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

TÓM TẮT LUẬN VĂN THẠC SĨ

HÀ NỘI – 2020

Luận văn được hoàn thành tại:

HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

Người hướng dẫn khoa học: TS. Đỗ Thị Bích Ngọc

Phản biện 1: ……

Phản biện 2: ……..……..

Luận văn sẽ được bảo vệ trước Hội đồng chấm luận văn thạc sĩ tại Học viện Công nghệ Bưu chính Viễn thông

Vào lúc: ..... giờ ..... ngày ..... tháng .... năm .....

Có thể tìm hiểu luận văn tại:

- Thư viện của Học viện Công nghệ Bưu chính Viễn thông

MỞ ĐẦU

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 ĐTĐ dự báo tăng 55% vào năm 2040, với chi phí y tế toàn cầu cho ĐTĐ lên tới 802 tỷ USD.

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

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. 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.

Vì vậy việc khai phá dữ liệu về bệnh án từ đó hỗ trợ các bác sĩ có thể đưa ra các chẩn đoán chính xác hơn, khách quan 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".

Nội dung luận văn

Chương 1: Tổng quan về hệ chuyên gia, trình bày cấu trúc chính và nguyên tắc hoạt động của hệ chuyên gia

Chương 2: 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

Chương 3: Thử nghiệm và lựa chọn thuật toán, Báo cáo đánh giá kết quả.

Mặc dù có nhiều cố gắng nhưng thời gian và năng lực còn hạn chế nên luận văn không tránh khỏi những khiếm khuyết. Kính mong thầy cô và đồng nghiệp thông cảm, cho ý kiến đóng góp.

Trân trọng cảm ơn !

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

1.1. Giới thiệu chung

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

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

1.2.1. Học máy và khám phá tri thức

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.

1.2.2. Học có giám sát.

Học có giám sát (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 $X_1, X_2, \ldots X_n$ tới một vài lớp bằng cách xem xét một số mẫu dữ liệu - kết quả của hàm đó.

1.2.3. Học không có giám sát.

Học không có giám sát (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.

1.2.4. Học giám sát một phần.

Học nửa giám sát (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.

1.2.5. Học tăng cường.

Học tăng cường (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.

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.

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 luật trích chọn không chỉ cần độ chính xác cao mà còn phải đơn giản và dễ hiểu.

Mục tiêu của luận văn: Đưa ra một model có tỷ lệ dự đoán bệnh nhân dương tính với bệnh Đái tháo đường tuýp 2.

Input hệ thống là : 8 thuộc tính và 2 class (0 tương ứng với âm tính, 1 tương ứng với dương tính).

Bảng 1: 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

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. Mô tả input và output của 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

Bệnh đái tháo đường của hồ sơ bệnh nhân được tính bằng cách sử dụng cây quyết định theo hai giai đoạn: xử lý trước dữ liệu trong đó các thuộc tính được xác định và thứ hai là mô hình dự đoán bệnh đái tháo đường được xây dựng bằng cách áp dụng các thuật toán sử dụng cây quyết định.

Cây quyết định là một cấu trúc cây, ở dạng sơ đồ. Nó được sử dụng như một phương pháp để phân loại và dự đoán với sự xuất hiện bằng cách sử dụng các nút và nút lá. Nút gốc và nút bên trong là các trường hợp thử nghiệm được sử dụng để phân tách các thể hiện với các tính năng khác nhau. Các nút nội bộ là kết quả của các trường hợp kiểm tra thuộc tính. Các nút lá biểu thị biến lớp.

Cây quyết định cung cấp một kỹ thuật mạnh mẽ để phân loại và dự đoán trong chẩn đoán bệnh đái tháo đường. Các thuật toán cây quyết định khác nhau có sẵn để phân loại dữ liệu, bao gồm ID3, C4.5, C5, J48, CART, CHAID.... Trong bài luận văn này, các thuật toán cây quyết định như J48 đã được chọn để thiết lập mô hình. Mỗi nút cho decisiontree được tìm thấy bằng cách tính mức tăng thông tin cao nhất cho tất cả các thuộc tính và nếu một thuộc tính cụ thể đưa ra một kết quả tường minh(phân loại rõ ràng thuộc tính lớp), nhánh của thuộc tính này kết thúc và giá trị cuối cùng được gán cho nó.

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).

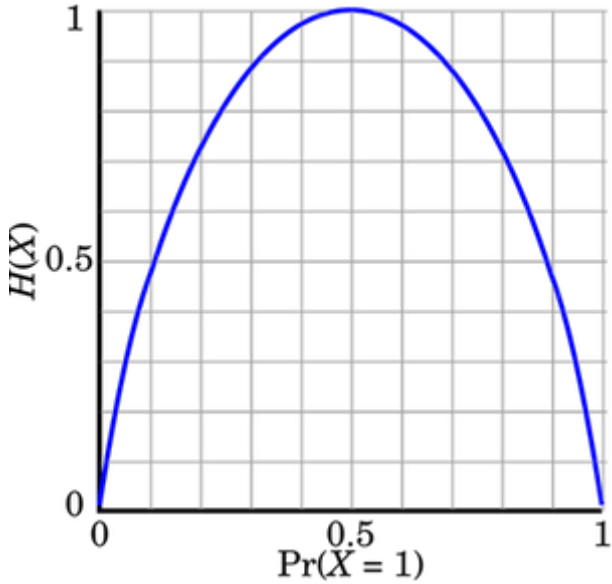
Hàm số Entropy

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 sau:



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.

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:

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

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:

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

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.

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

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.

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.

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 như sau:

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

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.

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.

Bảng 1: Bộ dữ liệu được sử dụng để thử nghiệm

Tập dữ liệu	**Số thuộc tính**	**Số bản ghi**
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	8	768

3.2. Tiền xử lý dữ liệu.

Tôi 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.

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.

Bộ dữ liệu được chia thành 10 phần. Trong đó 90% được lựa chọn làm bộ traning, 10% được chọn làm bộ test.

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 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.

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

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

Bảng 2: 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)	14.20 % (98)
3	82.0 % (566)	17.9 % (124)
4	84.78 % (585)	15.21 % (105)
5	83.62 % (577)	16.37 % (113)
6	84.20 % (581)	15.79 % (109)
7	80.53 % (556)	18.41 % (134)
8	80.57 % (556)	19.42 % (134)
9	84.63 % (584)	15.36 % (106)
10	80.87 % (558)	19.13 % (132)

Từ **Bảng 2** 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%.

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

Bảng 3: 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)	24.3 % (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)	25.22 % (26)
8	84.61 % (66)	15.38 % (12)
9	71.79 % (56)	28.20 % (22)
10	76.92 % (60)	23.08 % (18)

Từ **Bảng 3** 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%.

3.3.2. Đánh giá thuật toán SVM

Trong phần mềm weka thì thuật toán SVM có ký hiệu là SMO.

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

Bảng 4: 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 4** 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%.

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

Bảng 5: 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	78.20 % (61)	21.80 % (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 5** 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%.

3.3.3. Đánh giá thuật toán Naïve Bayes

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

Bảng 6: 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)	(164)

Từ **Bảng 6** 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 77,53% 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%.

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

Bảng 7: 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)

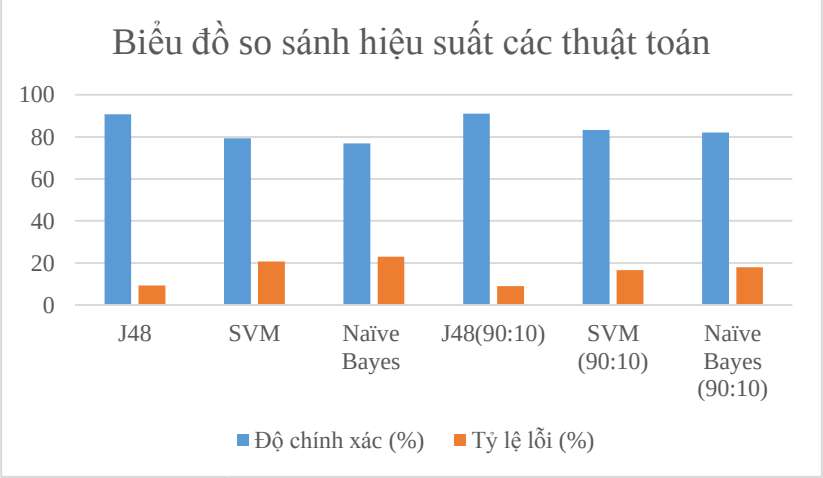
Từ **Bảng 7** ta có thể thấy được với lần chạy thứ 8 thì tỷ lệ dự đoán chính xác là tốt nhất với 78 trường hợp.

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.

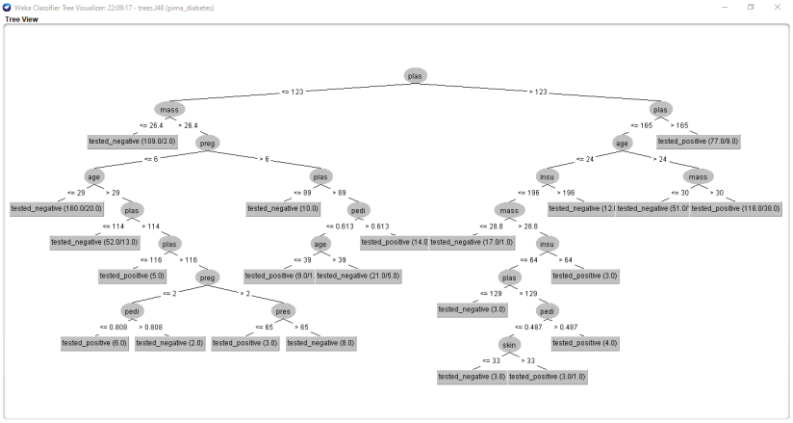
Tỷ lệ dự đoán không chính xác là 14 mẫu với tỷ lệ 17,95%.

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 2: 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

Hệ hỗ trợ chẩn đoán bệnh đái tháo đường là một vấn đề y tế quan trọng trong thực tế. Phát hiện bệnh đái tháo đường ở giai đoạn đầu là chìa khóa để điều trị một cách triệt để. Luận văn này cho thấy Cây quyết định được sử dụng như thế nào để mô hình chẩn đoán bệnh đái tháo đường phục vụ cho việc chẩn đoán, cũng như với việc tìm hiểu về bệnh đái tháo đường và các thuật toán áp dụng vào khai phá dữ liệu dựa trên bộ dữ liệu bệnh án đái tháo đường.

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.