây dựng mô hình
* Chương 4: Chương này sẽ đánh giá kết quả dự đoán mô hình đưa ra các chỉ số đồng thời kết luận được mô hình nào tốt nhất để dưa vào sử dụng.
* Kết luận

# CHƯƠNG 1 TỔNG QUAN VỀ BỆNH GAN

Gan là cơ quan lớn thứ hai trong cơ thể sau da, có vai trò quan trọng đối với sức khỏe. Khi gan bị tổn thương, các chức năng sẽ bị suy giảm, lâu ngày gây ra nhiều biến chứng nguy hiểm đến tính mạng như xơ gan, suy gan, ung thư gan. Vì vậy, tầm soát bệnh gan sớm là cách chủ động phòng ngừa căn bệnh âm thầm nhưng để lại nhiều hậu quả khôn lường hiệu quả nhất.

## Bệnh gan là gì?

Là các bệnh lý ở gan do di truyền hoặc nhiễm virus, uống rượu... gây ra với biểu hiện đặc trưng là sự tổn thương ở các tế bào gan. Theo thời gian, tổn thương tạo thành các dải xơ (sẹo), dẫn đến suy gan, đe dọa tính mạng người bệnh.

Theo thống kê của Tổ chức Y tế thế giới, hiện nay Việt Nam có khoảng 7,8 triệu trường hợp mắc [viêm gan B](https://careplusvn.com/vi/tiem-phong-viem-gan-b-lich-tiem-va-dia-chi-chung-ngua-uy-tin-o-tphcm) mạn tính, hơn 13.000 đối tượng bị xơ gan mất bù, gần 6.000 người mắc ung thư tế bào gan và hơn 6.400 trường hợp tử vong do bệnh gan.

## Các bệnh lý nguy hiểm về gan và triệu chứng sớm nhất

### Gan nhiễm mỡ

Gan nhiễm mỡ là tình trạng mỡ tích tụ quá nhiều (hơn 5% trọng lượng gan). Bệnh lý này xuất hiện do thói quen ăn uống không hợp lý, lối sống sinh hoạt không khoa học, [béo phì](https://careplusvn.com/vi/thua-can-beo-phi-o-tre-em-va-nhung-bien-chung-nguy-hiem) hay do căng thẳng kéo dài...

Gan nhiễm mỡ diễn tiến âm thầm với các triệu chứng không rõ ràng nên rất khó phát hiện. Nếu điều trị không kịp thời, các tổn thương ở gan có thể phát triển thành xơ gan, ung thư gan.

### Viêm gan

Viêm gan là bệnh lý có thể xảy ra ở mọi đối tượng, do 3 nguyên nhân chính:

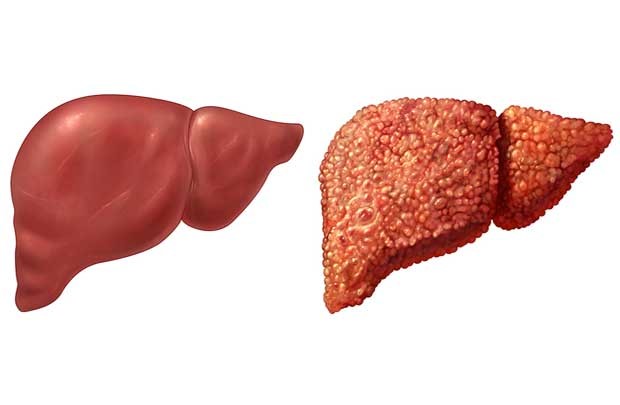
* Viêm gan do nhiễm virus, ký sinh trùng: virus viêm gan A, B, C, D, E, G, ký sinh trùng Plasmodium falciparum hay một số loại amip ký sinh tại gan.
* Viêm gan tự miễn: hệ thống miễn dịch của cơ thể tấn công gan.
* Viêm gan do nhiễm độc: rượu bia, thuốc.

Khi có dấu hiệu viêm, nhu mô gan bị tổn thương khiến các chức năng của cơ quan này suy yếu dần. Khi bệnh đã diến tiến đến giai đoạn nặng dễ dấn đến xơ gan không hồi phục, ung thư gan.

### Xơ gan

Là một trong những bệnh lý nguy hiểm về gan, đặc trưng bởi các tế bào gan dần được thay thế bằng các dải xơ, mô xẹo khiến cấu trúc tiểu thùy gan thay đổi thành những nốt tân sinh không có chức năng.

Khi mới khởi phát, [xơ gan](https://careplusvn.com/vi/nguyen-nhan-xo-gan-va-cach-phong-ngua) thường không có dấu hiệu rõ ràng. Ở giai đoạn nặng, triệu chứng của bệnh có thể là não gan, phù nề... rất nguy hiểm.



Hình 1.1 Hình ảnh gan khỏe mạnh (bên trái) và gan bị xơ (bên phải)

### Ung thư gan

Ung thư gan xảy ra khi các tế bào ác tính phát sinh tại các mô trong gan. Đây là một bệnh lý nguy hiểm nhưng khó phát hiện vì không có dấu hiệu đặc trưng. Khi phát hiện, người bệnh chỉ sống được từ 3 – 6 tháng và chỉ 1% là có thể sống được 5 năm.

### Áp xe gan

Là tình trạng xuất hiện một hoặc nhiều ổ mủ khác nhau trong tổ chức gan. Nguyên nhân gây bệnh có thể là nấm, các loại amip hoặc vi khuẩn. Bệnh có tỉ lệ tử vong cao nếu không điều trị kịp thời.

## Đối tượng có nguy cơ mắc bệnh gan cao

Đối tượng có nguy cơ mắc bệnh gan cao thường bao gồm những người có tiền sử gia đình mắc các bệnh về gan, đặc biệt là ung thư gan và xơ gan. Những người thường xuyên sử dụng rượu bia cũng nằm trong nhóm nguy cơ cao, vì việc tiêu thụ quá mức có thể gây tổn thương cho gan. Ngoài ra, những người có thói quen ăn đồ ăn sống, sử dụng chung kim tiêm, bàn chải đánh răng hay dao cạo cũng dễ bị lây nhiễm các virus gây bệnh gan. Những người xăm mình hoặc xỏ khuyên tai mà không đảm bảo vệ sinh cũng có nguy cơ cao.

Tiếp xúc với các chất dịch cơ thể của người khác, quan hệ tình dục không an toàn, và tiếp xúc với hóa chất độc hại cũng là những yếu tố làm tăng nguy cơ mắc bệnh gan. Hơn nữa, những người lạm dụng thuốc kháng sinh, mắc bệnh tiểu đường, béo phì hoặc thừa cân, và có nồng độ triglycerid trong máu cao cũng cần được chú ý, vì họ có thể dễ dàng phát triển các vấn đề về gan. Tất cả những yếu tố này đều góp phần làm tăng nguy cơ mắc bệnh gan, đòi hỏi sự chú ý và can thiệp kịp thời để bảo vệ sức khỏe.

## Biến chứng nguy hiểm

Bệnh gan có thể dẫn đến nhiều biến chứng nguy hiểm, ảnh hưởng nghiêm trọng đến sức khỏe và chất lượng cuộc sống của người bệnh. Một trong những biến chứng nghiêm trọng nhất là bệnh não gan, hay còn gọi là hôn mê gan, khi chức năng gan suy giảm dẫn đến tích tụ độc tố trong máu, gây rối loạn tâm thần và hôn mê.

Ngoài ra, giãn tĩnh mạch thực quản là một biến chứng phổ biến, xảy ra khi áp lực trong tĩnh mạch cửa tăng cao, có thể dẫn đến chảy máu nghiêm trọng. Hội chứng gan phổi cũng là một biến chứng đáng lo ngại, trong đó chức năng hô hấp bị suy giảm do sự tương tác giữa gan và phổi.

Suy thận là một biến chứng khác có thể xảy ra, khi gan không còn khả năng loại bỏ độc tố, dẫn đến tổn thương thận. Rối loạn đông máu cũng là một vấn đề nghiêm trọng, vì gan đóng vai trò quan trọng trong việc sản xuất các yếu tố đông máu. Cuối cùng, nhiễm trùng có thể xảy ra do hệ miễn dịch suy yếu, làm tăng nguy cơ mắc các bệnh nhiễm trùng nghiêm trọng. Những biến chứng này không chỉ đe dọa tính mạng mà còn yêu cầu sự can thiệp y tế kịp thời để giảm thiểu tác động tiêu cực đến sức khỏe.

## Chẩn đoán và điều trị bệnh hiệu quả

### Chẩn đoán

Thăm khám lâm sàng, trao đổi tiền sử bệnh lý trước đây.Xét nghiệm máu: công thức máu, tăng men gan AST (Aspartate amionotransferase), ALT (Alanine aminotransferase), tăng men GGT (Gamma glutamyl transferase) và ALP (Alkaline Phosphatase), tăng bilirubin, mức albumin thấp.

Xét nghiệm hình ảnh: X-quang bụng, chụp cắt lớp vi tính (CT scan), chụp cộng hưởng từ (MRI), nội soi mật tụy ngược dòng (ERCP).

Phân tích mô (sinh thiết gan): lấy một mẫu tế bào gan, kiểm tra dưới kính hiển vi để tìm tế bào ung thư, vi khuẩn trong gan.



Hình 1. Xét nghiệm máu và men gan

### Điều trị

Tùy thuộc vào nguyên nhân và diễn tiến của bệnh mà Bác sĩ sẽ chỉ định phương pháp điều trị phù hợp.

Cụ thể, gan bị tổn thương do nhiễm độc thuốc, cần ngưng sử dụng, tham vấn ý kiến Bác sĩ để thay thế bằng loại thuốc khác. Nếu nhiễm siêu vi, người bệnh cần điều trị theo phác đồ. Bệnh gan do bia rượu và một số bệnh lý về gan khác có thể điều trị bằng cách khắc phục các triệu chứng, xây dựng chế độ ăn uống, nghỉ ngơi khoa học.

Trường hợp, gan bị tổn thương nặng, gây ra các biến chứng nguy hiểm có thể phải phẫu thuật. Ví dụ, suy gan giai đoạn cuối cần tiến hành ghép gan. Vì vậy, để xác định chính xác nguyên nhân và có phương án điều trị bệnh kịp thời, bạn nên thăm khám định kỳ 6 tháng/lần hoặc theo chỉ định của Bác sĩ.

## Cách phòng ngừa bệnh gan

Thực hiện ăn chín uống sôi.Không sử dụng chung kim tiêm, dao cạo, đồ dùng cá nhân như bàn chải đánh răng...  Sử dụng các biện pháp an toàn khi quan hệ.Tiêm ngừa vaccine (viêm gan siêu vi A và B) đầy đủ theo [lịch tiêm chủng](https://careplusvn.com/vi/lich-tiem-phong-cho-tre-so-sinh-va-tre-nho-moi-nhat-nam-2020) hoặc theo chỉ định của Bác sĩ.Thăm khám định kì và tuân thủ điều trị theo phác đồ nếu có bệnh.Xây dựng chế độ ăn uống khoa học, đầy đủ chất [dinh dưỡng](https://careplusvn.com/vi/thap-dinh-duong-cho-nguoi-viet-hieu-dung-de-ap-dung-hieu-qua), lối sống lành mạnh, nghỉ ngơi hợp lý.Tham vấn ý kiến Bác sĩ trước khi sử dụng các loại thuốc giảm đau, kháng sinh. Tuyệt đối không tự ý mua thuốc uống, kết hợp các loại thuốc Đông Tây Y.Duy trì cân nặng luôn ổn định, tránh béo phì.Hạn chế sử dụng rượu bia và các chất kích thích.Tập thể dục ít nhất 30 phút mỗi ngày để nâng cao sức đề kháng, giúp cơ thể khỏe mạnh hơn.

## Người bệnh gan nên ăn gì và kiêng gì?

Chế độ dinh dưỡng ở mỗi giai đoạn của bệnh gan là khác nhau. Tuy nhiên, việc ăn uống cần phải có sự cân đối giữa các nhóm thực phẩm như: chất đạm, chất béo, đường, vitamin và khoáng chất. Theo đó, người bệnh gan nên ăn và kiêng một số thực phẩm như:

Nên **ăn:**Thức ăn giàu đạm: thịt nạc, ngũ cốc, cá, trứng, sữa, đậu phụ...Thức ăn thanh, nhẹ, dễ tiêu hóa.Các loại đậu.Rau củ quả: các loại rau có màu xanh đậm, củ quả màu cam, đỏ như cà rốt, cà chua, bí đỏ rất giàu vitamin và khoáng chất, rất tốt cho người bệnh gan.Dầu thực vật, dầu đậu nành, dầu mè.



Hình 1.3. Các loại rau củ quả là thực phẩm rất tốt cho người bệnh gan

**Kiêng ăn:** Gia vị:muối, ớt, cay, thực phẩm nhiều muối: thịt lợn xông khói, xúc xích, thực phẩm gói sẵn, đóng hộp.Chất lỏng: nướp ép, nước ngọt, súp vì bổ sung nhiều nước có thể gây phù nề, rượu, bia, thức ăn nhiều dầu mỡ.

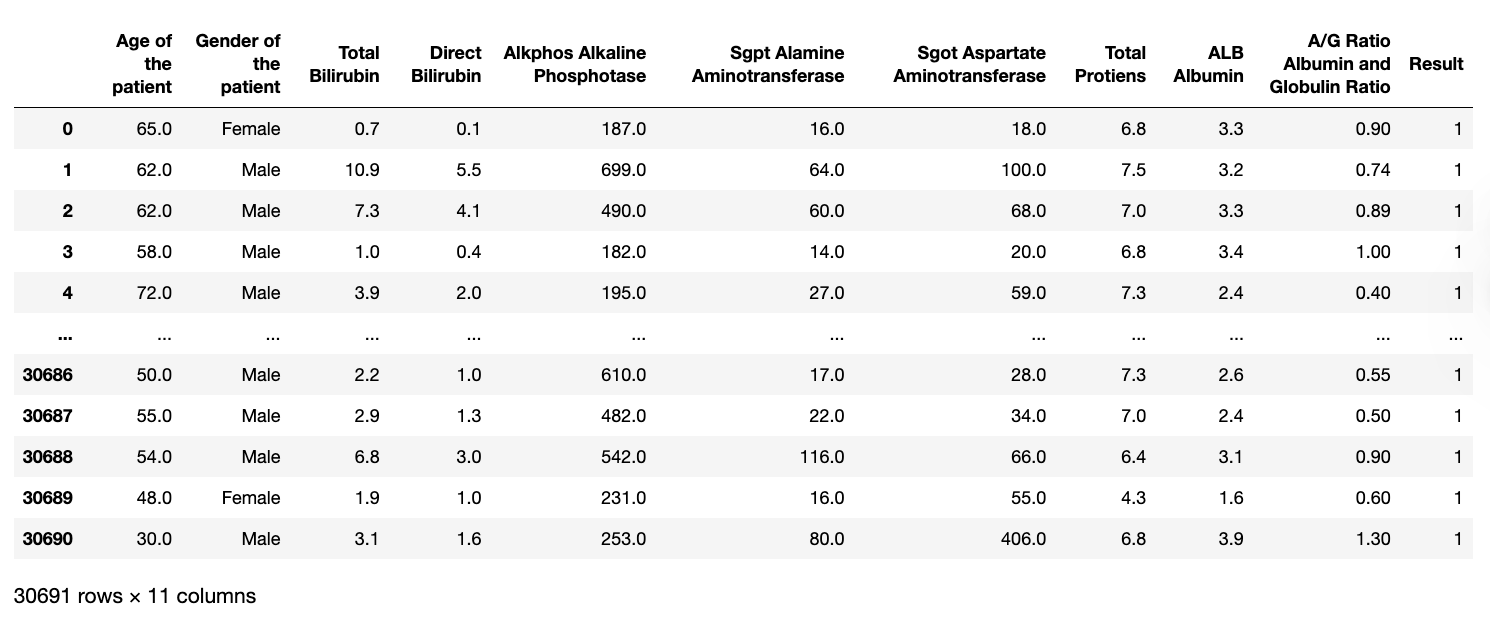
Bệnh gan là một bệnh lý nguy hiểm bởi hầu hết các trường hợp bệnh gan đều không có biểu hiện lâm sàng ở giai đoạn đầu. Triệu chứng chỉ rõ ràng khi bệnh ở giai đoạn nặng, khó điều trị. Vì vậy, tầm soát bệnh gan, kiểm tra sức khỏe theo định kỳ rất cần thiết nhằm phát hiện và điều trị bệnh sớm nhất.

# CHƯƠNG 2 TÌM HIỂU VÀ XỬ LÝ DỮ LIỆU

## Tìm hiểu dữ liệu

### Nguồn gốc

Bộ dữ liệu được tìm thấy trên openscience.vn. Được đăng tải bởi tổ chức Cục thông tin khoa học và Công nghệ Quốc gia. Của hơn 30.000 bệnh nhân được dự đoán mắc bệnh gan.



Hình 2.. Thông tin dữ liệu

Bộ dữ liệu này bao gồm 10 biến là: tuổi, giới tính, bilirubin toàn phần, bilirubin trực tiếp, tổng protein, albumin, tỉ lệ A/G, SGPT, SGOT và phosphatase kiềm.

### Ý nghĩa và thực tế của các biến

Age (Tuổi của bệnh nhân):Tuổi là một yếu tố quan trọng ảnh hưởng đến sức khỏe, bao gồm cả chức năng gan. Một số bệnh gan phổ biến hơn ở người lớn tuổi (ví dụ, xơ gan, ung thư gan).

Gender (Giới tính của bệnh nhân):Giới tính có thể ảnh hưởng đến nguy cơ mắc các bệnh gan

Bilirubin là sản phẩm phân hủy của hemoglobin từ hồng cầu già. Nồng độ bilirubin cao có thể chỉ ra vấn đề về gan (như vàng da), ống mật hoặc máu.

DB (Bilirubin trực tiếp):Bilirubin trực tiếp là dạng hòa tan trong nước và được gan bài tiết. Mức DB cao thường liên quan đến tắc nghẽn ống mật hoặc tổn thương gan.

Alkphos (Phosphatase kiềm): Một enzyme có mặt trong gan, xương và các mô khác. Mức Alkphos cao có thể chỉ ra bệnh lý liên quan đến mật, tổn thương gan hoặc bệnh về xương.

Sgpt (Alanine Aminotransferase - ALT): Enzyme này tăng cao khi gan bị tổn thương, ví dụ như trong viêm gan, gan nhiễm mỡ hoặc xơ gan.

Sgot (Aspartate Aminotransferase - AST):Enzyme này không chỉ đặc hiệu cho gan, mà còn được tìm thấy ở tim và cơ. Tuy nhiên, tỷ lệ AST/ALT cao có thể chỉ ra bệnh xơ gan hoặc viêm gan do rượu.

TP (Tổng protein):Bao gồm albumin và globulin. Tổng protein thấp có thể liên quan đến bệnh gan mạn tính hoặc suy dinh dưỡng.

ALB (Albumin):Một loại protein chính do gan sản xuất. Nồng độ albumin thấp có thể chỉ ra suy gan hoặc bệnh gan mạn tính.

A/G Ratio (Tỉ lệ Albumin và Globulin):Tỷ lệ này thường thấp khi gan bị tổn thương. Mức tỷ lệ bất thường cũng có thể liên quan đến bệnh tự miễn hoặc các bệnh lý khác.

Result:Đây là nhãn phân loại dựa trên chẩn đoán của chuyên gia y tế với các giá trị 1- bệnh nhân mắc bệnh gan,2-bệnh nhân không mắc bệnh gan.Nhãn này thường được sử dụng làm biến mục tiêu (target variable) để huấn luyện mô hình phân loại.

## Xử lý dữ liệu

Quá trình làm sạch dữ liệu là một bước quan trọng trong phân tích dữ liệu, giúp đảm bảo rằng dữ liệu được sử dụng là chính xác và đáng tin cậy. Đầu tiên, nhóm cần **loại bỏ các giá trị trùng lặp hoặc không liên quan**, điều này giúp giảm thiểu sự nhiễu trong dữ liệu và cải thiện độ chính xác của các phân tích sau này. Tiếp theo, việc **sửa lỗi cấu trúc** là cần thiết để đảm bảo rằng dữ liệu được định dạng đúng cách, từ đó giúp dễ dàng hơn trong việc xử lý và phân tích. Sau đó, nhóm cần **xử lý các dữ liệu bị thiếu,** bằng cách điền giá trị hoặc loại bỏ các hàng không đầy đủ thông tin. Cuối cùng, việc **lọc các ngoại lệ không mong muốn** sẽ giúp loại bỏ những dữ liệu không phù hợp, đảm bảo rằng chỉ những thông tin có giá trị được giữ lại. Thực hiện đầy đủ các bước này sẽ giúp nâng cao chất lượng dữ liệu và tối ưu hóa kết quả phân tích.

A diagram of a diagram

Description automatically generated

Hình 2.. Các bước xử lý dữ liệu

### Kiểm tra giá trị trùng lặp

Trong quá trình kiểm tra và làm sạch dữ liệu, nhóm đã phát hiện ra rằng có **18589 bản ghi trùng lặp** trong tập dữ liệu. Sự tồn tại của các bản ghi trùng lặp này có thể gây ra những sai lệch trong phân tích và ảnh hưởng đến độ chính xác của các mô hình học máy. Do đó, việc xử lý và loại bỏ các bản ghi trùng lặp là rất cần thiết để đảm bảo rằng dữ liệu được sử dụng là chính xác và đáng tin cậy. Sau khi loại bỏ các bản ghi này, nhóm sẽ có một tập dữ liệu sạch hơn, giúp cải thiện hiệu suất và độ tin cậy của các phân tích tiếp theo.

A screenshot of a table

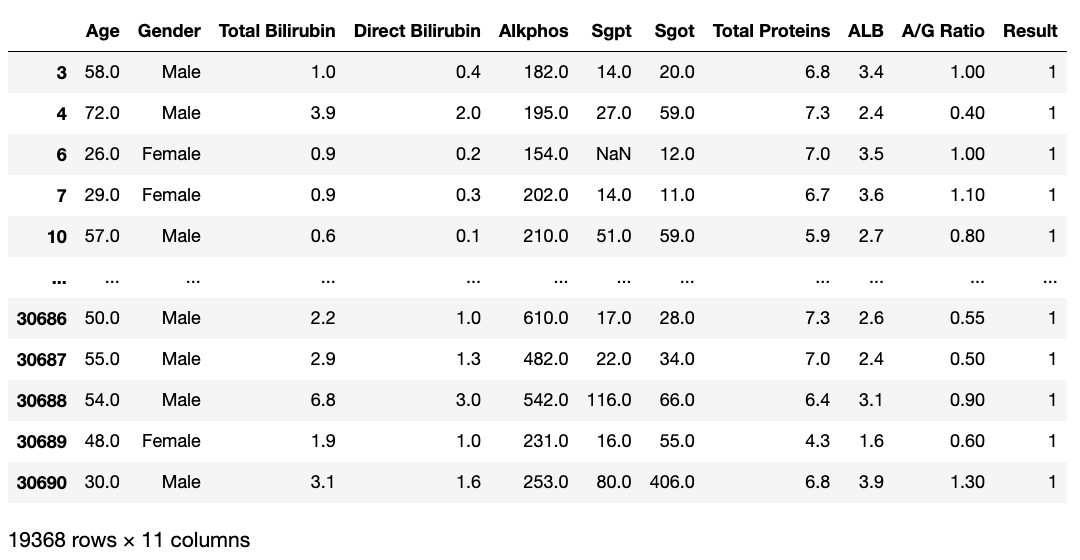
Description automatically generated

A blue and red pie chart

Description automatically generated

Hình 2.. Các giá trị trùng lặp

Sau khi phân tích và đánh giá tình hình dữ liệu, nhóm đã quyết định xử lý các bản ghi trùng lặp bằng cách xóa bỏ tất cả các bản ghi trùng lặp và giữ lại bản ghi cuối cùng. Quyết định được đưa ra nhằm đảm bảo rằng tập dữ liệu cuối cùng phản ánh chính xác nhất thông tin cần thiết, đồng thời giảm thiểu sự nhiễu và sai lệch có thể xảy ra do các bản ghi trùng lặp. Việc giữ lại bản ghi cuối cùng cũng giúp nhóm duy trì thông tin mới nhất, điều này là quan trọng trong các bài toán phân tích dữ liệu và xây dựng mô hình học máy. Sau khi thực hiện bước này, nhóm sẽ có một tập dữ liệu sạch hơn và đáng tin cậy hơn, sẵn sàng cho các bước phân tích tiếp theo. Sau khi xoá trùng lặp nhóm còn 19368 bản ghi



Hình 2. Bản ghi còn lại sau khi xoá trùng lặp

### Kiểm tra dữ liệu thiếu

Trong quá trình chuẩn bị dữ liệu cho mô hình học máy, việc kiểm tra dữ liệu thiếu là một bước quan trọng để đảm bảo rằng các đặc trưng được sử dụng là đầy đủ và chính xác.

A screenshot of a computer screen

Description automatically generated

Hình 2.. Kiểm tra các thuộc tính bị thiếu

Trong quá trình kiểm tra dữ liệu, nhóm nhận thấy có rất nhiều giá trị thiếu, đặc biệt là trong các giá trị đo đạc thực tế trong lĩnh vực y học. Việc điền các giá trị như độ lệch chuẩn hoặc trung vị vào những giá trị thiếu này không phải là một lựa chọn hợp lý, vì điều này có thể dẫn đến việc làm sai lệch thông tin và ảnh hưởng đến tính chính xác của các phân tích. Do đó, nhóm đã quyết định xóa bỏ các bản ghi có giá trị thiếu. Quyết định này nhằm đảm bảo rằng tập dữ liệu cuối cùng phản ánh chính xác nhất các thông tin y học thực tế, từ đó giúp nâng cao độ tin cậy của các mô hình học máy và phân tích sau này. Sau khi xoá các bản ghi bị thiếu nhóm còn lại 16389 bản ghi.

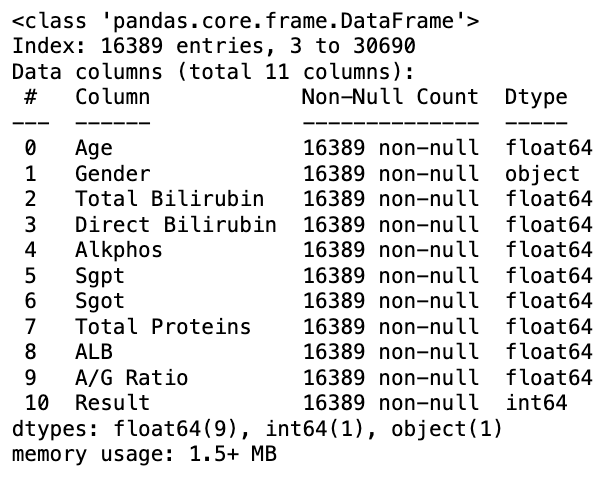
A table with numbers and letters

Description automatically generated

Hình 2.. Các bản ghi sau khi xử lý dữ liệu thiếu

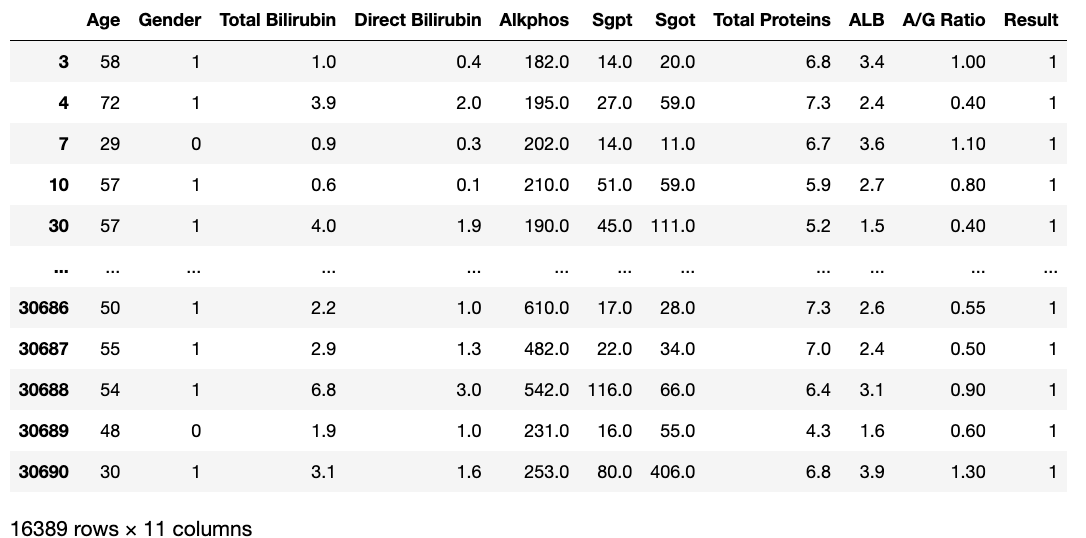
### Kiểm tra lỗi cấu trúc

Sau khi đã loại bỏ các bản ghi có giá trị thiếu, bước tiếp theo trong quy trình làm sạch dữ liệu là kiểm tra và sửa lỗi cấu trúc. Việc này rất quan trọng để đảm bảo rằng dữ liệu được định dạng đúng cách và có thể dễ dàng xử lý trong các bước phân tích tiếp theo. Nhóm sẽ xem xét các vấn đề như định dạng kiểu dữ liệu, sự nhất quán trong cách ghi chú và các quy tắc đặt tên cột. Nếu phát hiện bất kỳ lỗi nào, chẳng hạn như kiểu dữ liệu không phù hợp hoặc tên cột không rõ ràng, nhóm sẽ tiến hành điều chỉnh để đảm bảo rằng dữ liệu có cấu trúc hợp lý. Bằng cách này, nhóm không chỉ cải thiện chất lượng dữ liệu mà còn tạo điều kiện thuận lợi cho việc áp dụng các thuật toán học máy và phân tích sâu hơn trong tương lai.

******

Hình 2.. Kiểm tra lỗi cấu trúc

Trong quá trình kiểm tra và sửa lỗi cấu trúc dữ liệu, nhóm đã xác định rằng biến Age cần được đổi thành kiểu số nguyên để đảm bảo tính chính xác trong các phép toán và phân tích sau này. Bên cạnh đó, biến Gender, là một biến phân loại với các giá trị là "Male" và "Female", cũng cần được chuyển đổi thành các giá trị số 0 và 1 trước khi xây dựng mô hình. Việc này giúp mô hình dễ dàng xử lý và hiểu được thông tin phân loại. Các biến còn lại trong tập dữ liệu đều đã có kiểu dữ liệu chính xác, do đó không cần thực hiện thêm bất kỳ điều chỉnh nào. Những bước này sẽ giúp tối ưu hóa chất lượng dữ liệu và nâng cao hiệu suất của mô hình học máy trong các phân tích tiếp theo.



Hình 2.. Dữ liệu sau khi sửa lỗi cấu trúc

### Kiểm tra giá trị ngoại lai

Cuối cùng, bước quan trọng trong quy trình làm sạch dữ liệu là kiểm tra ngoại lệ. Việc này giúp nhóm phát hiện và xử lý những giá trị không bình thường hoặc không hợp lý trong tập dữ liệu, có thể gây ảnh hưởng tiêu cực đến kết quả phân tích và mô hình học máy. Nhóm sẽ xem xét các giá trị nằm ngoài khoảng giá trị hợp lý, chẳng hạn như tuổi không hợp lệ hoặc các chỉ số y tế không thực tế. Nếu phát hiện bất kỳ ngoại lệ nào, nhóm sẽ quyết định xem có nên loại bỏ hay điều chỉnh về giá trị hợp lý hơn. Bằng cách thực hiện bước kiểm tra này, nhóm không chỉ nâng cao chất lượng dữ liệu mà còn đảm bảo rằng các phân tích và mô hình được xây dựng trên nền tảng dữ liệu chính xác và đáng tin cậy.

A screenshot of a table

Description automatically generated

Hình 2.. Đặc trưng thống kê

Độ tuổi trong tập dữ liệu có khoảng biến thiên rộng từ 4 đến 90 tuổi, cho thấy rằng bệnh gan có thể ảnh hưởng đến mọi lứa tuổi, từ trẻ em đến người cao tuổi. Tuy nhiên, phân phối của bệnh lại tập trung chủ yếu ở nhóm trung niên và người lớn tuổi, với trung vị là 45 và giá trị trung bình là 43.77, cho thấy rằng những người trong độ tuổi này có nguy cơ cao hơn. Đặc biệt, không có giá trị ngoại lai nào trong biến này, điều này càng củng cố tính chính xác của dữ liệu và cho phép nhóm rút ra những kết luận đáng tin cậy về sự phân bố độ tuổi liên quan đến bệnh gan.

Trong y học, các mức chỉ số Total Bilirubin được xác định như sau: trẻ sơ sinh có mức dưới 10 mg/dL, trẻ trên 1 tháng tuổi từ 0.3 – 1.2 mg/dL, và người lớn từ 0.2 – 1.0 mg/dL. Đáng chú ý, chỉ số Total Bilirubin cao nhất được ghi nhận trên thế giới là 45 mg/dL. Tuy nhiên, trong tập dữ liệu của c nhóm lại xuất hiện giá trị Total Bilirubin lên đến 75 mg/dL, điều này cho thấy đây có thể là một giá trị ngoại lai và cần được kiểm tra kỹ lưỡng để xác định tính hợp lệ của nó.

Biến Direct Bilirubin (DB) trong tập dữ liệu có tổng số quan sát là 16,389, với giá trị nhỏ nhất là 0.1 mg/dL và giá trị lớn nhất lên đến 19.7 mg/dL. Giá trị trung bình của DB là 1.53 mg/dL, cho thấy rằng mặc dù 25% số quan sát có giá trị nhỏ hơn hoặc bằng 0.3 mg/dL (trong ngưỡng bình thường), nhưng vẫn có nhiều trường hợp có bilirubin trực tiếp tăng cao, đặc biệt khi Total Bilirubin (TB) cũng ở mức cao. Giá trị lớn nhất 19.7 mg/dL vượt xa ngưỡng bình thường, có thể là dấu hiệu của các bệnh lý nghiêm trọng. Do DB là một phần của TB, việc giữ lại các ngoại lai của DB là cần thiết để phân tích các trường hợp bệnh lý nặng hoặc hiếm gặp, từ đó cung cấp thông tin chi tiết về chức năng gan và nhận diện các bất thường trong quá trình bài tiết bilirubin.

Trong tập dữ liệu với 16,389 quan sát, giá trị Alkaline Phosphatase (Alkphos) có giá trị nhỏ nhất là 63 IU/L và lớn nhất là 2,110 IU/L, với giá trị trung bình là 290.83 IU/L; phân phối cho thấy 25% số quan sát nhỏ hơn hoặc bằng 175 IU/L, trung vị là 209 IU/L, và 75% nhỏ hơn hoặc bằng 298 IU/L, trong khi các giá trị rất cao (>1200 IU/L) thường liên quan đến các bệnh lý nghiêm trọng như tắc nghẽn đường mật, ung thư gan hoặc đường mật, và bệnh xương nghiêm trọng như bệnh Paget hoặc ung thư xương.

A table with numbers and numbers

Description automatically generated

Hình 2.. Hiển thị chỉ số Bilirubin toàn phần của các bệnh nhân trên 45mg/dl

A screenshot of a graph

Description automatically generated

Hình 2.. Bệnh nhân có chỉ số Alkphos cao hơn 1200

A screenshot of a graph

Description automatically generated

Hình 2.. Bệnh nhân có chỉ số Sgpt cap hơn 56

A table with numbers and a number of numbers

Description automatically generated with medium confidence

Hình 2.. Bênh nhân có chỉ số Sgot cao

A screenshot of a table

Description automatically generated

Hình 2.. Bệnh nhân có chỉ số ALB cao

A table with numbers and letters

Description automatically generated

Hình 2.. Bệnh nhân có A/G Ratio cao

**Bilirubin trực tiếp (DB):**Với 16.389 quan sát, giá trị Direct Bilirubin dao động từ 0.1 mg/dL đến 19.7 mg/dL, với trung bình là 1.53 mg/dL. Phân vị thứ 25 cho thấy 25% bệnh nhân có DB ≤ 0.3 mg/dL, phù hợp với ngưỡng bình thường, trong khi trung vị là 0.2 mg/dL. Tuy nhiên, phân vị thứ 75 cho thấy 75% bệnh nhân có DB ≤ 1.3 mg/dL, cao hơn mức bình thường. Những số liệu này cho thấy phần lớn các trường hợp có mức DB thấp, nhưng có một số bệnh nhân với giá trị DB cao bất thường, có thể liên quan đến các bệnh lý nghiêm trọng. Khi kết hợp với Total Bilirubin, DB giúp cung cấp cái nhìn sâu hơn về chức năng gan và xác định các trường hợp bệnh lý nặng.

**Alkphos (ALP):**Chỉ số ALP, với 16.389 mẫu, dao động từ 63 IU/L đến 2.110 IU/L, trung bình là 290.83 IU/L. Phân vị thứ 25, 50 và 75 lần lượt là 175 IU/L, 209 IU/L và 298 IU/L. Mức ALP cao, đặc biệt là trên 1200 IU/L, thường gợi ý các bệnh lý nghiêm trọng như tắc nghẽn đường mật, ung thư gan hoặc xương. Mặc dù hầu hết giá trị nằm trong khoảng chấp nhận, nhưng cần chú ý đến các trường hợp ngoại lai để phát hiện kịp thời các vấn đề sức khỏe nghiêm trọng.

SGPT **(ALT):**SGPT là enzyme đánh giá tổn thương gan, với 16.389 quan sát có giá trị từ 10 IU/L đến 2000 IU/L và trung bình là 80.15 IU/L. Phân vị thứ 25 là 23 IU/L, trung vị là 35 IU/L và phân vị thứ 75 là 62 IU/L. Giá trị trung bình vượt quá mức bình thường (7–56 IU/L), cho thấy có nhiều bệnh nhân có dấu hiệu tổn thương gan. Những giá trị SGPT rất cao, như 2000 IU/L, có thể là dấu hiệu của viêm gan cấp tính hoặc tổn thương gan nặng, đòi hỏi can thiệp y tế và kiểm tra thêm.

SGOT **(AST):**Với 16.389 quan sát, chỉ số SGOT dao động từ 10 IU/L đến 4.929 IU/L với giá trị trung bình là 111.37 IU/L. Phân vị thứ 25 là 25 IU/L, trung vị là 42 IU/L và phân vị thứ 75 là 88 IU/L. Mặc dù phần lớn các giá trị nằm trong khoảng bình thường hoặc tăng nhẹ, nhưng giá trị trung bình cao hơn ngưỡng bình thường (<40 IU/L) cho thấy nhiều bệnh nhân có thể gặp tổn thương gan. Một số trường hợp ngoại lai với giá trị SGOT vượt trội lên tới 4.929 IU/L cần được kiểm tra kỹ lưỡng vì có thể là dấu hiệu của các tình trạng y khoa nghiêm trọng.

Total **Proteins:**Chỉ số Tổng Protein trong máu với 16.389 mẫu có giá trị từ 6.1 g/dL đến 7.7 g/dL, trung bình là 6.85 g/dL. Phân vị thứ 25 là 6.6 g/dL, trung vị là 6.8 g/dL và phân vị thứ 75 là 7.0 g/dL. Các giá trị này nằm trong phạm vi bình thường (6.0–8.3 g/dL), cho thấy chức năng tổng hợp protein của gan ở đa số bệnh nhân không bị ảnh hưởng nghiêm trọng.

**Albumin (ALB):**Với 16.389 mẫu, nồng độ Albumin dao động từ 2.0 g/dL đến 3.5 g/dL, trung bình là 2.85 g/dL. Phân vị thứ 25 là 2.4 g/dL, trung vị là 2.8 g/dL và phân vị thứ 75 là 3.0 g/dL. Các giá trị Albumin thấp hơn mức bình thường (3.4–5.4 g/dL) cho thấy suy giảm chức năng gan trong việc tổng hợp albumin hoặc có thể do mất protein từ các bệnh lý khác. Mặc dù không có ngoại lai lớn, nhưng xu hướng Albumin thấp đòi hỏi sự theo dõi kỹ càng và đánh giá thêm nguyên nhân.

Tỷ **lệ A/G (Albumin/Globulin):**Trong 16.389 quan sát, tỷ lệ A/G dao động từ 0.4 đến 0.9 với trung bình là 0.7. Phân vị thứ 25 là 0.5, trung vị là 0.7 và phân vị thứ 75 là 0.8, đều thấp hơn ngưỡng bình thường (1.0–2.1). Tỷ lệ thấp có thể do giảm Albumin hoặc tăng Globulin, liên quan đến các tình trạng viêm mãn tính, suy gan hoặc các bệnh lý liên quan. Mặc dù một số tỷ lệ A/G cao bất thường có thể chỉ ra các trường hợp hiếm gặp quan trọng, nhưng đa số các trường hợp đều cho thấy xu hướng A/G giảm, cần theo dõi và phân tích chi tiết hơn khi kết hợp với các chỉ số khác.

Cuối cùng, sau khi xem xét kỹ lưỡng các giá trị ngoại lai, nhóm đưa ra quyết định chia thành hai hướng giải quyết khác nhau: xoá ngoại lai và không xoá ngoại lai. Mỗi hướng sẽ được áp dụng để xây dựng các mô hình riêng biệt, nhằm đánh giá ảnh hưởng của việc xử lý ngoại lai đến kết quả cuối cùng.

Hướng 1: Xoá ngoại lai - Trong trường hợp này, nhóm sẽ loại bỏ các giá trị ngoại lai đã được xác định là không hợp lý hoặc không phản ánh đúng thực tế. Việc này giúp tạo ra một tập dữ liệu sạch hơn, có thể cải thiện độ chính xác của mô hình và giảm thiểu sự nhiễu trong quá trình phân tích.

Hướng 2: Không xoá ngoại lai - Đối với hướng này, nhóm sẽ giữ lại tất cả các giá trị ngoại lai, với giả định rằng chúng có thể chứa thông tin quan trọng hoặc phản ánh các tình huống hiếm gặp trong y học. Việc này cho phép mô hình học máy học hỏi từ toàn bộ dữ liệu, bao gồm cả những trường hợp đặc biệt, từ đó có thể cải thiện khả năng tổng quát và độ tin cậy của mô hình.

Bằng cách xây dựng và so sánh hai mô hình này, nhóm có thể đánh giá được tác động của việc xử lý ngoại lai đến hiệu suất của mô hình, từ đó đưa ra quyết định tốt hơn cho các ứng dụng thực tế trong lĩnh vực y học.

# CHƯƠNG 3 QUY TRÌNH XÂY DỰNG MÔ HÌNH DỰ ĐOÁN BỆNH GAN

## Các kĩ thuật và mô hình sử dụng

### Kĩ thuật phân loại

Phân loại (Classification) là một trong những kỹ thuật quan trọng trong lĩnh vực khoa học dữ liệu, được sử dụng để gán nhãn hoặc dự đoán nhóm cho các đối tượng dữ liệu dựa trên các thuộc tính của chúng. Đây là kỹ thuật được ứng dụng rộng rãi trong các bài toán thực tế như dự đoán bệnh lý, nhận diện hình ảnh, phân tích cảm xúc, và phát hiện gian lận.

Sử dụng các thước đo như độ chính xác (Accuracy), độ nhạy (Recall), độ đặc hiệu (Specificity), và điểm F1 (F1-score).

Dùng các phương pháp như Cross-validation để kiểm tra độ ổn định.Dựa vào tính chất của mô hình là dự đoán xem bệnh nhân có mắc bệnh hay không nên nhóm sẽ lựa chọn kĩ thuật này để xây mô hình.

### Mô hình RandomForest

Random Forest là một trong những mô hình học máy mạnh mẽ và linh hoạt, thường được chọn trong các bài toán phân loại (Random Forest Classifier) và hồi quy (Random Forest Regressor). Vì kết quả dự đoán của mô hình là dạng phân loại nên nhóm sẽ chọn Random Forest Classifier.

Mô hình Random Forest được xây dựng từ một tập hợp các cây quyết định (Decision Trees). Các đặc điểm của mô hình này bao gồm:

Ensemble Learning: Random Forest là một mô hình học máy theo phương pháp kết hợp nhiều cây quyết định để đưa ra dự đoán.

Bootstrap Aggregating (Bagging): Các cây trong Random Forest được huấn luyện trên các mẫu ngẫu nhiên từ tập huấn luyện, giúp giảm thiểu hiện tượng overfitting.

Tính ngẫu nhiên: Random Forest sử dụng phương pháp lấy mẫu ngẫu nhiên cho dữ liệu và ngẫu nhiên chọn các đặc trưng tại mỗi nút trong cây quyết định.

### Mô hình Autogluon

AutoGluon là một thư viện học máy tự động (AutoML) được phát triển bởi AWS (Amazon Web Services), giúp người dùng xây dựng và triển khai các mô hình học máy mà không cần phải hiểu sâu về các thuật toán học máy hoặc tối ưu siêu tham số (hyperparameter tuning). AutoGluon cung cấp một cách tiếp cận tự động hóa cho việc huấn luyện và lựa chọn mô hình, giúp tiết kiệm thời gian và công sức cho các nhà phân tích dữ liệu và nhà nghiên cứu.

AutoGluon có thể xử lý các dữ liệu phức tạp như dữ liệu thiếu, dữ liệu không chuẩn, và các kiểu dữ liệu khác nhau (dữ liệu phân loại, số, văn bản, hình ảnh, v.v).

Việc tự động lựa chọn mô hình tốt nhất sau khi huấn luyện của AutoGluon sẽ giúp các nhà phân tích lựa chọn mô hình thích hợp với tập dữ liệu nghiên cứu.

Áp dụng Autogluon vào thử nghiệm để kiểm tra xem ngoài RF thì còn mô hình nào tốt hơn để thử nghiệm với tập dữ liệu đang nghiên cứu không.

### Các chỉ số để đánh giá mô hình

Phân loại nhị phân cung cấp tên cụ thể cho các danh mục dự đoán chính xác và dự đoán không chính xác. Vì vậy, công thức độ chính xác cho phương pháp phân loại nhị phân là như sau:

A black text with black letters

Description automatically generated

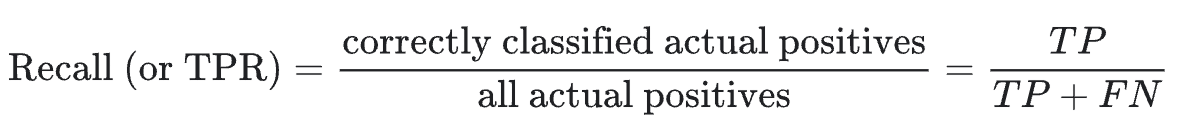
trong đó:

* TP là số [dương tính thật](https://developers.google.com/machine-learning/glossary?hl=vi#TP) (dự đoán chính xác).
* TN là số [dương tính giả](https://developers.google.com/machine-learning/glossary?hl=vi#TN) (dự đoán chính xác).
* FP là số lượng [dương tính giả](https://developers.google.com/machine-learning/glossary?hl=vi#FP) (dự đoán không chính xác).
* FN là số lượng [**dương tính giả**](https://developers.google.com/machine-learning/glossary?hl=vi#FN) (dự đoán không chính xác).

So sánh và đối chiếu độ chính xác với [độ chính xác](https://developers.google.com/machine-learning/glossary?hl=vi#precision) và [độ chính xác](https://developers.google.com/machine-learning/glossary?hl=vi#recall).

**Tỷ lệ dương tính thật (TPR),** hay tỷ lệ phần trăm tất cả các kết quả dương tính thực tế được phân loại chính xác là dương tính, còn được gọi là Recall.

**Recall** được định nghĩa về mặt toán học như sau:



Kết quả âm tính giả là kết quả dương tính thực tế bị phân loại nhầm thành kết quả âm tính, đó là lý do tại sao chúng xuất hiện trong mẫu số. Trong ví dụ về việc phân loại thư rác, giá trị gợi nhắc đo lường tỷ lệ phần trăm email rác được phân loại chính xác là email rác. Đây là lý do tại sao tên gọi khác của tỷ lệ thu hồi là **xác suất phát hiện:** tỷ lệ này trả lời câu hỏi "Mô hình này phát hiện được bao nhiêu phần trăm email rác?"

Một mô hình hoàn hảo theo giả định sẽ không có âm tính giả và do đó có giá trị gợi lại (TPR) là 1, 0, tức là tỷ lệ phát hiện là 100%.

Trong một tập dữ liệu mất cân bằng, trong đó số lượng ví dụ dương tính thực tế rất thấp, giả sử tổng cộng có 1-2 ví dụ, thì chỉ số hồi quy sẽ ít có ý nghĩa và ít hữu ích hơn.

Độ chính xác 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:

A black text on a white background

Description automatically generated

Trong ví dụ về việc phân loại thư rác, độ chính xác đo lường *tỷ lệ phần trăm email được phân loại là thư rác thực sự là thư rác.*

Một mô hình hoàn hảo giả định sẽ không có kết quả dương tính giả nào, do đó có độ chính xác là 1.0.

Trong một tập dữ liệu mất cân bằng, trong đó số lượng ví dụ thực tế là rất thấp, giả sử tổng cộng có 1-2 ví dụ, độ chính xác sẽ ít có ý nghĩa và ít hữu ích hơn dưới dạng một chỉ số.

Độ chính xác sẽ cải thiện khi kết quả dương tính giả giảm, còn khả năng thu hồi sẽ cải thiện khi giá trị âm tính giả giảm. Nhưng như đã thấy trong phần trước, việc tăng ngưỡng phân loại có xu hướng làm giảm số lượng dương tính giả và tăng số lượng âm tính giả, trong khi việc giảm ngưỡng sẽ gây ra tác động ngược lại. Do đó, độ chính xác và mức độ truy lại thường cho thấy mối quan hệ nghịch đảo, trong đó việc cải thiện một trong số này sẽ làm xấu đi yếu tố còn lại.

Đường cong ROC  là thông tin trình bày trực quan về hiệu suất của mô hình trên tất cả các ngưỡng. Phiên bản dài của tên, đặc điểm hoạt động của bộ thu, là một mã tạm thời từ việc phát hiện ra-đa trong Thế chiến thứ hai.

Đường cong ROC được vẽ bằng cách tính tỷ lệ dương thực (TPR) và tỷ lệ dương tính giả (FPR) ở mọi ngưỡng có thể xảy ra (trên thực tế, ở mức khoảng thời gian đã chọn), sau đó vẽ biểu đồ TPR trên FPR. Một người mẫu hoàn hảo, mà ở một số ngưỡng có TPR là 1.0 và FPR là 0.0, có thể được biểu thị bằng một điểm tại (0, 1) nếu tất cả các ngưỡng khác bị bỏ qua hoặc bằng cách xác định như sau:

Hình 1. Đồ thị của TPR (trục y) so với FPR (trục x) biểu thị
            hiệu suất của một mô hình hoàn hảo: một dòng từ (0,1) đến (1,1).

Hình 3.. ROC và AUC của một mô hình hoàn hảo theo giả định.

Diện tích dưới đường cong (AUC) : Vùng dưới đường cong ROC biểu thị xác suất mà mô hình đó nếu cho một ví dụ âm và dương được chọn ngẫu nhiên, sẽ xếp hạng dương cao hơn giá trị âm.

Mô hình hoàn hảo ở trên, chứa một hình vuông với các cạnh có độ dài là 1, có giá trị diện tích dưới đường cong (AUC) bằng 1,0. Điều này có nghĩa là xác suất 100% mô hình sẽ xếp hạng chính xác một ví dụ khẳng định được chọn ngẫu nhiên cao hơn ví dụ phủ định được chọn ngẫu nhiên. Nói cách khác, khi xem xét sự lan truyền của các điểm dữ liệu dưới đây, AUC cho biết xác suất mà mô hình sẽ đặt hình vuông được chọn ngẫu nhiên ở bên phải hình tròn được chọn ngẫu nhiên, không phụ thuộc vào nơi đặt ngưỡng.

Nói một cách cụ thể hơn, thuật toán phân loại thư rác sử dụng AUC 1.0 luôn chỉ định một email rác ngẫu nhiên thì khả năng bị so với một email hợp lệ ngẫu nhiên. Cách phân loại thực tế của mỗi email tuỳ thuộc vào ngưỡng mà bạn chọn.

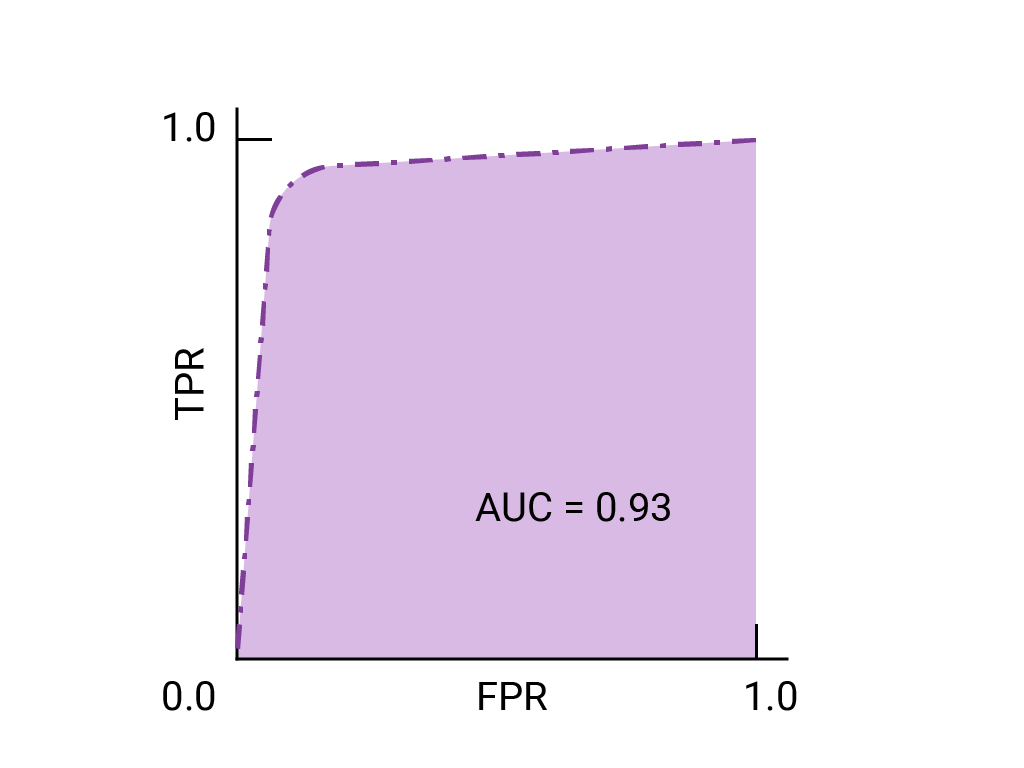
Đối với thuật toán phân loại nhị phân, một mô hình thực hiện chính xác như các dự đoán ngẫu nhiên hoặc tung đồng xu có ROC là một đường chéo từ (0,0) đến (1,1). AUC là 0, 5 thể hiện xác suất 50% xác suất xếp hạng chính xác kết quả dương tính ngẫu nhiên và ví dụ phủ định.

Trong ví dụ về thuật toán phân loại thư rác, thuật toán phân loại spam có AUC là 0,5 sẽ gán email rác ngẫu nhiên có xác suất là thư rác cao hơn so với email ngẫu nhiên email hợp lệ chỉ được một nửa thời gian.

Hình 2. Đồ thị của TPR (trục y) so với FPR (trục x) biểu thị
            hiệu suất của một người đoán ngẫu nhiên 50-50: một đường chéo từ (0,0)
            đến (1,1).

Hình 3.. ROC và AUC của những phỏng đoán hoàn toàn ngẫu nhiên.

AUC và ROC để chọn mô hình và ngưỡngAUC là thước đo hữu ích để so sánh hiệu suất của hai mô hình khác nhau, miễn là tập dữ liệu đã gần cân bằng.



Hình 3.3. ROC và AUC của hai mô hình giả định. Đường cong trên phải, có AUC lớn hơn, đại diện cho hai mô hình tốt hơn

Các điểm trên đường cong ROC gần nhất với (0,1) biểu thị cho ngưỡng hoạt động tốt nhất cho mô hình cụ thể. Như đã thảo luận trong [Ngưỡng](https://developers.google.com/machine-learning/crash-course/classification/thresholding?hl=vi), [Ma trận nhầm lẫn](https://developers.google.com/machine-learning/crash-course/classification/thresholding?hl=vi#confusion_matrix) và [Lựa chọn chỉ số và lựa chọn đánh đổi](https://developers.google.com/machine-learning/crash-course/classification/accuracy-precision-recall?hl=vi#choice_of_metric_and_tradeoffs) ngưỡng mà bạn chọn sẽ tuỳ thuộc vào chỉ số nào quan trọng nhất đối với trường hợp sử dụng cụ thể. Hãy xem xét các điểm A, B và C trong các trường hợp sau mỗi biểu đồ đại diện cho một ngưỡng:

Hình 4. Đường cong ROC có AUC=0,84 cho thấy 3 điểm trên
            phần lồi của đường cong gần với (0,1) nhất được dán nhãn A, B, C theo thứ tự.

Hình 3.. Điểm được gắn nhãn đại diện cho các ngưỡng.

Nếu dương tính giả (cảnh báo giả) có mức hao tổn cao, thì bạn nên chọn một ngưỡng đưa ra FPR thấp hơn, như ngưỡng ở điểm A, ngay cả khi TPR bị giảm. Ngược lại, nếu dương tính giả có giá rẻ và âm tính giả (bỏ lỡ dương tính thật) rất tốn kém, ngưỡng cho điểm C, mà tối đa hoá TPR, có thể phù hợp hơn. Nếu chi phí gần tương đương, điểm B có thể mang lại mức cân bằng tốt nhất giữa TPR và FPR.

## Xây dựng mô hình

### Quy trình xây dựng mô hình

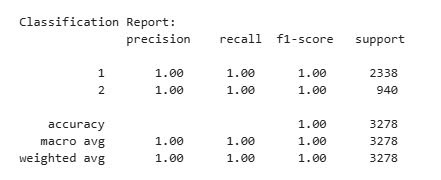
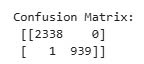
Bước đầu tiên trong quy trình xây dựng mô hình Random Forest là gán biến cho mô hình, trong đó biến phụ thuộc được xác định là ‘Result’. Sau khi gán biến, dữ liệu sẽ được chia thành hai tập: tập huấn luyện (train) và tập kiểm tra (test).

Tiếp theo, dữ liệu trong tập huấn luyện sẽ được chuẩn hóa bằng kỹ thuật StandardScaler. Kỹ thuật này giúp biến đổi các đặc trưng sao cho có trung bình bằng 0 và độ lệch chuẩn bằng 1, từ đó giúp cải thiện hiệu suất của mô hình. Việc chuẩn hóa này rất quan trọng, đặc biệt khi các biến có đơn vị hoặc thang đo khác nhau, vì nó giúp mô hình học tốt hơn và giảm thiểu ảnh hưởng của các biến có giá trị lớn.

Cuối cùng, sau khi dữ liệu đã được chuẩn hóa, mô hình Random Forest sẽ được xây dựng và huấn luyện trên tập dữ liệu đã chuẩn bị. Quy trình này đảm bảo rằng mô hình có thể hoạt động hiệu quả và chính xác trong việc dự đoán biến phụ thuộc ‘Result’.

Sau khi xây dựng mô hình Random Forest, bước tiếp theo là hiển thị kết quả để đánh giá hiệu suất của mô hình. Kết quả này thường bao gồm các chỉ số quan trọng như độ chính xác, độ nhạy, độ đặc hiệu và điểm F1.

Đầu tiên, nhóm sẽ so sánh các dự đoán của mô hình với giá trị thực tế trong tập kiểm tra. Bảng confusion matrix có thể được sử dụng để trực quan hóa số lượng dự đoán đúng và sai, từ đó giúp hiểu rõ hơn về cách mà mô hình hoạt động.

A close-up of a number

Description automatically generated

Hình 3.5.Kết quả các chỉ số

Mô hình hiện tại đang cho thấy kết quả tốt, với độ chính xác đạt mức 100%. Tuy nhiên, để đảm bảo tính ổn định và độ tin cậy của mô hình, việc thực hiện các thử nghiệm khác là rất cần thiết.

Các thử nghiệm này có thể bao gồm việc thay đổi các tham số của mô hình, sử dụng các kỹ thuật khác nhau để chuẩn hóa dữ liệu, hoặc thử nghiệm với các tập dữ liệu khác nhau để kiểm tra khả năng tổng quát của mô hình. Việc này không chỉ giúp xác định xem kết quả 100% có phải là một sự trùng hợp ngẫu nhiên hay không, mà còn giúp phát hiện ra các vấn đề tiềm ẩn như hiện tượng overfitting.

Bằng cách thực hiện các thử nghiệm bổ sung, nhóm có thể củng cố niềm tin vào hiệu suất của mô hình và đảm bảo rằng nó có thể hoạt động hiệu quả trong các tình huống thực tế khác nhau.

### fffXây dựng và huấn luyện mô hình với các lần kiểm thử khác

Quy trình thử nghiệm các trường hợp của mô hình Random Forest bao gồm bốn bước chính, mỗi bước nhằm kiểm tra và đánh giá hiệu suất của mô hình trong các điều kiện khác nhau.

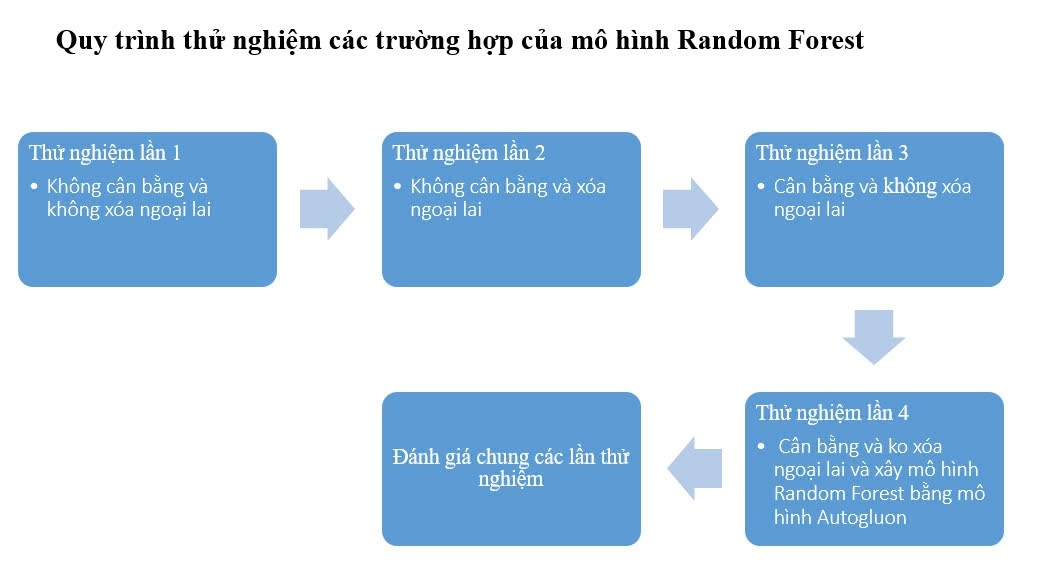
Thử nghiệm lần 1: Trong thử nghiệm này, dữ liệu không được cân bằng và không có bước xóa các giá trị ngoại lai. Điều này giúp đánh giá hiệu suất của mô hình trong điều kiện dữ liệu thực tế mà không qua xử lý.

Thử nghiệm lần 2: Tương tự như lần 1, dữ liệu vẫn không được cân bằng và không xóa ngoại lai. Việc lặp lại thử nghiệm này giúp xác nhận tính nhất quán của kết quả.

Thử nghiệm lần 3: Ở bước này, dữ liệu được cân bằng nhưng vẫn không xóa các giá trị ngoại lai. Điều này cho phép đánh giá tác động của việc cân bằng dữ liệu đến hiệu suất của mô hình.

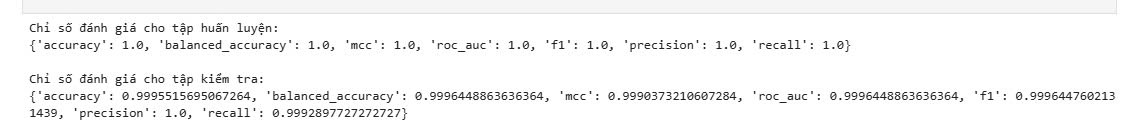
Thử nghiệm lần 4: Cuối cùng, dữ liệu được cân bằng và không xóa ngoại lai, đồng thời mô hình Random Forest được xây dựng bằng công cụ Autogluon. Thử nghiệm này nhằm kiểm tra khả năng của mô hình trong việc xử lý dữ liệu đã được chuẩn hóa.

Sau khi hoàn thành các thử nghiệm, kết quả sẽ được đánh giá chung để xác định hiệu suất tổng thể của mô hình trong các điều kiện khác nhau, từ đó đưa ra những điều chỉnh cần thiết cho các bước tiếp theo.



Hình 3.. Hình ảnh quy trình các thử nghiệm

Với lần thử nghiệm 1 là xây dựng mô hình và kết qủa như trên nhóm đã trình bày. Tiếp theo là các lần thử nghiệm còn lại.

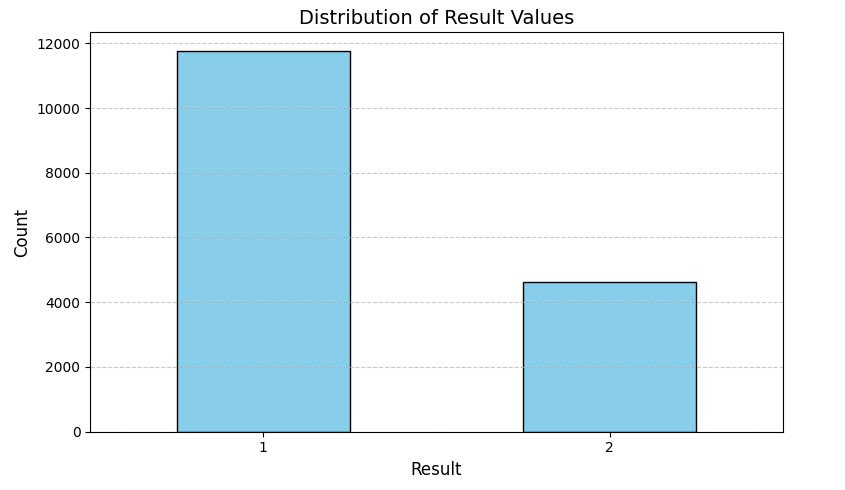


A graph with blue and orange bars

Description automatically generated

Hình 3.. Các chỉ số dự đoán kiểm thử lần 2

Nhìn vào chỉ số của lần thử nghiệm 2, nhóm thấy rằng các chỉ số đánh giá cho tập dữ liệu vẫn giữ nguyên, cho thấy mô hình không bị ảnh hưởng bởi việc không cân bằng dữ liệu và không xóa các giá trị ngoại lai. Tiếp tục đi đến huân sluyeejn thử nghiệm 3.



A graph with blue and orange bars

Description automatically generated

Hình 3.. Trước cân bằng và sau cân bằng

A screenshot of a graph

Description automatically generated

A graph with a red line

Description automatically generatedA graph with blue and orange bars

Description automatically generated

Hình 3.. Các chỉ số đánh giá cho mô hình thử nghiệm 3

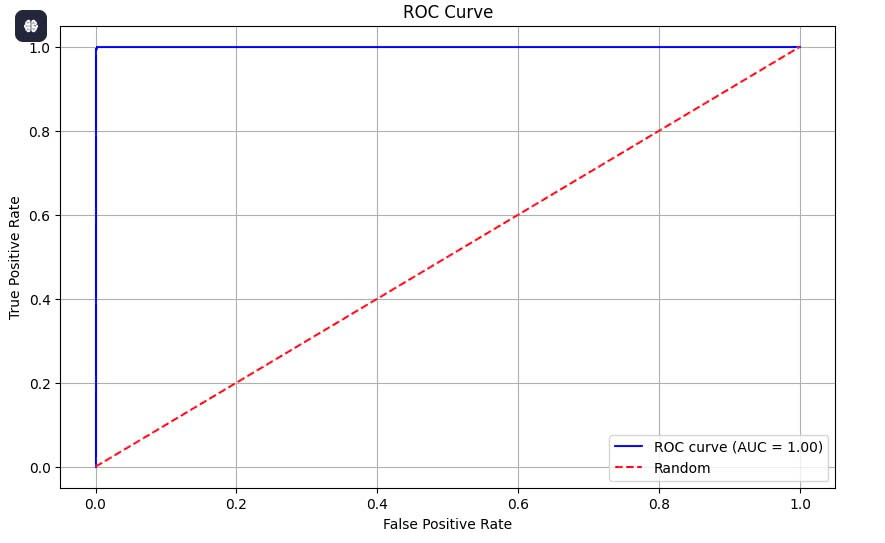
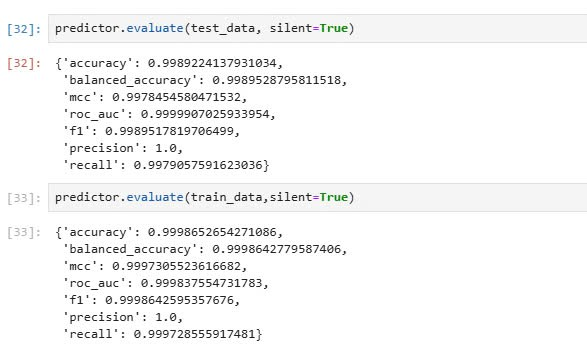
Mô hình hiện tại đã cho thấy kết quả tốt trong hai lần thử nghiệm trước, với các chỉ số đánh giá ổn định và khả năng xử lý dữ liệu không cân bằng cũng như các giá trị ngoại lai. Dựa trên những kết quả khả quan này, nhóm sẽ tiến hành thực nghiệm lần 4 bằng cách sử dụng mô hình Autogluon.

Mục tiêu của thử nghiệm này là tìm ra mô hình tốt nhất có thể cải thiện hiệu suất dự đoán. Autogluon là một công cụ mạnh mẽ cho phép tự động hóa quá trình xây dựng và tối ưu hóa mô hình, giúp tiết kiệm thời gian và công sức trong việc lựa chọn các thuật toán phù hợp.

Bằng cách áp dụng Autogluon, chúng nhóm có thể thử nghiệm nhiều cấu hình khác nhau, từ đó tìm ra mô hình tối ưu nhất cho tập dữ liệu hiện tại. Kết quả từ thử nghiệm sẽ cung cấp cái nhìn sâu sắc hơn về khả năng của mô hình và giúp đưa ra quyết định chính xác hơn trong việc triển khai mô hình trong thực tế.

A screenshot of a computer

Description automatically generated

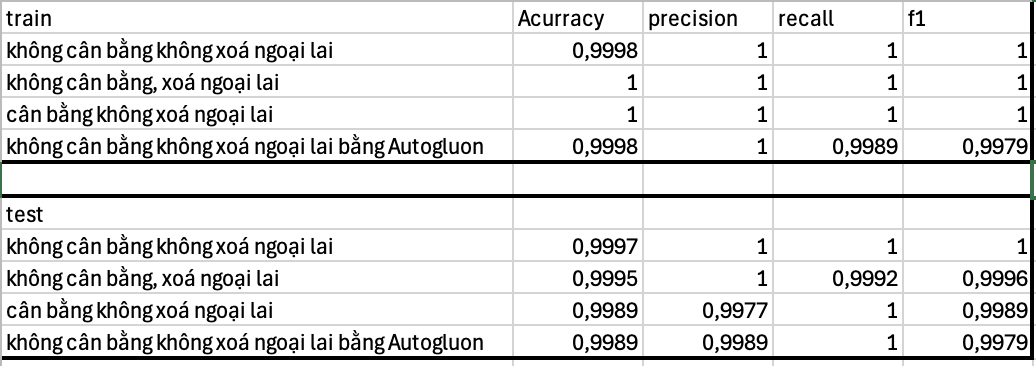


A blue squares with white numbers

Description automatically generated

Hình 3..Mô hình và chỉ số từ mô hình Autogluon

Nhận xét tổng kết cho 4 lần thử nghiệm:



Hình 3.. Bảng chỉ số của các thử nghiệm

Hiệu suất mô hình nhất quán: Trong cả 4 lần thử nghiệm, sự tương đồng giữa dữ liệu thực tế và dự đoán rất cao.Điều này cho thấy mô hình có tính ổn định và độ chính xác cao trên các tập dữ liệu khác nhau.

Độ cân bằng dữ liệu: Ở thử nghiệm lần 4, dữ liệu đã được cân bằng tốt giữa các lớp, giúp cải thiện độ chính xác và giảm nguy cơ thiên lệch mô hình.

Sự phù hợp giữa thực tế và dự đoán:Ở mỗi lần thử nghiệm (1-4), các biểu đồ cho thấy số lượng mẫu dự đoán gần như trùng khớp hoàn toàn với dữ liệu thực tế ở cả hai nhóm lớp, đặc biệt rõ ràng ở lần 3 và 4.

Tính tái lập kết quả:Mô hình duy trì hiệu suất cao qua các lần thử nghiệm, chứng minh khả năng tái lập kết quả tốt ngay cả khi thay đổi dữ liệu hoặc thiết lập.

Mô hình đạt hiệu suất gần như hoàn hảo và ổn định qua 4 lần thử nghiệm.Dữ liệu cân bằng và kết quả thử nghiệm lần 4 khẳng định mô hình đã được tối ưu và phù hợp cho việc áp dụng vào các bài toán thực tế.Từ 4 thử nghiệm cho thấy quá trình xây dựng mô hình được thực hiện chính xác và các chỉ số cao của mô hình là hoàn toàn hợp lý.

# KẾT LUẬN

Dựa vào kết quả đánh giá của mô hình sau nhiều lần thử nghiệm, có thể thấy rằng mô hình đã đạt được độ chính xác cao khoảng 98%, vượt qua mục tiêu ban đầu mà nhóm đề ra. Kết quả này không chỉ cho thấy khả năng dự đoán chính xác của mô hình mà còn phản ánh sự hiệu quả trong việc xử lý dữ liệu, bao gồm cả việc cân bằng và loại bỏ các giá trị ngoại lai.

Độ chính xác cao này chứng tỏ rằng mô hình có khả năng tổng quát tốt và có thể áp dụng hiệu quả trong các tình huống thực tế. Tuy nhiên, để đảm bảo tính bền vững của kết quả, việc tiếp tục theo dõi và đánh giá mô hình trong các điều kiện khác nhau là rất cần thiết. Nhìn chung, thành công này mở ra cơ hội cho việc triển khai mô hình trong các ứng dụng thực tế, đồng thời khẳng định giá trị của quy trình thử nghiệm và tối ưu hóa mà nhóm đã thực hiện.

# TÀI LIỆU THAM KHẢO

Các tài liệu từ Internet

1. Tập dữ liệu : <https://www.kaggle.com/datasets/abhi8923shriv/liver-disease-patient-dataset/data>
2. Bệnh gan: triệu chứng, diều trị va cach phong ngừa hữu hiệu : <https://careplusvn.com/vi/benh-gan-trieu-chung-dieu-tri-va-cach-phong-ngua-huu-hieu>
3. Kĩ thuật phân loại : https://s.net.vn/beNM
4. Phân loại : Độ chính xác, mức độ ghi nhớ, độ chính xác và các chỉ số liên quan  <https://developers.google.com/machinelearning/crashcourse/classification/accuracy-precision-recall?hl=vi>
5. Mô hình autogluon: <https://auto.gluon.ai/stable/index.html>
6. Mô hình randomforest :

<https://machinelearningcoban.com/tabml_book/ch_model/random_forest.html>

# PHỤ LỤC