TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN

KHOA TOÁN - CƠ - TIN HỌC

TIỂU LUẬN HỌC PHẦN
MÔ HÌNH TOÁN SINH THÁI

ĐỀ TÀI: MÔ HÌNH DỰ ĐOÁN CÁC GIAI ĐOẠN CỦA BỆNH ĐỘNG MẠCH VÀNH

Giảng viên: TS.Nguyễn Trọng Hiếu

Nhóm sinh viên thực hiện:

Họ và tên	MSV
Nguyễn Phương Bích	22001237
Đỗ Thảo Giang	22001253
Nguyễn Thị Phương Thảo	22001286

Hà Nội, 2024

Thông tin thành viên nhóm

Thành viên	Mã sinh viên	Công việc
Nguyễn Phương Bích	22001237	- Tìm hiểu lý thuyết bệnh động mạch vành. - Thực thi bài toán. - Viết báo cáo.
Đỗ Thảo Giang	22001253	- Tìm hiểu lý thuyết bệnh động mạch vành. - Thực thi bài toán. - Viết báo cáo.
Nguyễn Thị Phương Thảo	22001286	- Tìm hiểu lý thuyết bệnh động mạch vành. - Thực thi bài toán. - Viết báo cáo.

Mục lục

Danh sách các hình ảnh

Danh sách các bảng

Lời nói đầu

Hệ tim mạch bao gồm tim và hệ mạch máu, đóng vai trò nền tảng trong việc duy trì sự sống của cơ thể. Tim là một khối cơ mạnh mẽ, hoạt động như một máy bơm, đảm bảo máu được lưu thông qua hệ mạch máu. Hệ động mạch, một phần quan trọng của hệ tuần hoàn, bao gồm động mạch chủ, động mạch vành và mạng lưới động mạch nhỏ, có nhiệm vụ vận chuyển máu giàu oxy từ tim đến các cơ quan, cung cấp dưỡng chất và duy trì các chức năng sinh lý. Trong đó, hệ động mạch vành đặc biệt quan trọng khi trực tiếp nuôi dưỡng cơ tim, giúp tim hoạt động hiệu quả.

Khi chức năng của hệ động mạch bị suy giảm, đặc biệt là trong trường hợp tắc nghẽn hoặc xơ vữa động mạch vành, nguy cơ xảy ra các biến cố nghiêm trọng như nhồi máu cơ tim, suy tim, và đột tử gia tăng đáng kể. Bệnh động mạch vành (ĐMV) hiện là nguyên nhân hàng đầu gây tử vong trên toàn cầu, đặt ra nhu cầu cấp thiết về các phương pháp hiệu quả để phát hiện và dự đoán sớm bệnh.

Tuy nhiên, thực tế hiện nay, việc chẩn đoán bệnh tim mạch vẫn còn đối mặt với nhiều khó khăn. Phần lớn bệnh nhân thường đến bệnh viện khi bệnh đã tiến triển vào giai đoạn nghiêm trọng, kèm theo các triệu chứng nặng nề. Quá trình phân tích và chẩn đoán đòi hỏi sự tham gia của các chuyên gia giàu kinh nghiệm, trong khi số lượng bệnh nhân ngày càng gia tăng gây áp lực lớn lên hệ thống y tế, đặc biệt tại các cơ sở y tế tuyến cơ sở với nguồn lực hạn chế. Những yếu tố này không chỉ làm gia tăng nguy cơ chẩn đoán chậm trễ mà còn ảnh hưởng đến hiệu quả điều trị và cơ hội phục hồi của người bệnh. Điều này làm nổi bật tầm quan trọng của việc áp dụng các công nghệ hiện đại để hỗ trợ quá trình chẩn đoán, nhằm đảm bảo tính kịp thời và chính xác.

Bài báo này sử dụng hai phương pháp học máy phổ biến, Gaussian Naive Bayes và hồi quy Ordinal Logistic, để dự đoán các giai đoạn của bệnh động mạch vành, dựa trên bộ dữ liệu UCI Heart Disease dataset. Đây là bộ dữ liệu được thu thập từ các trung tâm y tế lớn, bao gồm các chỉ số quan trọng như tuổi, giới tính, huyết áp, cholesterol, và các yếu tố nguy cơ khác.

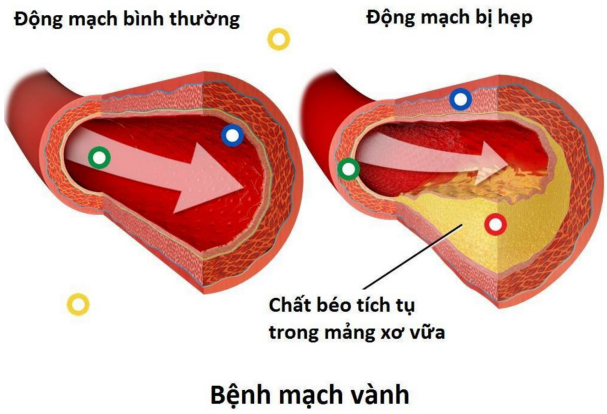
Phương pháp Gaussian Naive Bayes, dựa trên định lý Bayes, được sử dụng để phân loại các giai đoạn bệnh dựa trên các yếu tố nguy cơ. Trong khi đó, mô hình hồi quy Ordinal Logistic được triển khai để xử lý dữ liệu thứ tự, nhằm đánh giá mức độ nghiêm trọng của bệnh. Việc so sánh hiệu quả giữa hai mô hình trên cùng một bộ dữ liệu giúp đánh giá khả năng ứng dụng của từng phương pháp, góp phần cung cấp công cụ hữu ích cho các bác sĩ trong việc chẩn đoán và đưa ra quyết định điều trị.

Hy vọng rằng bài báo sẽ không chỉ góp phần cải thiện khả năng dự đoán bệnh lý động mạch vành mà còn mở rộng tiềm năng ứng dụng của các phương pháp học máy trong lĩnh vực y học, đặc biệt là trong việc phát triển các giải pháp công nghệ tiên tiến hỗ trợ chăm sóc sức khỏe.

Tổng quan

1 Khái niệm và thống kê bệnh động mạch vành

Bệnh động mạch vành (Coronary artery disease - CAD) hay còn gọi là bệnh tim thiếu máu cục bộ là thuật ngữ dùng để chỉ tình trạng thiếu máu của cơ tim, nguyên nhân do xơ vữa động mạch hoặc tắc nghẽn xơ vữa động mạch của động mạch vành (Xơ vữa động mạch xảy ra khi động mạch bị tắc nghẽn bởi các mảng bám, được hình thành từ các chất béo, cholesterol, canxi và các chất khác tích tụ trong thành động mạch).



Hình 1: Bệnh động mạch vành.
Nguồn ảnh: bauman.com, 2021

Các bệnh lý liên quan đến tim mạch ảnh hưởng rất lớn đến đời sống sức khỏe của bệnh nhân và tạo ra gánh nặng rất lớn cho toàn xã hội. Theo WHO, các bệnh tim mạch là nguyên nhân hàng đầu gây ra tử vong trên toàn cầu. Ước tính có 17,9 triệu người tử vong do bệnh tim mạch vào năm 2019 trong đó phần lớn là bệnh tim mạch do xơ vữa, chiếm 32% tổng số ca tử vong trên toàn cầu. Trong đó, hơn 75% số ca tử vong do bệnh tim mạch xảy ra ở các nước thu nhập trung bình hoặc thấp.

Tại Việt Nam, theo thống kê của Bộ Y tế, mỗi năm có khoảng 200.000 người tử vong vì bệnh tim mạch, chiếm 33% ca tử vong. Theo thống kê của Viện Tim Mạch Việt Nam qua các năm từ 2000 đến năm 2015, tỷ lệ tăng huyết áp ở người trưởng thành tăng khoảng 1% mỗi năm và đã chiếm 25%, vậy cứ 4 người trưởng thành thì có một người tăng huyết áp. Tăng huyết áp làm tăng nguy cơ tử vong do bệnh lý mạch lên gấp 3 lần so với người không mắc bệnh.

Ở các nước phát triển, bệnh động mạch vành là một căn bệnh phổ biến. Số người mắc bệnh động mạch vành ở Mỹ mỗi năm khoảng 13.200.000 bệnh nhân, tỷ lệ này ở châu Âu cũng đạt tới 3,5-4%.

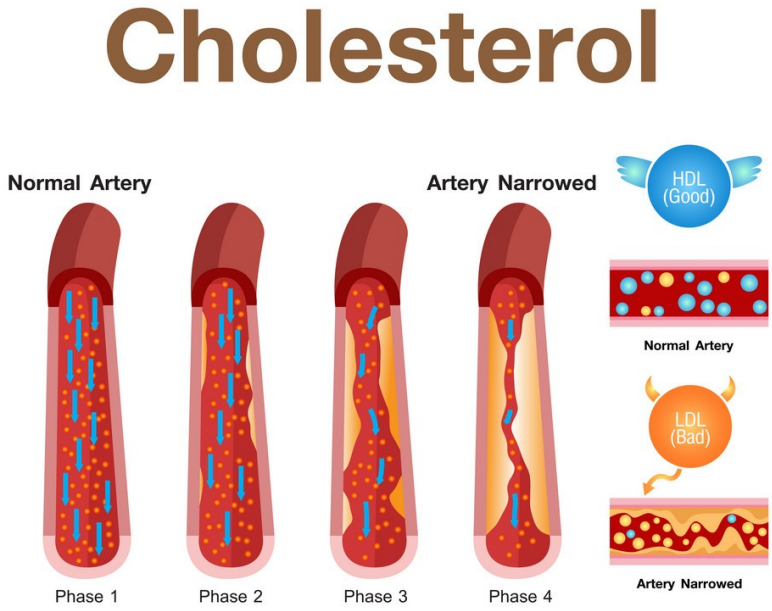
Ở Việt Nam cũng như các nước đang phát triển, bệnh động mạch vành đang có xu hướng gia tăng mạnh, tỷ lệ tử vong do bệnh này chiếm 11-36%. Bệnh dần trở thành gánh nặng cho sức khỏe cộng đồng của quốc gia và là một trong những bệnh gây tử vong cao.

2 Tổng quan về bệnh động mạch vành

2.1 Sự hình thành mảng xơ vữa

Sự hình thành các mảng xơ vữa có liên quan trực tiếp đến cholesterol.

Cholesterol không tan trong máu nên được vận chuyển bởi chất mang có tên là Lipoprotein để tuần hoàn đi khắp cơ thể. Có 3 loại Lipoprotein là LDL, VLDL và HDL. Cholesterol được vận chuyển trong máu với 3 loại Lipoprotein trên được ký hiệu là: LDL-c, VLDL-c, HDL-c.



Hình 2: LDL, HDL và sự hình thành mảng xơ vữa.
Nguồn ảnh: vectorstock.com, 2020

LDL-c ở nồng độ cao có khả năng thấm qua lớp nội mạc, trải qua quá trình oxy hóa hoặc các biến đổi khác sau đó thu hút bạch cầu vào lớp nội mạc của mạch vành, dẫn đến sự hình thành vệt mỡ và phát triển thành các mảng xơ vữa động mạch. Nên LDL-c còn được gọi là cholesterol xấu hay mỡ xấu.

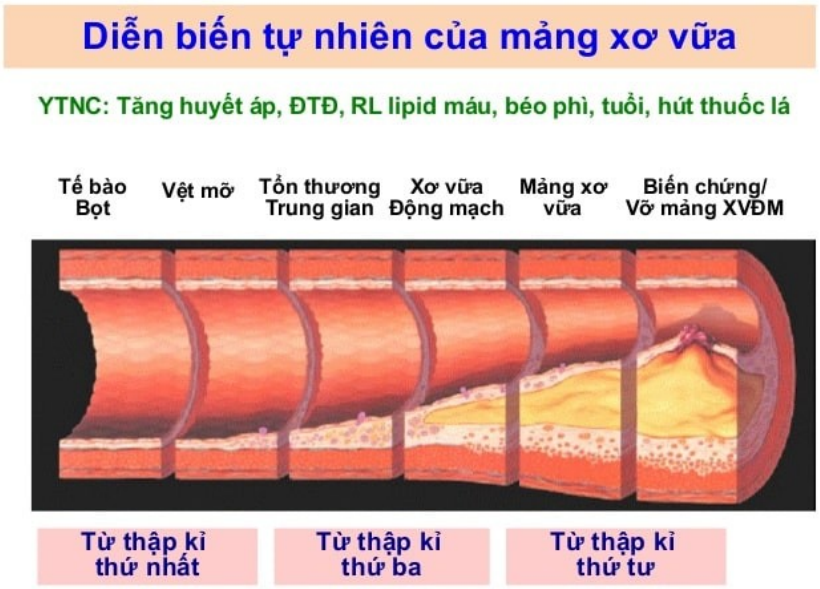
HDL-c là một dạng cholesterol có lợi cho cơ thể, chúng chống lại quá trình xơ mỡ động mạch bằng cách mang cholesterol dư thừa, ứ đọng từ trong thành mạch máu trở về gan . Nên HDL-c còn được gọi là cholesterol tốt hay mỡ tốt.

Chỉ số	Bình thường	Cao
1/Cholesterol toàn phần	<200 mg/dL (< 5,2 mmol/L)	> 240 mg/dL (> 6,2 mmol/L)
2/ LDL-c	< 130 mg/dL (< 3,3 mmol/L)	> 160 mg/dL (> 4,1 mmol/L)
3/HDL-c	> 50 mg/dL (> 1,3 mmol/L)	< 40 mg/dL (< 1 mmol/L)

Bảng 1: Chỉ số các loại cholesterol ở mức thường và cao.

Mảng xơ vữa động mạch hình thành qua nhiều giai đoạn. Ban đầu, lớp nội mạc động mạch bị tổn thương do các yếu tố như LDL cao, cao huyết áp, hút thuốc hoặc viêm nhiễm. LDL thấm vào thành động mạch và bị đại thực bào "nuốt" tạo thành tế bào bọt. Các tế bào bọt tích tụ, hình thành vệt mỡ – dấu hiệu sớm nhất của xơ vữa, có thể xuất hiện từ 8-10 tuổi.

Theo thời gian, tế bào bọt chết đi, kích thích phản ứng viêm và hình thành mảng xơ vữa. Mảng này phát triển với lõi cholesterol mềm và lớp vỏ xơ, có thể thu hẹp lòng động mạch. Nếu lớp vỏ xơ vỡ, quá trình đông máu xảy ra, tạo cục máu đông gây tắc nghẽn, dẫn đến nhồi máu cơ tim hoặc đột quỵ.



Hình 3: Diễn biến tự nhiên mảng xơ vữa.
Nguồn ảnh: Pepine CJ. Am J Cardio, 1998

2.2 Các yếu tố nguy cơ gây bệnh

Các yếu tố nguy cơ của bệnh lý mạch vành rất đa dạng và phức tạp, các yếu tố này bao gồm:

- Các yếu tố nguy cơ không thay đổi được:

  Tuổi càng cao nguy cơ càng tăng lên, ví dụ ở tuổi 70 trở đi, có đến 15% nam giới và 9% nữ giới có bệnh động mạch vành có triệu chứng và tăng lên 20% ở tuổi 80. Giới tính và tình trạng mãn kinh cũng là một yếu tố khách quan. Bệnh động mạch vành thường phổ biến và khởi phát sớm hơn ở nam giới. Tỷ lệ mắc bệnh động mạch vành ở nữ tăng nhanh sau tuổi mãn kinh và sau 65 tuổi ngang bằng với nam giới do suy giảm hormone sinh dục. Đây là nguyên nhân gây tử vong cao nhất ở nữ giới.

  Tiền sử gia đình ở bệnh nhân có xơ vữa động mạch cũng là yếu tố quan trọng khi bệnh xơ vữa động mạch xuất hiện ở thế hệ thứ nhất với nam giới trước tuổi 55 và nữ giới trước tuổi 65.

  Một yếu tố khác đó là yếu tố chủng tộc: Tỷ lệ mắc bệnh động mạch vành thấp

hơn ở nhóm người da đen và có xu hướng gia tăng mạnh ở một số quần thể Đông Á. Tỷ lệ tử vong do bệnh động mạch vành theo tuổi ở nhóm người gốc Nam Á ca hơn 50% so với nhóm người da trắng bản địa ở các nước phát triển.

- Các yếu tố nguy cơ có thể thay đổi được:

  Gia tăng căng thẳng trong công việc, thiếu sự hỗ trợ xã hội, cuộc sống cô đơn và trầm cảm là những yếu tố quan trọng làm tăng nguy cơ xơ vữa động mạch. Nguy cơ mắc bệnh động mạch vành tăng khoảng 50%, kèm theo tỷ lệ tử vong cao hơn 60%, đặc biệt lên đến 85% ở nhóm người nghiện thuốc lá. Hút thuốc lá thụ động cũng làm tăng nguy cơ mắc bệnh khoảng 25%.

  Ngoài ra, béo phì, được đo lường qua chỉ số khối cơ thể (BMI) cao, là nguyên nhân của 25 - 49% trường hợp mắc bệnh động mạch vành tại các quốc gia phát triển. Lối sống ít vận động cũng góp phần đáng kể, trong khi những người thường xuyên hoạt động thể chất lại có nguy cơ mắc bệnh thấp hơn. Việc nghiện rượu bia có mối liên quan chặt chẽ với nguy cơ gia tăng các bệnh tim mạch, tương tự như tình trạng tăng huyết áp - một yếu tố nguy cơ độc lập đối với bệnh động mạch vành, bệnh động mạch ngoại biên và bệnh thận mạn tính.

  Bên cạnh đó, rối loạn lipid máu thể hiện mối liên quan mạnh mẽ, liên tục và độc lập giữa nồng độ cholesterol toàn phần (TC) hoặc cholesterol LDL (LDL-c) với các biến cố tim mạch do xơ vữa. Đặc biệt, đái tháo đường là một yếu tố nguy cơ chính đối với bệnh lý tim mạch do xơ vữa, làm tăng gấp đôi nguy cơ xảy ra biến cố tim mạch (bao gồm bệnh động mạch vành, đột quỵ và tử vong liên quan đến mạch máu), đồng thời vẫn giữ vai trò độc lập so với các yếu tố nguy cơ khác.

2.3 Dấu hiệu và chẩn đoán bệnh động mạch vành

Đau thắt ngực là triệu chứng lâm sàng điển hình của bệnh động mạch vành, triệu chứng này có thể xảy ra lúc vận động mạnh hoặc ngay cả lúc nghỉ ngơi tùy theo mức độ nặng của bệnh. Tuy nhiên có một số bệnh nhân không biểu hiện triệu chứng này (thiếu máu cơ tim thầm lặng). Một số triệu chứng khác đi kèm với đau thắt ngực có thể là hụt hơi, khó thở, buồn nôn, đổ mồ hôi.

Sự bất thường trong nhịp tim và huyết áp giúp đánh giá các yếu tố nguy cơ có thể dẫn đến bệnh mạch vành, đồng thời bước đầu xác định vị trí tổn thương của tim. Ví dụ, thiếu máu cơ tim ở thành dưới sẽ làm chậm nhịp tim do nút nhĩ thất không được cung cấp máu đầy đủ.

Việc chẩn đoán xác định bệnh động mạch vành cần phải dựa rất nhiều vào kết quả cận lâm sàng. Các xét nghiệm cận lâm sàng cung cấp thông tin nhằm chẩn đoán xác định bệnh động mạch vành bao gồm:

- Điện tâm đồ : đo điện tâm đồ lúc nghỉ ngơi và khi vận động được khuyến cáo cho tất cả các trường hợp nghi ngờ mắc bệnh động mạch vành. Điện tâm đồ cung cấp thông tin về sự phì đại cơ tim.

- Chụp cắt lớp động mạch vành có cản quang : dùng máy chụp cắt lớp nhiều dãy để

dựng lại hình ảnh động mạch vành tim, giúp kiểm tra mức độ tắc nghẽn và vôi hóa mạch vành. Chụp động mạch vành qua da : dùng ống thông đưa qua động mạch ở tay, chân đến tận động mạch vành để chụp, phát hiện hoặc can thiệp luôn nếu cần.

- Các test gắng sức : Đo nhịp tim khi đi/chạy trên máy. Điều này giúp xác định xem tim hoạt động như thế nào khi nó phải bơm nhiều máu hơn.

Ngoài các triệu chứng lâm sàng và các kỹ thuật cận lâm sàng trên, việc điều tra về tiền sử bệnh, các bệnh nền và chế độ sinh hoạt của bệnh nhân cũng góp phần rất quan trọng trong chẩn đoán bệnh động mạch vành.

2.4 Cách phòng chống

- Thay đổi lối sống - Khuyến cáo chung trong bệnh lý động mạch vành:

  Thay đổi lối sống là yếu tố quan trọng trong phòng ngừa và kiểm soát bệnh lý động mạch vành. Các khuyến cáo bao gồm loại bỏ các yếu tố gây khởi phát cơn đau tim như lạnh, xúc động, căng thẳng, gắng sức. Người bệnh cần bỏ hút thuốc lá chủ động và tránh hít phải khói thuốc lá thụ động. Một chế độ ăn lành mạnh, cân đối là điều cần thiết, kết hợp với việc hạn chế rượu bia, không uống quá một ly mỗi ngày.

  Bên cạnh đó, cần kiểm soát cân nặng, duy trì chỉ số khối cơ thể (BMI) trong khoảng 18,5 - 22,9 kg/m^2. Tập luyện thể dục thường xuyên, ví dụ như đi bộ nhanh ít nhất 30 phút mỗi ngày, cũng là một biện pháp hữu hiệu. Người bệnh cần tránh tiếp xúc với môi trường ô nhiễm, tiêm phòng cúm định kỳ hàng năm, điều trị các rối loạn tâm lý nếu có, tránh căng thẳng tinh thần và kiểm soát stress.

  Việc phục hồi chức năng cũng rất quan trọng, giúp giảm các yếu tố nguy cơ tim mạch thông qua các chương trình tập luyện phục hồi sau đột quỵ, sau can thiệp động mạch vành, hoặc ở bệnh nhân suy tim có phân suất tống máu dưới 40%.

- Chế độ ăn uống hợp lý:

  Chế độ ăn uống đóng vai trò quan trọng trong việc quản lý và ngăn ngừa bệnh lý tim mạch. Khuyến cáo bao gồm việc bổ sung nhiều rau xanh, trái cây tươi và cá trong khẩu phần ăn hàng ngày. Người bệnh nên cung cấp đủ omega-3 với liều lượng khuyến cáo mỗi ngày từ 850 đến 1000 mg.

  Việc hạn chế mỡ động vật, axit béo bão hòa và lượng cholesterol tiêu thụ hàng ngày dưới 300 mg là điều cần lưu ý. Ngoài ra, cần giảm lượng muối ăn, đảm bảo không tiêu thụ quá 2-3 g muối natri mỗi ngày.

Đối tượng nghiên cứu

1 Đối tượng nghiên cứu

Đối tượng nghiên cứu trong báo cáo này là bộ dữ liệu UCI Heart Disease dataset 2, một bộ dữ liệu công khai được sử dụng rộng rãi trong các nghiên cứu liên quan đến bệnh tim mạch. Bộ dữ liệu bao gồm thông tin từ 920 bệnh nhân được thu thập tại các trung tâm y tế lớn như:

- Phòng khám Cleveland.
- Viện tim mạch Hungary.
- Phòng khám Long Beach.
- Các bệnh viện đại học tại Zurich và Basel, Thụy Sĩ (Switzerland).

Tổng cộng, bộ dữ liệu này bao gồm thông tin của 920 bệnh nhân với 76 thuộc tính được thu thập cho từng mẫu. Trong số đó, chỉ có 14 thuộc tính quan trọng được lựa chọn để sử dụng trong chẩn đoán. Các thuộc tính này chứa các thông tin y khoa quan trọng, giúp xây dựng các mô hình dự đoán và đánh giá tình trạng sức khỏe của bệnh nhân.

Cụ thể, bộ dữ liệu phân loại các bệnh nhân thành hai nhóm chính:

- 411 mẫu mang nhãn 0: Không bị bệnh.
- 509 mẫu mang nhãn từ 1 đến 4: Được chẩn đoán mắc bệnh động mạch vành với các mức độ nặng khác nhau.

Bộ dữ liệu	**Tổng số bệnh nhân**	**Có bệnh**	**Bình thường**
Cleveland	304	165	139
Hungary	293	187	106
Switzerland	123	8	115
VA Long Beach	200	51	149
Tổng	920	411	509

Bảng 2: Thống kê số bệnh nhân trong từng bộ dữ liệu (Đơn vị: người)

2 Mẫu và phương pháp lấy mẫu

Bộ dữ liệu được xây dựng từ các bệnh nhân đến khám tại các trung tâm y tế khác nhau, đảm bảo tính đa dạng về nhân khẩu học, giới tính và tình trạng sức khỏe. Phương pháp thu thập dữ liệu được thực hiện một cách hệ thống thông qua các thiết bị y tế hiện đại như máy đo huyết áp, thiết bị xét nghiệm máu, và các kết quả chẩn đoán hình ảnh. Quá trình lấy mẫu đảm bảo độ chính xác cao, bao gồm các thông số lâm sàng quan trọng như tuổi, giới tính, mức cholesterol, nhịp tim, và kết quả điện tâm đồ.

Dưới đây là các thuộc tính được sử dụng để hỗ trợ trong chẩn đoán, cùng với ý nghĩa của chúng:

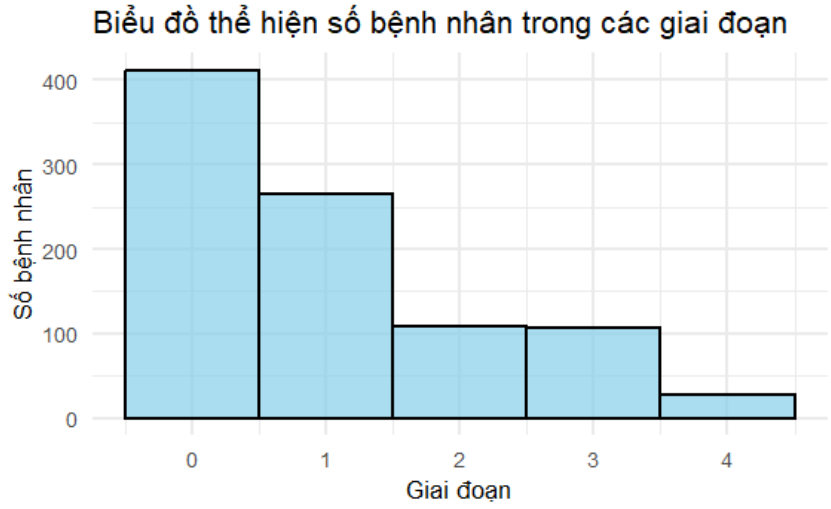
- Age (Tuổi): Tuổi của bệnh nhân.
- Sex (Giới tính): 1 = Male, 0 = Female.
- Cp (Phân loại đau ngực):
  - Typical angina (Đau thắt ngực điển hình) = 0.
  - Atypical angina (Đau thắt ngực không điển hình) = 1.
  - Non-anginal pain (Đau không liên quan đến thắt ngực) = 2.
  - Asymptomatic (Không có triệu chứng) = 3.
- Trestbps (Huyết áp lúc nghỉ ngơi): Đo bằng mmHg.
- Chol (Nồng độ cholesterol): Đo trong huyết thanh (mg/dl).
- Fbs (Lượng glucose huyết tương lúc đói >120mg/dl): 1 = True, 0 = False.
- Restecg (Kết quả điện tâm đồ lúc nghỉ ngơi):
  - Normal (Bình thường) = 0.
  - Having ST-T wave abnormality (Bất thường sóng ST-T) = 1.
  - Showing probable or definite left ventricular hypertrophy (Phì đại thất trái)=2.
- Thalach (Nhịp tim tối đa): Nhịp tim tối đa đo được.
- Exang (Đau ngực khi vận động): 1 = có, 0 = không.
- Oldpeak (Đoạn ST chênh xuống): So với lúc nghỉ ngơi.
- Slope (Độ dốc của đoạn ST):
  - Upsloping (Dốc lên) = 0.
  - Flat (Bằng phẳng) = 1.
  - Downsloping (Dốc xuống) = 2.
- Ca (Số mạch chính): Phát hiện qua nội soi huỳnh quang mạch vành (giá trị từ 0-3).
- Thal (Xạ hình tưới máu cơ tim):
  - Normal (Bình thường) = 0.
  - Fixed defect (Khiếm khuyết cố định) = 1.

  - Reversible defect (Khiếm khuyết có thể hồi phục) = 2.
  - Not described (Không mô tả) = 3.

- Num (Các giá trị phân loại): Thể hiện mức độ nghiêm trọng của bệnh.
  - Giai đoạn 0 : Không bị bệnh
  - Giai đoạn 1 : Giai đoạn nhẹ (Xuất hiện mảng xơ vữa nhỏ, chưa ảnh hưởng nghiêm trọng đến dòng máu)
  - Giai đoạn 2 : Giai đoạn trung bình (Lòng động mạch hẹp đáng kể, gây hạn chế lưu lượng máu đến tim)
  - Giai đoạn 3 : Giai đoạn nặng (Lòng động mạch bị hẹp nghiêm trọng, có nguy cơ cao gây thiếu máu cơ tim)
  - Giai đoạn 4 : Giai đoạn nguy hiểm (Động mạch bị tắc nghẽn gần như hoàn toàn, dẫn đến nhồi máu cơ tim hoặc suy tim)

Bộ dữ liệu này không chỉ cung cấp thông tin đa dạng về bệnh động mạch vành mà còn tạo cơ hội lớn để nghiên cứu, áp dụng các mô hình máy học trong y tế. Qua đó, có thể hỗ trợ chẩn đoán và cải thiện hiệu quả điều trị cho bệnh nhân.

3 Trực quan hóa bộ dữ liệu

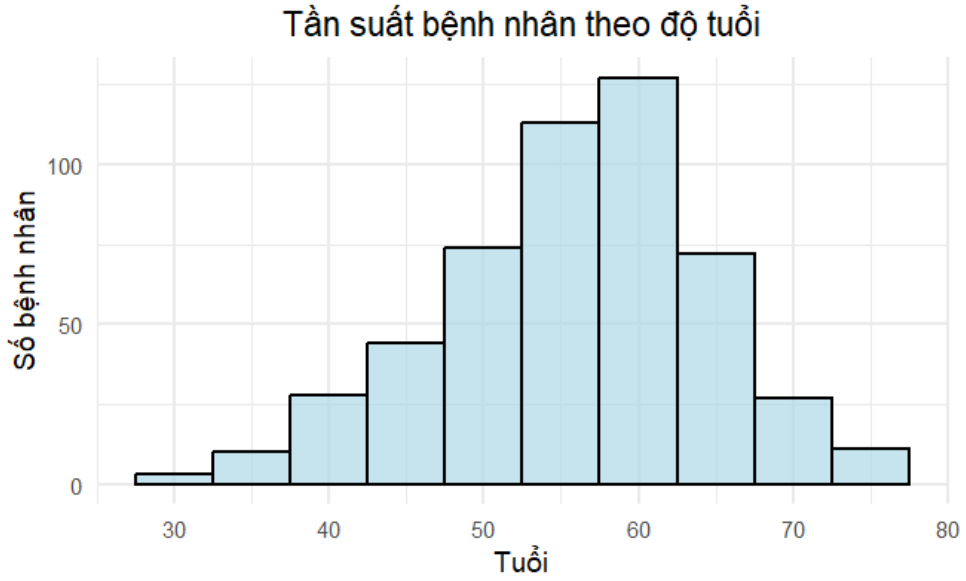


Hình 4: Biểu đồ thể hiện số bệnh nhân trong các giai đoạn

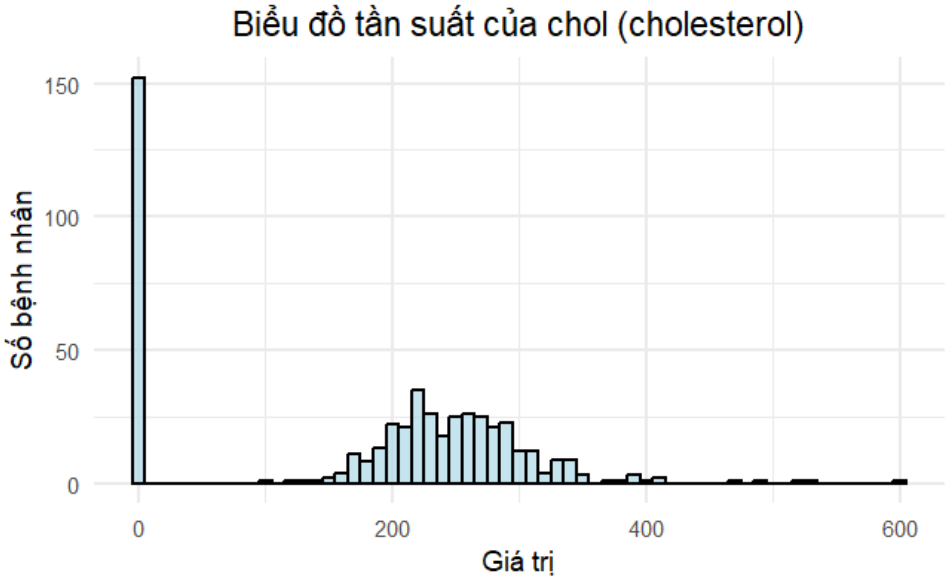
Hình 4 thể hiện số lượng bệnh nhân trong mỗi giai đoạn. Có thể thấy sự rõ ràng sự mất cân bằng trong bộ dữ liệu, với giai đoạn ít bệnh nhân nhất chỉ có 28 mẫu chênh lệch khá nhiều với giai đoạn có nhiều bệnh nhân nhất là 411 mẫu. Và việc mất cân bằng trong dữ liệu này sẽ ảnh hưởng đến khả năng dự đoán của mô hình.

Hình 5: Biểu đồ so sánh tỷ lệ mắc bệnh và không mắc bệnh theo giới tính

Hình 5 cho thấy số lượng bệnh nhân nam nhiều hơn rất nhiều so với bệnh nhân nữ, điều này cho việc mắc bệnh động mạch vành phụ thuộc nhiều vào giới tính (Giới tính nam sẽ có nguy cơ mắc bệnh cao hơn nữ).



Hình 6: Biểu đồ thể hiện số bệnh nhân theo các độ tuổi



Hình 7: Biểu đồ thể hiện tần suất của chol (cholesterol)

Hình 7 cho thấy *chol* có phân phối tương đối chuẩn tuy nhiên có xuất hiện giá trị bất thường $chol = 0$, điều này là không phù hợp với thực tế (Nguyên nhân có thể đến từ việc nhập liệu có sự sai sót hoặc quá trình xét nghiệm chưa có sự giám sát) và cần có biện pháp thay thế các giá trị này (Sẽ được trình bày ở phần sau).

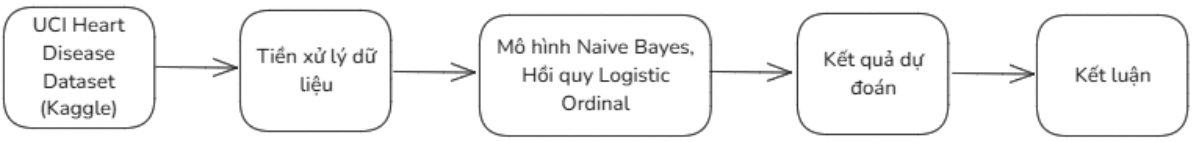
4 Nội dung nghiên cứu

Nội dung nghiên cứu tập trung vào việc phân tích và xây dựng các mô hình dự đoán khả năng mắc bệnh động mạch vành dựa trên 14 thuộc tính y khoa quan trọng được cung cấp trong bộ dữ liệu. Những thuộc tính này bao gồm các yếu tố lâm sàng như tuổi, giới tính, mức cholesterol, nhịp tim tối đa, kết quả đo điện tâm đồ và các dấu hiệu khác liên quan đến sức khỏe tim mạch. Việc lựa chọn 14 thuộc tính quan trọng giúp nghiên cứu tập trung vào các yếu tố có ảnh hưởng lớn đến tình trạng bệnh, từ đó nâng cao hiệu quả của các mô hình dự đoán.

Quy trình thực hiển gồm các bước sau:

- Thu thập dữ liệu: Bộ dữ liệu được lấy tự bộ dữ liệu UCI Heart Disease datase trên trang Kaggle.
- Xử lý dữ liệu: Bộ dữ liệu này bao gồm 14 thuộc tính được dùng để phân loại và chẩn đoán. Các thuộc tính được chia thành hai nhóm chính: dữ liệu dạng phân loại và dữ liệu dạng số.
- Xây dựng mô hình: Các thuật toán được triển khai bằng ngôn ngữ lập trình Python. Dữ liệu phân loại được xử lý bằng mô hình Naive Bayes, một phương pháp thống kê dựa trên định lý Bayes để đưa ra dự đoán. Đồng thời, nghiên cứu sử dụng mô hình hồi quy Ordinal Logistic để xử lý dữ liệu, qua đó so sánh hiệu quả của hai mô hình trên cùng bộ dữ liệu nhằm đánh giá độ chính xác và khả năng ứng dụng của từng phương pháp.

- Kết luận: Đánh giá kết quả của mô hình và bàn luận về ưu, nhược điểm còn tồn tại, hướng phát triển trong tương lai.



Hình 8: Quy trình thực hiện

Phương pháp nghiên cứu

1 Tiền xử lý dữ liệu

1.1 Chuẩn hóa dữ liệu

Dữ liệu tuân theo phân phối chuẩn rất có lợi cho việc xây dựng mô hình, làm cho việc tính toán dễ dàng hơn. Các mô hình như Gaussian Naive Bayes. Logistic Regression, Linear Regression,... được tính toán rõ ràng từ giả định rằng phân phối là chuẩn hai biến hoặc đa biến.

Nhiều hiện tượng trên thế giới tuân theo phân phối log-chuẩn, chẳng hạn như dữ liệu tài chính và dữ liệu dự báo. Bằng cách áp dụng các kỹ thuật biến đổi, chúng ta có thể chuyển đổi dữ liệu thành phân phối chuẩn. Ngoài ra, nhiều quá trình tuân theo tính chuẩn, chẳng hạn như nhiều lỗi đo lường trong thực nghiệm, vị trí của một hạt trải qua quá trình lan tỏa,...

Cho nên, trước khi đi vào việc xây dựng mô hình, việc quan trọng là cần khám phá dữ liệu kỹ càng và kiểm tra các phân phối tiềm ẩn cho mỗi biến.

2 Các phương pháp và cơ sở lý thuyết

2.1 Phân lớp Naive Bayes

Phân lớp Naive Bayes (NB) có cơ sở dựa trên định lý Bayes, hoạt động như một bộ phân lớp xác suất đơn giản có giả thuyết độc lập mạnh mẽ. Nhìn chung, bộ phân lớp Naive Bayes giả định rằng tất cả các thuộc tính là độc lập với nhau, tức là sự thay đổi giá trị của một thuộc tính không ảnh hưởng đến các giá trị thuộc tính còn lại. Các thuộc tính đưa vào mô hình được xem là có ảnh hưởng ngang nhau đối với đầu ra mục tiêu.

Tuy nhiên, trong thực tế, các thuộc tính thuộc lĩnh vực y tế như triệu chứng bệnh và trạng thái sức khỏe (nhịp tim, huyết áp, đường huyết,...) đều có mối tương quan với nhau. Chính vì hai giả thiết gần như không tồn tại trong thực tế nên thuật toán này được gọi là "ngây thơ" (naive).

Thuật toán

- Bước 1: Xét bài toán phân lớp với m class $(C_1, C_2, \ldots, C_m)$. Giả sử có một điểm dữ liệu $X \in \mathbb{R}^n$ $(X = (x_1, x_2, \ldots, x_n))$.
- Bước 2: Tính xác suất để điểm dữ liệu X rơi vào class c.
  - Theo định lý Bayes:

$$P(c|X) = \frac{P(X|c) \cdot P(c)}{P(X)} \tag{1}$$

– Với giả định các thuộc tính là độc lập:

$$P(X|c) = P(x_1|c) \cdot P(x_2|c) \cdots P(x_n|c) = \prod_{i=1}^{n} P(x_i|c) \tag{2}$$

- Bước 3: Sử dụng công thức:

$$c = \arg \max_{1 \le i \le m} P(C_i|X) \tag{3}$$

hoặc tương đương:

$$c = \arg \max_{1 \le i \le m} P(X|C_i) \cdot P(C_i) \tag{4}$$

Mặc dù giả thiết các thuộc tính của dữ liệu độc lập với nhau nếu biết C_i là rất chặt và ít khi tìm được dữ liệu mà các thành phần hoàn toàn độc lập, phân lớp Naive Bayes vẫn hoạt động khá hiệu quả trong nhiều bài toán thực tế và mang lại kết quả tốt.

Gaussian Naive Bayes Mô hình này được sử dụng chủ yếu cho dữ liệu mà các thành phần là các biến liên tục có phân phối chuẩn với kỳ vọng μ_c và phương sai σ_c^2. Công thức xác suất có dạng:

$$P(x_i|c) = \frac{1}{\sqrt{2\pi\sigma_c^2}} \exp\left(-\frac{(x_i - \mu_c)^2}{2\sigma_c^2}\right) \tag{5}$$

Trong đó:

- x_i: Giá trị của đặc trưng thứ i.
- C: Nhãn của lớp.
- μ_{ci}: Giá trị trung bình của đặc trưng x_i trong lớp C.
- σ_{ci}^2: Phương sai của đặc trưng x_i trong lớp C.

Multinomial Naive Bayes Có công thức xác suất được tính theo:

$$P(x_i|C_j) = \frac{N_{ci}}{N_C} \tag{6}$$

Trong đó:

- N_{ci} là số lần xuất hiện của x_i với nhãn C_j.
- N_C là tổng số lần xuất hiện C_j.

Nếu có một x_i chưa từng xuất hiện với nhãn c, công thức trên sẽ bằng 0, dẫn đến xác suất $P(c|X)$ bằng 0 bất kể các giá trị còn lại lớn hay nhỏ. Điều này dẫn đến kết quả không chính xác. Để giải quyết vấn đề này, một kỹ thuật gọi là Laplace smoothing được áp dụng:

$$P(x_i|C_j) = \frac{N_{ci} + \alpha}{N_C + n\alpha} \tag{7}$$

Với α là một số dương, thường bằng 1, để tránh trường hợp $P(c|X)$ bằng 0. Mẫu số được cộng thêm n để đảm bảo $\sum_{i=1}^{n} P(x_i|C_j) = 1$.

Trong dự án này, nhóm đã lựa chọn mô hình Gaussian Naive Bayes thay vì Multinomial Naive Bayes để phân tích và xử lý dữ liệu. Lý do chính xuất phát từ đặc điểm của bộ dữ liệu, trong đó các biến đầu vào chủ yếu là các giá trị số thực. Gaussian Naive Bayes được thiết kế để làm việc hiệu quả với dữ liệu số liên tục, dựa trên giả định rằng các đặc trưng đầu vào tuân theo phân phối chuẩn. Điều này phù hợp với bản chất của bộ dữ liệu hiện tại.

Ngược lại, Multinomial Naive Bayes thường được áp dụng cho dữ liệu rời rạc, ví dụ như số lần xuất hiện của từ trong tài liệu hoặc dữ liệu đếm. Mô hình này giả định rằng các giá trị đầu vào là số nguyên không âm, do đó không phù hợp để xử lý trực tiếp các giá trị liên tục trong bộ dữ liệu này.

Việc sử dụng Gaussian Naive Bayes không chỉ phù hợp với đặc điểm dữ liệu mà còn đảm bảo khai thác tối đa các thông tin từ các biến số thực, giúp cải thiện hiệu quả của mô hình trong các bước phân tích và dự đoán tiếp theo.

2.2 Mô hình hồi quy Ordinal Logistic

- **Khái niệm odds và logit**

  Odds được định nghĩa là tỷ lệ giữa xác suất xảy ra một sự kiện so với xác suất không xảy ra sự kiện đó

$$odds = \frac{P}{1-P} \tag{8}$$

  Tương ứng trong hồi quy Ordinal Logistic: Giả sử Y là kết quả thứ tự với J các danh mục, khi đó odds được xác định là:

$$odds = \frac{P(Y \leq j)}{P(Y > j)} = \frac{P(Y \leq j)}{1 - P(Y \leq j)} \qquad \text{với } j = 1, ..., J-1 \tag{9}$$

  Logit được xác định là *log* cơ số tự nhiên của odds (để mô hình hóa xác suất xảy ra một sự kiện)

$$log\frac{P(Y \leq j)}{P(Y > j)} = logit(P(Y \leq j)) \tag{10}$$

- **Kiểm định Wald**

  Được sử dụng để đánh giá xem từng hệ số của biến trong mô hình có ý nghĩa đối với mô hình hay không

$$\text{Giả thuyết: } \begin{cases} H_0 : \hat{\beta}_i = 0 \\ H_1 : \hat{\beta}_i \neq 0 \end{cases}$$

  Tiêu chuẩn kiểm định:

$$W = \frac{\hat{\beta}_i}{SE(\hat{\beta}_i)} \tag{11}$$

Trong đó:

- $\hat{\beta}$ là hệ số ước lượng của biến.
- $SE(\hat{\beta})$ là sai số chuẩn của hệ số ước lượng.

Giá trị $p-value$ (kiểm định 2 đuôi) được xác định là:

$$\begin{aligned} p-value &= 2\cdot P(Z>|W|)\sim\mathcal{N}(0,1)\\ &= 2\cdot P(Z<-W)\\ &= 2\cdot P(Z>W) \end{aligned} \tag{12}$$

Với mức ý nghĩa $\alpha = 0.05$:

Ta có :

$$\begin{aligned} p-value = 2\cdot P(Z>|W|) &< 0.05\\ \iff \left[\begin{array}{l} 2\cdot P(Z<-W)<0.05\\ 2\cdot P(Z>W)<0.05 \end{array}\right. &\iff \left[\begin{array}{l} P(Z<-W)<0.025\\ P(Z>W)<0.025 \end{array}\right.\\ \iff \left[\begin{array}{l} 1-P(Z<-W)<0.025\\ 1-P(Z>W)<0.025 \end{array}\right.&\\ \iff \left[\begin{array}{l} P(Z<-W)>0.975\\ P(Z>W)>0.975 \end{array}\right.&\\ \iff \left[\begin{array}{l} W<-1.96\\ W>1.96 \end{array}\right. &\text{ (Tra bảng phân phối chuẩn tắc)} \end{aligned}$$

Ta có miền bác bỏ giả thuyết:

- Với $W<-1.96$ hoặc $W>1.96$ ứng với $p-value<\alpha$: Bác bỏ H_0. Tức biến có ý nghĩa đối với mô hình).

- Với $-1.96<W<1.96$ ứng với $p-value>\alpha$: Chấp nhận H_0. Tức biến không có ý nghĩa đối với mô hình).

- **Hồi quy Ordinal Logistic**

Hồi quy Ordinal Logistic được sử dụng để xác định mối quan hệ giữa một tập hợp các biến dự báo và một biến phụ thuộc vào các yếu tố có thứ tự. Được định nghĩa là:

$$logit(P(Y\leq j)) = \beta_{j_0} - \beta_{j_1}x_1 - ... - \beta_{j_p}x_p \tag{13}$$

Trong đó: $\beta_{j_0}, \beta_{j_1}, ..., \beta_{j_p}$ là các hệ số mô hình

- β_{j_0} được gọi là điểm chặn

– $\beta_{j_1}, ..., \beta_{j_p}$ được gọi là độ dốc

Do các điểm chặn là khác nhau đối với mỗi loại nhưng độ dốc không đổi giữa các loại nên phương trình 13 trở thành:

$$logit(P(Y \leq j)) = \beta_{j_0} - \beta_1 x_1 - ... - \beta_p x_p \tag{14}$$

Xây dựng mô hình theo tất cả các biến

Biến	Hệ số	Độ lệch chuẩn	Giá trị W
age	0.28150	0.09364	3.0063
sex	1.03861	0.23271	4.4631
cp	0.62407	0.10278	6.0716
trestbps	0.05056	0.08744	0.5783
chol	-0.55104	0.08275	-6.6594
fbs	0.48154	0.21271	2.2638
restecg	0.20164	0.10444	1.9306
thalch	-0.27037	0.09617	-2.8113
exang	0.66870	0.18294	3.6554
oldpeak	0.60682	0.08746	6.9385
slope	0.18849	0.12095	1.5584
ca	0.75244	0.12489	6.0250
thal	0.10002	0.11742	0.8518

Bảng 3: Hệ số của các biến

Hệ số chặn	Giá trị	Độ lệch chuẩn	Giá trị W
α_0	2.3353	0.3555	6.5694
α_1	4.3003	0.3835	11.2123
α_2	5.2951	0.3960	13.3731
α_3	7.3956	0.4505	16.4181

Bảng 4: Hệ số chặn

Kiểm định hệ số theo kiểm định Wald sẽ loại bỏ một số biến không có ý nghĩa (Biến có giá trị $-1.96 < W < 1.96$) đối với mô hình để làm giảm mức độ phức tạp của mô hình.

Theo quan sát trong bộ dữ liệu có một số bất thường của biến *chol* (Cholesterol) xuất hiện 172 mẫu có giá trị $chol = 0$, điều này không phù hợp trong thực tế (Có thể do việc nhập liệu có sự sai sót). Do số lượng mẫu không quá lớn nên việc bỏ những mẫu có $chol = 0$ gây lãng phí thông tin mà những biến khác mang lại. Vì vậy nhóm sẽ thay thế những giá trị $chol = 0$ bằng cách sinh từ những biến có mối quan hệ với biến *chol*.

Trong quá trình thực hiện, nhóm thấy rằng biến *chol* có mối quan hệ tuyến tính với các biến $sex, trestbps, thalch, ca$ với phương trình là :

$$chol = 0.399 + 0.216 \cdot thalch + 0.2 \cdot ca + 0.112 \cdot trestbps - 0.386 \cdot sex \tag{15}$$

Do đó nhóm sẽ sử dụng phép biểu diễn15 để thay thế những vị trí có $chol = 0$ hoặc $chol = null$.

Xây dựng mô hình sau khi sinh dữ liệu cho biến *chol*

Biến	Hệ số	Độ lệch chuẩn	Giá trị W
age	0.30598	0.09264	3.3030
sex	0.99725	0.22944	4.3465
cp	0.74810	0.10621	7.0438
trestbps	0.03782	0.08382	0.4511
chol	-0.49580	0.08450	-5.8678
fbs	0.38682	0.21484	1.8005
restecg	0.22179	0.10264	2.1609
thalch	-0.26132	0.09610	-2.7193
exang	0.63449	0.18310	3.4653
oldpeak	0.58718	0.08785	6.6842
slope	0.13030	0.11899	1.0951
ca	0.76913	0.12500	6.1530
thal	0.17412	0.11785	1.4775

Bảng 5: Hệ số của các biến

Hệ số chặn	Giá trị	Độ lệch chuẩn	Giá trị W
α_0	2.5959	0.3608	7.1953
α_1	4.6898	0.3928	11.9385
α_2	5.7099	0.4051	14.0956
α_3	7.6564	0.4545	16.8464

Bảng 6: Hệ số chặn

3 Chỉ số đánh giá mô hình

	Dự đoán: Positive	Dự đoán: Negative
Thực tế: Positive	TP (True Positive)	FN (False Negative)
Thực tế: Negative	FP (False Positive)	TN (True Negative)

Bảng 7: Ma trận nhầm lẫn

Ý nghĩa của các chỉ số TP, TN, FP, FN là:

- **TP**: Tổng số trường hợp dự báo khớp Positive
- **TN**: Tổng số trường hợp dự báo khớp Negative
- **FP** : Tổng số trường hợp dự báo các quan sát thuộc nhãn Negative thành Positive
- **FN** : Tổng số trường hợp dự báo các quan sát thuộc nhãn Positive thành Negative

3.1 Accuracy

Accuracy giúp đánh giá hiệu quả dự báo của mô hình trên một bộ dữ liệu. *Accuracy* càng cao thì mô hình càng chính xác.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{16}$$

Trong đánh giá mô hình phân loại thì *Accuracy* khá được ưa chuộng vì nó có công thức tường minh và dễ diễn giải ý nghĩa. Tuy nhiên hạn chế của nó là đo lường trên tất cả các nhãn mà không quan tâm đến độ chính xác trên từng nhãn.

Không phù hợp để đánh giá những tác vụ mà tầm quan trọng của việc dự báo các nhãn không còn như nhau. Hay nói cách khác, việc phát hiện đúng một người mắc bệnh quan trọng hơn việc phát hiện đúng một người thông thường. Do đó cần kết hợp với các chỉ số khác để đánh giá mô hình như *Precision*, *Recall*,...

3.2 Precision

Precision trả lời cho câu hỏi trong các trường hợp được dự báo là positive thì có bao nhiêu trường hợp đúng? Và *Precision* càng cao thì mô hình càng tốt trong việc phân loại.

$$Precision = \frac{TP}{TP + FP} \tag{17}$$

3.3 Recall

Recall đo lường tỷ lệ dự báo chính xác các trường hợp positive trên toàn bộ các mẫu thuộc nhóm positive.

$$Recall = \frac{TP}{TP + FN} \tag{18}$$

3.4 F_1-score

Sự đánh đổi giữa *Precision* và *Recall* khiến cho kết quả của mô hình thường là *Precision* cao, *Recall* thấp hoặc *Precision* thấp, *Recall* cao. Khi đó rất khó để lựa chọn đâu là mô hình tốt vì không biết rằng đánh giá trên *Precision* hay *Recall* sẽ phù hợp hơn. Do đó chỉ số $F_1 - score$ sẽ là sự kết hợp giữa cả *Precision* và *Recall*

$F_1 - score$ là trung bình điều hòa giữa *Precision* và *Recall*. Do đó nó đại diện hơn trong việc đánh giá độ chính xác trên đồng thời *Precision* và *Recall*

$$F_1 - score = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{19}$$

Trong trường hợp $Precision = 0$ hoặc $Recall = 0$ thì $F_1 - score = 0$. $F_1 - score$ luôn nằm trong khoảng của *Precision* và *Recall*.

$$F_1 - score = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \leq \frac{2 \cdot Precision \cdot Recall}{2 \cdot \min(Precision, Recall)} = \max(Precision, Recall)$$

Tương tự:

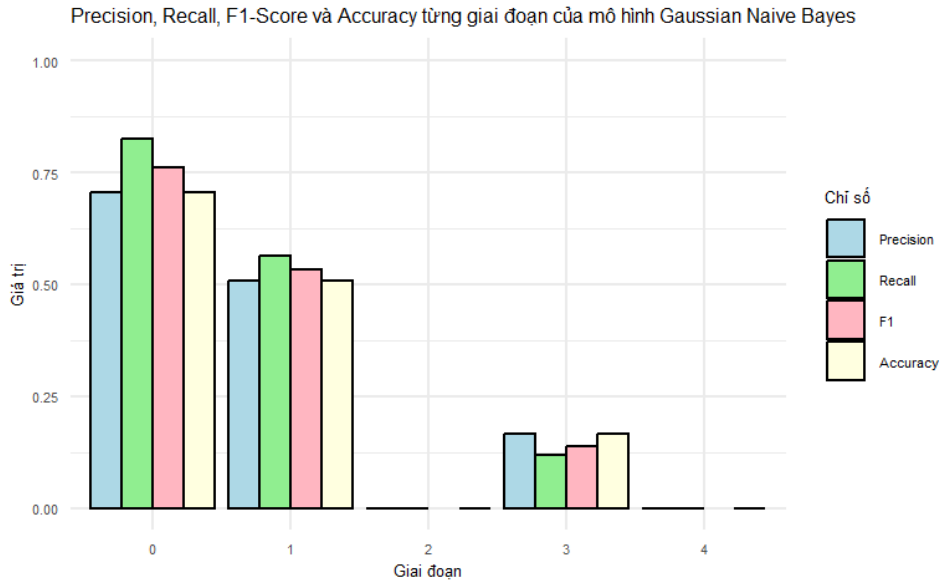
$$F_1 - score = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \geq \frac{2 \cdot Precision \cdot Recall}{2 \cdot max(Precision, Recall)} = min(Precision, Recall)$$

Do đó đối với những trường hợp mà $Precision$ và $Recall$ quá chệnh lệch thì $F_1 - score$ sẽ cân bằng cả hai chỉ số này và giúp đưa ra một đánh giá khách quan hơn.

Kết quả

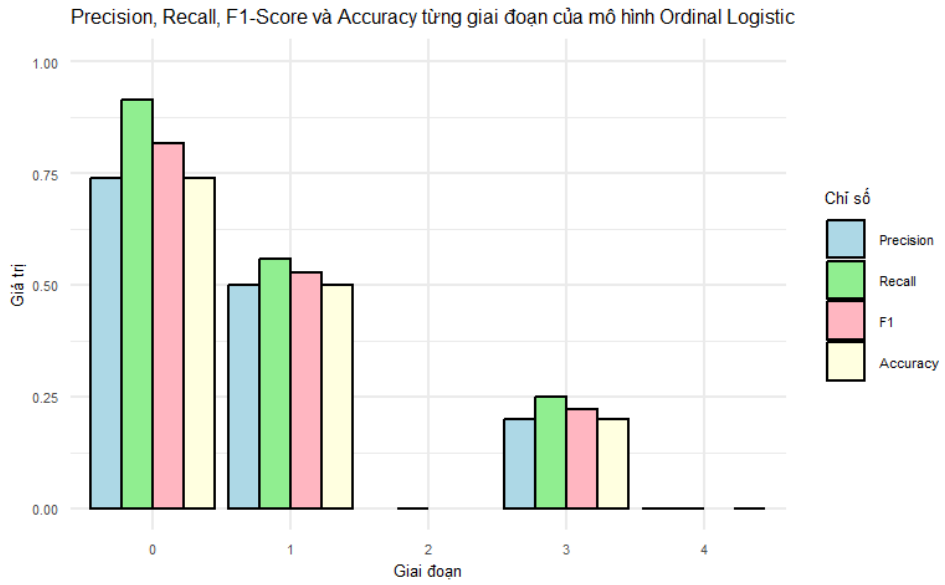
1 Trước khi sinh dữ liệu cho biến chol

1.1 Mô hình Gaussian Naive Bayes



Hình 9: Precision, Recall, F1-score và Accuracy từng giai đoạn của mô hình Gaussian Naive Bayes

1.2 Mô hình Ordinal Logistic



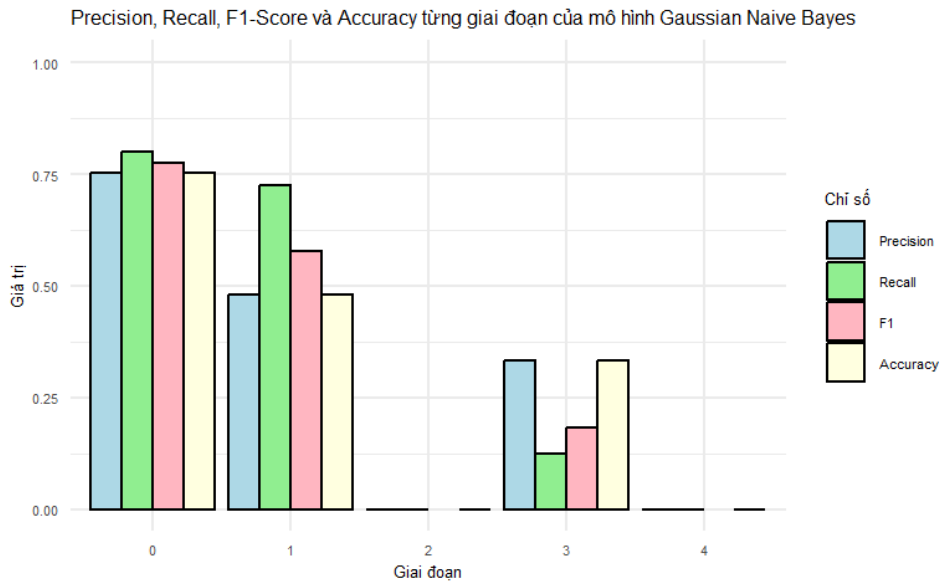
Hình 10: Precision, Recall, F1-score và Accuracy từng giai đoạn của mô hình hồi quy Ordinal Logistic

Từ hình 9 và 10 có thể thấy mô hình Ordinal Logistic có khả năng dự đoán tốt hơn mô hình Gaussian Naive Bayes. Tuy nhiên cả hai mô hình đều không có khả năng dự

đoán giai đoạn 2 (trung bình) và giai đoạn 4(nguy hiểm) có thể do sự mất cân bằng của dữ liệu ban đầu.

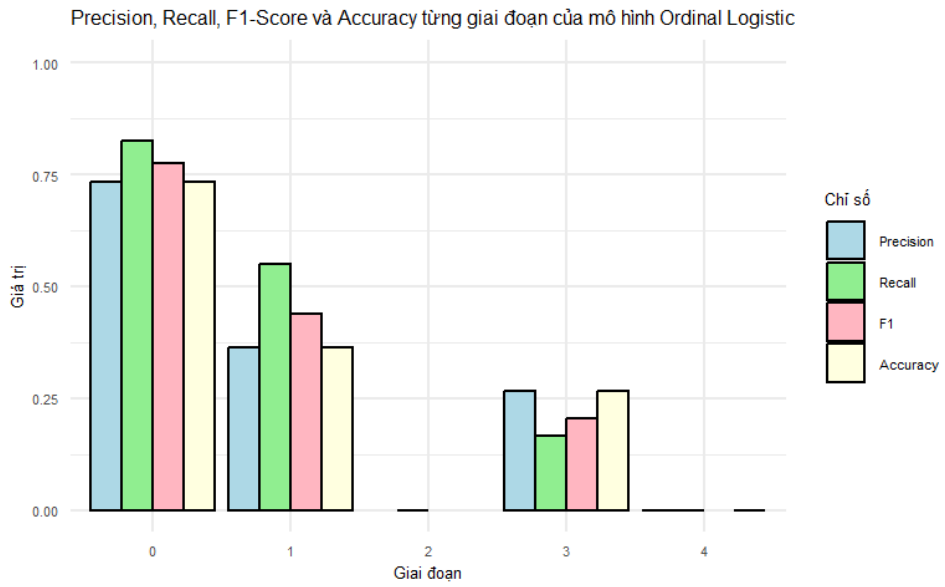
2 Sau khi sinh dữ liệu cho biến chol

2.1 Mô hình Gaussian Naive Bayes



Hình 11: Precision, Recall, F1-score và Accuracy từng giai đoạn của mô hình Gaussian Naive Bayes

2.2 Mô hình Ordinal Logistic



Hình 12: Precision, Recall, F1-score và Accuracy từng giai đoạn của mô hình hồi quy Ordinal Logistic

Từ hình 11 và 12 có thể thấy việc sinh thêm dữ liệu cho biến *chol* giúp cải thiện khả năng dự đoán các giai đoạn của cả hai mô hình, tuy nhiên có sự giảm nhẹ về các chỉ số

đánh giá đối với gian đoạn 0 (không bị bệnh) nhưng đối với các giai đoạn 1 (nhẹ) và 3 (nặng) thì các chỉ số lại tăng, làm cho mô hình tăng được khả năng dự đoán bị bệnh của bệnh nhân. Việc sinh thêm dữ liệu này lại làm cho khả năng dự đoán của mô hình Gaussian Naive Bayes tốt hơn mô hình Ordinal Logistic, ngược lại so với trước dữ liệu ban đầu.

Khả năng dữ đoán của cả hai mô hình trước là sau khi sinh thêm dữ liệu đề chưa có khả năng dự đoán các giai đoạn 2 (trung bình) và 4 (nguy hiểm) có thể do bộ dữ liệu sử dụng là bộ dữ liệu thu thập từ nguồn dữ liệu mở, gồm các bệnh nhân mắc đơn bệnh, tuy nhiên trong thực tế các bệnh nhân có thể có nhiều bệnh cùng lúc, nên từ đó có thể dẫn đến hạn chế trong việc dự đoán của các mô hình.

Kết luận

1 Nhận xét

Mô hình Gaussian Naive Bayes và mô hình Ordinal Logistic dựa vào lý thuyết xác suất Bayes không bỏ qua các dữ liệu khuyết thiếu khi đưa ra kết quả, do vậy ở bước tiền xử lý dữ liệu cần thay thế các dữ liệu bị thiếu bằng cách thay thế giá trị có tần xuất xuất hiện lớn nhất trong thuộc tính hoặc thay thế giá trị trung vị. Việc thay thế này có thể làm ảnh hưởng đến phân bố dữ liệu khiến độ chính xác của mô hình bị giảm đi

Nghiên cứu đã sử dụng bộ dữ liệu UCI Heart Disease dataset là tập hợp của 4 bộ dữ liệu con khác. Trên thế giới, phần lớn các nghiên cứu sử dụng bộ dữ liệu Cleveland thu được trong tập dữ liệu này vì tập dữ liệu Cleveland có ít dữ liệu khuyết thiếu, đây được xem là bước tiền xử lý dữ liệu. Đối với nghiên cứu này, nhóm sử dụng toàn bộ bộ dữ liệu UCI nhằm đánh giá tác động của các dữ liệu khuyết thiếu tới kết quả của mô hình đưa ra.

Nhóm nhận thấy trong nghiên cứu vẫn còn một số hạn chế như: Các mô hình xây dựng vẫn phải xử lý dữ liệu khuyết thiếu bằng cách thay thế như đối với mô hình học máy thông thường mà chưa có sự kiểm định về độ chính xác, phù hợp với thực thế. Việc xử lý mất cân bằng dữ liệu cũng chưa được thực hiện do sự thiếu hiểu biết chuyên môn về bệnh và cũng vì chưa có sự kiểm định về dữ liệu được sinh thêm của những người có chuyên môn.

2 Ý nghĩa

Nghiên cứu có thể được sử dụng và kết hợp để phát triển mô hình giúp phục vụ cho nhân viên ý tế trong việc đưa ra chẩn đoán và quyết dịnh điều trị cho bệnh nhân. Cung cấp công cụ hữu ích với tính cập nhật cao phục vụ trong y tế, xử lý được lượng thông tin lớn một cách nhanh chóng và chính xác.

Việc ứng dụng mô hình không chỉ giúp nâng cao độ chính xác trong chẩn đoán mà còn góp phần vào việc cải thiện chất lượng chăm sóc sức khỏe, giảm thiểu tỷ lệ tử vong và biến chứng liên quan đến bệnh động mạch vành. Trang bị kiến thức về bệnh tim mạch, cách ứng dụng các mô hình vào trong y học, cụ thể là chẩn đoán bệnh lý.

Lời cảm ơn

Chúng em xin chân thành cảm ơn thầy TS.Nguyễn Trọng Hiếu đã dành thời gian quý báu để đọc và đánh giá bài tiểu luận của nhóm chúng em. Tuy đã cố gắng để hoàn thiện bài tốt nhất có thể, song cũng không thể tránh khỏi những thiếu sót, vậy nên những nhận xét, góp ý của thầy không chỉ giúp chúng em hoàn thiện bài làm mà còn giúp chúng em hiểu sâu hơn về vấn đề nghiên cứu cũng như trau dồi thêm kiến thức và kỹ năng học thuật.

Sự hướng dẫn tận tình và sự quan tâm của thầy luôn là nguồn động viên to lớn để chúng em cố gắng học hỏi và phát triển hơn nữa. Một lần nữa, chúng em xin cảm ơn thầy về những bài học không chỉ về kiến trúc sách vở mà còn cả những kiến trúc làm người.

Trân trọng.

Tài liệu

[1] Forthofer, R. N., Lee, E. S., Hernandez, M. (2007). *Logistic and Proportional Hazards Regression.* In Biostatistics (pp. 387–419). https://doi.org/10.1016/b978-0-12-369492-8.50019-4

[2] (N.d.). Kaggle.com. Retrieved December 14, 2024, from https://www.kaggle.com/datasets/redwankarimsony/heart-disease-data/data

[3] James, G., Witten, D., Hastie, T., Tibshirani, R. (2021). *Statistical Learning. In: An Introduction to Statistical Learning. Springer Texts in Statistics*. Springer, New York, NY. https://doi.org/10.1007/978-1-0716-1418-1

[4] Brooks-Bartlett, J. (2018, January 3). *Probability concepts explained: Maximum likelihood estimation. Towards Data Science.* https://towardsdatascience.com/probability-concepts-explained-maximum-likelihood-estimation-c7b4342fdbb1

[5] Ranya N. Sweis, MD - MS - Northwestern University Feinberg School of Medicine; Arif Jivan, MD - PhD - Northwestern University Feinberg School of Medicine (February, 2024). *Overview of coronary artery disease.* https://www.msdmanuals.com/vi/professional/r%E1%BB%91i-lo%E1%BA%A1n-tim-m%E1%BA%A1ch/b%E1%BB%87nh-%C4%91%E1%BB%99ng-m%E1%BA%A1ch-v%C3%A0nh/t%E1%BB%95ng-quan-b%E1%BB%87nh-%C4%91%E1%BB%99ng-m%E1%BA%A1ch-v%C3%A0nh