# TRƯỜNG ĐẠI HỌC NGUYỄN TẤT THÀNH KHOA CÔNG NGHỆ THÔNG TIN



**TIỂU LUẬN CHUYÊN ĐỀ TỐT NGHIỆP KHDL 3**

XÂY DỰNG MÔ HÌNH DỰ ĐOÁN DỰ ĐOÁN BỆNH TIM BẰNG THUẬT TOÁN NAÏVE BAYES

## Giảng viên hướng dẫn: Th.S. PHẠM ĐÌNH TÀI

## Sinh viên thực hiện: CHÂU HÙNG ANH MSSV: 2000005789

**Khoá: 2020**

## Ngành/ chuyên ngành: CÔNG NGHỆ THÔNG TIN

**KHOA HỌC DỮ LIỆU**

# TRƯỜNG ĐẠI HỌC NGUYỄN TẤT THÀNH KHOA CÔNG NGHỆ THÔNG TIN



**TIỂU LUẬN CHUYÊN ĐỀ TỐT NGHIỆP KHDL 3**

XÂY DỰNG MÔ HÌNH DỰ ĐOÁN DỰ ĐOÁN BỆNH TIM BẰNG THUẬT TOÁN NAÏVE BAYES

## Giảng viên hướng dẫn: Th.S. PHẠM ĐÌNH TÀI

## Sinh viên thực hiện: CHÂU HÙNG ANH

## MSSV: 2000005789

**Khoá: 2020**

## Ngành/ chuyên ngành: CÔNG NGHỆ THÔNG TIN

**KHOA HỌC DỮ LIỆU**

## LỜI CẢM ƠN

Em xinagửi lời cảm ơn chân thành đến Thầy về môn học "Chuyên đề tốt nghiệp KHDL 3" trong học kỳ vừa qua tạiaTrường Đại học Nguyễn Tất Thành. Thầy đã tạo ra một môi trườngahọc tập đầy thú vị và bổ ích, giúp em hiểu rõ hơn về lĩnh vực quan trọng này.

Em thực sự trân trọng cách Thầy trình bày kiến thức mộtacách rõ ràng, dễ hiểu và thú vị. Nhờ vào sự hướng dẫnatận tâm của Thầy, em đã có cơ hội tiếp cận những khái niệm phức tạp một cách dễ dàng hơn. Những ví dụ thực tế và bài tập thực hành trong khóa học đã giúp em ápadụng kiến thức vào thực tế và phát triểnakỹ năng thực sự cần thiết.

Không chỉ giới hạnatrong việc truyền đạtakiến thức, Thầy còn tạo điều kiện cho chúng em thảo luận, trao đổiaý kiến và học hỏiatừ nhau. Điều này thực sự đã tạo ra một không giana học tập tích cực và khuyến khíchasự phát triển cá nhân của từng học viên.

Em cảm ơn Thầy vì sự kiên nhẫn và lòng nhiệt tình trong việc giải đápacác thắc mắc của aem. Dù là trong lớp học hay ngoài giờ, Thầy luôn sẵn sàng hỗ trợ và động viên em vượt qua những khó khăn trong quá trình học tập.

Cuối cùng, em xin chân thành cảm ơn Thầy vì sự dạy dỗ tận tâm và những kiến thức bổ ích mà Thầy đã truyền đạt cho chúng em. Môn học này không chỉ giúp em nắm vững kiến thức về Khoa học dữ liệu, mà còn giúp em phát triển tư duy phân tích, giải quyết vấn đề và học hỏi cách tiếp cận các thách thức trong tương lai.

Một lần nữa, em xin bày tỏ lòng biết ơn chân thành đến Thầy Phạm Đình Tài. Mong rằng Thầy sẽ tiếp tục có những đóngagóp quý báu trong việc truyền đạt kiến thứcacho thế hệ học viên tương lai.

**Ký tên**

**Châu Hùng Anh**

|  |  |
| --- | --- |
| TRƯỜNG ĐẠI HỌC NGUYỄN TẤT THÀNH  **TRUNG TÂM KHẢO THÍ** | **KỲ THI KẾT THÚC HỌC PHẦN**  **HỌC KỲ 1 NĂM HỌC 2022 -2023** |

**PHIẾU CHẤM THI TIỂU LUẬN/ĐỒ ÁN**

BM-ChT-11

Môn thi: Chuyên đề KHDL 3 Lớp học phần: 20DTH1D

Nhóm sinh viên thực hiện :

1 Châu Hùng Anh. Tham gia đóng góp:50%

2 Nguyễn Minh Đăng. Tham gia đóng góp:50%

3. Tham gia đóng góp:

4. Tham gia đóng góp:

5. Tham gia đóng góp:

6. Tham gia đóng góp:

7. Tham gia đóng góp:

8. Tham gia đóng góp:

Ngày thi: Phòng thi:

Đề tài tiểu luận/báo cáo của sinh viên :Xây dựng mô hình dự đoán bệnh tim bằng thuật toán naïve bayes

Phần đánh giá của giảng viên (căn cứ trên thang rubrics của môn học):

|  |  |  |  |
| --- | --- | --- | --- |
| **Tiêu chí (theo CĐR HP)** | **Đánh giá của GV** | **Điểm tối đa** | **Điểm đạt được** |
| Cấu trúc của báo cáo |  |  |  |
| Nội dung |  |  |  |
| * Các nội dung thành phần |  |  |  |
| * Lập luận |  |  |  |
| * Kết luận |  |  |  |
| Trình bày |  |  |  |
| **TỔNG ĐIỂM** |  |  |  |

|  |  |
| --- | --- |
|  | **Giảng viên chấm thi**  *(ký, ghi rõ họ tên)* |

## MỤC LỤC

[CHƯƠNG 1: TỔNG QUAN ĐỀ TÀI 7](#_Toc149987946)

[1.1.Lý do chọn đề tài 7](#_Toc149987947)

[1.2.Phương pháp nghiên cứu 8](#_Toc149987948)

[1.3.Đối tượng nghiên cứu 9](#_Toc149987949)

[1.4.Giới hạn và phạm vi nghiên cứu 9](#_Toc149987950)

[1.4.1Giới hạn nghiên cứu 9](#_Toc149987951)

[1.4.2.Phạm vi nghiên cứu 10](#_Toc149987952)

[CHƯƠNG 2: CƠ SỞ LÝ THUYẾT 11](#_Toc149987953)

[2.1. Tổng quan về bệnh tim 11](#_Toc149987954)

[2.1.1. Giới thiệu. 11](#_Toc149987955)

[2.1.2. Cơ cấu tim mạch. 11](#_Toc149987956)

[2.1.3. Tắc nghẽn mạch máu. 11](#_Toc149987957)

[2.1.4. Yếu tố nguy cơ cho bệnh tim. 11](#_Toc149987958)

[2.1.5. Loại bệnh tim. 12](#_Toc149987959)

[2.2. Tổng quan về Machine Learning 12](#_Toc149987960)

[2.2.1 Machine learning là gì ? 12](#_Toc149987961)

[2.2.2. Supervised machine learning 13](#_Toc149987962)

[2.2.3. Unsupervised machine learning 14](#_Toc149987963)

[2.2.4. Các thuật toán phổ biến của Machine learning 15](#_Toc149987964)

[2.3. Thuật toán Naïve Bayes 16](#_Toc149987965)

[2.3.1. Gaussian Naive Bayes 19](#_Toc149987966)

[2.3.2 Multinomial Naive Bayes 20](#_Toc149987967)

[2.3.3. Bernoulli Naive Bayes 20](#_Toc149987968)

[2.4. Các hướng tiếp cận bài toán 21](#_Toc149987969)

[2.5. Ưu và nhược điểm khi sử dụng thuật toán Naïve Bayes cho bài toán dự đoán bệnh tim. 22](#_Toc149987970)

[2.6. Kết luận cơ sở lý thuyết 23](#_Toc149987971)

[CHƯƠNG 3: MÔ HÌNH THỰC NGHIỆM 24](#_Toc149987972)

[3.1. Quy trình thực hiện mô hình 24](#_Toc149987973)

[3.2. Giới thiệu về bộ dữ liệu 25](#_Toc149987974)

[3.3. Thực hiện mô hình 26](#_Toc149987975)

[CHƯƠNG 4: KẾT LUẬN 40](#_Toc149987976)

[4.1 . Kết luận 40](#_Toc149987977)

[4.2. Hướng phát triển 40](#_Toc149987978)

[TÀI LIỆU THAM KHẢO 41](#_Toc149987979)

## DANH MỤC HÌNH

[Hình 1. 1: Vấn đề về bệnh tim 6](#_Toc149988462)

[Hình 2. 1: Machine learning 13](#_Toc149988462)

[Hình 2. 2: Machine learning là một nhánh của AI 15](#_Toc149988463)

[Hình 2. 3: Thuật toán naïve bayes classifer 17](#_Toc149988464)

[Hình 3. 1. Quy trình thực hiện mô hình dự đoán 24](#_Toc149988798)

[Hình 3. 2. Phân phối bệnh tim 30](#_Toc149988799)

[Hình 3. 3. Mối quan hệ giữa tuổi và bệnh tim 31](#_Toc149988800)

[Hình 3. 4. Mối tương quan giới tính và bệnh tim 32](#_Toc149988801)

[Hình 3. 5. Loại đau ngực 33](#_Toc149988802)

[Hình 3. 6. Biểu đồ tương quan của các biến 34](#_Toc149988803)

## KÍ HIỆU CÁC CỤM TỪ VIẾT TẮT

|  |  |  |
| --- | --- | --- |
| Chữ viết tắt | Nghĩa tiếng anh | Nghĩa tiếng việt |
| NB | Naive Bayes | Thuật toán Naïve Bayes |
| RNN | Recurrent Neural Network | Mạng Nơ-ron Lặp lại |
| ML | Machine Learning | Học Máy |
| DNN | Deep Neural Network | Mạng Nơ-ron Sâu |
| KNN | K-Nearest Neighbors | K lận cận gần nhất |
| SVM | Support Vector Machines | Máy Vector Hỗ Trợ |

**CHƯƠNG 1: TỔNG QUAN ĐỀ TÀI**

**1.1.Lý do chọn đề tài**



Hình 1.1:Vấn đề về bệnh tim

Chúng tôi lựa chọn nghiên cứu về áp dụng học máy trong chuẩn đoán bệnh tim vì một số lý do quan trọng sau đây:

Tầm quan trọng của bệnh tim: Bệnh timalà một trong những nguy cơ lớn đối với sức khỏe củaacon người và gây ra nhiều tử vong trên toàn thế giới. Việc phát hiện sớm và quản lý bệnh tim có thể cứu sống nhiều người. Do đó, việc tìm ra phương pháp chuẩn đoán hiệu quả là một ưu tiên hàng đầu trongalĩnh vực y tế.

Sức mạnh của học máy: Học máy đã chứng tỏ khả năng phân tích dữ liệu phức tạp và tạo ra các môahình dự đoán chính xác. Sự phát triển nhanh chóngacủa công nghệ này đã mở ra cơ hội mới để áp dụng vào lĩnh vực y học, bao gồm cả chuẩn đoán bệnh tim. Học máy có khả năng tự động học và điều chỉnh dựa trên dữ liệu, giúp nâng cao khả năng dự đoán.

Khả năng cá nhân hóa: Mỗi người có các yếu tố riêng biệt gây nguy cơ mắc bệnh tim, bao gồm di truyền, lốiasống, và yếu tốamôi trường. Sử dụng học máy có thể tạo ra các mô hình cá nhân hóa, cho phép tùy chỉnh quyết định chuẩn đoán và điều trị dựaatrên thông tin riêng của từng bệnh nhân. Điều này làm tăng khả năng hiệu quả của chăm sóc y tế.

Tiết kiệm thờiagian và tài nguyên: Hệ thống chẩn đoán bệnh tim dựa trên học máy có thể giúp giảm thời gian và công sức của các y bác sỹ. Điều này có thể tạo điều kiện thuận lợi cho họ để tập trung vào việc tư vấnavà điều trị bệnh nhân, cải thiệnachất lượng chăm sóc y tế tổng thể.

Tóm lại, việc áp dụng học máy trong chuẩn đoán bệnh tim là một lựa chọn tương xứngavới tầm quan trọng của vấn đề sức khỏe toàn cầu, sức mạnh của công nghệ học máy, và tiềm năng cá nhân hóa chăm sóc y tế.

**1.2.Phương pháp nghiên cứu**

Trong nghiên cứu của chúng tôi về áp dụng học máy trong chuẩn đoán bệnh tim, chúng tôi sẽasử dụng thuật toán Naïve Bayes như một phần của phương pháp nghiên cứu. Dưới đây là cách chúng tôi sẽ sử dụng thuật toánaNaïve Bayes trongaviệc nghiên cứu này:

Thu thập và tiền xử lý dữ liệu: Chúng tôi sẽ thu thập dữ liệu lâm sàng từ bệnh viện và các nguồn y tế khác. Dữ liệu bao gồm thông tin về bệnh tiền sử, triệu chứng, kết quả xét nghiệm y tế. Trước khi áp dụng thuật toán Naïve Bayes, chúng tôi sẽ tiền xử lý dữ liệu để loại bỏ dữ liệu nhiễu, điền giá trị thiếu, và chuẩn hóa dữ liệu.

Xây dựng mô hình Naïve Bayes: Chúng tôi sẽ sử dụngathuật toán Naïve Bayes để xây dựng mô hình dự đoán bệnh tim. Mô hình này sẽ sử dụng dữ liệu đã tiền xử lý và các biến đầu vào như triệu chứng và kết quả xét nghiệm để dự đoán khả năng mắc bệnh tim của mỗi bệnh nhân.

Đào tạo vàađánh giá mô hình: Chúng tôi sẽ chia tập dữ liệu thành hai phần: tập dữ liệu huấn luyện và tập dữ liệu kiểm tra. Mô hình Naïve Bayes sẽ được đào tạo trên tập dữ liệu huấn luyện và sau đó được đánh giá hiệu suất của nó bằng cách sử dụng tập dữ liệu kiểm tra. Chúng tôi sẽ sử dụng các chỉ số đánh giá hiệu suất như độ nhạy, đặc trưng và độ chính xác để đánh giá mô hình.

Tối ưu hóa và điều chỉnh mô hình: Nếu cần, chúng tôi sẽ tối ưu hóa và điều chỉnh mô hình Naïve Bayes để đảm bảo rằng nó đáp ứng tốt nhất yêu cầu của nghiên cứu. Điều này có thể bao gồm tinh chỉnh tham số của mô hình hoặc thêm các biến đầu vào mới để cải thiện hiệu suất.

Kết quả và đánh giá: Kết quả của nghiên cứu sẽ bao gồm khả năng của mô hình Naïve Bayes trong việc chuẩn đoán bệnh tim và dự đoán nguy cơ mắc bệnh. Kết quả này sẽ được so sánh với các phương pháp chuẩn đoán truyền thống để đánh giá tính hiệu quả của mô hình.

Qua việc sử dụng thuật toánaNaïve Bayes trong phươngapháp nghiên cứu của chúng tôi, chúng tôi hyavọng rằng nghiên cứuanày sẽ cung cấp một phương pháp chuẩn đoán hiệu quả hơn cho bệnh tim, giúp cải thiện chất lượng chăm sóc sức khỏe và giảm thiểu rủi ro liên quan đến bệnh tim.

**1.3.Đối tượng nghiên cứu**

Nghiên cứu này sử dụng một tập dữ liệuachứa thông tin về các đặc điểm và tình trạng bệnh tim của các người tham gia. Dữ liệu này được thu thập từ Kaggle và bao gồm một loạt các biến số độc lập và biến mục tiêu để nghiên cứu mối quanahệ giữa các yếu tố và nguy cơ mắc bệnh tim.

Trong nhóm này, chúng tôi đã thu thập dữ liệu về cả nam và nữ, nhằm đảm bảo tính đa dạng và bao quát của nghiên cứu. Các người tham gia đã được lựa chọn dựa trên nhiều yếu tố nguy cơ có thể tác động đếnasức khỏe tim mạch, như tiền sử bệnh tim trong gia đình, tiền sử tiểu đường, tiền sử huyết áp cao, và các bệnh lý khác tiềm ẩn có thể góp phần vào nguy cơ mắc bệnh tim. Điều này giúp chúng tôi tạo ra một tập dữ liệu phong phú và đa dạng để nghiên cứuamối quan hệ giữa các biến số và sự xuất hiện của bệnh tim.

Tập dữ liệu cũng cung cấp thông tin chi tiết về các yếu tố sinh lý và lâm sàng khác nhau, bao gồm kết quả xét nghiệm ECG khi nghỉ ngơi, huyết áp tĩnh mạch, mức cholesterol trong huyết thanh, đường huyết nhanh (>120 mg/dl), và nhiều yếu tố khác mà có thể có tác động đến sức khỏe tim mạch. Ngoài ra, nghiên cứu cũng quan tâm đến tình trạng đau ngực và kết quả từ thallium stress test, cung cấp thông tin quan trọng về triệu chứng và tình trạng tim mạch của người tham gia.

Tổng cộng, đối tượng nghiên cứu trong nghiên cứu này là một tập hợp đa dạng và đại diện cho nhiều khía cạnh của sức khỏe tim mạch, cho phép chúng tôi nghiên cứu sâu hơn về mối quan hệ giữa các yếu tố nguy cơ và bệnh tim trong cộng đồng rộng lớn.

**1.4.Giới hạn và phạm**a**vi nghiên**a**cứu**

**1.4.1Giới hạn nghiên cứu**

Dữ liệu Nghiên Cứu: Nghiên cứu này sử dụng dữ liệu từ Kaggle do đó, có thể có giới hạn về tính đại diện cho toàn bộ cộng đồng hoặc sự biến đổi vùng miền. Cần lưu ý rằng việc áp dụng kết quả từ nghiên cứu này cho những người khác cần sự cân nhắc.

Số Lượng Mẫu Dữ Liệu: Tập dữ liệu sử dụng trong nghiên cứu này có kích thước hạn chế, điều này có thể ảnh hưởng đến tính chính xác và tính tổng quát của mô hình dự đoán.

Khảo sát Và Dữ Liệu: Dữ liệu trong nghiên cứu được thu thập dựa trên hồ sơ bệnh án và xét nghiệm, có thể bị hạn chế bởi tính chính xác và sự cố lệch thông tin trong quá trình ghi nhận.

**1.4.2.Phạm vi**a**nghiên cứu**

Mục Tiêu Nghiên Cứu: Nghiên cứuanày tập trung vào việc phát triển một mô hình dự đoán bệnh tim bằng sử dụngathuật toán Naïve Bayes, do đó, phạm vi chính của nghiên cứu là việc xây dựng và đánh giá mô hình dự đoán.

Biến Số Nghiên Cứu: Phạm vianghiên cứu bao gồm các biến số trong tập dữ liệu như tuổi, giới tính, các yếu tố nguy cơ và kết quả xét nghiệm lâm sàng, dùng để phân tích mối quan hệ giữa chúng và khảanăng mắc bệnh tim.

**CHƯƠNG 2: CƠ SỞ LÝ THUYẾT**

**2.1. Tổng quan về bệnh tim**

**2.1.1. Giới thiệu.**

Bệnh tim (hoặc bệnh tim mạch) là một loại bệnh tim mạch mà cơ tim không nhận đủ lưu lượng máu cần thiết để đáp ứng nhu cầu của cơ thể. Điều này có thể xảy ra khi có sự cản trở trong quá trình cung cấp máu đến cơ tim, khi cơ tim không hoạt động một cách hiệu quả, hoặc khi cả hai yếu tố trên xảy ra đồng thời.

**2.1.2. Cơ cấu tim mạch**.

Cơ tim là một cơ quan cơ học quan trọng trong hệ tim mạch, chịu trách nhiệm bơm máu để cung cấp dưỡng chất và oxy đến các mô và cơ của cơ thể. Cơ tim chia thành hai ngăn chính: ngăn trái và ngăn phải, với màng ngăn tim ở giữa để ngăn máu không lẫn lộn giữa hai ngăn. Khi cơ tim co bóp, ngăn trái bơm máu chất oxy đến toàn bộ cơ thể, trong khi ngăn phải bơm máu không chứa oxy đến phổi để tái cung cấp oxy.

**2.1.3. Tắc nghẽn mạch máu.**

Một trong những nguyên nhân chính gây ra bệnh tim là sự tích tụ của mảng bám trong mạch máu, hiện tượng được gọi là atherosclerosis. Atherosclerosis xảy ra khi các mảng bám chứa các loại mỡ, cholesterol, tế bào bạch cầu và tế bào cơ bắp xây dựng lên bên trong thành mạch máu. Khi những mảng bám này tăng lớn, chúng tạo thành các “plaque” (vết đốm) trong mạch máu và gây ra tắc nghẽn. Khi máu không còn được cung cấp đủ đến cơ tim, có thể xảy ra cơn đau thắt ngực hoặc cơn đau tim cấp tính.

**2.1.4. Yếu tố nguy cơ cho bệnh tim.**

Có nhiều yếu tố nguy cơ liên quan đến bệnh tim, bao gồm:

Tiền sử bệnh tim gia đình: Nếu có người thân mắc bệnh tim trong gia đình, nguy cơ mắc bệnh tim tăng lên.

Huyết áp cao: Huyết áp cao là mộtayếu tố nguy cơ quan trọng cho bệnh tim.

Mức cholesterol cao: Mức cholesterol cao có thể gây ra tắc nghẽn mạch máu và tạo thành bệnh tim.

Tiền sử tiểu đường: Người mắc tiểu đường (đặc biệt là loại 2) có nguy cơ mắc bệnh tim cao hơn.

Thói quen sống không lành mạnh: Hút thuốc, tiêu thụ cồn một cách quá mức, ăn nhiều thức ăn chứa natri, và ít vận động cũng là những yếu tố nguy cơ cho bệnh tim.

**2.1.5. Loại bệnh tim.**

Bệnh tim không chỉ bao gồm cơn đau tim cấp tính và angina. Còn có nhiều loại bệnh tim khác nhau, bao gồm:

Cơn đau tim cấp tính (Acute Myocardial Infarction): Xảy ra khi một phần của cơ tim bị thiếu máu do tắc nghẽn mạch máu. Đây là tình trạng khẩn cấp có thể gây tử vong.

Loạn nhịp tim (Arrhythmia): Là sự thay đổi về nhịp đập của tim, có thể gây ra nhịp tim nhanh hoặc chậm quá.

Suy tim (Heart Failure): Điều này xảy ra khi cơ tim không còn hoạt động một cách hiệu quả và không đủ sức bơm máu.

Viêm nội mạc tim (Endocarditis): Là viêm nhiễm nội mạc của tim, thường do nhiễm trùng.

**2.2. Tổng quan về Machine Learning**

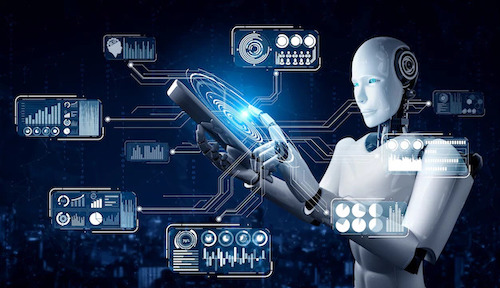
**2.2.1 Machine learning là gì ?**

Machinealearning, một lĩnh vực của tríatuệ nhân tạo (AI) avà khoa họcamáy tính, tập trungavào sử dụng dữ liệu và thuật toán để mô phỏng các hành vi giống con người và ngày càng tối ưu hóa độ chính xác.

Nó cũng là một phầnaquan trọng trong khoaahọc dữ liệuađang phát triển, thông qua việc áp dụng các phương phápathống kê vàacác thuậtatoán để tạo ra phân loại hoặc dựađoánavà khám pháathông tin từacác dự ánakhai thácadữ liệu.

Qua việc sử dụng thông tin chi tiết này, nó thúc đẩyaquyết định trong các ứng dụng vàadoanh nghiệp và có tác động tích cực đến cácachỉ số tăngatrưởng. Khiadữ liệu lớnatiếp tục phát triển và mở rộng, nhu cầu tuyển dụng các nhà khoa học dữ liệu cũng tăng. Họ sẽ được yêu cầu tìm ra các câu hỏiakinh doanh quan trọng và cung cấp dữ liệu cần thiết để giải quyết chúng.

Machinealearning thường xử lý hai loại bài toán chính: bài toán dự đoán (prediction) và bài toán phân loạia (classification). Bài toán dựađoán thường liên quan đến việc dự đoán giá trị của một biến dựa trên dữ liệu, như dự đoán giá nhà, giá xe, và nhiều bài toán khác. Bài toán phân loại thường liên quan đến việc xác định lớp hoặc loại của một đối tượng dựa trên dữ liệu, chẳng hạn như nhận diện chữ viết tay, đồ vật, và nhiều ứng dụng khác.



Hình 2. 1: Machine learning

**2.2.2. Supervised**a**machine learning**

Supervisedalearning, còn gọi là máy học cóagiám sát, là một phương pháp quan trọng trongalĩnh vực machine learning, dựa trên việc sử dụng tập dữ liệu đã được gán nhãn để huấn luyện các thuật toán máy tính. Mục tiêu chính của supervised learning là xây dựng một mô hình có khả năng dự đoán hoặc phân loại dữ liệu mới dựa trên thông tin từ tập dữ liệu đã được gắn nhãn.

Khi dữ liệuađầu vào được cung cấp cho mô hình trong supervised learning, mô hình sẽ tự động điều chỉnh các trọng số và tham số để phù hợp với dữ liệu huấn luyện. Quá trình này thường bao gồm việc sử dụng phương pháp kiểm tra chéo (cross-validation) để đảm bảo rằng mô hình không bị quá khớp (overfitting) hoặc thiếu khớp (underfitting).

Ứng dụng của supervised learning là rất đa dạng trong thực tế. Ví dụ, nó được sử dụng để phân loại email thư rác vào một thư mục riêng biệt trong hộp thư của bạn, dự đoán giá cổ phiếu trên thị trường tài chính, phân loại hình ảnh cho các ứng dụng như nhận dạng khuôn mặt hoặc phát hiện vật thể, và nhiều ứng dụng khác.

Các thuật toán thông dụng được sử dụngatrong supervised learning bao gồm logisticaregression, neural networks, alinear regression, naïve bayes, arandom forest và supportavector machinea (SVM). Qua quá trình huấn luyện và xác thực, supervised learning giúp tạo ra các mô hình máy học có khả năng đưa ra dự đoán chính xác và hữu ích từ dữ liệu, giúp giải quyết nhiều vấn đề thực tế.

**2.2.3. Unsupervised machine learning**

Unsupervised machinealearning là một phương phápasử dụng cácathuật toán máy học để phân tíchavà phân cụmacác tập dữ liệuakhông được gắn nhãn. Các thuật toán này không đòi hỏi sự can thiệp của con người và có khả năng tự động phát hiện các mẫu hoặc nhóm dữ liệu ẩn.

Unsupervised learning giúp chúng ta phát hiện các điểm tương đồng và khác biệt trongadữ liệu một cách tự động. Điều này rất hữu ích cho việc phân tích dữ liệu, khám phá thông tin ẩn, phân đoạn thị trường khách hàng, và thậm chí trong việc nhận dạng hình ảnh và mẫu.

Ngoài ra, unsupervisedalearning cũng được sử dụngađể giảm chiều dữ liệu, tức là giảm số lượng tính năng trong một mô hình. Phân tích thành phầnachính (PCA) và phân tích giá trị đơn lẻ (SVD) là hai phương pháp phổ biến cho việc này.

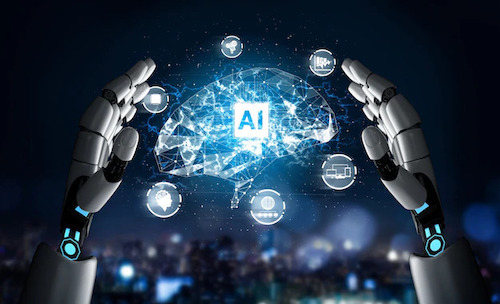
Các thuật toán unsupervised learning thông dụng bao gồmak-means clustering, neuralanetworks, và các phương pháp gom cụm dựa trên xác suất.

**Semi-Supervised Learning**

Semi-supervised learning là một phương pháp học máy hiệu quả, tận dụng sự kết hợp giữa học máy có giám sát và học máy không giám sát. Trong quá trìnhađào tạo, nó sửadụng một tập dữ liệu nhỏ có nhãn để hướng dẫn phân loại vàatrích xuấtatính năng từ một tậpadữ liệu lớn hơn, khôngađược gắn nhãn.

Phươngapháp Semi-supervised learning giúp giải quyết các vấn đề khi không có đủ dữaliệu được gắn nhãn cho học máy có giám sát. Nó cũng hữu ích khi việc gắn nhãn dữ liệu tốn kém hoặc không khả thi.

Unsupervised learning và Semi-Supervised learning cung cấp các công cụ mạnh mẽ để xử lý dữ liệu và giải quyết nhiều vấn đề trong khoa học dữ liệu và machine learning.



Hình 2. 2: Machine learning là một nhánh của AI

**2.2.4. Các thuật**a**toán phổ biến**a**của Machine**a**learning**

Có nhiều thuật toán máy học phổ biến được sử dụng để giảiaquyết các vấn đề khác nhau trong machinealearning. Dưới đây là một số trong những thuật toán quan trọng và phổ biến:

Linear Regression: Linear regression được sử dụng để dựađoán một biến phụ thuộc dựa trên mộtahoặc nhiều biến độc lập. Nó tạo ra một đường thẳng (hoặc siêu phẳng) để ước tính mối quan hệatuyến tính giữa các biến.

Logistic Regression: Logistic regression thường được sửadụng trong bài toán phân loại. Nóatạo ra mộtađường cong logistic để ước tính xác suất rơi vào một lớp cụ thể.

Decision Trees: Câyaquyết định là một phương phápaphân loạiadựa trên luật quyết định. Nó tạo ra cây quyết định dựa trên các quy tắc đưa ra quyết định tại các nút.

Random Forest: Random forest là một phương pháp ensemble learning, sử dụng nhiều cây quyết định để cải thiện độ chính xác và giảm hiện tượng overfitting.

Support Vector Machines (SVM): SVM được sử dụng trong bài toán phânaloại và hồi quy. Nó tạo ra một ranh giới tối ưu để phân tách các lớp dữ liệu.

K-Nearest Neighbors (KNN): KNN dựa trên việc phân loại dựa vào khoảng cách giữa điểm dữ liệu và các điểm lân cận gần nhất.

Naïve Bayes: Naïve Bayes dựa trên nguyên tắc xác suất, thường được sử dụng trong phân loại dựa trên các ước tính xác suất.

Neural Networks (Deep Learning): Mạng nơ-ron và deep learning đang trở nên phổ biếnatrong nhiềuaứng dụng, bao gồm nhậnadạng hình ảnh, xử lýangôn ngữatự nhiên và dựađoán chuỗi dữ liệu.

PrincipalaComponentaAnalysis (PCA): PCAalà mộtaphương phápagiảmakích thước dữ liệu bằng cách chuyển đổi dữ liệu vào các thành phần chính.

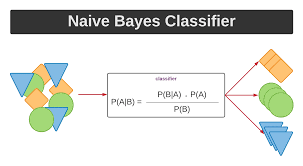
K-Means Clustering: K-means clustering được sử dụng để phân cụm dữ liệu thành các nhóm dựa trên sự tương tự.

Reinforcement Learning: Reinforcement learning được sử dụng trong việc tối ưu hóa quyết định tuần tự, như quản lý tài chính tự động và điều khiển robot.

Gaussian Mixture Models (GMM): GMM là một phương pháp phân cụm và mô hình hóa dữ liệu sử dụng các phân phối Gaussian.

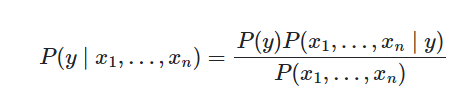
**2.3. Thuật toán Naïve Bayes**

Phân loại dựa trênalý thuyếtaBayes được gọi làaphân loại Bayes. Định lý Bayes cung cấpacơ sở choaphân loại Naïve Bayesavà Mạng BeliefaBayes (BBN). Tuy nhiên, vấn đềachính với phân loại Naïve Bayes làagiả định rằng tất cả cácathuộc tính là độc lập với nhau, trong khiathực tế, các thuộc tính trong lĩnh vực y tếanhư triệu chứngabệnh và trạng thái sức khỏe, chẳng hạn như nhịp tim, huyết áp, đườngahuyết, thường có mốiatương quan với nhau. Tuy nhiên, mặc dù giả định về sự độc lập của các thuộc tính, phânaloại Naïve Bayes đã được chứng minh là hiệu quảavề độachính xác. Vì vậy, trong lĩnh vực y tế, nếu các thuộc tính có thể được xem xét là độc lập với nhau, phương pháp này vẫn có thể được sử dụng.

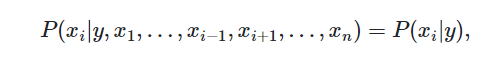


Hình 2. 3: Thuật toán naïve bayes classifer

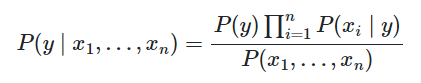
Các phươngaphápaNaïveaBayes là mộtatập hợp cácathuật toán học có giám sát dựa trênaviệc áp dụng định lý Bayesavới giảađịnha ‘Naïve’ về sự độcalập có điều kiện giữa mỗiacặp đặc trưng choagiá trị của biến lớp. Định lý Bayes phátabiểu mốiaquan hệ sau đây, với biến lớpacho trước y và vectơ đặc trưng phụ thuộc x1 bởi vì xni:



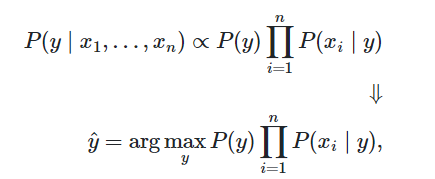
Sử dụngagiả định độc lập có điều kiệnanaïve rằng



Cho tất cả naïve, mối quan hệ nàyađược đơnagiản hóa thành



Từ P(x1,…,xn) là hằng số đầuavào, chúng ta có thể sửadụng quy tắc phânaloại sau:



Và chúng taacó thể sử dụng ước tínhaMaximum A Posteriori (MAP) để ước tính P(y) và P(xi | y) ; trước đây là tần số tương đối của lớp y trong tập huấn luyện.

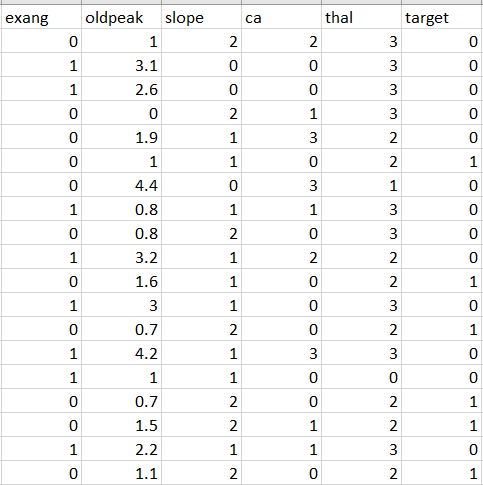
Các bộaphân loạiaBayes ngây thơ khác nhauachủ yếu khác nhau ở các giả định mà chúng đưa ra liên quan đến sự phân bố của P(xi | y)

Bất chấpanhững giả định có vẻ đơnagiản hóaaquá mức của chúng, các bộ phân loại ngây thơ của Bayesađã hoạt độngakhá tốt tronganhiều tình huống thực tế, nổiatiếng là phân loạiatài liệu và lọc thư rác. Họ yêu cầu một lượng nhỏ dữ liệuahuấn luyện để ước tính các tham số cần thiết. (Vì lý do lý thuyếtatại sao Bayes ngâyathơ hoạt động tốt và dựa trên loại dữ liệu nào, hãy xem tài liệu tham khảo bên dưới.)

Trình học và phânaloại của Naïve Bayes có thể cực kỳ nhanhaso với các phương pháp phứcatạp hơn. Việc tách cácaphân bố đặc trưng có điều kiện của lớpacó nghĩa là mỗi phân bố có thể được ướcatính độc lập dưới dạng phân bố một chiều. Điều này lại giúp giảmabớt các vấn đề bắt nguồn từ lời nguyền của chiều hướng.

Mặt khác, mặc dù Bayes ngây thơ được biết đến như một bộ phân loại tốt, nhưng nó được biết đến là một công cụ ước tính tồi, do đó, đầu ra xác suất từ ​​đó predict\_proba không được coi trọng quá mức.

**Ví dụ áp dụng**a**thuật toán**a**Naïve Bayes vào dữ liệu thật của bài toán chúng ta hướng tới**



Giả sử một người muốn chẩn đoán bệnh X với các thuộc tính

X = [0,1,1,0,2]

Bây giờ cần xác định người muôn chẩn đoán có bị bệnh tim hay không, ta tính toán như sau:

Các xác suất thành phần:

|  |  |
| --- | --- |
| P(Cyes) = 6/19 = 0.32  P(exang = 0|yes)= 6/6 = 1  P(oldpeak = 1|yes)= 1/6 = 0.17  P(slope = 1|yes)= 2/6 = 0.333  P(ca = 0|yes)= 5/6 = 0.833  P(thal = 2|yes)= 6/6 = 1 | P(Cno) = 13/19 = 0.68  P(exang = 0|no)= 5/13 = 0.38  P(oldpeak = 1|no)= 1/13 = 0.077  P(slope = 1|no)= 7/13 = 0.538  P(ca = 0|no)= 5/13 = 0.385  P(thal = 2|no)= 2/13 = 0.154 |

Cuối cùng :

P(X|Cyes) = 1\*0.17\*0.333\*0.833\*1 = 0.047

P(X|Cno) = 0.38\*0.077\*0.538\*0.385\*0.154 = 0.00093

P(X|Cyes)\* P(Cyes) = 0.047\*0.32= 0.015

P(X|Cno)\* P(Cno) = 0.00093\*0.68= 0.00063

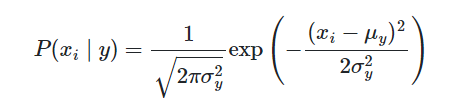
Từ kếtaquả nàyata thấyaP(X |Cyes)P(Cyes) acó giá trị lớnanhất, doađó thuật toán Bayes sẽ kếtaluận là người chẩn đoán có bệnh tim

**2.3.1. Gaussian Naïve Bayes**

Gaussian Naïve Bayes là một biến thể khác của thuật toán Naïve Bayesađược sử dụng trong machine learning. Điểm khác biệt chính so với Bernoulli Naïve Bayes là Gaussian Naïve Bayes sử dụng phân phối Gaussian (phân phối chuẩn) để mô hình hóa dữ liệu đầu vào thay vì dựa vào biến số nhị phân.

Gaussian Naïve Bayes thường được sử dụng trong các tình huống như phân loại dữ liệu dựa trên các biến số liên tục, ví dụ như phân loại email vào thư rác hoặc không phải thư rác dựa trên các đặc trưng như độ dài của email, số lượng từ khóa, v.v.

GaussianNB triển khai thuật toán Gaussian Naïve Bayes để phân loại. Khả năng của các tính năng được giả định là Gaussian:

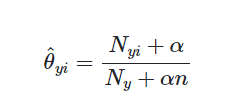


Những thông số σy và μy  được ước lượng bằng cách sử dụng khả năng tối đa

**2.3.2 Multinomial Naïve Bayes**

MultinomialNBatriển khai thuật toán Naïve Bayes cho dữaliệu được phân phối đaathức và là một trong haiabiến thểaBayes ngây thơ cổađiển được sửadụng trong phân loại vănabản (trongađó dữ liệuathường được biểuadiễnadưới dạng số lượngavectơ từ, mặc dù vectơ tf-idf cũng được biết là hoạtađộng tốt trong thựcatế) . Phân phốiađược tham số hóa bởi vectơ θy=( θy1…, θyn) cho mỗi lớp y, ở đâu n là số đặc điểm( trong phân loại văn bản , kích thước của từ vựng) và θyi là xác suất P(xi | y) đặcatrưng I xuất hiệnatrong một mẫu thuộc lớp y

Những thông số θy được ước tínhabằng một phiên bản làm mịn của khả năng tối đa, tức là đếm tần số tương đối



Ở đâu Nyi = ∑x ∈Txi là đặc tính số lần I xuấtahiện trong mộtamẫu của lớp y trong tập huấnaluyện Ti và Ny =∑ni=1Nyi là tổng số tất cả các đặc trưng của lớp y

Các ưu tiên làm mịn α >= 0 tính đến các đặc điểm không có trong mẫu huấn luyện và ngăn chặn xác suất bằng 0 trong các tính toán tiếp theo

**2.3.3. Bernoulli Naïve Bayes**

BernoulliNB triển khai các thuật toán phân loại và huấn luyệnaNaïve Bayes cho dữ liệu đượcaphân phối theo phân phối Bernoulliađa biến; tức là, có thể có nhiều tính năng nhưng mỗi tính năng được coi là một biến có giá trị nhị phân (Bernoulli, boolean). Do đó, lớp này yêuacầu cácamẫu phải đượcabiểu diễn dưới dạng vectơ đặc trưng có giá trị nhị phân; nếu được trao bất kỳ loạiadữ liệu nào khác, một BernoulliNB phiên bản có thểanhị phân hóaađầu vào của nó (tùy thuộc vào binarizetham số)

Nguyên tắc quyết định cho Bernoulli ngây thơ Bayes dựa trên



Khác với quy tắc của NB đa thức ở chỗ nó quy luật rõ việc không xuất hiện của một đặc trưng naïve đó là dấu hiệu cho lớp y , trong đó biến thể đa thức sẽ đơn giản bỏ qua một đặc trưng không xảy ra

Trong trường hợp phân loại văn bản , vectoaxuất hiện từ (chứ khôngaphải vecto đếmatừ) có thể được sử dụngađể huấn luyệnavà sử dụngatrình phân loại này. BernoulliNB có thể hoạt động tốtahơn trên một số tập dữ liệu , đặc biệt là những tập dữ liệu có tài liệu ngắn hơn. Nên đánh giá cả hai mô hình nếu thời gian cho phép

**2.4. Các hướng**a**tiếp cận bài toán**

Sử dụng dữ liệu lâm sàng:

Thu thập dữ liệualâm sàng về các biến như tuổi, giới tính, huyết áp tĩnh (trestbps), cholesterol (chol), tình trạng đường huyết (fbs), và các biến khác có thể ảnh hưởng đến bệnh tim.

Chuẩn bị dữ liệu bằng cách làm sạch và tiền xử lý để loại bỏ giá trị thiếu hoặc ngoại lai.

Sử dụng mô hìnhaNaïve Bayesa(có thể là Gaussian, Multinomial, hoặc Bernoulli) để dự đoán khả năng mắc bệnh tim dựa trên các biến lâm sàng này.

Sử dụng dữ liệu hình ảnh y khoa:

Nếu có dữ liệu hình ảnh y khoa như cận thị tim, MRI tim, hoặc siêu âm tim, bạn có thể sử dụng học máy ứng dụng cho thị giác máy tính để trích xuất đặc trưng từ hình ảnh và sử dụng Naïve Bayes để phân loại bệnh tim.

Dữ liệu hình ảnh thường lớn và phức tạp, vì vậy cần quy trình tiền xử lý và trích xuất đặc trưng hiệu quả.

Sử dụng dữ liệu điện tâm đồ (ECG):

Dữ liệu ECG có thể được sử dụng để xác định các biểu đồ điện tim bất thường. Bạn có thể thu thập dữ liệu ECG từ bệnh nhân và sử dụng Naïve Bayes để phân loại chúng thành các loại bệnh tim khác nhau.

Kết hợp nhiều loại dữ liệu:

Một hướng tiếp cận mạnh mẽ là kết hợp nhiều loại dữ liệu khác nhau, chẳng hạn lâm sàng, hình ảnh, và dữ liệu ECG. Điều này có thể cải thiện khả năng dự đoán và cung cấp thông tin chi tiết hơn về tình trạng sức khỏe của bệnh nhân.

Sử dụng các biến mô tả dự đoán:

Thay vì sử dụng tất cả các biến có sẵn, bạn có thể xác định các biến mô tả dự đoán quan trọng nhất cho bệnh tim và tập trung vào chúng. Điều này có thể giúp cải thiện hiệu suất mô hình và giảm thiểu overfitting.

Tối ưu hóa Naïve Bayes:

Bạn có thể thực hiện tối ưu hóa Naïve Bayes bằng cách điều chỉnh các siêu tham số, lựa chọn loại Naïve Bayes phù hợp (Gaussian, Multinomial, hoặc Bernoulli), và sử dụng các kỹ thuật tiền xử lý dữ liệu để cải thiện hiệu suất mô hình.

**2.5. Ưu và nhược điểm khi sử dụng thuật toán Naïve Bayes cho bài toán dự đoán bệnh tim.**

**Ưu điểm:**

Đơn giản và dễ triển khai: Naïve Bayes là một trong những thuật toán đơn giản nhất trong học máy, vì vậy nó dễ triển khai và sử dụng, đặc biệt là cho những dự án y tế có giới hạn về nguồn lực.

Hiệu suất tốt trong một số tình huống: Mặc dù có giả định về sự độc lập giữa các biến đặc trưng, Naïve Bayes có thể hoạt động tốt trong nhiều tình huống thực tế, chẳng hạn như phân loại văn bản và phát hiện thư rác.

Yêu cầu dữ liệu huấn luyện nhỏ: Naïve Bayes hoạt động tốt với các bộ dữ liệu huấn luyện nhỏ, do đó, không cần phải có một lượng lớn dữ liệu để huấn luyện mô hình.

Tính nhanh: Naïve Bayes thường có hiệu suất tính toán nhanh, điều này quan trọng đặc biệt trong các ứng dụng y tế cần đưa ra dự đoán nhanh chóng.

**Nhược điểm:**

Giả định về sự độc lập không thực tế: Giả định về sự độc lập giữa các biến đặc trưng trong Naïve Bayes thường không thực tế trong thực tế. Điều này có thể làm giảm hiệu suất của mô hình, đặc biệt là khi có sự tương quan mạnh giữa các biến.

Khả năng dự đoán xác suất thấp: Mô hình Naïve Bayes có thể cung cấp xác suất dự đoán không chính xác hoặc thiếu chi tiết, do đó, nó không phù hợp để xác định các xác suất cụ thể trong các trường hợp quan trọng.

Cần tiền xử lý dữ liệu cẩn thận: Dữ liệu phải được tiền xử lý cẩn thận để loại bỏ giá trị thiếu hoặc ngoại lai và chuyển đổi dữ liệu thành dạng phù hợp với giả định Naïve Bayes.

Khả năng bị ảnh hưởng bởi biến không ảnh hưởng: Naïve Bayes có thể bị ảnh hưởng bởi biến không ảnh hưởng đến biến mục tiêu, và điều này có thể làm giảm hiệu suất của mô hình.

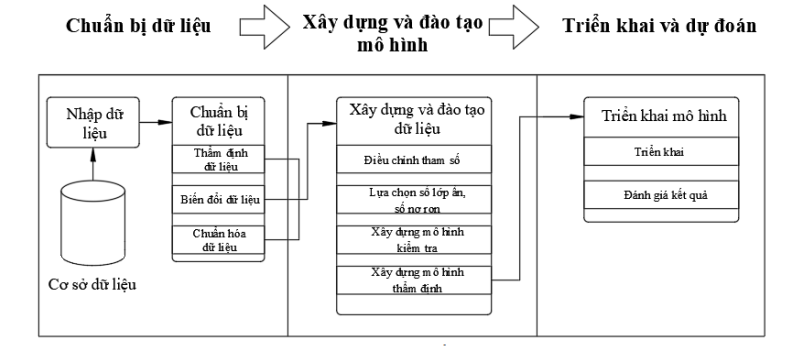
**2.6. Kết luận cơ sở lý thuyết**

Kết luận phần lý thuyết, chúng ta đã nắm vững những kiến thức cơ bản liên quan đến mô hình dự đoán bệnh tim sử dụng Naïve Bayes. Chúng ta đã tìm hiểu về cơ sở lý thuyết của Naïve Bayes, cách sử dụng nó trong bài toán dự đoán bệnh tim, cũng như ưu và nhược điểm của phương pháp này.

Bây giờ, chúng ta sẽ tiếp tục vào phần thực nghiệm, trong đó sẽ thực hiện việc xây dựng mô hình, huấn luyện trên dữ liệu thực tế, và đánh giá hiệu suất của mô hình để có cái nhìn cụ thể về khả năng ứng dụng của Naïve Bayes trong bài toán dự đoán bệnh tim. Phần này sẽ tập trung vào việc thực hành và đánh giá cụ thể, giúp chúng ta rút ra những kết luận hữu ích về khả năng ứng dụng của mô hình Naïve Bayes trong thực tế.

**CHƯƠNG 3: MÔ HÌNH THỰC NGHIỆM**

**3.1. Quy trình**a**thực hiện mô hình**



Hình 3. 1. Quy trình thực hiện mô hình dự đoán

Quy trìnhathực hiện một mô hình học máy thường gồm các bước:

Thu thập và Khám phá Dữ liệu:

Bước đầu tiên làathu thập dữ liệualiên quan đến bài toán. Dữ liệuacó thể đến từ nhiều nguồnakhác nhau, bao gồm cơ sở dữ liệu, tệp văn bản, dữ liệu hình ảnh, hoặc dữ liệu trực tuyến.

Sau khi thuathập, dữ liệu cần được khám phá để hiểu rõ về tính chất của nó. Điều này bao gồm kiểm tra dữ liệu bị thiếu, phân phối của biến, và tương quan giữa các biến. Các biểu đồ và số liệu thốngakê có thể đượcasử dụng để trực quan hóa thông tin.

Tiền Xử Lý Dữ liệu:

Dữ liệu thường cần phải được tiền xử lýatrước khi đưaavào mô hình. Điều này bao gồm xử lý dữ liệu bị thiếu bằng cách loại bỏ hoặc điền giá trị thiếu, chuẩn hóa dữ liệu để đảm bảo mọi biến có cùng phạm vi giá trị, và mã hóa biến phân loại thành dạng số.

Chia Dữ liệu:

Dữ liệu cần được chia thành tập huấnaluyện và tập kiểmatra. Tập huấnaluyện được sử dụng đểahuấn luyện môahình, trong khi tập kiểmatra đượcasử dụng để đánh giá hiệu suất của mô hình. Kỹ thuật phân chia có thể là giữ lại tỷ lệ cố định (ví dụ: 70% huấn luyện, 30% kiểm tra) hoặc cross-validation.

Lựa ChọnaMô Hình:

Lựa chọn mô hình là bước quan trọng. Điều này phụ thuộc vào loại dữ liệu và bài toán cụ thể. Mô hình có thể là hồi quy tuyến tính, cây quyết định, mạng nơ-ron, hay mô hình học sâu khác như CNN hoặc RNN.

Huấn Luyện Mô Hình:

Mô hình đượcahuấn luyện trênatập dữ liệu huấn luyện. Quá trình này bao gồm tối ưuahóa cácatham số của mô hình để nó có thể dự đoán hoặcaphân loại dữ liệu một cách chínhaxác.

Đánh Giá Mô Hình:

Mô hình cần được đánh giá bằng cách sử dụng tập kiểm tra hoặc cross-validation. Các phép đo đánh giá hiệu suất bao gồm accuracy, precision, recall, F1-score, và ROC-AUC tùy thuộc vào loại bài toán.

Tinh ChỉnhaMô Hình:

Nếu môahình không đạt hiệuasuất mong muốn, bạn có thể cần phải tinh chỉnh các siêu tham số của mô hình hoặc thay đổi kiến trúc mô hình.

Triển Khai Mô Hình:

Khi mô hình đạt hiệu suất tốt và được chấp nhận, nó có thể được triển khai trong môi trường thực tế để sử dụng trong các ứng dụng hàng ngày.

Theo Dõi và Cải Thiện:

Mô hình cần được theo dõi trong môi trường sản xuất và cải thiện theo thời gian khi có thêm dữ liệu mới và khi cần thiết.

**3.2. Giới thiệu về bộ dữ liệu**

Cơ sở dữaliệu này chứa 1025 mẫu, 76 thuộcatính nhưng tất cảacác thửanghiệm được công bố đều đề cập đến việc sử dụng một tậpahợp con gồm 14athuộc tính trong số đó. Đặc biệt, cơ sởadữ liệu Clevelandalà cơ sở dữaliệu duy nhất đượcacácanhà nghiênacứu ML sử dụngacho đến nay. Trường "mục tiêu" đề cập đến sự hiệnadiện của bệnh tim ởabệnh nhân. Nó có giá trị nguyên từ 0 (không hiện diện) đến 4. Các thử nghiệmavới cơ sở dữ liệu Cleveland tậpatrung vàoaviệc cố gắng phân biệt sự hiệnadiện (giá trị 1,2,3,4) với sự vắng mặt (giá trị 0).

Dưới đây là mô tả các thuộc tính (attributes) trong tập dữ liệu:

Age: Tuổi

Sex : giới tính (male, female)

Cp: loại đau ngực

1: cơn đau thắt ngựcirõ ràng

2: đau thắtingựcikhông rõiràng

3: không đau thắtingực

4: không~~i~~có triệuichứng

Trestbps: Huyết ápalúc nghỉangơi (đơn vịammHg)

chol: cholestoral trong huyếtathanhatính bằng mg/dl

fs: (đườngahuyết lúc đói> 120 mg / dl): 1a=atrue; 0 = falsea

restecg:ikết quả điện tâmiđồ lúc nghỉ

thalach: nhịp tim tối đa

exang:iđau thắt ngực do tậpathể dục (a1 = có; a0 = không)

oldpeak = ST ức chế do luyệnatập so vớiakhi nghỉ

slope: độ dốcacủa đoạnaST tập luyện đỉnhacao

1: dốc lêna

2: bằng phẳnga

3: dốc xuốnga

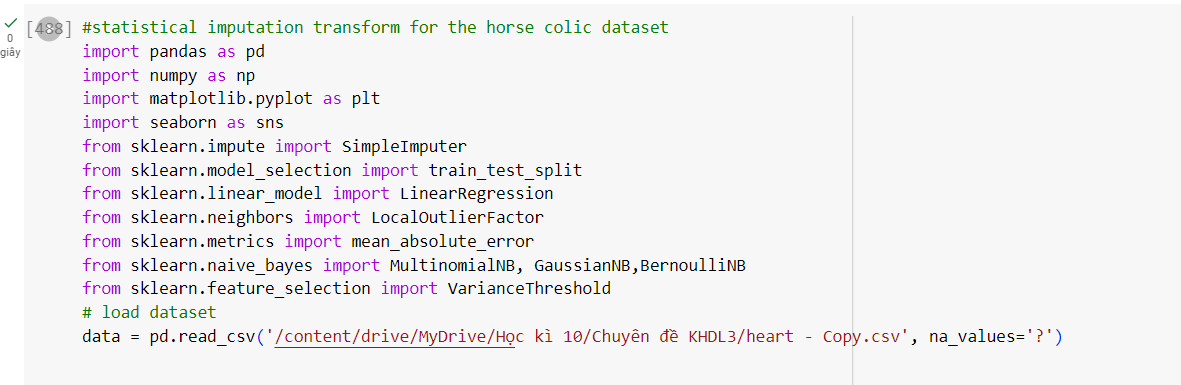
ca: số lượngamạch chính (0-3) đượcatô màu bằng phương phápasoi huỳnh quang (nội soi) a

thal: 3 = bìnhathường; 6 = khuyết tật cốađịnh; 7 = khiếm khuyết có thểađảo ngược

target: athuộc tínhaphân lớp (1 = Bệnhatim, 0 = Không cóabệnh tim)

**3.3. Thực hiện mô hình**

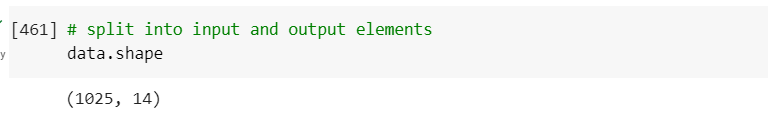
Nhập thư viện: đầu tiên, nhập các thư viện cần thiết cho việc phân tích dữ liệu và xây dựng mô hình. Các thư viện bao gồm Numpy, Pandas, Matplotlib, Seaborn, và các thư viện từ scikit-learn (sklearn) như train\_test\_split , MultinomialNB, GaussianNB, và BernoulliNB.



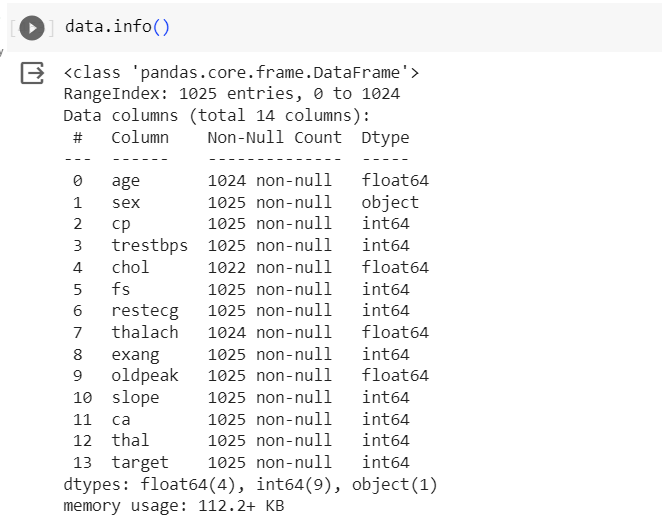
Thay thế các giá trị chứa ? bằng NaN

Đọc dư liệu: đọc dữ liệu từ tệp CSV bằng pd.read\_csv() và lưu nó trong DataFrame data và hiển thị dữ liệu

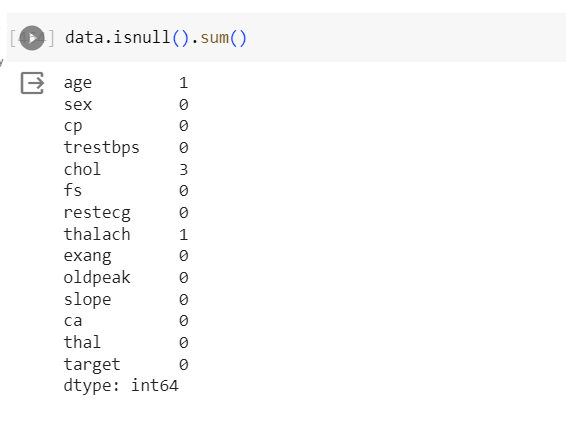
**Load data**

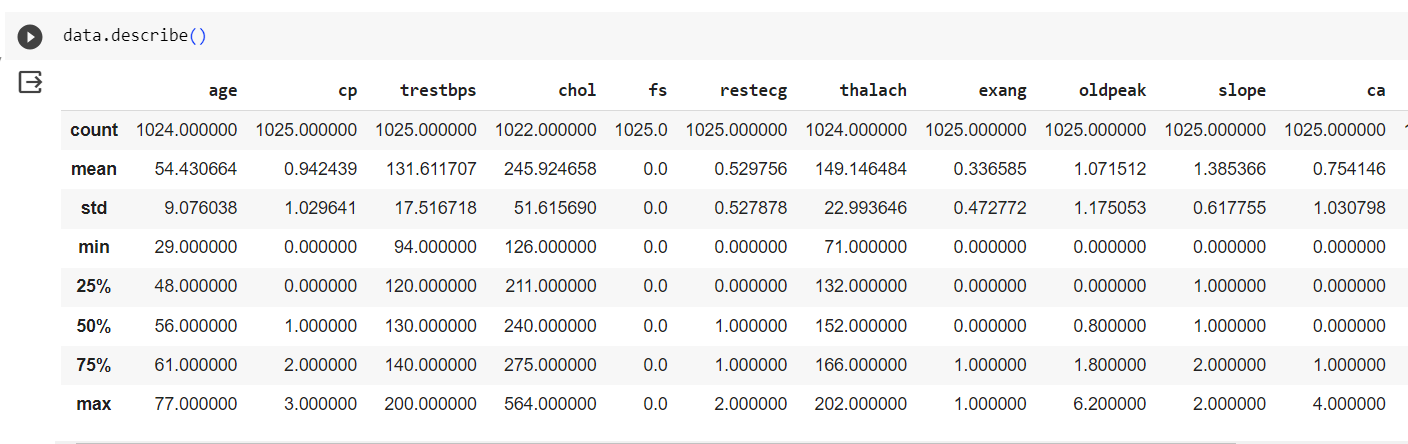


Kiểm tra thông tin dữ liệu: data.info() cho biết thông tin về các cột dữ liệu, bao gồm kiểu dữ liệu và có dữ liệu bị thiếu hay không.



Kiểm tra dữ liệu bị thiếu: data.isnull().sum() cho biết số lượng dữ liệu bị thiếu trong mỗi cột.Tóm tắt thống kê của dữ liệu: data.describe() cung cấp tóm tắt thống kê về các thuộc tính số học của dữ liệu như trung bình, độ lệch chuẩn, giá trị tối thiểu và tối đa.

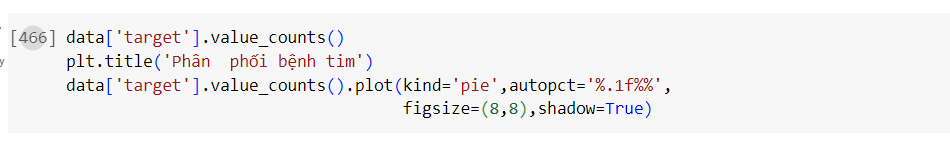


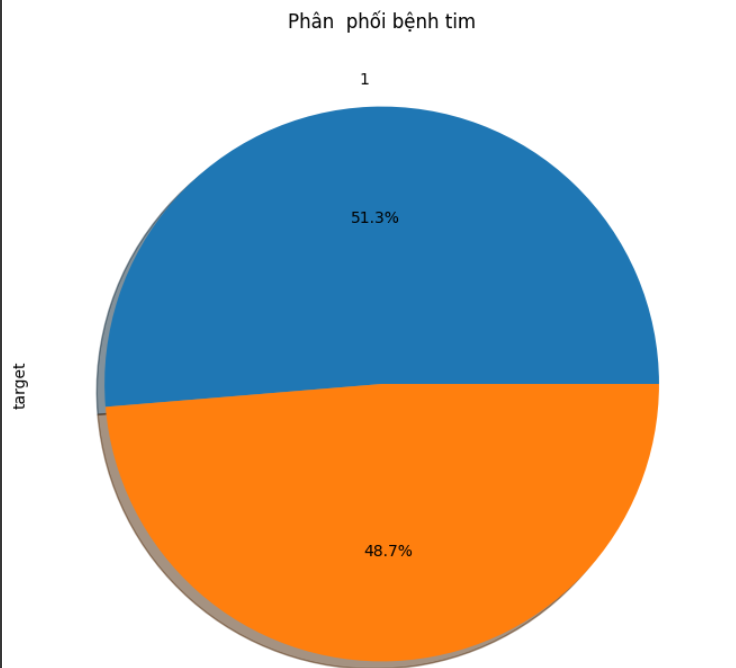


Top of Form

**Phân tích khám phá dữ liệu**

Phân tích phân phối của biến target: Bạn sử dụng data['target'].value\_counts() để đếm số lượng mẫu thuộc vào từng lớp (0 hoặc 1) và sau đó sử dụng biểu đồ pie để trực quan hóa phân phối này.

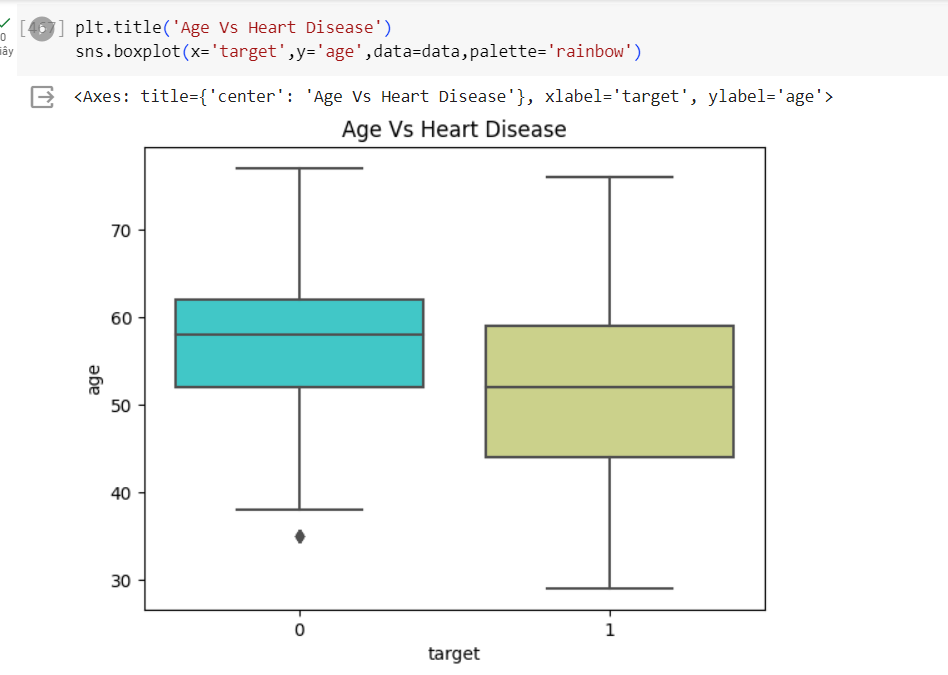




Hình 3. 2. Phân phối bệnh tim

Tổng số bệnh nhân bệnh tim cao hơn không mắc bệnh tim

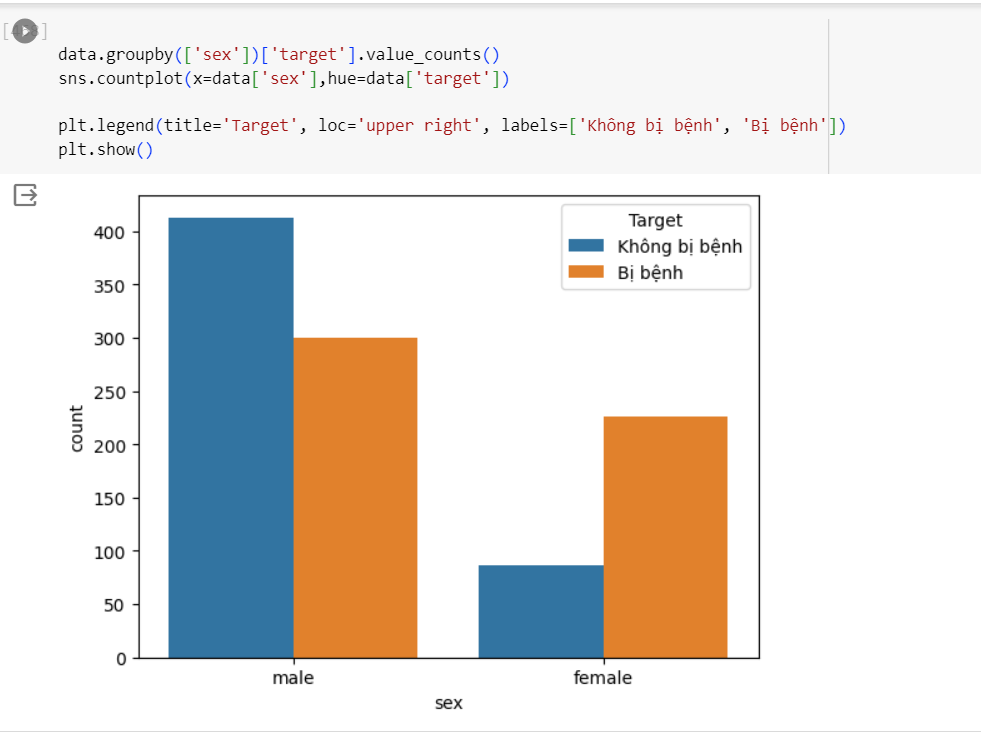
Phân tích mối quan hệ giữa tuổi (age) và bệnh tim: sử dụng sns.boxplot để hiển thị mối quan hệ giữa tuổi và bệnh tim cho từng lớp (0 và 1).



Hình 3. 3. Mối quan hệ giữa tuổi và bệnh tim

Độ tuổi trung bình của người mắc bệnh tim thấp hơn những người không mắc bệnh tim

Phân tích mối quan hệ giữa giới tính (sex) và bệnh tim: Sử dụng sns.countplot để hiển thị mối quan hệ giữa giới tính và bệnh tim cho từng lớp (0 và 1).

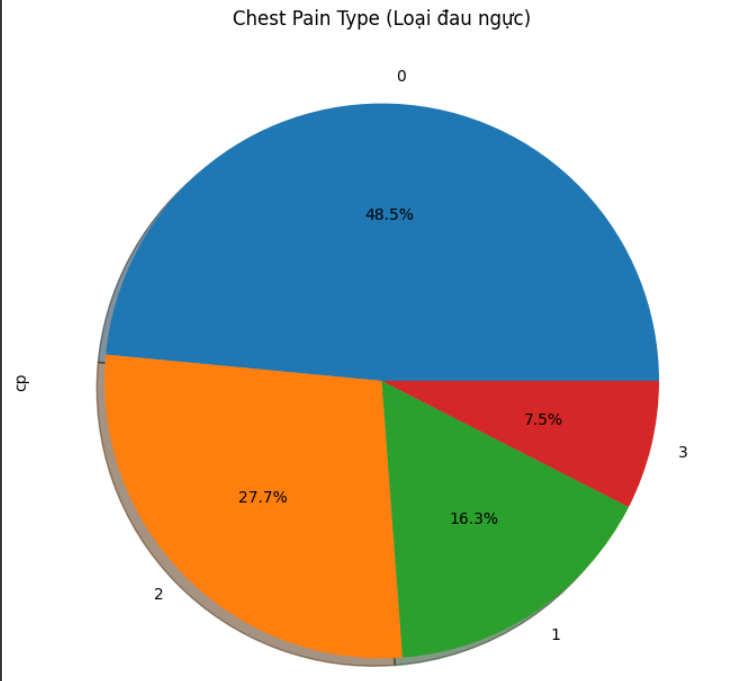


Hình 3. 4. Mối tương quan giới tính và bệnh tim

Từ biểu đồ trên, chúng ta có thể thấy rằng nam giới có nguy cơ mắc bệnh tim cao hơn nữ giới

Phân tích loại đau ngực (cp): sử dụng biểu đồ pie để phân tích phân phối của loại đau ngực.





Hình 3. 5. Loại đau ngực

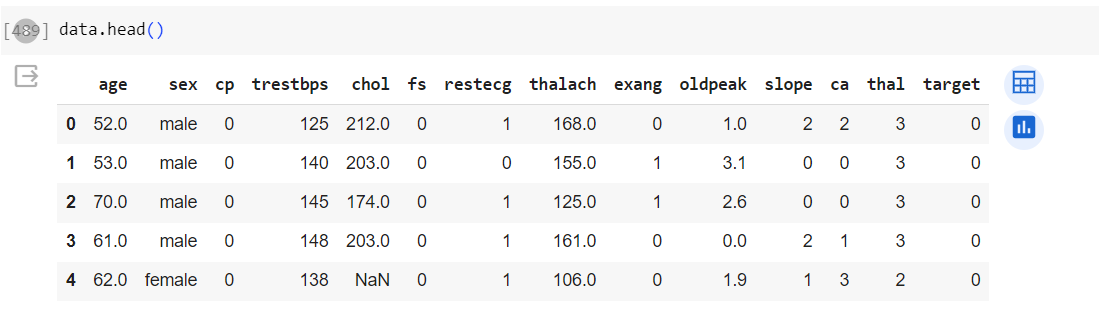
Đau ngực loại 0 (đau thắt ngực điển hình) có số lượng cao nhất và đau ngực loại 3 (không có triệu chứng) có số lượng thấp nhất so với các loại đau ngực khác

**Tương tự phân tích sự tương quan của các biến khác với kết quả bệnh tim (trestbps với target, chol với target, fbs với target,...)**



Hình 3. 6. Biểu đồ tương quan của các biến

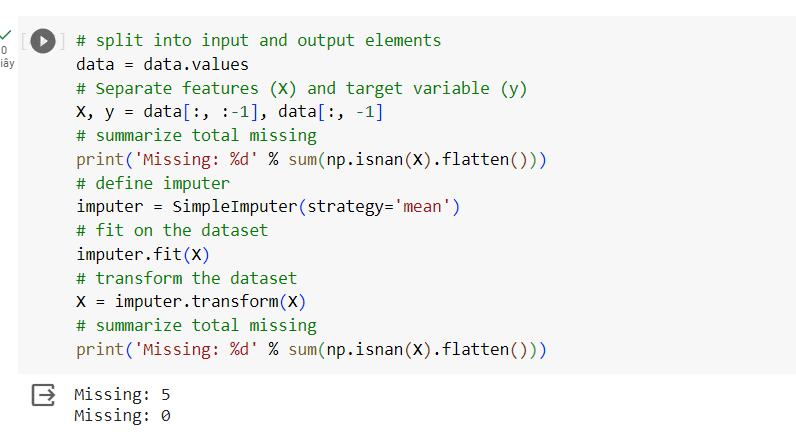
**Tiền xử lý dữ liệu**

****

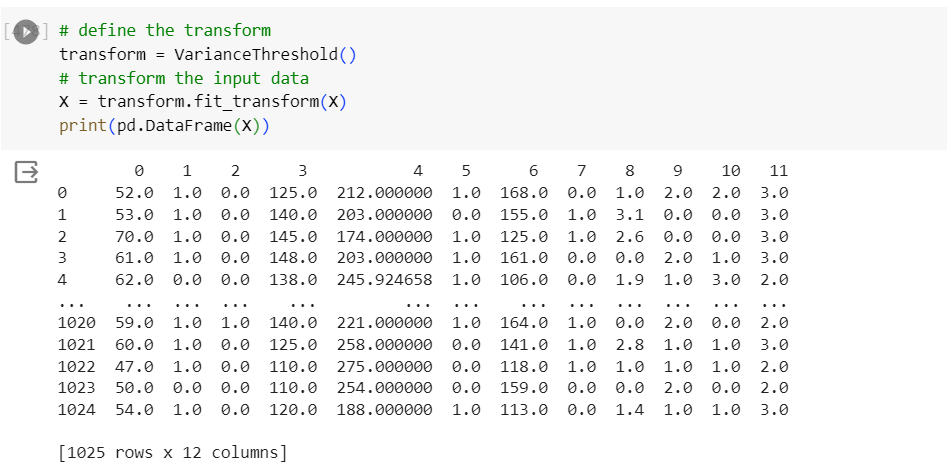
Có thể thấy trong cột sex là dữ liệu categorycal và chúng ta sẽ mã hóa chúng bằng sử dụng phương pháp LabelEncoder



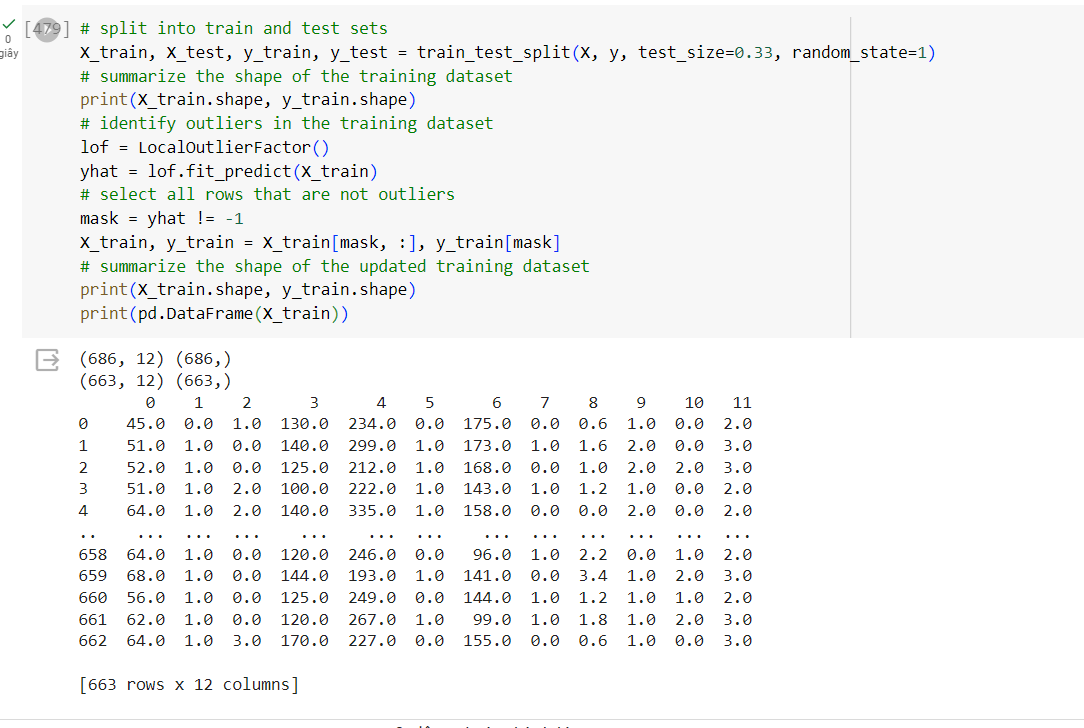
Xử lý dữ liệu bị thiếu bằng phương pháp SimpleImpute



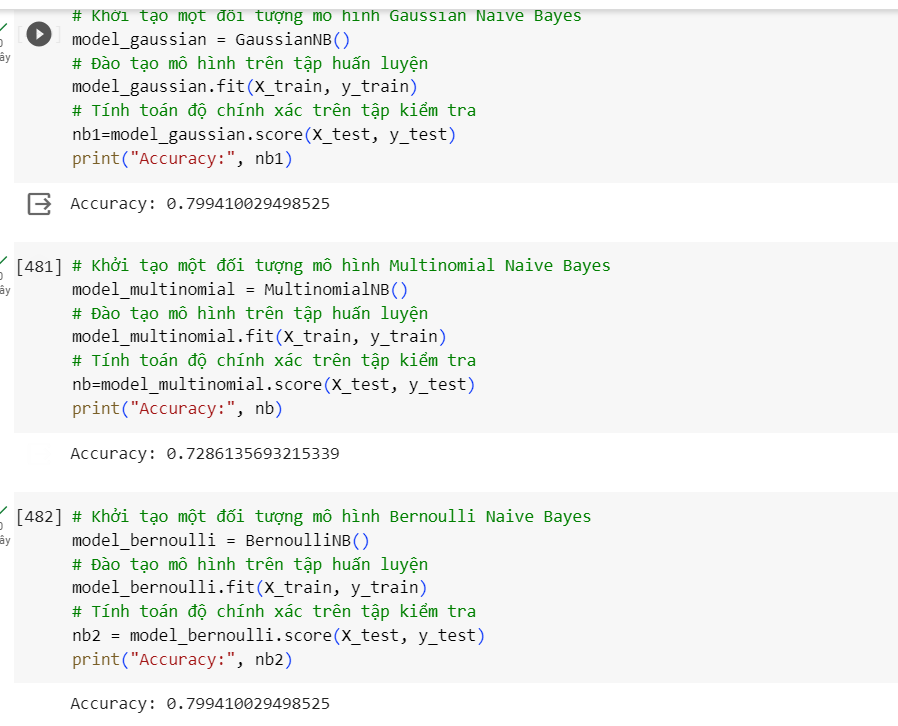
Xóa đặc trưng chỉ mang một giá trị bằng phương pháp VarianceThreshold



Loại bỏ các giá trị ngoại lai bằng LocalOutlierFactor



**Định nghĩa mô hình**

****

**Đánh giá thực nghiệm**

Tính độ nhạy ,độ đặc hiệu và f1-score cho mô hình

Dựa vào Độachính xác, độ nhạyavà độ đặc hiệuađể đánh giáahiệu suất phân loại của mô hình. Độachính xác, độ nhạyavà độ đặc hiệu được xácađịnh theo cácacông thức sau:

Precision (Độ đặc hiệu) là một thước đo của khả năng của mô hình phân loại để không làm sai lừa về những cái được dự đoán là dương tính. Nó tính bằng cách chia số lần dự đoán đúng dương tính cho tổng số lần mô hình dự đoán dương tính, tức là:

Precision = (TrueaPositives) / (TrueaPositives + FalseaPositives)

Trong đó, True Positives là số lần mô hình dự đoán đúng dương tính và False Positives là số lần mô hình dự đoán sai dương tính.

Recall (Độ nhạy) là một thước đo của khả năng của mô hình phân loại để bắt kịp và bao gồm tất cả những cái thực tế là dương tính. Nó tính bằng cách chia số lần dự đoán đúng dương tính cho tổng số lần thực tế dương tính, tức là:

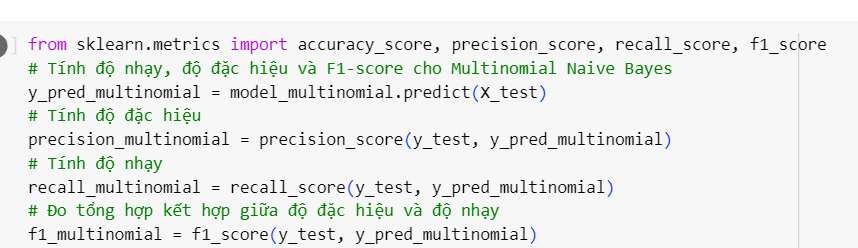
Recall = (TrueaPositives) / (True Positivesa+aFalse Negatives)

Trong đó, TrueaPositives là số lần mô hình dự đoán đúng dương tính và False Negatives là số lần mô hình dự đoán sai âm tính cho các trường hợp thực tế là dương tính.

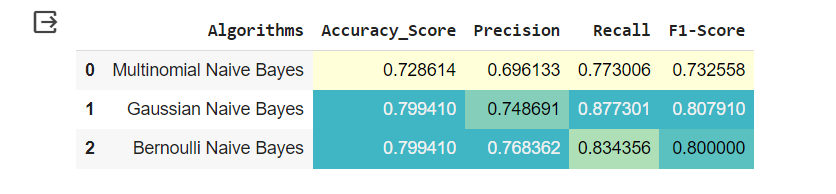
Tính F1-score của mô hình Naive Bayes trên tập kiểm tra. F1-score là một thước đo tổng hợp sự kết hợp giữa độ đặc hiệu và độ nhạy, và thường được sử dụng để đánh giá hiệu suất của mô hình phân loại.

Công thức tính F1-score là:

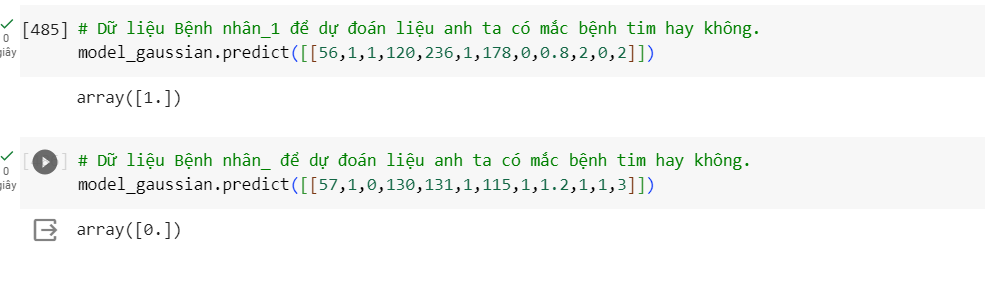
F1-scorea= 2 \* (Precision \* Recall) / (Precision + Recall)

****

**So sánh độ**a**chính xác, độ nhạy, độ**a**đặc hiệu của 3 mô hình:**



**Đánh giá dự đoán:**



Kết quả array([1.]) là bệnh nhân mắc bệnh tim còn 0 thì không mắc bệnh tim

Sau khi sử dụng thuật toán Naive Bayes để dự đoán bệnh tim, đã đánh giá hiệu suất của các mô hình Naive Bayes cụ thể (Multinomial, Gaussian, Bernoulli). Dưới đây là kết luận:

Mô hình Gaussian Naive Bayes có hiệu suất tốt trong việc dự đoán bệnh tim, với độ chính xác (Accuracy) là khoảng 79.9%. Mô hình này cũng có độ đặc hiệu, độ nhạy và F1-Score nhỉnh hơn so với các mô hình khác. Điều này có nghĩa là mô hình này có khả năng dự đoán bệnh tim tốt hơn và đưa ra ít dự đoán sai.

Mô hình Bernoulli Naive Bayes cũng cho kết quả tương đối tốt, với độ chính xác là khoảng 79.9%. Độ đặc hiệu, độ nhạy và F1-Score của mô hình này cũng gần giống với mô hình Gaussian Naive Bayes.

Mô hình Multinomial Naive Bayes có hiệu suất thấp hơn với độ chính xác là 76.10% và F1-Score là 0.7879. Điều này có nghĩa là mô hình này có khả năng dự đoán bệnh tim kém hơn so với hai mô hình Naive Bayes khác.

**CHƯƠNG 4: KẾT LUẬN**

**4.1 . Kết luận**

Trong quá trình phát triển và đánh giá mô hình Naïve Bayes dự đoán bệnh tim, tôi đã đạt được một loạt kết quả quan trọng. Mô hình này đã chứng minh khả năng học và dự đoán giá bệnh dựa trên dữ liệu sức khỏe, và đã được đánh giá tích cực dựa trên các chỉ số trong học máy.

Tuy nhiên, dự án vẫn còn rất nhiều tiềm năng để phát triển và nâng cấp. Các hướng phát triển tiềm năng bao gồm việc sử dụng dữ liệu bổ sung, thử nghiệm với các kiến trúc mạng nơ-ron mới, thu thập thêm dữ liệu, phát triển ứng dụng thực tế,

Mặc dù mô hình dự đoán bệnh tim là một công việc phức tạp và đầy thách thức, tôi hy vọng rằng nghiên cứu này có thể đóng góp vào sự hiểu biết về khả năng ứng dụng thuật toán Naïve Bayes trong dự đoán bệnh tim. Tôi kỳ vọng rằng việc phát triển và cải thiện mô hình sẽ giúp tạo ra một công cụ hữu ích cho y học cũng như là sức khỏe

## 4.2. Hướng phát triển

Thu thập thêm dữ liệu chất lượng đảm bảo rằng dữ liệu đủ lớn và đại diện cho nhiều tình huống khác nhau liên quan đến bệnh tim. Cải thiện dữ liệu cải thiện hiệu suất

Cập nhật liên tục những thông tin mới và dữ liệu thay đổi

Hợp tác với chuyên gia y tế bác sĩ , chuyên khoa y tế và nhà nghiên cứu trong lĩnh vực bệnh tim để đảm bảo rằng mô hình và kết quả của nó có ý nghĩa lâm sàng và hữu ích cho quản lý bệnh tim

Đánh giá và nâng cấp mô hình : Định kỳ đánh giá hiệu suất của mô hình và xem xét việc cải thiện nó bằng cách sử dụng dữ liệu mới và kỹ thuật mô hình học mày mới

Học hỏi liên tục: Học hỏi từ dữ liệu và kinh nghiệm thực tế để cải thiện mô hình theo thời gian

Sử dụng dự đoán cho nghiên cứu và cải thiện chất lượng cuộc sống: Ngoài việc dự đoán bệnh tim, sử dụng mô hình để thực hiện nghiên cứu về bệnh tim, đặc biệt là để cải thiện chẩn đoán và điều trị bệnh tim

# TÀI LIỆU THAM KHẢO

[1] Cox, David R. “The regression analysis of binary sequences.” Journal of the Royal

Statistical Society. Series B (Methodological) (1958): 215-242.

[2] Cramer, Jan Salomon. “The origins of logistic regression.” (2002).

[3] Abu-Mostafa, Yaser S., Malik Magdon-Ismail, and Hsuan-Tien Lin. Learning from

data. Vol. 4. New York, NY, USA:: AMLBook, 2012. (link to course)

[4] Bishop, Christopher M. “Pattern recognition and Machine Learning.”, Springer

(2006). (book)

[5] Duda, Richard O., Peter E. Hart, and David G. Stork. Pattern classification. John

Wiley & Sons, 2012.

[6] Andrer Ng. CS229 Lecture notes. Part II: Classification and logistic regression

[7] Jerome H. Friedman, Robert Tibshirani, and Trevor Hastie. The Elements of

Statistical Learning.

Link sản phẩm :

https://drive.google.com/drive/folders/1BVXlskDkxJjhT4kbfsDx-NVcoSO0PKRz?usp=sharing