TRƯỜNG ĐẠI HỌC ĐIỆN LỰC

KHOA CÔNG NGHỆ THÔNG TIN

BÁO CÁO MÔN HỌC

KHAI PHÁ DỮ LIỆU

ĐỀ TÀI: KHAI PHÁ DỮ LIỆU CHUẨN ĐOÁN BỆNH TIỂU ĐƯỜNG BẰNG NAIVE BAYES

Sinh viên thực hiện	: ĐẶNG THỊ NGỌC LINH
	ĐẶNG KHÁNH LINH
	NGUYỄN THỊ HUYỀN
Giảng viên hướng dẫn	: VŨ VĂN ĐỊNH
Ngành	: CÔNG NGHỆ THÔNG TIN
Chuyên ngành	: HỆ THỐNG THƯƠNG MẠI ĐIỆN TỬ
Lớp	: D13HTTMDT1

Hà Nội, tháng 03 năm 2021

PHIẾU CHẤM ĐIỂM

Họ và tên	Chữ ký	Ghi chú
Đặng Thị Ngọc Linh		
Đặng Khánh Linh		
Nguyễn Thị Huyền		

Sinh viên thực hiện:

Giảng viên chấm:

Họ và tên	Chữ ký	Ghi chú
Giảng viên chấm 1:		

Giảng viên chấm 2:		

MỤC LỤC

LỜI CẢM ƠN

Qua bài tập lớn này, chúng em xin gửi lời cảm ơn tới thầy cô khoa công nghệ thông tin, đặc biệt là thầy Vũ Văn Định rất cảm ơn cô đã cho chúng em có cơ hội được tìm hiểu một góc kiến thức mới, hay và bổ ích cùng với đó là sự tận tâm dạy dỗ chúng em, giúp chúng em có thể hoàn thiện đề tài này. Trong quá trình tìm hiểu và hoàn thiện, đề tài sẽ không thể tránh khỏi những sai sót, khuyết điểm. Vì vậy, nhóm thực hiện chúng em hy vọng nhận được sự đánh giá và đóng góp nhiệt tình từ phía thầy và các bạn để bài của nhóm chúng em được hoàn thiện hơn.

Qua bài tập lớn này, chúng em xin cảm ơn các bạn bè lớp D13HTTMDT1 đã giúp đỡ chúng em trong quá trình học tập và làm bài tập lớn, đã chia sẻ kinh nghiệm kiến thức của các bạn đã tạo nên nền tảng kiến thức cho chúng em.

Cuối cùng, chúng em xin gửi lời cảm ơn gia đình đặc biệt là cha mẹ đã tạo điều kiện tốt nhất cho con có đủ khả năng thực hiện bài tập lớn này, trang trải học phí, động viên tinh thần cho em để học tập trong môi trường đại học tuyệt vời này.

Chúng em xin chân thành cảm ơn!

Nhóm sinh viên thực hiện

Đặng Thị Ngọc Linh

Đặng Khánh Linh

Nguyễn Thị Huyền

TÓM TẮT

Ngành y tế và giáo dục luôn là vấn đề sống còn của bất kỳ quốc gia nào trên thế giới. Trong những năm gần đây, chính phủ Việt nam đặc biệt đầu tư cho hai ngành mũi nhọn này thông qua các chính sách , nguồn vốn dành cho trang thiết bị hạ tầng và nghiên cứu khoa học. Trong lĩnh vực kho học, càng ngày càng có nhiều công trình khoa học trong y tế. Tuy nhiên các nghiên cứu khoa học về ứng dụng công nghệ thông tin để giải quyết bài toán về y tế là không nhiều. Do tình hình sức khỏe và cách sinh hoạt của người dân Việt Nam rất bất ổn nên đã tạo ra nhiều căn bệnh, đặc biệt là bệnh tiểu đường, vì vậy đề tài nghiên cứu chuẩn đoán bệnh tiểu đường tại Việt Nam bằng kỹ thuật kai phá dữu liệu. Dựa trên các triệu chứng lâm sàng và cận lâm sàng có thể phân lớp bệnh của bệnh nhân nhằm giúp các bác sĩ chuẩn đoán và điều trị tốt hơn cho bệnh nhân.

Ngiên cứu tiến hành theo 4 bước chính:

(1) Tìm hiểu nghiệp vụ y tế liên quan đến bệnh tiểu đường.

(2) Thu nhập và tiền xử lý dữ liệu.

(3) Tìm hiểu bài toán phân lớp trong khai phá dữ liệu, lựa chọn thuật toán phù hợp với yêu cầu bài toán đặt ra và dữ liệu thu nhập được.

(4) Hiện thực chương trình máy tính và đánh giá ý nghĩa thực tiễn.

DANH SÁCH CÁC BẢNG

DANH SÁCH CÁC HÌNH

CHƯƠNG 1: TỔNG QUAN ĐỀ TÀI

1.1 Đặt vấn đề

Ứng dụng công nghệ thông tin vào việc lưu trữ và xử lý thông tin ngày nay đuược áp dụng hầu hết trong lĩnh vực, điều này đã tạo ra một lượng lớn dữ liệu đuược lưu trữ với kích thước tăng lên không ngừng. Đay chính là điều kiện tốt cho việc khai thác kho dữ liệu để đemlại tri thức có ích với các công cụ truy vấn, lập việc khai thác kho dữ liệu để đem lại tri thức có ích với các công cụ truy vấn, lập bẳng biểu và khai phá dữ liệu.

Khai phá dữ liệu là một kỹ thuật dựa trên nền tảng của nhiều lý thuyết như xác xuất, thống kê, máy học nhằm tìm kiếm các tri thức tiềm ẩn trong các kho dữ liệu có kích thước lớn mà người dùng khó có thể nhận biết bằng những kỹ thuật thông thường. Nguồn dữ liệu y khoa rất lớn, nếu áp dụng khai phá dữ liệu trong lĩnh vực này sẽ mang lại nhiều ý nghĩa cho ngành y tế. Nó sẽ cung cấp nững thông tin quý giá nhằm hỗ trợ trong việc chuẩn đoán và điều trị sớm giúp bệnh nhân thoát đuược nhiều căn bệnh hiểm nghèo.

Trong lĩnh vực y khoa Việt Nam, hiện nay các tuyến y tế phường, xã, vùng sâu, vùng xa còn thiếu nhân lực y tế có trình độ chuyên môn và thiếu các trang thiết bị cần thiết trong chuẩn đoán bệnh. Vì vậy xây dựng hệ thống chuẩn đoán rất cần thiết cho ngành y tế hiện nay ở Việt Nam. Hệ hỗ trợ sẽ kết hợp với cán bộ y tế giúp chuẩn đoán sớm một số bệnh phát hiện sớm được những bệnh nguy hiểm và giảm gánh nặng kinh tế cho gia đình bệnh nhân và xã hội. Để minh chứng cho những lợi ích mà việc chuẩn đoán mang lại, đề tài chọn bộ dữ liệu bệnh tiểu đường để thử nghiệm và đánh giá.

Ứng dụng kỹ thuật phân lớp dữu liệu trong khai phá dữ liệu nhằm xây dựng hệ thống chuẩn đoán là một trong những hướng nghiên cứu chính của đề tài. Sau khi phân tích một số thuật toán cũng như đặc điểm của dữu liệu thu nhập đưuọc về bệnh tiểu đưuòng, đề tài đề xuất ứng dụng mô hình phân lớp bằng cây quyết định với thuật toán Naive bayes để tìm ra qui luật tìm ẩn trong dữ liệu.

1.2 Cơ sở hình thành đề tài

Theo thống kê năm 2019 từ tổ chức Y tế Thế giới(WHO), bệnh đái tháo đưuòng(tiểu đường) đang ảnh hưởng đến 732 triệu người trên toàn cầu. Nếu không có sự tăng cường nhận thức và can thiệp kịp thời, đái tháo đưuòng sẽ trở thành một trong bảy nguyên nhân hàng đầu gây chết người vào năm 2030.

Tỷ lệ mắc bệnh gấp 4 lần so với năm 1980, mỗi năm có 3.7 triệu người chết mỗi năm, tại Việt Nam có 50% dân số chưa được chuẩn đoán. Bộ Y tế Việt Nam luôn quan tâm đến những nhiệm vụ trọng tâm của chương trình quốc gia phòng trống bệnh tiểu đường. Vì vậy xây dựng hệ thống chuẩn đoán tiểu đường để góp phần chuẩn đoán và phát hiện sớm những nguy cơ dịch bệnh là vấn đề quan tâm nhất của gia đình và xã hội. Đề tài áp dụng công nghệ thông tin xây dựng chuẩn đoán bệnh với bộ dữ liệu thu nhập được từ bệnh tiểu đường.

1.3 Một số kết quả thực nghiệm trong và ngoài nước

1.3.1 Kết quả thực nghiệm thế giới

Trên thế giới dã cho ra nhiều ứng dụng từ hệ hỗ trợ chuẩn đoán nhanh và điều trị bệnh tốt hơn như hệ thống chuẩn đoán y tế Caduceus của Harry Pope; hệ thống chuyên gia y tế Diagnosipro; MYCIN hệ hỗ trợ chuẩn đoán bệnh mất ngủ; BI-RADS(2007) chuẩn đoán ung thu vú; PSG-Expert(2000) chuẩn đoán bệnh mất ngủ; Naser xây dựng hệ thống chuẩn

đoán bệnh về da, Comete quản lý bệnh nhân tăng huyết áp, bệnh mãn tính,…

1.3.2 Kết quả thực nghiệm trong nước

Ở Việt Nam tình hình ứng dụng công nghệ thông tin bắt đầu phát triển, nhiều ứng dụng công nghệ thông tin đã được áp dụng vào y khoa, vào năm cuối 1980 những nghiên cứu hệ hỗ trợ bác sĩ chuẩn đoán bệnh nội khoa, châm cứu và chuẩn đoán đông y, hệ hỗ trợ ra quyết định trong việc chuẩn đoán lâm sàng… tuy vậy những nghiên cứu chuẩn đoán y khoa nhằm xây dựng các hệ hỗ trợ quyết định vẫn còn hạn chế.

1.4 Mục tiêu đề tài

Đề tài tập chung vào nghiên cứu kỹ thuật phân lớp trong khai phá dữu liệu, từ đó nắm bắt được những giải thuật làm tiền đề cho nghiên cứu và xây dựng ứng dụng cụ thể. Ngoài ra, việc thu nhập dữ liệu bệnh của bệnh cụ thể cũng được quan tâm và đề tài đề xuất sửu dụng dữ liệu bệnh tiểu đường. Sau khi phân tích đặc điểm của dữu liệu thu nhập đuược và lựa chọn giải thuật phù hợp với dữ liệu, việc xây dựng và đánh giá chất lượng, độ hiệu quả của hệ thống cuẩn đoán cũng là mục tiêu chính của đề tài.

1.5 Đối tượng và phương pháp nghiên cứu

Đề tài tập chung vào nghiên cứu kỹ thuật phân lớp trong khai phá dữu liệu(cụ thể là nghiên cứu thuật toán Naive bayes) để áp dụng vào việc phân tích cơ sở dữ liệu bệnh tiểu đường. Luận văn thu nhập dữ liệu bệnh tiểu đường của tất cả bệnh nhân(không phân biệt tuổi, giới tính) đến khám vào điều trị tại bệnh viện Bạch Mai và Bệnh viện Nội tiết Trung ương. Sử dụng phương pháp và nghiên cứu hồi cứu với sự hỗ trợ chuyên môn của các bác sĩ chuyên khoa, đề tài tiến hành nghiên cứu trên cơ sở thuật toán phân lớp trong khai phá dữ liệu.

1.6 Ý nghĩa đề tài

1.6.1 Ý nghĩa khoa học

Với sự trợ giúp của máy tính, đề tài đóng góp một biện pháp thực hiện hỗ trợ các cán bộ y tế chuẩn đoán bệnh cho bệnh nhân. Kết quả, Kinh nghiệm thu đuược khi thực hiện đề tài này sẽ giúp các cán bộ y tế phát hiện sớm bệnh cho bệnh nhân, đồng thời mong muốn những người đang công tác trong lĩnh vực y khoa và Khoa học máy tính ngồi lại với nhau để tìm ra những giải pháp tốt hơn trong vấn đề chuẩn đoán và điều trị bệnh bằng cách kết hợp giữa 2 lịnh vực y học và khoa học máy tính.

1.6.2 Ý nghĩa thực tiễn

Chuẩn đoán bệnh và phát hiện bệnh là cả một quá trình, đòi hỏi các cán bộ y tế không những phải thật vững chuyên môn mà còn có đầy đủ các trang thiết bị y tế mới có thể chuẩn đoán chính xác bệnh cho bệnh nhân. Nếu chuẩn đoán sai bệnh sẽ đưa đến điều trị sai, không phát hiện sớm bệnh cho bệnh nhân,…

1.7 Bố cục đề tài

Đề tài được chia thành các phần:

Chương 1: Tổng quan đề tài

Chương 2: Khai phá dữ liệu

Chương 3: Xây dựng mô hình dữu liệu sử dụng Naive bayes

Chương 4: Thực nghiệm và đánh giá

CHƯƠNG 2: KHAI PHÁ DỮ LIỆU

2.1 Tổng quan về kỹ thuật Khai phá dữu liệu(Data mining)

2.1.1 Khái niệm về khai phá dữ liệu

Khai phá dữ liệu (*data mining*) Là quá trình tính toán để tìm ra các mẫu trong các bộ dữ liệu lớn liên quan đến các phương pháp tại giao điểm của máy học, thống kê và các hệ thống cơ sở dữ liệu. Đây là một lĩnh vực liên ngành của khoa học máy tính. Mục tiêu tổng thể của quá trình khai thác dữ liệu là trích xuất thông tin từ một bộ dữ liệu và chuyển nó thành một cấu trúc dễ hiểu để sử dụng tiếp.Ngoài bước phân tích thô, nó còn liên quan tới cơ sở dữ liệu và các khía cạnh quản lý dữ liệu, xử lý dữ liệu trước, suy xét mô hình và suy luận thống kê, các thước đo thú vị, các cân nhắc phức tạp, xuất kết quả về các cấu trúc được phát hiện, hiện hình hóa và cập nhật trực tuyến. Khai thác dữ liệu là bước phân tích của quá trình "khám phá kiến thức trong cơ sở dữ liệu" hoặc KDD.

Khai phá dữ liệu là một bước của quá trình khai thác tri thức (*Knowledge Discovery Proces*s), bao gồm:

- Xác định vấn đề và không gian dữ liệu để giải quyết vấn đề (*Problem understanding and data understanding*).
- Chuẩn bị dữ liệu (*Data preparation*), bao gồm các quá trình làm sạch dữ liệu (*data cleaning*), tích hợp dữ liệu (*data integration*), chọn dữ liệu (*data selection*), biến đổi dữ liệu (*data transformation*).
- Khai thác dữ liệu (*Data mining*): xác định *nhiệm vụ khai thác dữ liệu* và lựa chọn *kỹ thuật khai thác dữ liệu*. Kết quả cho ta một *nguồn tri thức thô*.
- Đánh giá (*Evaluation*): dựa trên một số tiêu chí tiến hành *kiểm tra* và *lọc* nguồn tri thức thu được.
- Triển khai (*Deployment*).

Quá trình khai thác tri thức không chỉ là một quá trình tuần tự từ bước đầu tiên đến bước cuối cùng mà là một quá trình lặp và có quay trở lại các bước đã qua.

2.1.2 Quy trình khai phá dữ liệu

2.1.2.1 Nghiên cứu lĩnh vực

Ta cần nghiên cứu lĩnh vực cần sử dụng Data mining để xác định được những tri thức ta cần chất lọc, từ đó định hướng để tránh tốn thời gian cho những tri thức không cần thiết .

2.1.2.2 Tạo tập tin dữ liễu đầu vào

Ta xây dựng tập tin để lưu trữ các dữ liệu đầu vào để máy tính có thể lưu trữ và xử lý.

2.1.2.3 Tiền xử lý, làm sạch, mã hóa

Ở bước này ta tiến hành bỏ bớt những dữ liệu rườm rà, không cần thiết, tinh chỉnh lại cấu trúc của dữ liệu và mã hóa chúng để tiện cho quá trình xử lý .

2.1.2.4 Rút gọn chiều

Thông thường một tập dữ liệu có chiều khá lớn sẽ sinh ra một lượng dự liệu khổng lồ,ví dụ với n chiều ta sẽ có 2^n nguyên tổ hợp .Do đó , đây là một bước quan trọng giúp giảm đáng kể hao tổn hề tài nguyên trong quá trình xử lý tri thức.Thông thường ta sẽ dung Rough set (http://en.wikipedia.org/wiki/Rough_set) để giảm số chiều.

2.1.2.5 Chọn tác vụ khai thác dữ liệu

Để đạt được mục đích ta cần, ta chọn được tác vụ khai thác dữ liệu sao cho phù hợp.Thông thường có các tác vụ sau:

- Đặc trưng(feature)

- Phân biệt(discrimination)
- Kết hợp(association)
- Phân lớp(classification)
- Gom cụm(clusterity)
- Xu thế(trend analysis)
- Phân tích độ lệch
- Phân tích độ hiếm

2.1.2.6 Chọn các thuật giải khai thác dữ liệu

2.1.2.7 Khai thác dữ liệu: Tìm kiếm tri thức

Sau khi tiến hành các bước trên thì đây là bước chính của cả quá trình , ta sẽ tiến hành khai thác và tìm kiếm tri thức.

2.1.2.8 Đánh giá mẫu tìm được

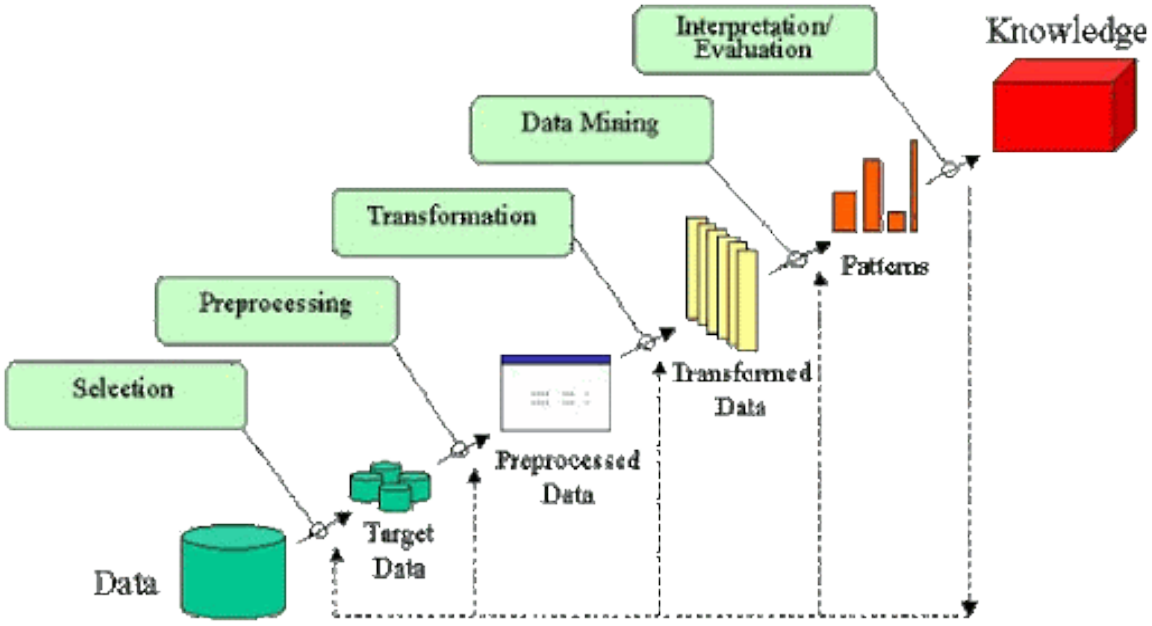
Ta cần đánh giá lại trong các tri thức tìm được , ta sẽ sử dụng được những tri thức nào , những tri thức nào dư thừa,không cần biết.

2.1.2.9 Biểu diễn tri thức

Ta biểu diễn tri thức vừa thu nhập được dưới dạng ngôn ngữ tự nhiên và hình thức sao cho người dùng có thể hiểu được những tri thức đó.

2.1.2.10 Sử dụng các tri thức vừa khám phá

Ta có thể tham khảo tiến trình KDD(Knowledge Discovery in Databases) để hiểu rõ hơn về khai phá dữu liệu:



Hình 2.1: Knowledge Discovery in Databases

Chuẩn bị dữ liệu (data preparation), bao gồm các quá trình làm sạch dữ liệu (data cleaning), tích hợp dữu liệu (data integration), chọn dữ liệu (data selection), biến đổi dữ liệu (data transformation).

Khai thác dữu liệu (data mining): xác định nhiệm vụ khai thác dữ liệu và lựa chọn kỹ thuật khai thác dữu liệu. Kết quả cho ta một nguồn tri thức thô.

Đánh giá (evaluation): dựa trên một tiêu chí tiến hành kiểm tra và lọc nguồn tri thức thu được.

Triển khai (deployment).

Quá trình khai thác tri thức không chỉ là một quá trình tuần tự từ bước đầu tiên đên bước cuối cùng mà là một quá trình lặp và có quay trở lại các bước đã qua.

2.1.3 Ứng dụng của khai phá dữ liệu

Kinh tế - ứng dụng trong kinh doanh, tài chính, tiếp thị bán hàng, bảo hiểm, thương mại, ngân hàng,.. Đưa ra các bản báo cáo giàu thông tin, phân tích rủi ro trước khi đưa ra các chiến lược kinh doanh, sản xuất, phân loại khách hàng từ đó phân định ra thi trường, thị phân:…

Khoa học: Thiên văn học - dự đoán đường đi các thiên thể, hành tinh,...; Công nghệ sinh học – tìm ra các gen mới, cây con giống mới,…

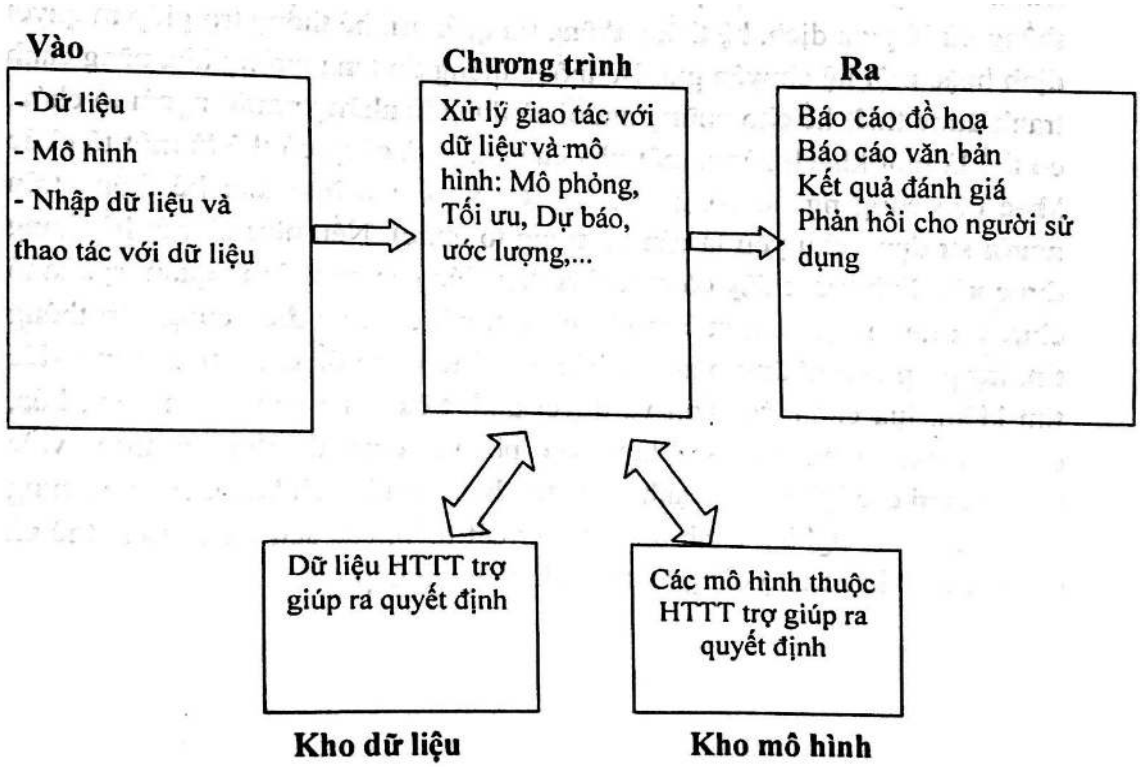
Web: các công cụ tìm kiếm.

2.2 Tổng quan về hệ hỗ trợ ra quyết định

Hệ hỗ trợ ra quyết định là một hệ thống thuộc hệ thống thông tin, có nhiệm vụ cung cấp các thông tin hỗ trợ cho việc ra quyết định để tham khảo và giải quyết vấn đề. Hệ hỗ trợ ra quyết định có thể dùng cho cá nhân hay tổ chức và có thể hỗ trợ gián tiếp hoặc trực tiếp.

Trong lĩnh vực y tế, hệ hỗ trợ ra quyết định dựa vào tri thức đã học sẽ cung cấp thông tin chuẩn đoán bệnh cho nhân viên y tế. Thông tin này đươợc trích lọc để cung cấp một cách thông minh có giá trị cho quá trình chuẩn đoán, theo dõi và điều trị bệnh hiệu quả hơn, từ đó ta thấy một số lợi ích của hệ hỗ trợ ra quyết định trong y tế như sau:

- Tăng cường chất lượng chuẩn đoán, chăm sóc bệnh nhân.
- Giảm nguy cơ sai sót để tránh các tình huống nguy hiểm cho bệnh nhân.
- Tăng cường hiệu quả ứng dụng công nghệ thông tin vào lĩnh vực y tế để giảm bớt những thủ tục giấy tờ không cần thiết.



Hình 2.2: Sơ đồ hệ hỗ trợ quyết định

2.3 Bài toán phân lớp trong khai phá dữ liệu

2.3.1 Khái niệm về phân lớp

Phân lớp là một hình thức phân tích dữ liệu nhằm rút ra những mô hình mô tả những lớp trong dữ liệu. Những mô hình này gọi là mô hình phân lớp (classifier hoặc classsification) được dùng để dự đoán những nhãn lớp có tính phân loại (categorical), rời rạc và không có thứ tự cho những đối tượng dữ liệu mới.

2.3.2 Quá trình phân lớp dữ liệu

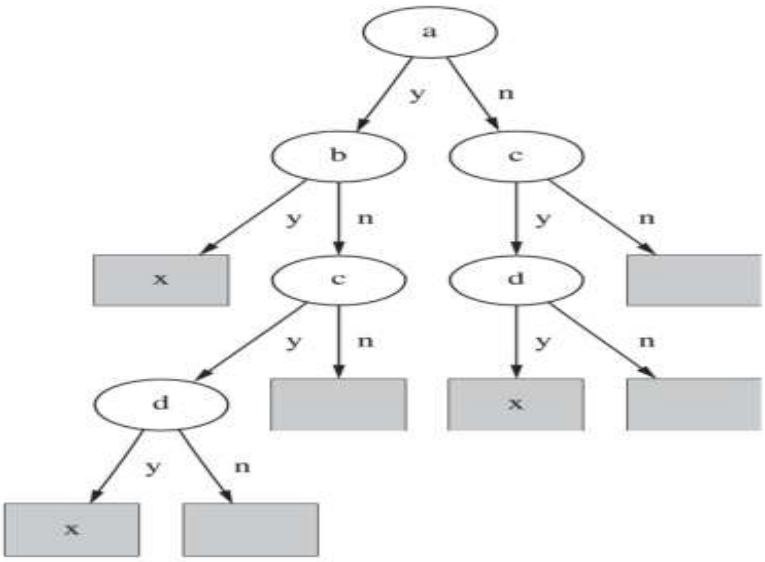
Một quá trình phân lớp dữ liệu gồm 2 bước:

- ❖ Bước thứ nhất: Học/Huấn luyện:

  Quá trình học nhằm xây dựng một mô hình phân lớp (Classifier) bao gồm các lớp dữ liệu đã đuuọc khái niệm trước từ tập dữ liệu đầu vào. Bước học (hay giai đoạnh huấn luyện) dùng một giải thuật phân lớp (Classification Algorithms) để phân lớp các bản ghi của dữ liệu huấn luyện. Trong đó tập huấn luyện là một tập dữ liệu có cấu trúc với các thuốc tính và bộ dữ liệu tương ứng với các thuộc tính.

- Bước thứ hai: Phân lớp (Classification)

  Ở bước thứ hai (Hình 2.3), mô hình tìm được ở bức thứ nhất sẽ được dùng cho việc phân loại những dữ liệu mới. Ta dùng một tập kiểm tra, bao gồm các bản ghi kiểm tra và nhãn lớp liên kết với chúng để so sánh kết quả đầu ra của bộ phân lớp. Các bản ghi kiểm tra này chưa đuuọc dùng để xây dựng mô hình phân lớp. Các bản ghi kiểm tra này chưa đuuọc dụng để xây dựng mô hình phân lớp ở bước 1. Kết quả mô hình phân lớp như sơ đồ sau:



Hình 2.3: Kết quả quá trình phân lớp

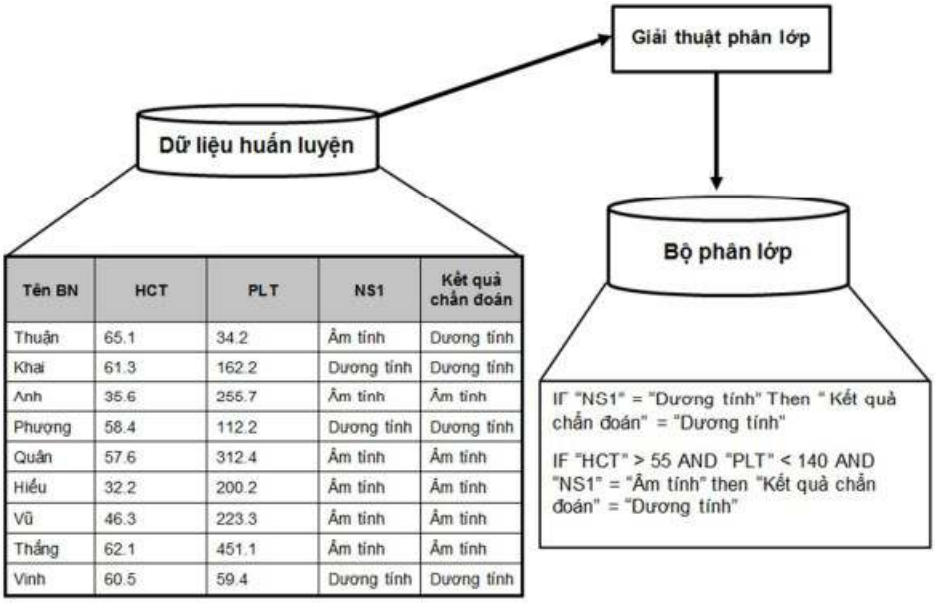
IF a = y and b = y then class x

IF a = n and c = y and d = y then class x

❖ **Ví dụ minh họa bài toán phân lớp:**

Bước 1: Xây dựng mô hình:

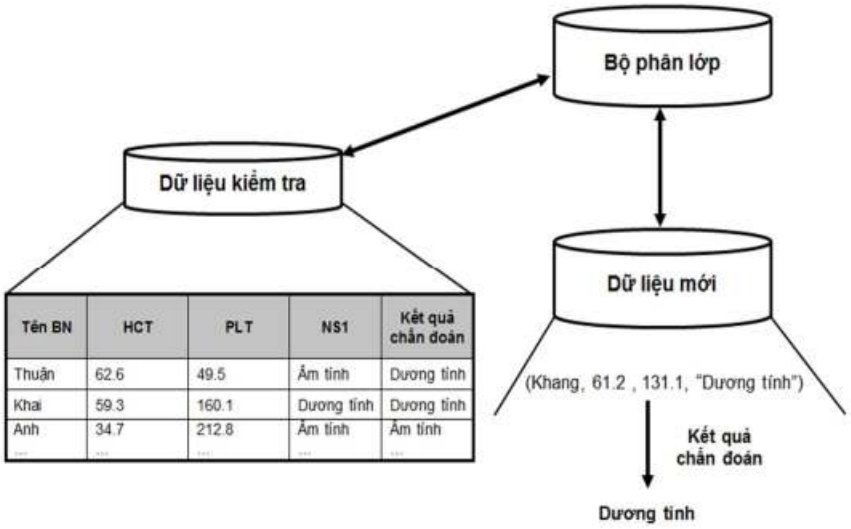
Mục đích: Phân lớp bệnh nhân vào 2 lớp: " Dương tính " và " Âm tính " trong bộ phận lớp có nhãn "KẾT QUẢ CHUẨN ĐOÁN". Mỗi bệnh nhân có các thuộc tính dùng để phân lớp như sau: HCL, PLT, NS1. Sau khi huấn luyện, ta được mô hình phân lớp.



Tên BN	HCT	PLT	NS1	Kết quả chẩn đoán
Thuận	65.1	34.2	Âm tính	Dương tính
Khai	61.3	162.2	Dương tính	Dương tính
Anh	35.6	255.7	Âm tính	Âm tính
Phượng	58.4	112.2	Dương tính	Dương tính
Quân	57.6	312.4	Âm tính	Âm tính
Hiếu	32.2	200.2	Âm tính	Âm tính
Vũ	46.3	223.3	Âm tính	Âm tính
Thắng	62.1	451.1	Âm tính	Âm tính
Vinh	60.5	59.4	Dương tính	Dương tính

Hình 2.4 : Xây dựng mô hình phân lớp

Bước 2: Phân lớp



Tên BN	HCT	PLT	NS1	Kết quả chẩn đoán
Thuận	62.6	49.5	Âm tính	Dương tính
Khai	59.3	160.1	Dương tính	Dương tính
Anh	34.7	212.8	Âm tính	Âm tính
...	...	...	...	...

Hình 2.5: Bước phân lớp

Đánh giá kết quả mô hình ở bước 1, ta dùng tập dữ liệu kiểm tra. Với một mẫu mới, dùng bộ phân lớp để phân lớp mẫu nay vào một trong các lớp đuược rút ra từ mô hình ở bước 1. Trong dữu liệu kiểm tra của hình 2.5, bệnh nhân khai thác có các giá trị: HCT

= 59.3; PTL = 160.1; NS1 = “Dương tính” thì mô hình sẽ phân lớp cho trường hợp này là kết “Kết quả chuẩn đoán” = “Dương tính” (hình 2.5).

Một số vấn đề cho bộ phân lớp cần quan tâm giải quyết:

- Độ chính xác: Độ tin cậy của một luật dựa vào độ chính xác khi phân lớp.
- Tốc độ: Trong một số tình huống, tốc độ phân lớp được xem như là một yếu tố quan trọng.
- Dễ hiểu: Một bộ phân lớp dễ hiểu sẽ tạo cho người sử dụng tin tưởng hơn vào hệ thống, đồng thời giúp cho người sử dụng tránh đươuọc việc hiểu lầm kết quả của một luật đươuọc đưa ra bởi hệ thống.
- Đơn giản: Kết quả đưa ra cây quyết định liên quan kích thước của nó.
- Thời gian để học: Khi hệ thống hoạt động trong môi trường thay đổi thường xuyên, điều đó yêu cầu hệ thống phải học rất nhanh một luật phân lớp hoạc nhanh chóng điều chỉnh một luật đã được học cho phù hợp với thực tế.

Các kỹ thuật phân lớp:

- Mô hình phân lớp dùng cây quyết định (Decision tree classification)
- Phân lớp dùng mạng Neural
- Phân lớp dùng mạng Bayes
- Phân lớp với K-nearest neighbor classifier
- Phân tích thống kê
- Các thuật toán di truyền
- Phương pháp tập thô (Rough set Approach)

2.4 Cơ sở dữ liệu Y khoa

2.4.1 Sơ lược bệnh Tiểu đường

Bệnh tiểu đường, theo y học còn gọi là bệnh đái tháo đường, là một rối loạn chuyển hóa mạn tính rất phổ biến. Khi mắc bệnh, cơ thể bạn mất đi khả năng sử dụng hoặc sản xuất ra hormone insulin một cách thích hợp.

Mặc bệnh này có nghĩa là bạn có lượng đường trong máu quá cao do nhiều nguyên nhân. Tình trạng này có thể gây ra các vấn đề nghiêm trọng cho cơ thể, bao gồm cả mắt, thận, thần kinh và tim.

2.4.2 Diễn biến lâm sàng bệnh Tiểu đường

❖ Phân loại

- Loại 1 (type 1, Juvenile *diabetes*)

  Khoảng 5-10% tổng số bệnh nhân bệnh tiểu đường thuộc loại 1 (*type 1 diabetes*), phần lớn xảy ra ở trẻ em và người trẻ tuổi (dưới 20 tuổi). Các triệu chứng thường khởi phát đột ngột và tiến triển nhanh nếu không điều trị.

  Bệnh tiểu đường type 1 do sự bất thường tế bào β đảo Langerhans làm giảm tiết hormone insulin (có chức năng kích thích tế bào hấp thụ, sử dụng glucose huyết và kích thích gan polymer hóa glucose thành glycogen, từ đó làm giảm lượng đường huyết) trong khi tế bào đích của insulin không có hiện tượng kháng insulin (*insulin resistance*), đặc trưng bởi sự giảm nhạy cảm hoặc hư hỏng thụ thể tiếp nhận insulin, *Insulin receptor*).

  Thông thường, bệnh đái tháo đường type 1 thường có nguyên nhân do di truyền. Nó thường xuất hiện đột ngột và diễn biến nhanh ở trẻ em. Tuy nhiên, cũng có một số trường hợp bệnh xuất hiện tương đối muộn, ở người trưởng thành, gọi là bệnh đái tháo đường tiềm ẩn tự miễn ở người trưởng thành LADA (*Latent autoimmune diabetes in adults*) hoặc bệnh đái tháo đường type 1.5. 80% người mắc bệnh LADA được chẩn đoán nhầm sang đái tháo đường type 2.

- Loại 2 (type 2)

  Bệnh tiểu đường loại 2 chiếm khoảng 90 - 95 % trong tổng số bệnh nhân bệnh tiểu đường, thường gặp ở lứa tuổi trên 40, nhưng gần đây xuất hiện ngày càng nhiều ở lứa tuổi 30, thậm chí cả lứa tuổi thanh thiếu niên. Bệnh nhân thường ít có triệu chứng và thường chỉ được phát hiện bởi các triệu chứng của biến chứng, hoặc chỉ được phát hiện tình cờ khi đi xét nghiệm máu trước khi mổ hoặc khi có biến chứng như nhồi máu cơ tim, tai biến mạch máu não; khi bị nhiễm trùng da kéo dài; bệnh nhân nữ hay bị ngứa vùng kín do nhiễm nấm âm hộ; bệnh nhân nam bị liệt dương.
- Bệnh tiểu đường do thai nghén

  Tỷ lệ bệnh tiểu đường trong thai kỳ chiếm 3 - 5 % số thai nghén; phát hiện lần đầu tiên trong thai kỳ.

❖ Bệnh sinh

Sự thiếu hụt insulin một cách tương đối hay tuyệt đối dẫn đến glucose không thể vận chuyển vào tế bào, làm rối loạn chuyển hóa các chất: glucid, lipid, protid, nước, điện giải,...

❖ Triệu chứng

Do tế bào không nhận được glucose nên tế bào hiểu rằng "cơ thể đang thiếu đường" do đó bằng đường liên hệ ngược, cơ thể buộc phải depolymer hóa glycogen thành glucose (*glycogenolysis*) để tăng lượng đường trong máu. Kết quả làm nồng độ glucose huyết cao và làm tăng áp suất thẩm thấu của máu. Điều này khiến nước theo gradient nồng độ khuếch tán vào máu làm tăng khối lượng máu và tăng huyết áp. Mặt khác, do nồng độ glucose cao nên tănghàm lượng glucose lắng đọng vào hemoglobin (tạo Hb1AC), vì thế người ta có thể xét nghiệm nồng độ Hb1AC để chẩn đoán đái tháo đường.

Tiểu nhiều: Do nồng độ glucose huyết cao, nên nồng độ glucose trong nước tiểu đầu cao. Nồng độ này vượt quá ngưỡng glucose thận nên một phần glucose không được tái hấp thu ở ống lượn gần (*proximal convoluted tubule*).

Ăn nhiều: Cơ thể không thể sử dụng đường để cung cấp năng lượng làm cho bệnh nhân nhanh đói chỉ sau bữa ăn một thời gian ngắn.

Uống nhiều: Mất nước làm kích hoạt trung tâm khát ở vùng hạ đồi, làm cho bệnh nhân có cảm giác khát và uống nước liên tục.

Gầy nhiều: Dù ăn uống nhiều hơn bình thường, nhưng do cơ thể không thể sử dụng glucose để tạo năng lượng, buộc phải tăng cường thoái hóa lipid và protid để bù trừ, làm cho bệnh nhân sụt cân, người gầy còm, xanh xao. Với bệnh nhân đái tháo đường loại 2 thường không có bất kỳ triệu chứng nào ở giai đoạn đầu và vì vậy bệnh thường chẩn đoán muộn khoảng 7 - 10 năm (chỉ có cách kiểm tra đường máu cho phép chẩn đoán được ở giai đoạn này).

2.4.3 Chuẩn đoán

❖ **Xét nghiệm máu**

- Đo nồng độ glucose trong máu lúc đói:

  Xác định tiểu đưuòng trong 2 lần xét nghiệm đều cho kết quả là nồng độ glucose trong máu lúc đói cao hơn 126 mg/dl. Khi kết quả xét nghiệm có nồng độ 110 và 126 mg/dl thì coi là tiền tiểu đường, báo hiệu nguy cơ bị tiểu đường type 2 với các biến chứng của bệnh.
- Đo nồng độ glucose sau khi ăn:

Nếu kết quả đo nồng độ glucose sau ăn cao hơn 200 mg/dl kèm các triệu chứng của bệnh (khát nhiều, đái nhiều và mệt mỏi) thì nghi ngờ bệnh tiểu đường.

- Đánh giá sự dung nạp sau khi uống glucose:

  Đôi khi bác sĩ muốn chuẩn đoán sớm bệnh đái tháo đường hơn nữa bằng cách cho uống glucose làm bộc lộ những trường hợp Đái tháo đưuòng nhẹ mà thửu máu theo cách thông thương không đủ tin cậy để chuẩn đoán. Cách đó gọi là "test dung nạp glucose bằng đưuòng uống". Xét nghiệm nồng độ glucose sau khi uống 2 giờ. Nếu kết quả xét nghiệm cho thấy nồng độ này vẫn cao hơn 200 mg/dl thì chuẩn đoán là bệnh tiểu đưuòng type 2.

- Tóm tắt:

  - Rối loạn hạ đường huyết

    Nếu kết quả đo mức đường máu lúc đói < 70 mg/dl là có rối loạn hạ đường huyết, như kết quả đo 53 mg/dl là có thể bị hôn mê do hạ đường huyết.

  - Tiền đái tháo đường

    Người có mức đường máu lúc đói từ >110 mg/dl được gọi là những người có "rối loạn dung nạp đường khi đói". Những người này tuy chưa được xếp vào nhóm bệnh nhân đái tháo đường, nhưng cũng không được coi là "bình thường" vì theo thời gian, rất nhiều người người "rối loạn dung nạp đường khi đói" sẽ tiến triển thành đái tháo đường thực sự nếu không có lối sống tốt. Mặt khác, người ta cũng ghi nhận rằng những người có "rối loạn dung nạp đường khi đói" bị gia tăng khả năng mắc các bệnh về tim mạch, đột quỵ hơn.

  - Đái tháo đường

    Đường máu lúc đói ≥ 126 mg/dl (≥ 7 mmol/l) thử ít nhất 2 lần liên tiếp.

Đường máu sau ăn hoặc bất kỳ ≥ 200 mg/dl (≥ 11,1 mmol/l).

- Định lượng HbA1C:

Ngoài các xét nghiệm này, HbA1C cũng là một xét nghiệm giúp việc chẩn đoán xác định bệnh tiểu đường mang lại kết quả chính xác. Glucose trong máu có thể gắn kết với hemoglobin (phần mang oxy) của hồng cầu để tạo nên một phức hợp gọi là HbA1C (Hemoglobin glycosylat). Một khi glucose gắn kết với hemoglobin, nó sẽ ở đó và tồn tại đến hết đời sống của hồng cầu kéo dài khoảng 3 tháng. Như vậy nếu nồng độ glucose trong máu càng cao thì lượng glucose gắn vào hemoglobin của hồng cầu càng nhiều, và như vậy nồng độ HbA1C cũng sẽ gia tăng. Định lượng HbA1C đánh giá hồi cứu tình trạng đường máu 2 - 3 tháng gần đây. Đường máu cân bằng tốt nếu HbA1C < 6,5%.

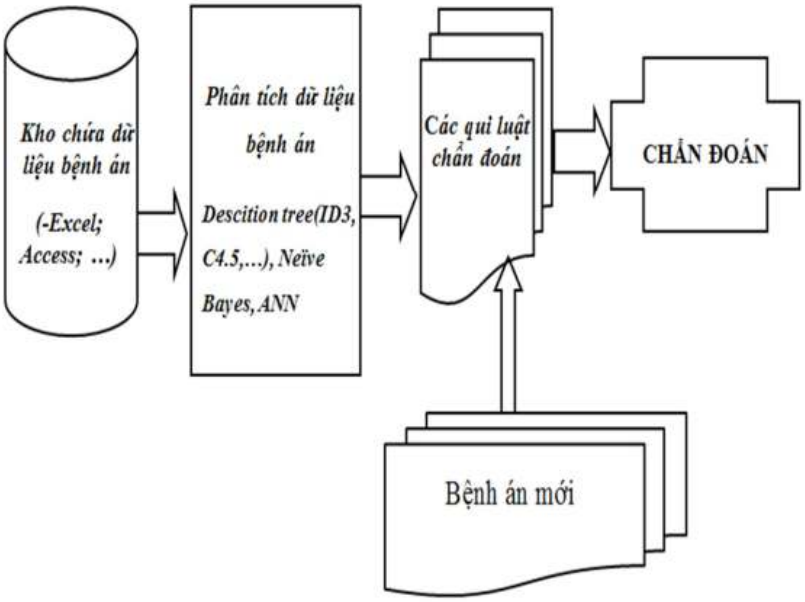
❖ **Các xét nghiệm bổ sung**

- Khám lâm sàng: kiểm tra cân nặng, huyết áp, bắt mạch ngoại biên và so sánh nhiệt độ da, khám bàn chân, khám thần kinh bao gồm thăm dò cảm giác sâu bằng âm thoa.
- Khám mắt: phát hiện và đánh giá tiến triển bệnh lý võng mạc.
- Xét nghiệm: đặc biệt lưu ý creatinin, mỡ máu, microalbumin niệu (bình thường < 30 mg/ngày) hoặc định lượng protein niệu. Đo điện tim nhằm phát hiện sớm các biểu hiện thiếu máu cơ tim. Soi đáy mắt..
- Fructosamin: cho biết đường máu trung bình 2 tuần gần đây, có nhiều lợi ích trong trường hợp người mắc đái tháo đường đang mang thai. Nếu đường máu cân bằng tốt, kết quả < 285 mmol/l.
- Peptid C (một phần của pro-insulin): cho phép đánh giá chức năng tế bào beta tụy.
- Thử đường trong nước tiểu.

CHƯƠNG 3: XÂY DỰNG MÔ HÌNH DỮ LIỆU SỬ DỤNG NAIVE BAYES

3.1 Cơ sở dữ liệu xây dựng mô hình

Sau khi thu thập dữ liệu ta cần xây dựng cơ sở dữ liệu, lưu trữ các thông tin cần thiết cho bộ điều khiển theo mô hình sau:



Hình 3.1: Mô hình xây dựng giải pháp hỗ trợ chuẩn đoán bệnh

3.2 Phương pháp Bayes sử dụng trong khai phá dữ liệu

3.2.1 Giới thiệu về phương pháp Bayes trong khai phá dữ liệu

Phân loại là việc gán một phần tử mới thích hợp nhất vào các tổng thể đã được biết trước dựa vào biến quan sát của nó. Đây là một hướng phát triển quan trọng của nhận dạng không được giám sát của thống kê. Bài toán phân loại được ứng dụng rộng rãi trong nhiều lĩnh vực khác nhau, đặc biệt trong xã hội, sinh học và y học. Hiện tại có ba phương pháp chính được đưa ra để giải quyết bài toán phân loại: phương pháp Fisher, phương pháp hồi quy logistic và phương pháp Bayes [2], [3], [10]. Phương pháp hồi quy logistic được sử dụng phổ biến nhất hiện nay, nhưng nó chỉ áp dụng cho dữ liệu rời rạc và chỉ phân loại cho hai tổng

thể. Phương pháp Fisher cũng áp dụng cho dữ liệu rời rạc, mặc dù có thể phân loại cho hai hay nhiều hơn hai tổng thể nhưng phải giả thiết ma trận hiệp phương sai của các tổng thể bằng nhau. Phương pháp Bayes có thể phân loại cho hai và nhiều hơn hai tổng thể, được xem có nhiều ưu điểm nhất vì nó đã đạt được mục tiêu về mặt lý thuyết cho bài toán phân loại. Các kết quả nghiên cứu mới trong những năm gần đây về bài toán phân loại chủ yếu tập trung xung quanh phương pháp Bayes. Một ưu điểm nổi bật của phương pháp này là tính được xác suất sai lầm trong phân loại mà nó được gọi là sai số Bayes. Sai số Bayes đã được chứng minh là xác suất sai lầm nhỏ nhất trong bài toán phân loại. Một số kết quả mới rất có ý nghĩa về phương pháp Bayes đã được trình bày trong những năm gần đây bởi các bài báo [6], [7], [8]. Một cản trở lớn của việc áp dụng thực tế bài toán phân loại bằng phương pháp Bayes trong những lĩnh vực cụ thể là vấn đề tính toán. Phương pháp Bayes dựa trên cơ sở hàm mật độ xác suất đã biết, tuy nhiên số liệu thực tế là số liệu rời rạc, vì vậy để phân loại bằng phương pháp Bayes có ý nghĩa thực tế việc đầu tiên là phải ước lượng hàm mật độ xác suất. Vấn đề tính sai số Bayes, phân loại một phần tử mới còn rất nhiều khó khăn khi gặp số liệu lớn của thực tế. Trong bài viết này, chúng tôi quan tâm đến lý thuyết tính toán các vấn đề liên quan đến phân loại bằng phương pháp Bayes từ số liệu rời rạc. Đặc biệt đưa ra một công thức tương đương của sai số Bayes mà nó rất thuận lợi cho việc tính toán. Các lý thuyết liên quan đến việc tính toán này sẽ được cụ thể hóa bằng các chương trình được viết trên phần mềm Matlab. Các chương trình này sẽ được sử dụng để áp dụng cho bài toán phân loại từ các số liệu rời rạc thực tế trong lĩnh vực sinh học và y học.

Phương pháp Bayes

- Cho X là một bộ dữ liệu được đo trên n thuộc tính khác nhau.
- Cho H là một bộ dữ liệu được đo trên n thuộc tính khác nhau.
- Đối với các bài toán phân lớp, chúng ta muốn xác định P(H|X) – là xác suất xảy ra H khi X đã xảy ra. Đây gọi là xác suất hậu nghiệm.

Ví dụ:

X được dùng để mô tả về bệnh nhân trên 2 thuộc tính là tuổi tác và nồng độ insulin. Và H là giả thuyết bệnh nhân sẽ bị tiểu đưuòng. Khi ấy P(H|X) biểu đạt xác suất bệnh nhân X sẽ bị bệnh tiểu đường khi đã biết tuổi tác và nồng độ insulin của bệnh nhân.

Ngược lại P(H) được gọi là xác suất tiêu nghiệm. Theo lý thuyết Bayes:

$$P(H|X) = \frac{P(X|H)P(H)}{P(X)}$$

Nguyên tắc hoạt động của bộ phân lớp Naïve Bayes

1. Cho D là tập dữ liệu huấn luyện cùng với các nhãn lớp tương ứng. Như thường lệ, mỗi bộ dữ liệu được mô tả bởi n thuộc tính và được diễn đạt dưới dạng vector n chiều X = (x1, x2,x3,…,xn).
2. Giả sử rằng có m nhãn lớp khác nhau gồm C1, C2,…, Cm. Cho một bộ dữ liệu X, bộ phân lớp sẽ dự đoán X thuộc về phân lớp có xác suất hậu nghiệm cao nhất.

$$P(C_i|X) > P(C_j|X) với\ 1 \le j \le m, j \ne i$$
$$P(C_i|X) = \frac{P(X|C_i)P(C_i)}{P(X)}$$

3. Do P(X) không đổi, nên ta chỉ cần cực đại hóa giá trị P(X|Ci)P(Ci)

Ví dụ:

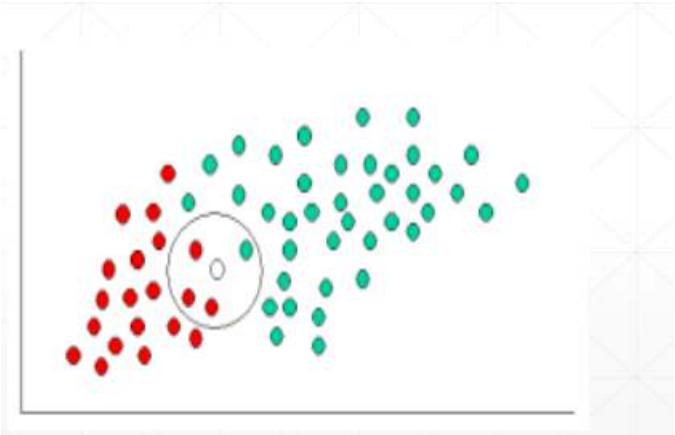
Dữ liệu được minh họa như hình:

Có 2 lớp: xanh và đỏ; N: tổng số đối tượng

P(xanh) = |xanh|/N = 40/60

P(đỏ) = |đỏ|/N = 20/60

Với các xác suất tiên nghiệm đã xác định ở trên: P(xanh) và P(đỏ) hãy xác định nhãn lớp cho các đối tượng x mới trên hình.



Lấy x làm tâm, vẽ vòng tròn giới hạn các đối tượng lân cận với x, tính:

P(x|xanh) = |xanh lân cận|/|xanh| = 1/40

P(x|đỏ) = |đỏ lân cận|/|đỏ| = 3/20

P(xanh|x) = P(x|xanh).P(xanh) = (1/40 *40/40) = 1/60

P(đỏ|x) = P(x|đỏ).P(đỏ) = (3/20 * 20/60) = 1/20

x được gán nhãn đỏ.

3.2.2 Thuật toán Bayes

3.2.2.1 Phân loại một phần tử mới

Cho k tổng thể w1, w2, ..., wk có biến quan sát với hàm mật độ xác suất được xác định là f1(x), f2(x), …, fk(x) và xác suất tiên nghệm cho các tổng thể lần lượt là , ,..., , q1 q2 qk q1 + q2 .+...+ qk =1. Ta có nguyên tắc phân loại một phần tử mới với biến quan sát x bằng phương pháp Bayes như sau: Nếu ( ) ( ) max g x q f x = j j thì xếp phần tử mới vào , wj (1) Trong đó: qi là xác suất tiên nghiệm của tổng thể thứ i, g (x) q f (x) i = i i và gmax (x) = max{g1 (x), g2 (x),...,gk (x)}.

3.2.2.2 Sai số Bayes

❖ **Trường hợp hai tổng thể**

Trong trường hợp không quan tâm đến xác suất tiên nghiệm q của w1 , ta có: 1 τ = P(w2|w1) = qf ( )dx n R x ∫ 2 1 : xác suất phân loại một phần tử vào w2 khi nó thuộc w1 .

τ = P(w1|w2) = q f ( )dx n R x ∫ - 1 2 (: xác su 1) ất phân loại một phần tử vào w1 khi nó thuộc w2 .

Trong đó: { } | ( ) (1) ( ) , 1 1 2 R x qf x q f x n = ≥ - R2 {x | qf1 (x) (1 q) f 2 (x)} n = < - .

Xác suất sai lầm trong phân loại Bayes được gọi là sai số Bayes và được xác định bởi công thức:

$$Pe = \tau 1 + \tau 2.$$

Khi quan tâm đến xác suất tiên nghiệm q của w1 thì 1 τ trở thành 1 τ) và 2 τ trở thành 2 τ) với

$$\bar{\tau}_1 = \int_{\bar{R}_2^n} q f_1(x)dx \text{ và } \bar{\tau}_2 = \int_{\bar{R}_1^n} (1-q) f_2(x)dx$$

Trong đó

$$\bar{R}_1^n = \{x \mid q f_1(x) \geq (1-q) f_2(x)\}, \quad \bar{R}_2^n = \{x \mid q f_1(x) < (1-q) f_2(x)\}.$$

Đặt (q) = (q, 1- q), khi đó sai số Bayes xác định bởi

$$Pe^{(q)} = \tau_1^* + \tau_2^*.$$

τ và 2 τ ; 1 τ) và 2 τ) được gọi chung là hai thành phần của sai số Bayes.

❖ **Trường hợp nhiều hơn hai tổng thể**

Sai số Bayes trong phân loại k tổng thể được định nghĩa bởi biểu thức

$$Pe_{1,2,\ldots,k}^{(q)} = \sum_{i=1}^{k} \int_{R^n \setminus R_i^n} q_i f_i dx$$

Để thuận lợi hơn trong tính sai số Bayes, người ta thường tính xác suất của sự phân loại đúng $Pe_{1,2,\ldots,k}^{(c)} = \sum_{i=1}^{k} \int_{R_i^n} q_i f_i dx$ khi đó sai số Bayes sẽ được tính bởi:

$$Pe_{1,2,\ldots,k}^{(q)} = 1 - Pc_{1,2,\ldots,k}^{(q)}.$$

3.3 Thuật toán Naive Bayes trong giải quyết bài toán chuẩn đoán bệnh tiểu đường

3.3.1 Thuật toán Bayes

Lý thuyết Bayes thì có lẽ không còn quá xa lạ nữa rồi. Nó chính là sự liên hệ giữa các xác suất có điều kiện. Điều đó gợi ý cho chúng ta rằng chúng ta có thể tính toán một xác suất chưa biết dựa vào các xác suất có điều kiện khác. Thuật toán ***Naive Bayes*** cũng dựa trên việc tính toán các xác suất có điều kiện đó. Nghe tên thuật toán là đã thấy gì đó ngây ngô rồi. Tại sao lại là **Naive** nhỉ. Không phải ngẫu nhiên mà người ta đặt tên thuật toán này như thế. Tên gọi này dựa trên một giả thuyết rằng các chiều của dữ liệu X=(x_1, x_2, ..., x_n)X=(x1,x2,....,xn) là độc lập về mặt xác suất với nhau.

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

Chúng ta có thể thấy rằng giả thuyết này có vẻ khá ngây thơ vì trên thực tế điều này có thể nói là không thể xảy ra tức là chúng ta rất ít khi tìm được một tập dữ liệu mà các thành phần của nó không liên quan gì đến nhau. Tuy nhiên, giả thiết ngây ngô này lại mang lại những kết quả tốt bất ngờ. Giả thiết về sự độc lập của các chiều dữ liệu này được gọi là Naive Bayes (xin phép không dịch). Cách xác định class của dữ liệu dựa trên giả thiết này có tên là Naive Bayes Classifier (NBC). Tuy nhiên dựa vào giả thuyết này mà bước training và testing trở nên vô cùng nhanh chóng và đơn giản. Chúng ta có thể sử dụng nó cho các bài toán large-scale. Trên thực tế, NBC hoạt động khá hiệu quả trong nhiều bài toán thực tế, đặc biệt là trong các bài toán phân loại văn bản, ví dụ như lọc tin nhắn rác hay lọc email spam. Trong bài viết này mình sẽ cùng với các bạn áp dụng lý thuyết về NBC để giải quyết một bài toán mới đó chính là bài toán chuẩn đoán bệnh tiểu đường

3.3.2 Tập dữ liệu tiểu đường

Tập dữ liệu này bao gồm dữ liệu của 768 tình nguyện viên bao gồm những người bị tiểu đường và những người không bị tiểu đường. Tập dữ liệu này bao gồm các thuộc tính như sau:

1. Số lần mang thai
2. Nồng độ glucose huyết tương trong 2 giờ xét nghiệm dung nạp
3. Huyết áp tâm tương (mmHg)
4. Triceps độ dày nếp gấp da (mm)
5. Insulin huyết thanh 2 giờ(mu U/ml)
6. Chỉ số khối cơ thể (cân nặng tính bằng kg / chiều cao (tính bằng m)^2)
7. Chức năng phả hệ tiểu đường

8. Tuổi

Với mỗi tình nguyện viện, dữ liệu bao gồm tập hợp các chỉ số kể trên và tình trạng bị bênh ***tức class 1*** hay không bị bệnh ***tức class 0***. Về bản chất đây là một bài toán phân loại 2 lớp và chúng ta có thể sử dụng các phương pháp phân loại khác như ***SVM, Random Forest, KNN…*** để phân loại cũng cho kết quả khá tốt. Nếu có dịp mình sẽ trình bày phương pháp này trong một dịp khác. Chúng ta có thể hình dung tập dữ liệu này thông qua biểu diễn dưới dạng file CSV như sau, trong đó cột cuối cùng chính là tình trạng bị bệnh của tình nguyện viên, các cột từ 1 đến 8 tương ứng với các chỉ số nếu trên

1	so_lan_ma	nong_do_g	huyet_ap_	triceps_do	insulin_hu	oi_co_the	ieu_duong	tuoi	bien_lop
2	6	148	72	35	0	33,6	0,627	50	1
3	1	85	66	29	0	26,6	0,371	31	0
4	8	183	64	0	0	23.3	0,672	32	1
5	1	89	66	23	94	28.1	0.167	21	0
6	0	137	40	35	168	43.1	2,288	33	1
7	5	116	74	0	0	25,6	0,201	30	0
8	3	78	50	32	88	31	0,248	26	1
9	10	115	0	0	0	35.3	0,125	29	0
10	2	197	70	45	543	30,5	0,125	53	1
11	8	125	96	0	0	0,0	0,232	54	1
12	4	110	92	0	0	37,6	0,191	30	0
13	10	168	74	0	0	38,0	0,537	34	1
14	10	139	80	0	0	27.1	1.441	57	0
15	1	189	60	23	846	30.1	0,388	59	1
16	5	166	72	19	175	25.8	0,587	51	1
17	7	100	0	0	0	30,0	0,484	32	1
18	0	118	84	47	230	45,8	0,551	31	1
19	7	107	74	0	0	29,6	0,254	31	1
20	1	103	30	38	83	43.3	0,183	33	0
21	1	115	70	30	96	34,6	0,529	32	1
22	3	126	88	41	235	39.3	0,704	27	0
23	8	99	84	0	0	35,4	0,388	50	0
24	7	196	90	0	0	39,8	0,471	41	1
25	9	119	80	35	0	29,0	0,223	29	1
26	11	143	94	33	146	36,6	0,254	51	1
27	10	125	70	26	115	31.1	0,205	41	1
28	7	147	76	0	0	39,4	0,257	43	1
29	1	97	66	15	140	23.2	0,487	22	0

Hình 3.2: Bảng dữ liệu dataset bệnh tiểu đường

Có một điều nhận thấy rằng giá trị của các chỉ số là một biến liên tục chứ không phải một giá trị rời rạc chính vì thế nên khi áp dụng thuật toán Naive Bayes chúng ta cần phải áp dụng một phân phối xác suất cho nó. Một trong những phân phối xác suất phổ biến được sử dụng trong phần này đó chính là phân phối Gaussian. Chúng ta cùng tìm hiểu qua một chút về nó nhé. Phải hiểu được bản chất thì mới có thể thực hành được.

3.3.3 Phân phối Gaussian

Với một một dữ liệu x_ixi thuộc một class c_ici chúng ta thấy x_ixi tuân theo một phân phối chuẩn với kì vọng \muμ và độ lệch chuẩn \sigmaσ. Khi đó hàm xác suất của x_ixi được xác định như sau:

$$p(x_i|c_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right)$$

Đây chính là cách tính của thư viện ***sklearn*** tuy nhiên trong bài viết này mình sẽ hướng dẫn các bạn cài đặt thủ công. Chính việc cài đặt thủ công này giúp cho chúng ta hiểu hơn về bài toán.