HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

Hoàng Văn Thắng

ỨNG DỤNG KHAI PHÁ DỮ LIỆU TRONG HỖ TRỢ CHẨN ĐOÁN BỆNH ĐÁI THÁO ĐƯỜNG TUÝP 2

LUẬN VĂN THẠC SĨ KỸ THUẬT

(*Theo định hướng ứng dụng*)

HÀ NỘI – 2020

HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

Hoàng Văn Thắng

ỨNG DỤNG KHAI PHÁ DỮ LIỆU TRONG HỖ TRỢ CHẨN ĐOÁN BỆNH ĐÁI THÁO ĐƯỜNG TUÝP 2

CHUYÊN NGÀNH : HỆ THỐNG THÔNG TIN

MÃ SỐ: 8.48.01.04

LUẬN VĂN THẠC SĨ KỸ THUẬT

(*Theo định hướng ứng dụng*)

NGƯỜI HƯỚNG DẪN KHOA HỌC

TS. ĐỖ THỊ BÍCH NGỌC

HÀ NỘI - 2020

LỜI CAM ĐOAN

Tôi cam đoan rằng luận văn này: "***Ứng dụng khai phá dữ liệu trong hỗ trợ chẩn đoán bệnh đái tháo đường tuýp 2***" là bài nghiên cứu của chính tôi. Ngoại trừ những tài liệu tham khảo được trích dẫn trong luận văn này, tôi cam đoan rằng toàn phần hay những phần nhỏ của luận văn này chưa từng được công bố hay được sử dụng để nhận bằng cấp ở những nơi khác.

Không có sản phẩm/nghiên cứu nào của người khác được sử dụng trong luận văn này mà không được trích dẫn theo đúng quy định.

Luận văn này chưa bao giờ được nộp để nhận bất kỳ bằng cấp nào tại các trường Đại học hoặc cơ sở đào tạo khác.

Hà Nội, ngày tháng 12 năm 2019

Tác giả luận văn

Hoàng Văn Thắng

LỜI CẢM ƠN

Trước hết, tôi xin được tỏ lòng biết ơn và gửi lời cám ơn chân thành đến TS. Đỗ Thị Bích Ngọc người trực tiếp hướng dẫn luận văn, đã tận tình chỉ bảo và hướng dẫn tôi tìm ra hướng nghiên cứu, tiếp cận thực tế, tìm kiếm tài liệu, xử lý và phân tích số liệu, giải quyết vấn đề nhờ đó tôi mới có thể hoàn thành luận văn cao học của mình.

Ngoài ra, trong quá trình học tập, nghiên cứu và thực hiện đề tài tôi còn nhận được nhiều sự quan tâm, góp ý, hỗ trợ quý báu của quý thầy cô, đồng nghiệp, bạn bè và người thân. Tôi xin bày tỏ lòng biết ơn sâu sắc đến:

Ban giám hiệu, Ban lãnh đạo Khoa Sau đại học, Ban lãnh đạo Khoa Công nghệ thông tin cùng các quý thầy cô – Học viện Công nghệ Bưu chính Viễn thông đã tạo điều kiện giúp tôi hoàn thành Luận văn này.

Ban giám đốc Học viện Y Dược học cổ truyền Việt Nam, Ban giám đốc Bệnh viện Tuệ Tĩnh và đội ngũ cán bộ, y bác sĩ, sinh viên và các bệnh nhân tại Bệnh viện Tuệ Tĩnh đã rất nhiệt tình tham gia trả lời phỏng vấn nghiên cứu cho đề tài.

Cuối cùng, chân thành cảm ơn Cha mẹ và những người thân trong gia đình đã hỗ trợ, tạo điều kiện thuận lợi cho tôi trong suốt thời gian qua và đặc biệt trong thời gian tôi theo học khóa thạc sỹ tại Học viện Công nghệ Bưu chính Viễn thông.

MỤC LỤC

DANH MỤC CÁC THUẬT NGỮ, CHỮ VIẾT TẮT

Viết tắt	Tiếng Anh	Tiếng Việt
LADA		Đái tháo đường tự miễn tiềm tàng ở người trưởng thành
FPG	Fasting Plasma Glucose	Lượng đường Glucose lúc đói
OGTT	Oral Glucose Tolerance Test	Lượng đường Glucose sau khi nạp đường
HbA1c	Glycated Hemoglobin	
DNA	Axit đêôxyribônuclêic	Chuỗi ADN
	Robot Locomotion	Cử động robot
	Supervised Learning	Học có giám sát
	Agent	Hành động
	Classification	Phân chia dữ liệu
	Input	Đầu vào
	Output	Đầu ra
	Maximum Margin Classifiers	Phân loại tối đa khoảng cách
NBC	Naive Bayes Classification	
	Training data	Dữ liệu huấn luyện
SMO	Sequential Minimal Optimization	
SVM	Support Vector Machines	
	Class	Lớp
CSDL		Cơ sở dữ liệu

DANH SÁCH BẢNG

DANH SÁCH HÌNH VẼ

MỞ ĐẦU

1. Lý do chọn đề tài

Đái tháo đường là một trong những vấn đề y tế toàn cầu cấp bách của của thế kỷ 21, là gánh nặng tài chính cho chăm sóc y tế cản trở quá trình đạt mục tiêu phát triển bền vững, đặc biệt ở các nước thu nhập thấp và trung bình. Trên toàn thế giới, năm 2015, có 415 triệu người mắc bệnh đái tháo đường, chi phí y tế toàn cầu cho điều trị đái tháo đường và các biến chứng là 673 tỷ USD. Số bệnh nhân mắc bệnh Đái tháo đường dự báo tăng 55% vào năm 2040, với chi phí y tế toàn cầu cho Đái tháo đường lên tới 802 tỷ USD[20].

Tại Việt Nam, năm 2015 có 3.5 triệu người mắc bệnh, chiếm 6% người lớn trong độ tuổi từ 20 tới 79[2]. Năm 2040, số người mắc bệnh có thể lên tới 6.1 triệu người. Chi phí y tế trên đầu người là 162.7 USD[2].

Theo điều tra năm 2015 của Bộ Y tế, tỉ lệ mắc đái tháo đường trong độ tuổi 50-69 là 7.7% và có xu hướng ngày càng trẻ hoá [2]. Chỉ có 31.1% bệnh nhân đái tháo đường được chẩn đoán. Do đó, việc phát hiện sớm sẽ giúp người bệnh tiết kiệm chi phí điều trị và hạn chế thấp nhất biến chứng.

Bệnh đái tháo đường tuýp 2 chiếm gần 90% các trường hợp đái tháo đường và thường được gọi là bệnh đái tháo đường khởi phát ở người lớn hoặc bệnh đái tháo đường không phụ thuộc insulin. Trong trường hợp này các cơ quan của cơ thể trở nên kháng insulin, và điều này làm tăng nhu cầu về insulin. Tại điểm này, tuyến tụy không tạo ra lượng insulin cần thiết. Để giữ loại này Bệnh đái tháo đường, bệnh nhân phải tuân theo chế độ ăn kiêng nghiêm ngặt, tập thể dục thường xuyên và theo dõi đường huyết. Béo phì, thừa cân, không hoạt động thể chất có thể dẫn đến Bệnh đái tháo đường loại 2. Ngoài ra khi lão hóa, nguy cơ phát triển bệnh đái tháo đường tăng theo thời gian. Phần lớn bệnh nhân đái tháo đường loại 2 mắc bệnh đái tháo đường ở biên hoặc Tiền đái tháo đường, một tình trạng nồng độ glucose trong máu cao hơn bình thường nhưng không cao bằng bệnh nhân đái tháo đường.

Những năm gầy đây công nghệ thông tin trong ngành Y tế được đẩy mạnh và có nhiều bước phát triển mạnh mẽ để trợ giúp đội ngũ bác sĩ và các bệnh nhân. Bệnh án điện tử đã và đang phát triển đưa tới tiềm năng khai thác dữ liệu về các bệnh án để hỗ trợ chẩn đoán.

Vì vậy việc khai phá dữ liệu về bệnh án từ đó hỗ trợ các bác sĩ có thể đưa ra các chẩn đoán bước đầu nhanh hơn, dễ dàng hơn. Xuất phát từ những nhu cầu thực tế trên và đó là những lý do học viên chọn đề tài "***Ứng dụng khai phá dữ liệu trong hỗ trợ chẩn đoán bệnh đái tháo đường tuýp 2***".

2. Tổng quan về vấn đề nghiên cứu

Xuất phát từ thực trạng các bác sĩ luôn trong tình trạng quá tải tại nhiều bệnh viện và các cơ sở khám chữa bệnh; Vì vậy cần nghiên cứu hệ thống hỗ trợ chẩn đoán bệnh trợ giúp công tác khám và chẩn đoán cho các Bác sĩ. Để hoàn thành đề tài nghiên cứu học viên thực hiện các định hướng nghiên cứu bao gồm:

- Tìm hiểu về khai phá dữ liệu và các thuật toán
- Phân tích và thu thập thông tin dữ liệu từ các bệnh án;
- Thử nghiệm và lựa chọn thuật toán phù hợp với bài toán hỗ trợ chuẩn đoán bệnh đái tháo đường tuýp 2.
- Báo cáo đánh giá kết quả.

3. Mục đích nghiên cứu

Nghiên cứu tìm hiểu các thuật toán trong chẩn đoán bệnh đái tháo đường, từ đó áp dụng và thử nghiệm hỗ trợ chẩn đoán bệnh đái tháo đường tuýp 2.

4. Đối tượng và phạm vi nghiên cứu

Đối tượng nghiên cứu: Nghiên cứu thông tin dữ liệu về các bệnh án đái tháo đường tuýp 2.

5. Phương pháp nghiên cứu

Nghiên cứu về khai phá dữ liệu và các thuật toán liên quan.

Phân tích dữ liệu các bệnh án, hỗ trợ chẩn đoán bệnh đái tháo đường.

Thử nghiệm các thuật toán và lựa chọn cho hỗ trợ chẩn đoán bệnh đái tháo đường tuýp 2.

CHƯƠNG 1: BÀI TOÁN HỖ TRỢ CHẨN ĐOÁN BỆNH ĐÁI THÁO ĐƯỜNG

1.1. Bệnh đái tháo đường là gì ?

Bệnh đái tháo đường là một bệnh mạn tính xảy ra khi tuyến tụy không sản xuất đủ insulin hoặc khi cơ thể không thể sử dụng hiệu quả insulin nó tạo ra.

1.1.1. Các loại bệnh đái tháo đường.

Bệnh đái tháo đường có thể được phân thành bốn loại chính sau đây:

1. Đái tháo đường loại 1 là một bệnh tự miễn mạn tính xảy ra khi hệ thống miễn dịch của chính cơ thể tấn công các tế bào beta sản xuất insulin của tuyến tụy. Đái tháo đường loại 1 chiếm khoảng 5-10% số những người bị đái tháo đường. Trong đái tháo đường loại 1, các yếu tố di truyền, biểu sinh, môi trường và miễn dịch phá hủy β tế bào của tụy nội tiết và dẫn đến thiếu hụt insulin. Đái tháo đường loại 1 thường xảy ra ở trẻ em và thanh thiếu niên, nhưng có thể phát triển ở người lớn, chẳng hạn như dạng đái tháo đường tự miễn tiềm ẩn ở người trưởng thành (LADA).
2. Đái tháo đường loại 2 là loại phổ biến nhất, chiếm khoảng 90% trong tất cả các trường hợp đái tháo đường. Đái tháo đường loại 2 là kết quả của sự kết hợp của các yếu tố di truyền, môi trường, lối sống, thừa cân, huyết áp cao và cholesterol cao. Đái tháo đường loại 2 là một rối loạn chuyển hóa trong một thời gian dài, được đặc trưng bởi glucose máu cao, kháng insulin và thiếu insulin tương đối.
3. Đái tháo đường thai kỳ xảy ra ở phụ nữ mang thai ở tuần 24-28. Đái tháo đường thai kỳ chiếm khoảng 3-5% số thai phụ, phổ biến nhất là đái tháo đường loại 2. Đái tháo đường thai kỳ hoàn toàn có thể điều trị được, nhưng cần có sự giám sát y tế cẩn thận trong suốt thai kỳ. Nếu được điều trị, thai và trẻ sơ sinh có thể khỏe mạnh.

4. Các loại đái tháo đường khác: các loại đái tháo đường này chỉ chiếm khoảng 2% trong tất cả các trường hợp đái tháo đường. Các loại đái tháo đường khác có thể được chia thành đái tháo đường đơn gen, đái tháo đường do bệnh tụy ngoại tiết, do bệnh nội tiết, do thuốc, đái tháo đường qua trung gian tự miễn và đái tháo đường liên quan đến các hội chứng di truyền.

1.1.2. Tiêu chuẩn chẩn đoán bệnh Đái tháo đường

Tiêu chuẩn chẩn đoán đái tháo đường của Bộ Y Tế [1] (theo Hiệp Hội Đái tháo đường Mỹ - ADA) dựa vào 1 trong 4 tiêu chuẩn sau đây:

a, Glucose huyết tương lúc đói (fasting plasma glucose: FPG) ≥ 126 mg/dL (hay 7 mmol/L). Bệnh nhân phải nhịn ăn (không uống nước ngọt, có thể uống nước lọc, nước đun sôi để nguội) ít nhất 8 giờ (thường phải nhịn đói qua đêm từ 8 -14 giờ), hoặc:

b, Glucose huyết tương ở thời điểm sau 2 giờ làm nghiệm pháp dung nạp glucose đường uống 75g (oral glucose tolerance test: OGTT) ≥ 200 mg/dL (hay 11,1 mmol/L).

c, Nghiệm pháp dung nạp glucose đường uống phải được thực hiện theo hướng dẫn của Tổ chức Y tế thế giới: Bệnh nhân nhịn đói từ nửa đêm trước khi làm nghiệm pháp, dùng một lượng glucose tương đương với 75g glucose, hòa tan trong 250-300 ml nước, uống trong 5 phút; trong 3 ngày trước đó bệnh nhân ăn khẩu phần có khoảng 150-200 gam carbohydrat mỗi ngày.

d, HbA1c[19] ≥ 6,5% (48 mmol/mol). Xét nghiệm này phải được thực hiện ở phòng thí nghiệm được chuẩn hóa theo tiêu chuẩn quốc tế.

Ở bệnh nhân có triệu chứng kinh điển của tăng glucose huyết hoặc mức glucose huyết tương ở thời điểm bất kỳ ≥ 200 mg/dL (hay 11,1 mmol/L).

Nếu không có triệu chứng kinh điển của tăng glucose huyết (bao gồm tiểu nhiều, uống nhiều, ăn nhiều, sụt cân không rõ nguyên nhân), xét nghiệm chẩn đoán

a, b, d ở trên cần được thực hiện lặp lại lần 2 để xác định chẩn đoán. Thời gian thực hiện xét nghiệm lần 2 sau lần thứ nhất có thể từ 1 đến 7 ngày.

Trong điều kiện thực tế tại Việt Nam, nên dùng phương pháp đơn giản và hiệu quả để chẩn đoán đái tháo đường là định lượng glucose huyết tương lúc đói 2 lần ≥ 126 mg/dL (hay 7 mmol/L). Nếu HbA1c[19] được đo tại phòng xét nghiệm được chuẩn hóa quốc tế, có thể đo HbA1c[19] 2 lần để chẩn đoán Đái tháo đường.

1.2. Khai phá dữ liệu trong hỗ trợ chẩn đoán bệnh đái tháo đường.

1.2.1. Học máy và khám phá tri thức

Sử dụng thông tin một cách có hiệu quả là một vấn đề rất quan trọng để dẫn đến thành công[7].

Điều đó có nghĩa là từ các dữ liệu sẵn có phải tìm ra những thông tin tiềm ẩn có giá trị mà trước đó chưa được phát hiện, phải tìm ra những xu hướng phát triển và những yếu tố tác động lên chúng. Thực hiện công việc đó chính là thực hiện quá trình phát hiện tri thức trong cơ sở dữ liệu (Knowledge Discovery in Database – KDD) mà trong đó kỹ thuật này cho phép ta lấy được các tri thức chính là pha khai phá dữ liệu (KPDL).

Quá trình xử lý KPDL bắt đầu bằng cách xác định chính xác vấn đề cần giải quyết. Sau đó sẽ xác định các dữ liệu liên quan dùng để xây dựng giải pháp. Bước tiếp theo là thu thập các dữ liệu có liên quan và xử lý chúng thành dạng sao cho giải thuật KPDL có thể hiểu được. Về lý thuyết thì có vẻ rất đơn giản nhưng khi thực hiện thì đây thực sự là một quá trình rất khó khăn, gặp phải rất nhiều vướng mắc như: các dữ liệu phải được sao ra nhiều bản (nếu được chiết xuất vào các tệp), quản lý tập các tệp dữ liệu, phải lặp đi lặp lại nhiều lần toàn bộ quá trình (nếu mô hình dữ liệu thay đổi),… Bước tiếp theo là chọn thuật toán KPDL thích hợp và thực hiện việc KPDL để tìm được các mẫu (pattern) có ý nghĩa dưới dạng biểu diễn tương

ứng với các ý nghĩa đó (thường được biểu diễn dưới dạng các luật xếp loại, cây quyết định, luật sản xuất, biểu thức hồi quy,…). Đặc điểm của mẫu phải là các mẫu mới (ít nhất là đối với hệ thống đó). Độ mới có thể được đo tương ứng với độ thay đổi trong dữ liệu (bằng cách so sánh các giá trị hiện tại với các giá trị trước đó hoặc các giá trị mong muốn), hoặc bằng tri thức (mối liên hệ giữa phương pháp tìm mới và phương pháp cũ như thế nào). Thường thì độ mới của mẫu được đánh giá bằng một hàm logic hoặc một hàm đo độ mới, độ đột phá của mẫu. Ngoài ra, mẫu còn phải có khả năng sử dụng mở rộng. Các mẫu này sau khi được xử lý và diễn giải phải dẫn đến những hành động có ích nào đó được đánh giá bằng một hàm chức năng. Mẫu khai thác được phải có giá trị đối với các dữ liệu mới với độ chính xác nhất định.

Bước thứ nhất: Tìm hiểu lĩnh vực ứng dụng và hình thành bài toán, bước này sẽ quyết định cho việc rút ra được các tri thức hữu ích và cho phép chọn các phương pháp khai phá dữ liệu thích hợp với mục đích ứng dụng và bản chất của dữ liệu.

Bước thứ hai: Thu thập và xử lý dữ liệu thô, còn được gọi là tiền xử lý dữ liệu nhằm loại bỏ nhiễu, xử lý việc thiếu dữ liệu, biến đổi dữ liệu và rút gọn dữ liệu nếu cần thiết, bước này chiếm khá nhiều thời gian trong toàn bộ quy trình khám phá tri thức.

Bước thứ ba: Khai phá dữ liệu, hay nói cách khác là trích ra các mẫu hoặc/và các mô hình ẩn dưới các dữ liệu.

Bước thứ tư: Hiểu tri thức đã tìm được, đặc biệt là làm sáng tỏ các mô tả và dự đoán. Các bước trên có thể lặp đi lặp lại một số lần, kết quả thu được có thể được lấy trung bình trên tất cả các lần thực hiện.

Bước thứ năm: Sử dụng tri thức đã được khai phá vào thực tế. Các tri thức phát hiện được tích hợp chặt chẽ trong hệ thống. Tuy nhiên để sử dụng được các tri thức đó đôi khi cần đến các chuyên gia trong các lĩnh vực quan tâm vì tri thức rút ra

có thể chỉ mang tính chất hỗ trợ quyết định hoặc cũng có thể được sử dụng cho một quá trình khám phá tri thức khác.

Mặc dù được tóm tắt thành năm bước nhưng thực chất quá trình xây dựng và thực hiện việc khám phá tri thức không chỉ tuân theo các bước cố định mà các quá trình này còn có thể được lặp đi lặp lại ở một hoặc một số giai đoạn trước và cứ tiếp tục như thế sẽ làm cho quá trình khai phá và tìm kiếm dữ liệu ngày càng hoàn thiện hơn.

Học máy có hiện nay được áp dụng rộng rãi bao gồm máy truy tìm dữ liệu, chẩn đoán y khoa, phát hiện thẻ tín dụng giả, phân tích thị trường chứng khoán, phân loại các chuỗi DNA[8], nhận dạng tiếng nói và chữ viết, dịch tự động, chơi trò chơi và cử động rô-bốt (robot locomotion).

Các thuật toán học máy được phân loại theo kết quả mong muốn của thuật toán. Các loại thuật toán thường dùng bao gồm:

1.2.2. Học có giám sát

Học có giám sát [7] (supervised learning) là một kỹ thuật của ngành học máy nhằm mục đích xây dựng một hàm f từ dữ tập dữ liệu huấn luyện (Training data). Dữ liệu huấn luyện bao gồm các cặp đối tượng đầu vào và đầu ra mong muốn. Đầu ra của hàm f có thể là một giá trị liên tục hoặc có thể là dự đoán một nhãn phân lớp cho một đối tượng đầu vào.

Trong đó, thuật toán tạo ra một hàm ánh xạ dữ liệu vào tới kết quả mong muốn. Một phát biểu chuẩn về một việc học có giám sát là bài toán phân loại: chương trình cần học (cách xấp xỉ biểu hiện của) một hàm ánh xạ một vector $X_1, X_2, \dots X_n$ tới một vài lớp bằng cách xem xét một số mẫu dữ liệu - kết quả của hàm đó.

Bước 1: Xác định loại của các dữ liệu huấn luyện: Trước tiên ta cần phải quyết định xem loại dữ liệu nào sẽ được sử dụng làm dữ liệu huấn luyện. Ta có thể

chọn dữ liệu một kí tự viết tay đơn lẻ, toàn bộ một từ viết tay, hay toàn bộ một dòng chữ viết tay, …

Bước 2: Thu thập tập dữ liệu huấn luyện. Khi thu thập tập dữ liệu huấn luyện cần phải đảm bảo được sự đặc trưng cho thực tế sử dụng của hàm chức năng. Do đó tập các dữ liệu đầu vào và đầu ra tương ứng phải được thu thập từ các chuyên gia hoặc từ việc đo đạc tính toán.

Bước 3: Xác định việc biểu diễn các đặc trưng đầu vào cho hàm mục tiêu cần tìm. Độ chính xác của mục tiêu phụ thuộc rất lớn vào các đối tượng đầu vào được biểu diễn như thế nào. Đa số các đối tượng đầu vào được chuyển đổi thành một véc tơ đặc trưng chứa các đặc trưng cơ bản của đối tượng đó. Chú ý số lượng các đặc trưng không được lớn quá, để tránh sự bùng nổ tổ hợp tuy nhiên nó phải đủ lớn để đảm bảo dự đoán chính xác đầu ra.

Bước 4: Xác định cấu trúc của hàm mục tiêu cần tìm và giải thuật học tương ứng. Ví dụ, ta có thể sử dụng mạng nơ-ron nhân tạo, cây quyết định, …

Bước 5: Hoàn thiện và thiết kế chương trình.

Tiến hành chạy giải thuật học với tập dữ liệu huấn luyện thu thập được. Ta có thể điều chỉnh các tham số của giải thuật học bằng cách tối ưu hóa hiệu năng trên một tập con của tập huấn luyện, (gọi là tập kiểm chứng -validation set) của tập huấn luyện hay thông qua kiểm chứng chéo (cross-validation). Sau đó ta tiến hành đo đạc hiệu năng của giải thuật trên một tập dữ liệu kiểm tra độc lập với tập huấn luyện.

1.2.3. Học không có giám sát

Học không có giám sát [7](unsupervised learning) là một phương pháp nhằm tìm ra một mô hình mà phù hợp với các quan sát. Trong học không có giám sát, một tập dữ liệu đầu vào được thu thập. Học không có giám sát thường đối xử với các đối tượng đầu vào như là một tập các biến ngẫu nhiên. Sau đó, một mô hình mật độ kết hợp sẽ được xây dựng cho tập dữ liệu đó.

Tất cả dữ liệu không được gắn nhãn và các thuật toán tìm hiểu cấu trúc vốn có từ dữ liệu đầu vào. Mô hình hóa một tập dữ liệu, không có sẵn các ví dụ đã được gắn nhãn.

Học không có giám sát có thể được dùng kết hợp với các thuật toán để cho ra xác suất có điều kiện (nghĩa là học có giám sát) cho bất kì biến ngẫu nhiên nào khi biết trước các biến khác.

Học không có giám sát cũng hữu ích cho việc nén dữ liệu: về cơ bản, mọi giải thuật nén dữ liệu hoặc là dựa vào một phân bố xác suất trên một tập đầu vào một cách tường minh hay không tường minh.

Một dạng khác của học không có giám sát là gom nhóm dữ liệu (data clustering), nó đôi khi không mang tính xác suất.

1.2.4. Học giám sát một phần

Học nửa giám sát [7] (semi-supervised learning) là một lớp của kỹ thuật học máy, sử dụng cả dữ liệu đã gán nhãn và chưa gán nhãn để huấn luyện - điển hình là một lượng nhỏ dữ liệu có gán nhãn cùng với lượng lớn dữ liệu chưa gán nhãn.

Học nửa giám sát đứng giữa học không giám sát (không có bất kì dữ liệu có nhãn nào) và có giám sát (toàn bộ dữ liệu đều được gán nhãn). Nhiều nhà nghiên cứu nhận thấy dữ liệu không gán nhãn, khi được sử dụng kết hợp với một chút dữ liệu có gán nhãn, có thể cải thiện đáng kể độ chính xác. Để gán nhãn dữ liệu cho một bài toán học máy thường đòi hỏi một chuyên viên có kĩ năng để phân loại bằng tay các ví dụ huấn luyện. Chi phí cho quy trình này khiến tập dữ liệu được gán nhãn hoàn toàn trở nên không khả thi, trong khi dữ liệu không gán nhãn thường tương đối rẻ tiền. Trong tình huống đó, học nửa giám sát có giá trị thực tiễn lớn lao.

Một số dữ liệu được dán nhãn nhưng phần lớn dữ liệu còn lại không có nhãn và một hỗn hợp các kỹ thuật có giám sát và không giám sát có thể được sử dụng.

Kết hợp các ví dụ có gắn nhãn và không gắn nhãn để sinh một hàm hoặc một bộ phân loại thích hợp.

Một ví dụ cho kỹ thuật học máy nửa giám sát là đồng huấn luyện (co-training), trong đó một hay nhiều bộ học được huấn luyện cùng một tập ví dụ nhưng mỗi bộ sử dụng một tập đặc trưng khác nhau, lý tưởng nhất là độc lập với nhau.

Một cách tiếp cận khác là mô hình hoá phân phối xác suất đồng thời của các đặc trưng và nhãn. Với dữ liệu chưa gán nhãn, có thể coi nhãn là "dữ liệu còn thiếu". Các kỹ thuật xử lý dữ liệu còn thiếu như là lấy mẫu Gibbs và tối ưu kỳ vọng có thể được sử dụng để ước lượng tham số.

1.2.5. Học tăng cường

Học tăng cường [7] (reinforcement learning) là một lĩnh vực con của học máy, nghiên cứu cách thức một agent trong một môi trường nên chọn thực hiện các hành động nào để cực đại hóa một khoản thưởng (reward) nào đó về lâu dài. Các thuật toán học tăng cường cố gắng tìm một chiến lược ánh xạ các trạng thái của thế giới tới các hành động mà agent nên chọn trong các trạng thái đó.

Trong đó, thuật toán học một chính sách hành động tùy theo các quan sát về thế giới. Mỗi hành động đều có tác động tới môi trường, và môi trường cung cấp thông tin phản hồi để hướng dẫn cho thuật toán của quá trình học.

Do đó, học tăng cường đặc biệt thích hợp cho các bài toán có sự được mất giữa các khoản thưởng ngắn hạn và dài hạn. Học tăng cường đã được áp dụng thành công cho nhiều bài toán, trong đó có điều khiển robot, điều vận thang máy, viễn thông, các trò chơi có tính may mắn hoặc có tính chiến thuật cao và cờ vua.

1.3. Bài toán hỗ trợ chẩn đoán bệnh đái tháo đường

Từ mục 1.1, chúng ta thấy khai phá dữ liệu là một lĩnh vực đa ngành, là sự kết hợp giữa học máy, thống kê, công nghệ phân tích dữ liệu và trí tuệ nhân tạo.

Khai phá dữ liệu đã được chứng minh là rất có lợi trong lĩnh vực phân tích y tế vì nó làm tăng độ chính xác chẩn đoán, giảm chi phí điều trị bệnh nhân và tiết kiệm nguồn nhân lực[5].

Một số phương pháp dự đoán cho đái tháo đường tuýp 2 dựa vào các kỹ thuật khai phá dữ liệu. Các luật để trích chọn thông tin cần được giải thích. Tuy nhiên, trong y tế, các luât trích chọn không chỉ cần độ chính xác cao mà còn phải đơn giản và dễ hiểu.

Mục tiêu của luận văn: Đánh giá thuật toán cho tỷ lệ tốt nhất để áp dụng vào bài toán dự đoán bệnh nhân dương tính với bệnh Đái tháo đường tuýp 2.

Input hệ thống là: Gồm các chỉ số của bệnh án trong hồ sơ bệnh nhân.

Output của hệ thống là: Bài toán hệ hỗ trợ chẩn đoán bệnh đái tháo đường phù hợp với học có giám sát vì đây là một bài toán dựa trên các thuộc tính có dạng số trong hồ sơ bệnh nhân, class quyết định có 2 class là 0 và 1. Đưa ra tỷ lệ dự đoán chính xác nhất với bộ dataset tương ứng.

Kết luận chương 1

Chương 1 đã nêu ra được chủ đề cần nghiên cứu, trình bày các khái niệm về bệnh đái tháo đường, trình bày các mô hình học máy được sử dụng để giải quyết bài toán.

CHƯƠNG 2: KHẢO SÁT MỘT SỐ THUẬT TOÁN CHO HỖ TRỢ CHẨN ĐOÁN BỆNH ĐÁI THÁO ĐƯỜNG TUÝP 2

2.1. Giới thiệu chung

Trên thế giới, đã có nhiều nghiên cứu về áp dụng khai phá dữ liệu trong chẩn đoán bệnh đái tháo đường:

Nilam Chandgude và giáo sư Suvarna[13] trình bày thuật toán phân loại được sử dụng để chẩn đoán bệnh tiểu đường. Tác giả đã sử dụng mạng nơ ron, Cây quyết định Naïve Bayes, SVM, ID3, C 4.5, Thuật toán CART và so sánh những thuật toán này. Kết quả là CART cho độ chính xác tốt hơn các thuật toán khác.

Thirumal P. C. và Nagarajan .N [14] đã trình bày các kỹ thuật khai phá dữ liệu khác nhau để dự đoán bệnh đái tháo đường. Bộ dữ liệu bệnh tiểu đường của người Pima Ấn Độ được sử dụng để phân tích. Sau khi tiền xử lý dữ liệu, các thuật toán như Naïve Bayes Classifier, thuật toán C4.5, SVM, KNN được áp dụng. Kết quả là thuật toán C4.5 cung cấp độ chính xác cao hơn và KNN cung cấp độ chính xác thấp hơn.

K.Rajalakshmi và Tiến sĩ S.S.Dhenakaran [15] đã phân tích các kỹ thuật dự đoán khai phá dữ liệu trong các hệ thống quản lý chăm sóc sức khỏe. Các kỹ thuật khai phá dữ liệu như Cây quyết định, Phân loại Bayes, Mạng nơ ron và SVM được trình bày. Các kỹ thuật khai phá dữ liệu khác nhau được so sánh dựa trên dự đoán bệnh khác nhau. Thuật toán SVM thực hiện tốt trong việc dự đoán bệnh đái tháo đường.

Agarwal, Amit kumar Dewangan [16] tập trung trong chẩn đoán bệnh tiểu đường Mellitus sử dụng các kỹ thuật khai phá dữ liệu. Các tác giả đã phân tích xác thực chéo, phương pháp phân loại, lớp K - láng giềng gần nhất [CKNN], Vector hỗ trợ Máy [SVM], Máy Vector hỗ trợ LDA và Chuyển tiếp mạng nơ ron, Mạng nơ ron nhân tạo, chuẩn hóa thống kê và phương pháp lan truyền ngược để chẩn đoán

bệnh Đái tháo đường. Và chỉ ra rằng, SVM cho độ chính xác tốt hơn về bệnh Đái tháo đường đường tập dữ liệu.

Qua phân this các nghiên cứu, chúng ta thấy các thuật toán như Decision tree, C4.5, Naïve Bayes, SVM,… cho những kết quả rất tốt. Vì vậy, phần tiếp theo sẽ trình bày các thuật toán sẽ áp dụng vào bài toán xây dựng hệ hỗ trợ chẩn đoán bệnh Đái tháo đường tuýp 2.

2.2. Khảo sát mô hình Decision tree

Cây quyết định (gọi tắt là DT) là mô hình đưa ra quyết định dựa trên các câu hỏi. Cây quyết định (Decision Tree) là một mô hình thuộc nhóm thuật toán Học có giám sát (Supervised Learning).

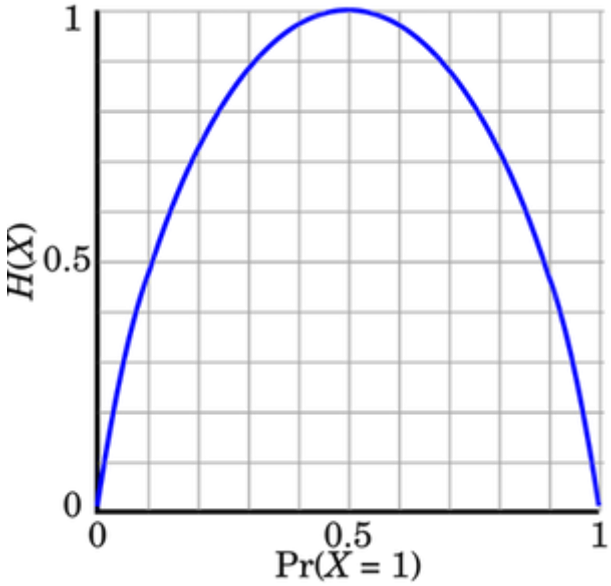
Hàm số Entropy

Cho một phân phối xác suất của một biến rời rạc x có thể nhận n giá trị khác nhau $x_1, x_2, \dots, x_n$. Giả sử rằng xác suất để x nhận các giá trị này là $p_i = p(x = x_i)$

Ký hiệu phân phối này là $p = (p_1, p_2, \dots, p_n)$.

Entropy của phân phối này là: $H(p) = -\sum_{i=1}^{n} p_i \log_2(p_i)$

Hàm Entropy được biểu diễn dưới dạng đồ thị như Hình 1:



Hình 1: Biểu đồ Entropy

Từ đồ thị ta thấy, hàm Entropy sẽ đạt giá trị nhỏ nhất nếu có một giá trị $p_i = 1$, đạt giá trị lớn nhất nếu tất cả các p_i bằng nhau.

Hàm Entropy càng lớn thì độ ngẫu nhiên của các biến rời rạc càng cao (càng không tinh khiết).

Với cây quyết định, ta cần tạo cây như thế nào để cho ta nhiều thông tin nhất, tức là Entropy là cao nhất.

Information Gain

Tại mỗi tầng của cây, cần chọn thuộc tính nào để độ giảm Entropy là thấp nhất.

Người ta có khái niệm Information Gain được tính bằng

$$Gain(S, f) = H(S) - H(f, S)$$

trong đó:

$H(S)$ là Entropy tổng của toàn bộ tập data set S.

$H(f, S)$ là Entropy được tính trên thuộc tính f.

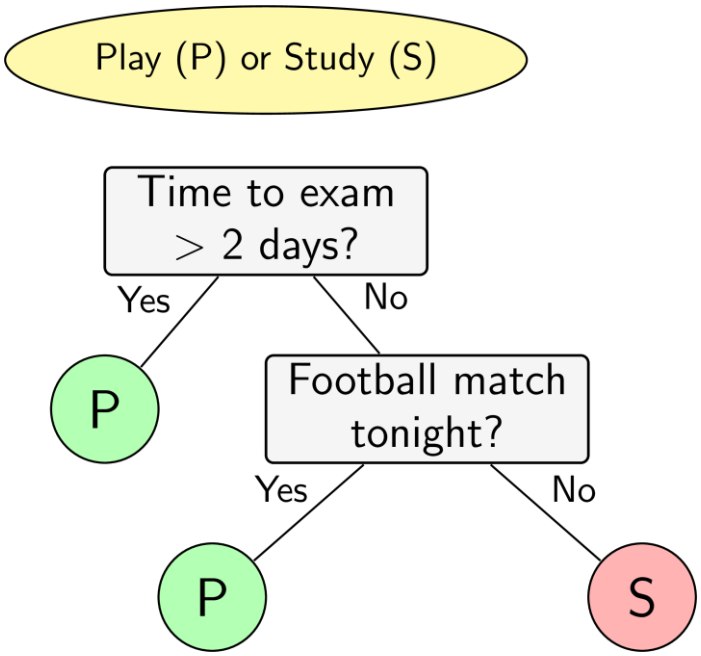
Do $H(S)$ là không đổi với mỗi tầng, ta chọn thuộc tính f có Entropy nhỏ nhất để thu được $Gain(S, f)$ lớn nhất.

Ví dụ minh hoạ:

Sắp đến kỳ thi, một cậu sinh viên tự đặt ra quy tắc **học** hay **chơi** của mình như sau. Nếu còn nhiều hơn hai ngày tới ngày thi, sinh viên đó ra sẽ đi chơi. Nếu còn không quá hai ngày và đêm hôm đó có một trận bóng đá, sinh viên đó sẽ sang nhà bạn chơi và cùng xem bóng đêm đó. Sinh viên đó sẽ chỉ học trong các trường hợp còn lại.

Việc ra quyết định của cậu sinh viên này có thể được mô tả trên sơ đồ trong Hình 2. Hình ellipse nền vàng thể hiện quyết định cần được đưa ra. Quyết định này

phụ thuộc vào các câu trả lời của các câu hỏi trong các ô hình chữ nhật màu xám. Dựa trên các câu trả lời, quyết định cuối cùng được cho trong các hình tròn màu lục (chơi) và đỏ (học). Sơ đồ trong Hình 2 còn được gọi là một cây quyết định.



Hình 2: Ví dụ về việc ra quyết định dựa trên các câu hỏi

Việc quan sát, suy nghĩ và ra các quyết định của con người thường được bắt đầu từ các câu hỏi. Machine learning cũng có một mô hình ra quyết định dựa trên các câu hỏi. Mô hình này có tên là cây quyết định (decision tree).

2.3. Khảo sát thuật toán C4.5

Phần lớn các hệ thống đều cố gắng để tạo ra một cây càng nhỏ càng tốt, vì những cây nhỏ hơn thì dễ hiểu hơn và dễ đạt được độ chính xác dự đoán co hơn. Do không thể đảm bảo được sự cực tiểu của cây quyết định, C4.5 dựa vào nghiên cứu tối ưu hóa, và sự lựa chọn cách phân chia mà có độ đo lựa chọn thuộc tính đạt giá trị cực đại.

Hai độ đo được sử dụng trong C4.5 là information gain và gain ratio. $RF(C_j, S)$ biểu diễn tần xuất (Relative Frequency) các case trong S thuộc về lớp C_j

$$\boldsymbol{RF(C_j, S) = {|S_j|}/{|S|}}$$

Với $|S_j|$ là kích thước tập các case có giá trị phân lớp là C_j. $|S|$ là kích thước tập dữ liệu đào tạo.

Chỉ số thông tin cần thiết cho sự phân lớp: I(S) với S là tập cần xét sự phân phối lớp được tính bằng:

$$\boldsymbol{I(S) = -\sum_{j=1}^{x} RF(C_j, S)\,\log(RF(C_j, S))}$$

Sau khi S được phân chia thành các tập con S_1, S_2,…, S_t bởi test B thì information gain được tính bằng:

$$\boldsymbol{G(S, B) = I(S) - \sum \frac{|S_i|}{|S|} I(S_i)}$$

Test $\boldsymbol{B}$ sẽ được chọn nếu có $\boldsymbol{G(S, B)}$ đạt giá trị lớn nhất.

Tuy nhiên có một vấn đề khi sử dụng $\boldsymbol{G(S, B)}$ ưu tiên test có số lượng lớn kết quả, ví dụ $\boldsymbol{G(S, B)}$ đạt cực đại với test mà từng $\boldsymbol{S_i}$ chủi chứa một case đơn. Tiêu chuẩn gain ratio giải quyết được vấn đề này bằng việc đưa vào thông tin tiềm năng của bản than mỗi phân hoạch.

$$\boldsymbol{P(S, B) = -\sum \frac{|S_i|}{|S|} \log(\frac{|S_i|}{|S|})}$$

Test $\boldsymbol{B}$ sẽ được chọn nếu có tỉ số giá trị gain ratio $= {\boldsymbol{G(S, B)}}/{\boldsymbol{P(S, B)}}$ lớn nhất.

Trong mô hình phân lớp C4.5, có thể dùng một trong hai loại chỉ số Information Gain hay Gain ratio để xác định thuộc tính tốt nhất. Trong đó Gain ratio là lựa chọn mặc định.

Ví dụ minh hoạ:

Bảng 1: Tập dữ liệu khách hàng mua máy tính

rid	age	income	student	ereclit_rating	Class: buys_computer
1	<30	high	no	fair	no
2	<30	high	no	excellent	no
3	30-90	high	no	fair	yes
4	>90	medium	no	fair	yes
5	>90	low	yes	fair	yes
6	>40	low	yes	excellent	no
7	30-40	low	yes	excellent	yes
8	<30	medium	no	fair	no
9	<30	low	yes	fair	yes
10	>90	medium	yes	fair	yes
11	<30	medium	yes	excellent	yes
12	30-90	medium	no	excellent	yes
13	30-40	high	yes	fair	yes
14	>90	medium	no	excellent	no

Trong tập dữ liệu Bảng 2: S1 là tập những bản ghi có giá trị phân lớp là yes, S2 là tập những bản ghi có giá trị phân lớp là no. Khi đó:

$$I(S) = I(S1, S2) = I(9,5) = -\frac{9}{14} * \log_2 \frac{9}{14} - \frac{5}{14} * \log_2 \frac{5}{14} = 0.940$$

Tính G(S,A) với A lần lượt là từng thuộc tính: A = age. Thuộc tính age đã được rời rạc hóa thành các giá trị <30, 30-40, và >40.

+ với age = "<30":

$$I(S1) = (S11, S21) = -\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5} = 0.971$$

+ với age = "30-40":

$$I(S1) = (S12, S22) = 0$$

+ với age = ">40":

$$I(S3) = (S13, S23) = 0.971$$

$$\sum \frac{|Si|}{|S|} * I(Si) = \frac{5}{14} * I(S1) + \frac{4}{14} * I(S2) + \frac{5}{14} * I(S3)$$

$$Gain(S, age) = I(S1.S2) - 0.694 = 0.246$$

Tính tương tự với các thuộc tính khác ta được:

– A = income: Gain (S, income) = 0.029

– A = student: Gain (S, student) = 0.151

– A = credit_rating: Gain (S, credit_rating) = 0.048

Thuộc tính age là thuộc tính có độ đo Information Gain lớn nhất. Do vậy age được chọn làm thuộc tính phát triển tại node đang xét.

2.4. Khảo sát thuật toán SVM

Support Vector Machine (SVM) là một thuật toán thuộc nhóm Supervised Learning (Học có giám sát) dùng để phân chia dữ liệu (Classification) thành các nhóm riêng biệt.

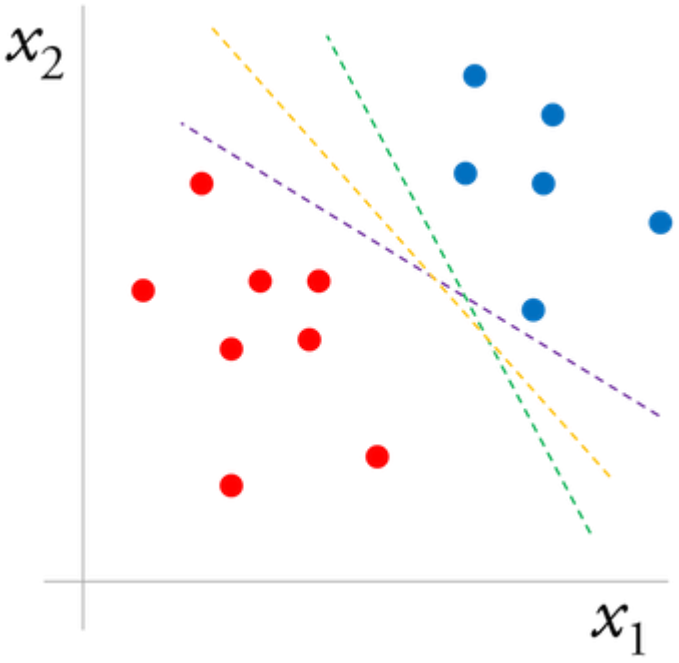
SVM là một bộ phương pháp học có giám sát liên quan được sử dụng trong chẩn đoán y khoa để phân loại và hồi quy. SVM đồng thời giảm thiểu lỗi phân loại thực nghiệm và tối đa hóa biên độ hình học. Vì vậy, SVM được gọi là Maximum Margin Classifiers.

SVM là một thuật toán chung dựa trên giới hạn xác suất được kế thừa của lý thuyết học thống kê gọi là nguyên tắc giảm thiểu rủi ro cấu trúc. SVM có thể thực hiện hiệu quả phân loại phi tuyến tính bằng cách sử dụng thủ thuật kernel, ánh xạ ngầm định các đầu vào của chúng vào các không gian đặc trưng chiều cao. Mô hình SVM là một đại diện của các ví dụ dưới dạng các điểm trong không gian, được ánh xạ sao cho các loại riêng biệt được chia cho một khoảng cách rõ ràng càng rộng càng tốt[9].

Ví dụ minh hoạ:

Hình dung ta có bộ data gồm các điểm xanh và đỏ đặt trên cùng một mặt phẳng.

Ta có thể tìm được đường thẳng để phân chia riêng biệt các bộ điểm xanh và đỏ như hình 3



Hình 3: Biểu đồ phân lớp dữ liệu

Nhìn bằng mắt thường ta có thể thấy, đường tối ưu là đường tạo cho ta có cảm giác 2 lớp dữ liệu nằm cách xa nhau và cách xa đường đó nhất.

Tuy nhiên tính toán sự tối ưu bằng toán học, trong SVM sử dụng thuật ngữ Margin.

Margin là khoảng cách giữa siêu phẳng (trong trường hợp không gian 2 chiều là đường thẳng) đến 2 điểm dữ liệu gần nhất tương ứng với 2 phân lớp.

Trong bài toán không gian 2 chiều, giả sử 2 đường thẳng đi qua các support vector của 2 lớp dữ liệu lần lượt là:

$$w_1 x_1 + w_2 x_2 + b = 1$$

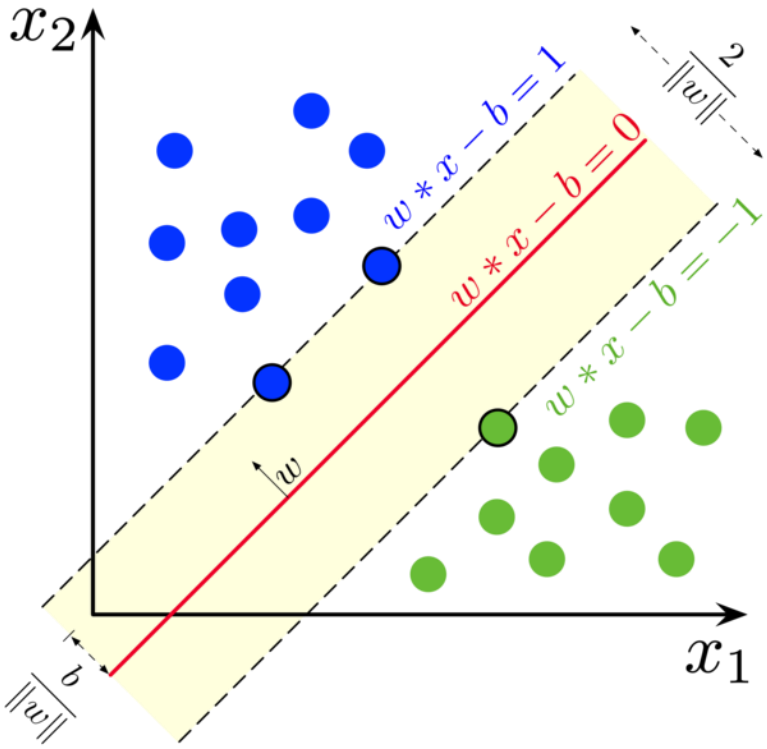
$$w_1 x_1 + w_2 x_2 + b = -1$$

Với không gian 2 chiều: Margin giữa 2 đường thẳng được tính bằng công thức:

$$margin = \frac{2}{\sqrt{w_1^2 + w_2^2}}$$

Với không gian nhiều chiều: Tổng quát lên không gian nhiều chiều, cần tìm phương trình siêu phẳng có phương trình: $w^T x + b = 0$

Margin sẽ được tính bằng công thức: $margin = \frac{2}{||w||}$



Hình 4: Siêu phẳng tối đa cho SVM được huấn luyện với các mẫu từ hai lớp

2.5. Khảo sát thuật toán Naïve Bayes

Naive Bayes Classification (NBC) là một thuật toán phân loại dựa trên tính toán xác suất áp dụng định lý Bayes

Thuật toán này thuộc nhóm Supervised Learning (Học có giám sát).

Theo định lý Bayes, ta có công thức tính xác suất ngẫu nhiên của sự kiện y khi biết x:

$$P(y|x) = \frac{P(x|y)P(y)}{P(x)}$$

Giả sử ta phân chia 1 sự kiện x thành n thành phần khác nhau $x_1, x_2, \dots, x_n$. Naive Bayes theo đúng như tên gọi dựa vào một giả thiết rằng $x_1, x_2, \dots, x_n$ là các thành phần độc lập với nhau. Từ đó ta có thể tính được:

$$P(x|y) = P(x_1 \cap x_2 \dots \cap x_n)|y = P(x_1|y)P(x_2|y) \dots P(x_n|y)$$

Do đó ta có:

$$P(x|y) \propto P(y) \prod_{i=1}^{n} P(x_i|y)$$

$| \propto$ là phép tỉ lệ thuận.

Trên thực tế thì ít khi tìm được dữ liệu mà các thành phần là hoàn toàn độc lập với nhau. Tuy nhiên giả thiết này giúp cách tính toán trở nên đơn giản, training data nhanh, đem lại hiệu quả bất ngờ với các lớp bài toán nhất định.

Cách xác định các thành phần (class) của dữ liệu dựa trên giả thiết này có tên là Naive Bayes Classifier.

Ví dụ minh hoạ:

Giả sử trong tập training có các văn bản d1, d2, d3, d4 như trong Bảng 3. Mỗi văn bản này thuộc vào 1 trong 2 classes: B(Bắc) hoặc N(Nam). Hãy xác định class của văn bản d5.

Bảng 2: Dữ liệu có dạng văn bản trong tập huấn luyện

	Document	**Content**	**Class**
Training	d1	hanoi pho chaolong hanoihanoi pho chaolong hanoi	B
	d2	hanoi buncha pho omaihanoi buncha pho omai	B
	d3	pho banhgio omaipho banhgio omai	B
	d4	saigon hutiu banhbo phosaigon hutiu banhbo pho	N
Test	d5	hanoi hanoi buncha hutiuhanoi hanoi buncha hutiu	?

Ta có thể dự đoán rằng d5 thuộc class *Bắc*.

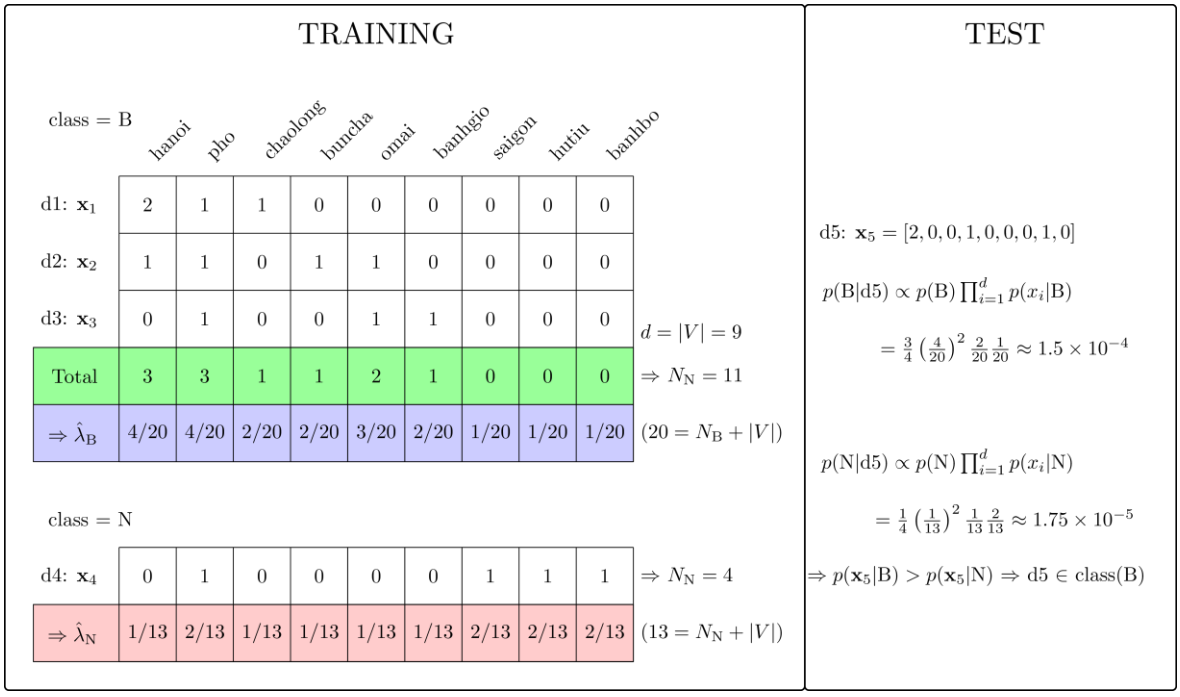
Bài toán này có thể được giải quyết bởi hai mô hình: Multinomial Naive Bayes và Bernoulli Naive Bayes. Việc mô hình nào tốt hơn phụ thuộc vào mỗi bài toán. Chúng ta có thể thử cả hai để chọn ra mô hình tốt hơn.

Nhận thấy rằng ở đây có 2 class B và N, ta cần đi tìm $p(B)$ và $p(N)$ là dựa trên tần số xuất hiện của mỗi class trong tập training. Ta sẽ có:

$$p(B) = \frac{3}{4}, \qquad p(N) = \frac{1}{4}$$

Tập hợp toàn bộ các từ trong văn bản, hay còn gọi là từ điển, là: V={hanoi, pho, chaolong, buncha, omai, banhgio, saigon, hutiu, banhbo}. Tổng cộng số phần tử trong từ điển là |V|=9.

Hình 5 minh hoạ quá trình Training và Test cho bài toán này khi sử dụng Multinomial Naive Bayes, trong đó có sử dụng Laplace smoothing với α=1.



Hình 5: Các bước trainning và test sử dụng dụng Multinomial Naive Bayes

Chú ý, hai giá trị tìm được 1.5×10^{-4} và 1.75×10^{-5} không phải là hai xác suất cần tìm mà chỉ là hai đại lượng tỉ lệ thuận với hai xác suất đó. Để tính cụ thể, ta có thể làm như sau:

$$p(B|d5) = \frac{1.5 \times 10^{-4}}{1.5 \times 10^{-4} + 1.75 \times 10^{-5}} \approx -0.8955, p(N|d5) = 1 - p(B|d5) \approx 0.1045$$

Bạn đọc có thể tự tính với ví dụ khác d6 = pho hutiu banhbo. Nếu bạn và tôi tính ra kết quả giống nhau, chúng ta sẽ thu được:

$$p(B|d6) \approx 0.29, \qquad p(N|d6) \approx 0.71$$

và suy ra d6 thuộc vào class *Nam*

Thuật toán Naïve Bayes là một thuật toán xác suất có tính chất tương đương, theo các bước thực hiện, phân loại, ước tính và dự đoán. Để tìm mối quan hệ giữa các bệnh, triệu chứng và thuốc, có nhiều giải pháp khai thác dữ liệu khác nhau, nhưng các thuật toán này có mô phỏng riêng, nhiều lần lặp lại, lập luận về các đối số liên tục, thời gian tính toán cao, v.v. ước lượng lặp phức tạp của tham số và có thể được áp dụng trên một tập dữ liệu lớn trong thời gian thực.

Kết luận chương 2

Chương 2 nghiên cứu một số thuật toán học máy, các thuật toán hỗ trợ bài toán đưa ra tỷ lệ dự toán trong bài toán chẩn đoán bệnh đái tháo đường. Từ đó sẽ áp dụng và đánh giá kết quả của từng thuật toán trong Chương 3.

CHƯƠNG 3: CÀI ĐẶT VÀ THỬ NGHIỆM

3.1. Khảo sát và lựa chọn bộ dữ liệu để thử nghiệm

Có rất nhiều bộ dữ liệu bệnh án về Đái tháo đường trên Thế giới.

Bộ dataset mà luận văn sử dụng là Cơ sở dữ liệu về bệnh đái tháo đường của người Ấn Độ thuộc Viện Tiểu đường và Bệnh tiêu hóa và Thận Hoa Kỳ[22](Bảng 3) vì bộ dữ liệu này đã được dùng trong các nghiên cứu về các bệnh Đái tháo đường[11][12].

Bộ dữ liệu gồm: 8 thuộc tính và 2 class (0 tương ứng với âm tính, 1 tương ứng với dương tính)[22] (Bảng 1).

Bảng 3: Bảng thuộc tính và gán nhãn giá trị

Thuộc tính	Gán nhãn giá trị
1. Số lần mang thai	preg
2. Nồng độ glucose trong máu	plas
3. Huyết áp (mm Hg)	pres
4. Độ dày nếp gấp da (mm)	skin
5. Insulin huyết thanh 2 giờ	insu
6. Chỉ số khối cơ thể (kg/m^2)	mass
7. Chức năng phả hệ tiểu đường	pedi
8. Tuổi (năm)	age
Biến lớp (0 hoặc 1) 268 trong 768 là 1, các biến khác là 0	class

3.2. Tiền xử lý dữ liệu

Học viên chọn bộ dữ liệu Pima Indians Diabetes vì nó là bộ dữ liệu thu thập các số liệu về các chỉ số y khoa của những người mắc và không mắc bệnh đái tháo đường đường trong vòng 5 năm tại Pima Indian.

Đây là một bài toán phân lớp nhị phân. Số lượng dữ liệu là 768 mẫu với 8 đặc trưng về các chỉ số y khoa và 1 thuộc tính nhãn lớp. Số lượng các quan sát cho các lớp là không đồng đều.

Theo kết quả quan sát được, bộ dữ liệu có 6 đặc trưng đầu tiên có giá trị nhỏ nhất là 0, điều này đồng nghĩa với việc 6 đặc trưng này có thể đã bị khuyết dữ liệu ở một số mẫu dữ liệu. Tuy nhiên, đặc trưng NoPregnant là đặc trưng về số lần mang thai, một người có thể đã mang thai hoặc chưa từng mang thai. Do đó giá trị 0 của đặc trưng này biểu thị cho những người chưa từng mang thai chứ không phải là bị khuyết dữ liệu. Các đặc trưng còn lại chứa giá trị 0 đang bị khuyết dữ liệu.

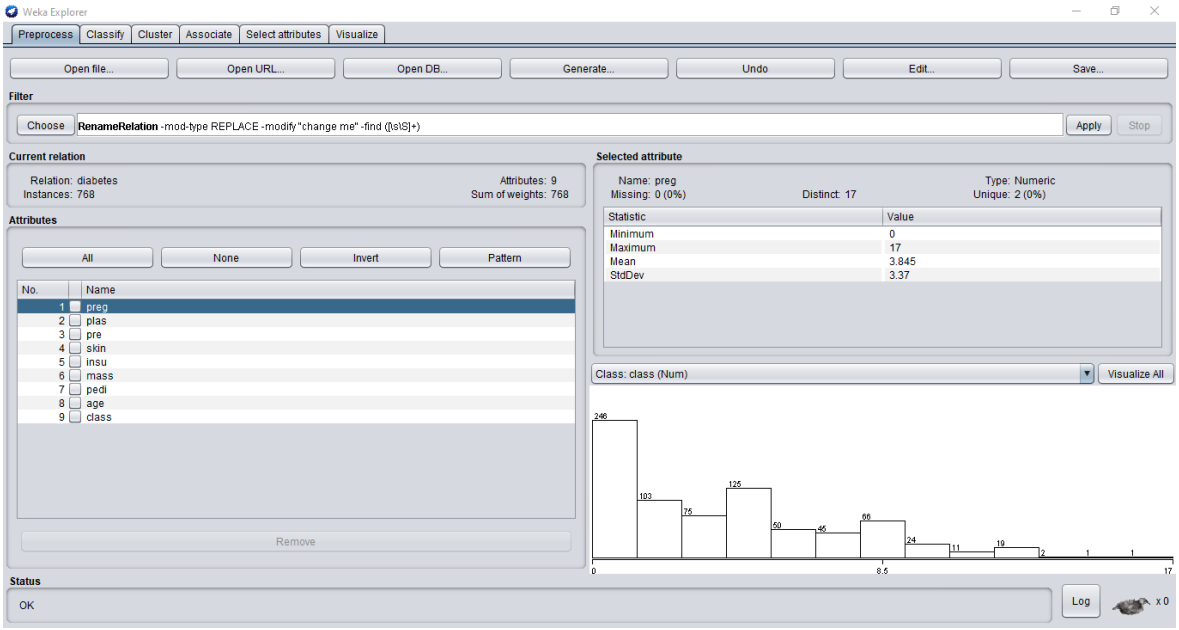
Bảng 4: Bảng thống kê số lượng mẫu bị khuyết của các đặc trưng

Đặc trưng	Số mẫu
Plas	5
Pres	35
Skin	227
Insu	374
Mass	11

Kết quả cho thấy 2 đặc trưng Skin và Insu có nhiều giá trị bị khuyết nhất.

Học viên sử dụng công cụ Weka[6] để khai phá dữ liệu, các chức năng được sử dụng trong công cụ như: Tiền xử lý dữ liệu, các giải thuật học máy và các phương pháp thí nghiệm đánh giá.

Giao diện Weka:



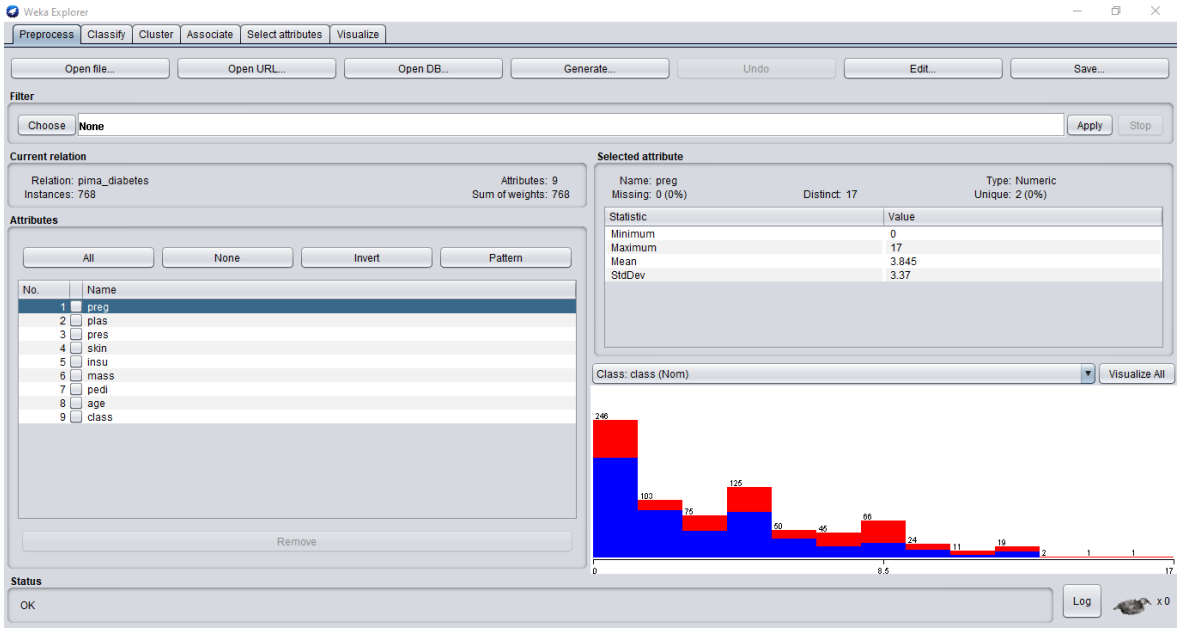
Hình 6: Giao diện công cụ Weka

Các bước xử lý bao gồm:

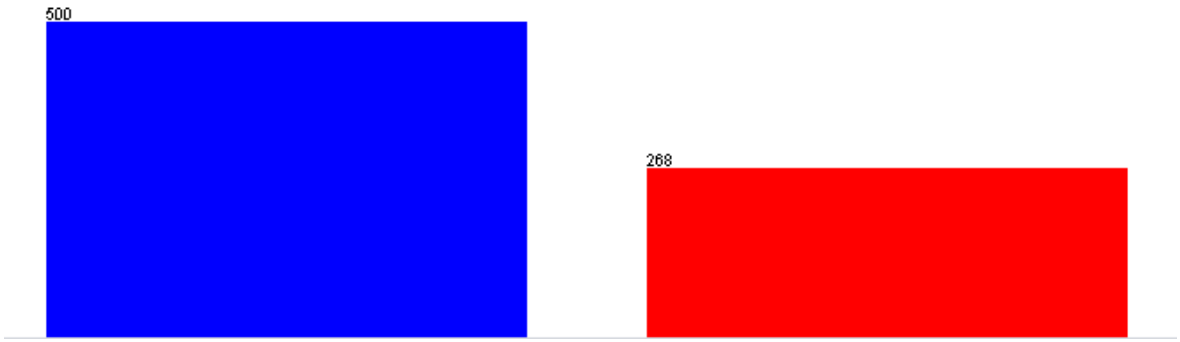
- Chuẩn hóa các thuộc tính số về đoạn [0, 1] bằng bộ lọc Normalize.

- Sau đó, dùng bộ lọc ReplaceMissingValue để thay thế tất cả các giá trị thiếu bằng giá trị trung bình của thuộc tính.

- Chuẩn hoá các giá trị bằng thuộc tính: Normalization.



Hình 7: Dữ liệu sau khi tinh chỉnh



Hình 8: Lớp thuộc tính phân lớp (class)

3.3. Thử nghiệm và đánh giá kết quả

Câu hỏi: Có dương tính với bệnh Đái tháo đường đường không?

Quyết định đưa ra dựa trên các yếu tố về các chỉ số của bệnh án: Pregnancies (Số lần mang thai), Glucose (nồng độ glucose trong 2 giờ sau khi xét nghiệm máu nạp glucose), BloodPressure (Huyết áp), SkinThickness (độ căng da), Insulin (Xét

nghiệm máu Insulin 2 giờ), BMI (Chỉ số khối cơ thể), DiabetesPedigreeFunction (chức năng tiểu đường phả hệ), Age.

Có rất nhiều thuật toán phân lớp như ID3, J48, C4.5, CART (Classification and Regression Tree), … Việc chọn thuật toán nào để có hiệu quả phân lớp cao tuy thuộc vào rất nhiều yếu tố, trong đó cấu trúc dữ liệu ảnh hưởng rất lớn đến kết quả của các thuật toán.

Với thuật toán ID3 và CART cho hiệu quả phân lớp rất cao đối với các trường dữ liệu số (quantitative value) trong khi đó các thuật toán như J48, C4.5 có hiệu quả hơn đối với các dữ liệu có giá trị định tính (ordinal, Binary, nominal).

Sau khi đã chuẩn hóa dữ liệu thì được bảng dữ liệu chỉ toàn kiểu Nominal, vì vậy ta nên sử dụng thuật toán J48 để đạt hiệu quả phân lớp cao.

Từ 768 mẫu trong bộ dữ liệu, chia thành 2 phần: 90% được sử dụng làm bộ training, 10% còn lại được làm bộ đánh giá (test). Mỗi lần chạy sẽ chọn 1 bộ dữ liệu train và test khác nhau.

3.3.1. Đánh giá thuật toán C4.5.

3.3.1.1. Phân loại đầu ra dựa trên tập huấn luyện toàn bộ

Trong phần mềm weka thì thuật toán C4.5 có ký hiệu là J48[21].

Ở **Bảng 5** là kết quả khi chạy với chế độ huấn luyện toàn bộ trên bộ dữ liệu training đã chia sau khi tiền xử lý dữ liệu.

Bảng 5: Kết quả sau khi chạy kiểm thử phân lớp n lần với thuật toán J48

K = 10 (n lần)	**Trường hợp phân lớp chính xác (Số trường hợp)**	**Trường hợp phân lớp không chính xác (Số trường hợp)**
1	**90.72 % (626)**	**9.28 % (64)**
2	85.79 % (592)	14.21 % (98)

3	82.1 % (566)	17.9 % (124)
4	84.78 % (585)	15.22 % (105)
5	83.63 % (577)	16.37 % (113)
6	84.21 % (581)	15.79 % (109)
7	80.53 % (556)	18.47 % (134)
8	80.57 % (556)	19.43 % (134)
9	84.63 % (584)	15.37 % (106)
10	80.87 % (558)	19.13 % (132)

Từ **Bảng 5** ta có thể thấy được với lần chạy đầu tiên thì tỷ lệ dự đoán chính xác là tốt nhất với 690 trường hợp.

Trong đó tỷ lệ dự đoán chính xác Dương tính với bệnh là 187 mẫu, âm tính là 439 mẫu. Có tỷ lệ chính xác đạt 90,72% đối với bộ dữ liệu.

Tỷ lệ dự đoán không chính xác là 64 mẫu với tỷ lệ 9,28%.

Kết quả có hiệu suất tốt nhất trong các lần chạy với chế độ huấn luyện toàn bộ từ tập dữ liệu:

Bảng 6: Kết quả thuật toán phân lớp J48

	Số trường hợp	**tỷ lệ**
Trường hợp phân lớp chính xác	626	90.72 %
Trường hợp phân lớp không chính xác	64	9.28 %

Các kết quả khác của thuật toán phân lớp J48:

Bảng 7: Kết quả khác của thuật toán phân lớp J48

Kappa statistic	0.7865
Mean absolute error	0.1541
Root mean squared error	0.2776
Relative absolute error	34.22 %
Root relative squared error	58.51 %
Total Number of Instances	690

Giải thích thuật ngữ:

+ Kappa statistic : Là một số liệu so sánh Độ chính xác được thấy với Độ chính xác dự kiến (cơ hội ngẫu nhiên).

+ Mean Absolute Error: Trung bình của lỗi tuyệt đối giữa giá trị được quan sát và dự báo.

+ Root Mean Squared Error: Đo lường sự khác biệt giữa giá trị (mẫu thử và giá trị tương đối) được dự đoán bởi một mô hình hoặc công cụ ước tính và các giá trị thực sự được quan sát.

+ Relative Absolute Error: Tỷ lệ sai số tuyệt đối của phép đo với phép đo được chấp nhận.

+ Total Number of Instances: Tổng số trường hợp.

Ma trận hỗn loạn:

Bảng 8: Ma trận hỗn loại thuật toán phân lớp J48

	A – Dương tính	**B - Âm tính**
A - Dương tính	187 (1)	49 (2)
B – Âm tính	15 (3)	439 (4)

Trong bảng, các giá trị thể hiện như sau:

(1) Số dự báo đúng mà trường hợp đã kiểm tra dương tính

(2) Số dự báo không chính xác mà phiên bản thử nghiệm âm tính

(3) Số dự báo không chính xác mà trường hợp đã kiểm tra dương tính

(4) Số dự báo đúng mà trường hợp đã kiểm tra âm tính

3.3.1.2. Phân loại đầu ra dựa trên tập tin huấn luyện (90:10)

Ở **Bảng 9** là kết quả khi chạy huấn luyện trên tập test 10% trên bộ dữ liệu training đã chia sau khi tiền xử lý dữ liệu.

Bảng 9: Kết quả sau khi chạy kiểm thử phân lớp n lần với thuật toán J48 (90:10)

K = 10 (n lần)	**Trường hợp phân lớp chính xác (Số trường hợp)**	**Trường hợp phân lớp không chính xác (Số trường hợp)**
1	71.43 % (55)	28.57 % (22)
2	75.64 % (59)	24.36 % (19)
3	69.23 % (54)	30.77 % (24)
4	56.41 % (44)	43.59 % (34)
5	80.77 % (63)	19.23 % (15)
6	**91.03 % (71)**	**8.97 % (7)**
7	74.74 % (53)	25.26 % (26)
8	84.61 % (66)	15.39 % (12)
9	71.79 % (56)	28.21 % (22)
10	76.92 % (60)	23.08 % (18)

Từ **Bảng 9** ta có thể thấy được với lần chạy thứ 6 thì tỷ lệ dự đoán chính xác là tốt nhất với 78 trường hợp.

Trong đó tỷ lệ dự đoán chính xác Dương tính với bệnh là 25 mẫu, âm tính là 46 mẫu. Có tỷ lệ chính xác đạt 91,03% đối với bộ dữ liệu.

Tỷ lệ dự đoán không chính xác là 7 mẫu với tỷ lệ 8,97%

Kết quả có hiệu suất tốt nhất trong các lần chạy với chế độ huấn luyện 10% từ tập dữ liệu:

Bảng 10: Kết quả thuật toán phân lớp J48 (90:10)

	Số trường hợp	**tỷ lệ**
Trường hợp phân lớp chính xác	71	91.03 %
Trường hợp phân lớp không chính xác	7	8.97 %

Các kết quả khác của thuật toán phân lớp J48:

Bảng 11: Kết quả khác của thuật toán phân lớp J48 (90:10)

Kappa statistic	0.8068
Mean absolute error	0.1837
Root mean squared error	0.2785
Total Number of Instances	78

Ma trận hỗn loạn:

Bảng 12: Ma trận hỗn loại thuật toán phân lớp J48 (90:10)

	A – Dương tính	**B - Âm tính**
A - Dương tính	25(1)	2 (2)
B – Âm tính	5(3)	46(4)

3.3.2. Đánh giá thuật toán SVM

3.3.2.1. Phân loại đầu ra dựa trên tập huấn luyện toàn bộ

Thuật toán SVM(SMO[10]) cho kết quả sau đây từ tập dữ liệu đã cho:

Ở **Bảng 13** là kết quả khi chạy với chế độ huấn luyện toàn bộ trên bộ dữ liệu training đã chia sau khi tiền xử lý dữ liệu.

Bảng 13: Kết quả sau khi chạy kiểm thử phân lớp n lần với thuật toán SMO

K = 10 (n lần)	**Trường hợp phân lớp chính xác (Số trường hợp)**	**Trường hợp phân lớp không chính xác (Số trường hợp)**
1	**79.27 % (547)**	**20.73 % (143)**
2	77.25 % (533)	22.75 % (157)
3	77.68 % (536)	22.32 % (154)
4	77.87 % (538)	22.13 % (152)
5	77.39 % (534)	22.61 % (156)
6	76.82 % (530)	23.18 % (160)
7	76.95 % (531)	23.05 % (159)
8	76.95 % (531)	23.05 % (159)
9	77.83 % (537)	22.17 % (153)
10	77.11 % (532)	22.89 % (158)

Từ **Bảng 13** ta có thể thấy được với lần chạy đầu tiên thì tỷ lệ dự đoán chính xác là tốt nhất với 690 trường hợp.

Trong đó tỷ lệ dự đoán chính xác Dương tính với bệnh là 128 mẫu, âm tính là 419 mẫu. Có tỷ lệ chính xác đạt 79,28% đối với bộ dữ liệu.

Tỷ lệ dự đoán không chính xác là 143 mẫu với tỷ lệ 20,72%

Kết quả có hiệu suất tốt nhất trong các lần chạy với chế độ huấn luyện toàn bộ từ tập dữ liệu:

Bảng 14: Kết quả thuật toán phân lớp SMO

	Số trường hợp	**tỷ lệ %**
Trường hợp phân lớp chính xác	547	79.27 %
Trường hợp phân lớp không chính xác	143	20.72 %

Các kết quả khác của thuật toán phân lớp SMO:

Bảng 15: Kết quả khác của thuật toán phân lớp SMO

Kappa statistic	0.5026
Mean absolute error	0.2072
Root mean squared error	0.4552
Relative absolute error	46.03 %
Root relative squared error	95.96 %
Total Number of Instances	690

Ma trận hỗn loạn:

Bảng 16: Ma trận hỗn loại thuật toán phân lớp SMO

	A – Dương tính	**B - Âm tính**
A - Dương tính	128	108
B – Âm tính	35	419

3.3.2.2. Phân loại đầu ra dựa trên tập tin huấn luyện (90:10)

Ở **Bảng 17** là kết quả khi chạy huấn luyện trên tập test 10% trên bộ dữ liệu training đã chia sau khi tiền xử lý dữ liệu.

Bảng 17: Kết quả sau khi chạy kiểm thử phân lớp n lần với thuật toán SMO (90:10)

K = 10 (n lần)	**Trường hợp phân lớp chính xác (Số trường hợp)**	**Trường hợp phân lớp không chính xác (Số trường hợp)**
1	67.53 % (52)	32.47 % (25)
2	**83.33 % (65)**	**16.67 % (13)**
3	75.64 % (59)	24.36 % (19)
4	70.51 % (55)	29.49 % (23)
5	78.2 % (61)	21.8 % (17)
6	79.49 % (62)	20.51 % (16)
7	**83.33 % (65)**	**16.67 % (13)**
8	76.22 % (54)	24.78 % (20)
9	71.79 % (56)	28.21 % (22)
10	79.49 %(62)	20.51 % (16)

Từ **Bảng 17** ta có thể thấy được với lần chạy thứ 2 và lần chạy thứ 7 thì tỷ lệ dự đoán chính xác là tốt nhất với 78 trường hợp.

Với lần chạy thứ 2: Trong đó tỷ lệ dự đoán chính xác Dương tính với bệnh là 13 mẫu, âm tính là 52 mẫu. Có tỷ lệ chính xác đạt 83,33% đối với bộ dữ liệu.

Tỷ lệ dự đoán không chính xác là 13 mẫu với tỷ lệ 16,67%

Với lần chạy thứ 7: Trong đó tỷ lệ dự đoán chính xác Dương tính với bệnh là 8 mẫu, âm tính là 57 mẫu. Có tỷ lệ chính xác đạt 83,33% đối với bộ dữ liệu.

Tỷ lệ dự đoán không chính xác là 13 mẫu với tỷ lệ 16,67%

Kết quả có hiệu suất tốt nhất trong các lần chạy trên tập dữ liệu:

Bảng 18: Kết quả thuật toán phân lớp SMO (90:10)

	Số trường hợp	**tỷ lệ**
Trường hợp phân lớp chính xác	65	83.33 %
Trường hợp phân lớp không chính xác	13	16.67 %

Các kết quả khác của thuật toán phân lớp SMO:

Bảng 19: Kết quả khác của thuật toán phân lớp SMO (90:10)

Kappa statistic	0.5603
Mean absolute error	0.1667
Root mean squared error	0.4082
Total Number of Instances	78

Ma trận hỗn loạn:

Bảng 20: Ma trận hỗn loại thuật toán phân lớp SMO (90:10)

	A – Dương tính	**B - Âm tính**
A - Dương tính	13(1)	3 (2)
B – Âm tính	10(3)	52(4)

3.3.3. Đánh giá thuật toán Naïve Bayes

3.3.3.1. Phân loại đầu ra dựa trên tập huấn luyện toàn bộ

Thuật toán Naïve Bayes cho kết quả sau đây từ tập dữ liệu đã cho:

Ở **Bảng 21** là kết quả khi chạy với chế độ huấn luyện toàn bộ trên bộ dữ liệu training đã chia sau khi tiền xử lý dữ liệu.

Bảng 21: Kết quả sau khi chạy kiểm thử phân lớp n lần với thuật toán Naïve Bayes

K = 10 (n lần)	**Trường hợp phân lớp chính xác (Số trường hợp)**	**Trường hợp phân lớp không chính xác (Số trường hợp)**
1	77.54 %(535)	**22.46 %(155)**
2	75.94 %(524)	24.06 %(166)
3	76.52 %(528)	23.48 %(162)
4	**76.96 %(531)**	23.04 %(159)
5	76.66 %(529)	23.34 %(161)
6	75.07 %(518)	24.93 %(172)

7	76.48 %(515)	23.32 %(162)
8	76.08 %(525)	23.92 %(165)
9	76.38 %(527)	23.62 %(163)
10	76.24 %(526)	23.76 %(164)

Từ **Bảng 21** ta có thể thấy được với lần chạy thứ 1 cho tỷ lệ không chính xác thấp nhất và lần chạy thứ 4 thì tỷ lệ dự đoán chính xác là tốt nhất với 690 trường hợp.

Với lần chạy đầu tiên: Trong đó tỷ lệ dự đoán chính xác Dương tính với bệnh là 147 mẫu, âm tính là 388 mẫu. Có tỷ lệ chính xác đạt 77,54% đối với bộ dữ liệu.

Tỷ lệ dự đoán không chính xác là 155 mẫu với tỷ lệ 22,46%

Với lần chạy thứ 4: Trong đó tỷ lệ dự đoán chính xác Dương tính với bệnh là 144 mẫu, âm tính là 387 mẫu. Có tỷ lệ chính xác đạt 76,96% đối với bộ dữ liệu.

Tỷ lệ dự đoán không chính xác là 159 mẫu với tỷ lệ 23,04%

Bảng 22: Kết quả thuật toán phân lớp Naïve Bayes

	Số trường hợp	**tỷ lệ %**
Trường hợp phân lớp chính xác	531	76.96 %
Trường hợp phân lớp không chính xác	159	23.04 %

Các kết quả khác của thuật toán phân lớp Naïve Bayes:

Bảng 23: Kết quả khác của thuật toán phân lớp Naïve Bayes

Kappa statistic	0.4749
Mean absolute error	0.276
Root mean squared error	0.4144
Relative absolute error	61.04 %
Root relative squared error	87.17 %
Total Number of Instances	690

Ma trận hỗn loạn:

Bảng 24: Ma trận hỗn loại thuật toán phân lớp Naïve Bayes

	A – Dương tính	**B - Âm tính**
A - Dương tính	144	94
B – Âm tính	65	387

3.3.3.2. Phân loại đầu ra dựa trên tập huấn luyện (90:10)

Ở **Bảng 25** là kết quả khi chạy huấn luyện trên tập test 10% trên bộ dữ liệu training đã chia sau khi tiền xử lý dữ liệu.

Bảng 25: Kết quả sau khi chạy kiểm thử phân lớp n lần với thuật toán Naïve Bayes (90:10)

K = 10 (n lần)	**Trường hợp phân lớp chính xác (Số trường hợp)**	**Trường hợp phân lớp không chính xác (Số trường hợp)**
1	67.53 %(52)	32.47 %(25)
2	80.77 %(63)	19.23 %(15)
3	75.64 %(59)	24.36 %(19)
4	71.79 %(56)	28.21 %(22)
5	73.08 %(57)	26.92 %(21)
6	76.92 %(60)	23.08 %(18)
7	80.77 %(63)	19.23 %(15)
8	**82.05 %(64)**	**17.95 %(14)**
9	74.36 %(58)	25.64 %(20)
10	75.64 %(59)	24.36 %(19)

Từ **Bảng 25** ta có thể thấy được với lần chạy thứ 8 thì tỷ lệ dự đoán chính xác là tốt nhất với 78 trường hợp.

Với lần chạy thứ 8: Trong đó tỷ lệ dự đoán chính xác Dương tính với bệnh là 45 mẫu, âm tính là 19 mẫu. Có tỷ lệ chính xác đạt 82,05% đối với bộ dữ liệu.

Tỷ lệ dự đoán không chính xác là 14 mẫu với tỷ lệ 17,95%.

Kết quả có hiệu suất tốt nhất trong các lần chạy trên tập dữ liệu:

Bảng 26: Kết quả thuật toán phân lớp Naïve Bayes (90:10)

	Số trường hợp	**tỷ lệ**
Trường hợp phân lớp chính xác	64	82.05 %
Trường hợp phân lớp không chính xác	14	17.95 %

Các kết quả khác của thuật toán phân lớp Naïve Bayes :

Bảng 27: Kết quả khác của thuật toán phân lớp Naïve Bayes (90:10)

Kappa statistic	0.5965
Mean absolute error	0.229
Root mean squared error	0.3423
Total Number of Instances	78

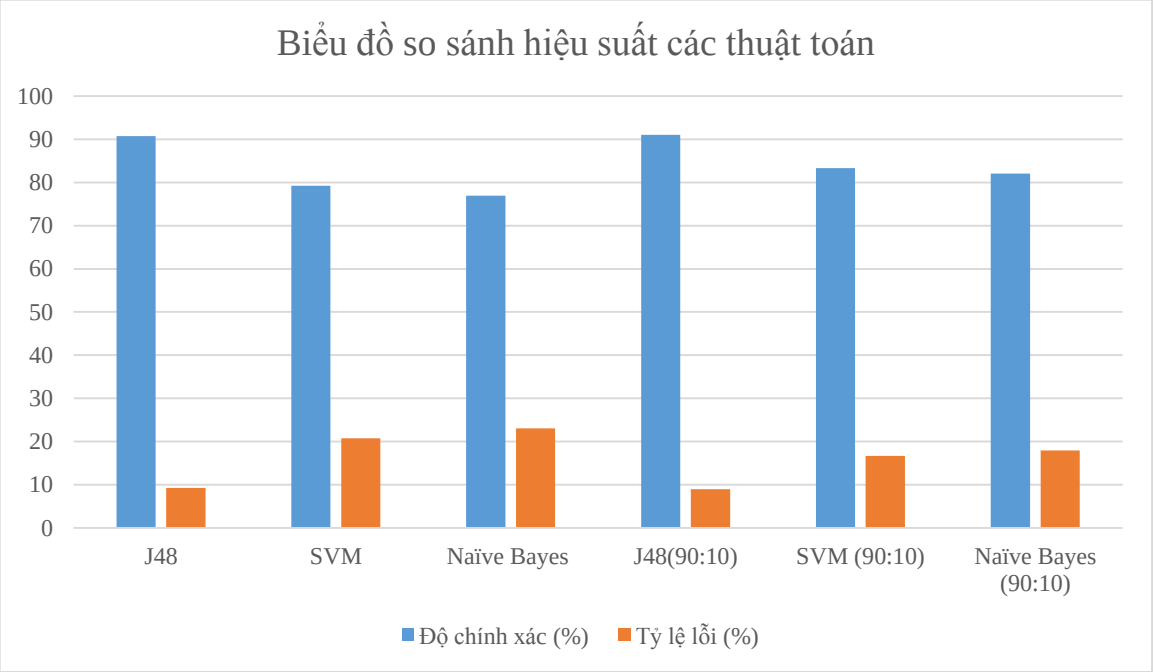
Ma trận hỗn loạn:

Bảng 28: Ma trận hỗn loại thuật toán phân lớp Naïve Bayes (90:10)

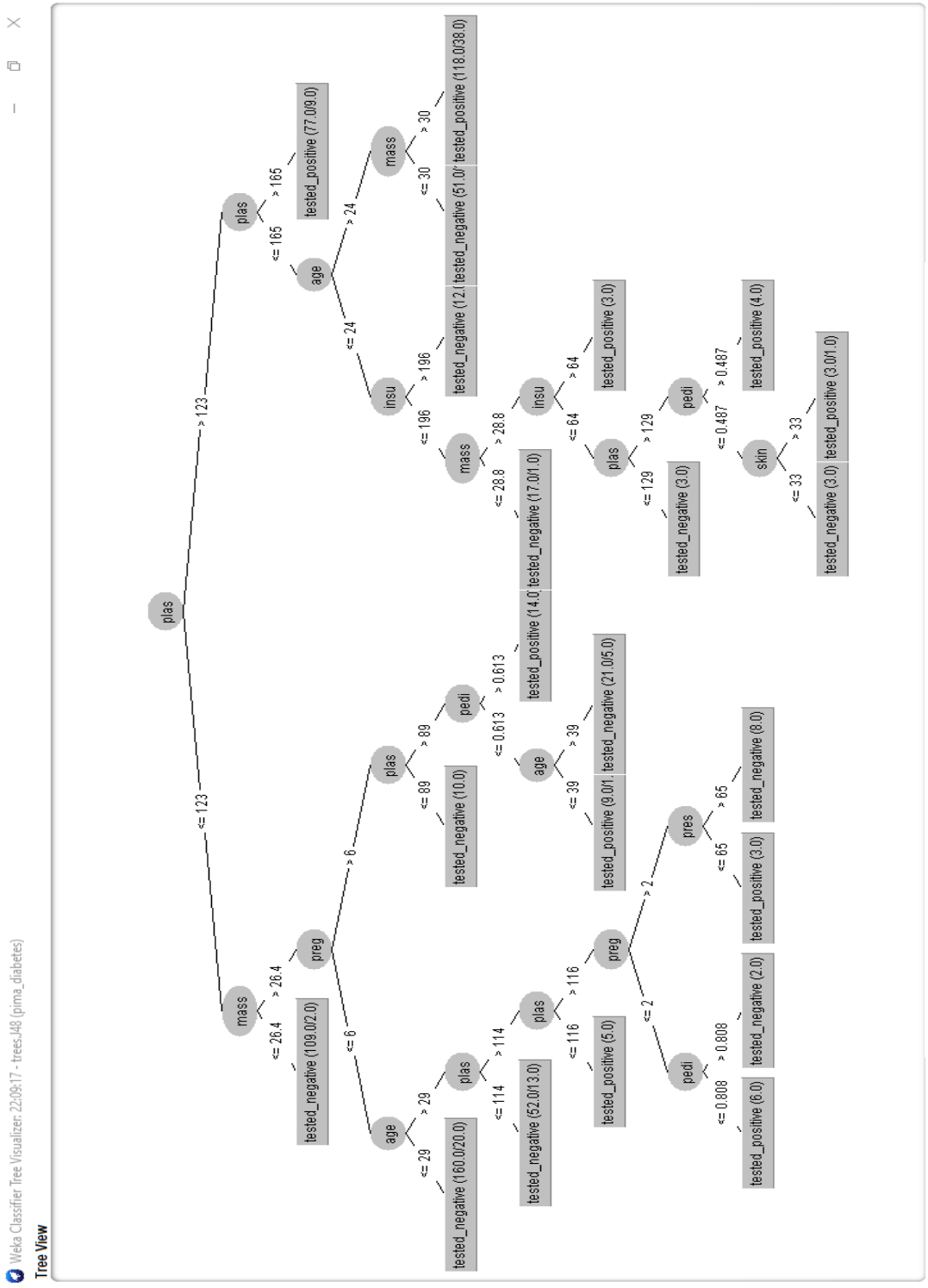
	A – Dương tính	**B - Âm tính**
A - Dương tính	19(1)	8 (2)
B – Âm tính	6(3)	45(4)

3.4. Đánh giá hiệu suất các thuật toán được áp dụng

Từ các kết quả ở mục 3.3 ta thấy được tỷ lệ dự đoán tốt nhất để áp dụng vào cho bài toán hệ hỗ trợ chẩn đoán bệnh Đái tháo đường thì thuật toán J48 cho ra kết quả với hiệu suất tốt nhất với độ chính xác cao nhất và tỷ lệ lỗi thấp nhất.



Xây dựng cây quyết định dựa trên thuật toán J48 từ bộ dữ liệu:



Hình 9: Cây quyết định được sinh ra bằng thuật toán J48.

Các luật sinh ra:

plas <= 123

| mass <= 26.4: tested_negative (109.0/2.0)

| mass > 26.4

| | preg <= 6

| | | age <= 29: tested_negative (160.0/20.0)

| | | age > 29

| | | | plas <= 114: tested_negative (52.0/13.0)

| | | | plas > 114

| | | | | plas <= 116: tested_positive (5.0)

| | | | | plas > 116

| | | | | | preg <= 2

| | | | | | | pedi <= 0.808: tested_positive (6.0)

| | | | | | | pedi > 0.808: tested_negative (2.0)

| | | | | | preg > 2

| | | | | | | pres <= 65: tested_positive (3.0)

| | | | | | | pres > 65: tested_negative (8.0)

| | preg > 6

| | | plas <= 89: tested_negative (10.0)

| | | plas > 89

| | | | pedi <= 0.613

| | | | | age <= 39: tested_positive (9.0/1.0)

| | | | | age > 39: tested_negative (21.0/5.0)

| | | | pedi > 0.613: tested_positive (14.0)

plas > 123

```
|   plas <= 165
|   |   age <= 24
|   |   |   insu <= 196
|   |   |   |   mass <= 28.8: tested_negative (17.0/1.0)
|   |   |   |   mass > 28.8
|   |   |   |   |   insu <= 64
|   |   |   |   |   |   plas <= 129: tested_negative (3.0)
|   |   |   |   |   |   plas > 129
|   |   |   |   |   |   |   pedi <= 0.487
|   |   |   |   |   |   |   |   skin <= 33: tested_negative (3.0)
|   |   |   |   |   |   |   |   skin > 33: tested_positive (3.0/1.0)
|   |   |   |   |   |   |   pedi > 0.487: tested_positive (4.0)
|   |   |   |   |   insu > 64: tested_positive (3.0)
|   |   |   insu > 196: tested_negative (12.0)
|   |   age > 24
|   |   |   mass <= 30: tested_negative (51.0/19.0)
|   |   |   mass > 30: tested_positive (118.0/38.0)
|   plas > 165: tested_positive (77.0/9.0)
Số lượng lá: 22

Kích thước của cây: 43
```

Kết luận chương 3

Sau khi áp dụng các thuật toán khai phá dữ liệu thì kết quả cho thấy thuật toán J48 cho kết quả khả quan nhất, có tỷ lệ chính xác cao nhất trong 3 thuật toán, và tỷ lệ lỗi cũng ít nhất. Trong khi đó thuật toán Naïve Bayes cho kết quả có tỷ lệ dự đoán chính xác thấp nhất so với các thuật toán còn lại.

Kết luận

Luận văn đã thực hiện được các công việc như tìm hiểu về bệnh Đái tháo đường, hướng điều trị bệnh Đái tháo đường theo tiêu chuẩn của Bộ Y tế. Học viên đã tìm hiểu về học máy, đặc biệt các thuật toán học có giám sát, áp dụng một số thuật toán học máy (Decision tree, C4.5, SVM, Naïve Bayes) vào bài toán hỗ trợ chẩn đoán bệnh Đái tháo đường. Thực nghiệm một số thuật toán và đánh giá dựa trên kết quả của các thuật toán.

Trong tương lai, hệ hỗ trợ chẩn đoán đái tháo đường sẽ có thêm giao diện để giao tiếp với người sử dụng và đưa ra một mô hình có độ chính xác tốt hơn để chẩn đoán bệnh đái tháo đường. Có thể tập trung vào việc thu thập thông tin từ bệnh án của bệnh nhân được theo dõi qua quá trình điều trị để đưa ra chẩn đoán bệnh một cách chính xác nhất. Đề tài này có thể được mở rộng và cải thiện hơn để tự động hóa phân tích bệnh đái tháo đường một cách chính xác nhất.

DANH MỤC TÀI LIỆU THAM KHẢO

[1] Hướng dẫn chẩn đoán và điều trị đái tháo đường típ 2. Quyết định số 3319/QĐ-BYT ngày 19 tháng 7 năm 2017 của Bộ trưởng Bộ Y tế.

[2] Điều tra quốc gia yếu tố nguy cơ bệnh không lây nhiễm Việt Nam, năm 2015.

[3] Hồ Tú Bảo (2017), Khoa học Dữ liệu và Cách mạng Công nghiệp lần thứ Tư.

[4] Lê Hữu Lập (2014), Bài giảng Phương pháp nghiên cứu khoa học, Học viện Công nghệ BCVT.

[6] Nguyễn Đức Cường, "Slide bài giảng môn học BI & DM: Bussiness Intellegent and Data Mining", 2011-2012.

[5] Từ Minh Phương (2011), Giáo trình trí tuệ nhân tạo, Học viện Công nghệ BCVT.

[7] Trần Đình Quế (2019), Bài giảng Khai phá dữ liệu (Data Mining) , Học viện Công nghệ BCVT.

[8] Arnold Berk, Harvey Lodish, Chris A. Kaiser, Monty Krieger, Anthony Bretscher (Bản dịch: Nhiều tác giả) (2012). "4". Molecular Cell Biology (Sinh học phân tử của tế bào). Tập 2. Di truyền học và sinh học phân tử (ấn bản 7). Hoa Kỳ (Bản dịch: Việt Nam): W. H. Freeman (Bản dịch: Nhà xuất bản Trẻ). tr 2. ISBN 9781429234139. Truy cập ngày 7 tháng 4 năm 2017.

[19] Bonora E, Calcaterra F, Lombardi S, Bonfante N, Formentini G, Bonadonna RC, Muggeo M: "Plasma glucose levels throughout the day and HbA1c interrelationships in type 2 diabetes: implications for treatment and monitoring of metabolic control". Diabetes Care 24:2023– 2029, 2001.

[17] Class for generating a pruned or unpruned C4.5 decision tree. For more information, see Ross Quinlan (1993). C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, San Mateo, CA.

[9] John C. Platt, Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines, Technical Report MSR-TR-98-14 April 21, 1998.

[12] Karegowda, Asha Gowda, A. S. Manjunath, and M. A. Jayaram. "Application of genetic algorithm optimized neural network connection weights for medical diagnosis of pima Indians diabetes." International Journal on Soft Computing 2.2 (2011): 15-23.

[15] K. Rajalakshmi, Dr. S. S. Dhenakaran, "Analysis of Datamining Prediction Techniques in Healthcare Management System", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 5, Issue 4, ISSN: 2277 128X, April 2015.

[11] Lekkas, Stavros and Ludmil Mikhailov. "Evolving fuzzy medical diagnosis of Pima Indians diabetes and of dermatological diseases." Artificial Intelligence in Medicine 50.2 (2010): 117-126.

[13] Ms. Nilam chandgude, Prof. Suvarna pawar, "A survey on diagnosis of diabetes using various classification algorithm", International Journal on Recent and Innovation Trends in Computing and Communication, Volume: 3 Issue: 12, ISSN: 2321-8169, 6706 – 6710, December 2015.

[16] Pragati Agrawal, Amit kumar Dewangan, "A Brief Survey on the Techniques used for the Diagnosis of Diabetes-Mellitus" International Research Journal of Engineering and Technology (IRJET), Volume: 02 Issue: 03, e-ISSN: 2395 - 0056, p-ISSN: 2395-0072, June 2015.

[10] T. Mitchell, Machine Learning and Data Mining, Communications of the ACM, Vol. 42 (1999), No. 11, pp. 30--36.s

[14] Thirumal P. C, Nagarajan N, ―Utilization of Data Mining Techniques for Diagnosis of Diabetes Mellitus- A Case Study", ARPN Journal of Engineering and Applied Sciences, VOL. 10, NO. 1, ISSN 1819-6608, January 2015.

[18] V. Anuja Kumari, R.Chitra "Classification Of Diabetes Disease Using Support Vector Machine", Vol. 3, Issue 2, March -April 2013, pp.1797-1801.

Website:

[21] Class J48 *http://weka.sourceforge.net/doc.dev/weka/classifiers/trees/J48.html*

[20] IDF. Diabetes Atlas, Seventh Edition, 2015. Available at:*http://www.diabetesatlas.org/component/attachments/?task=download&id=11*

[22] Pima-indians-diabetes *https://data.world/data-society/pima-indians-diabetes-database*

DỰ KIẾN KẾ HOẠCH THỰC HIỆN

Kế hoạch thực hiện luận văn thể hiện trong bản sau:

TT	Nội dung	Dự kiến thời gian thực hiện
	Nghiên cứu, chọn đề tài, xây dựng đề cương luận văn	Từ 07/05/2019 – 06/06/2019
	Nộp đề cương luận văn	07/06/2019
	Bảo vệ đề cương, sửa chữa hoàn thiện, nộp đề cương sau bảo vệ	Từ 11/06/2019 – 28/06/2019
	Nghiên cứu, viết, hoàn thiện luận văn	Từ 28/06/2019 – 18/11/2019
	Nộp quyển luận văn và hồ sơ bảo vệ luận văn	Từ 19/11/2019 – 30/11/2019

Ý KIẾN CỦA
NGƯỜI HƯỚNG DẪN KHOA HỌC

(Ký ghi rõ họ tên)

TS. Đỗ Thị Bích Ngọc

NGƯỜI LẬP ĐỀ CƯƠNG

(Ký ghi rõ họ tên)

Hoàng Văn Thắng

DUYỆT CỦA TRƯỞNG TIỂU BAN ĐÁNH GIÁ ĐỀ CƯƠNG

(Ký ghi rõ họ tên)