

ỨNG DỤNG KỸ THUẬT CÂY QUYẾT ĐỊNH XÂY DỰNG HỆ THỐNG DỰ ĐOÁN BỆNH ĐÁI THÁO ĐƯỜNG

APPLYING DECISION TREE TECHNIQUE TO BUILD A SYSTEM FOR PREDICTING DIABETES MELLITUS

Nguyễn Văn Chức¹, Trần Thị Kim Hằng²

¹*Trường Đại học Kinh tế, Đại học Đà Nẵng; Email: chuc1803@gmail.com;*

²*Lớp 36K14 Trường Đại học Kinh tế, Đại học Đà Nẵng; Email: hangtran.0709@gmail.com*

Tóm tắt - Hiện nay, bệnh đái tháo đường đang ngày càng trở nên phổ biến trên khắp thế giới, trong đó có Việt Nam. Bệnh gây ra rất nhiều biến chứng nguy hiểm nếu không được phát hiện và chữa trị kịp thời. Tuy nhiên, vấn đề phát hiện sớm bệnh đái tháo đường tại Việt Nam vẫn còn có gặp nhiều khó khăn, nhất là các vùng sâu, vùng xa, nơi không có đầy đủ các trang thiết bị y tế cần thiết [3]. Bài báo tập trung nghiên cứu kỹ thuật cây quyết định trong khai phá dữ liệu để xây dựng mô hình dự đoán nhằm giúp người dùng có thể tự kiểm tra nguy cơ mắc bệnh đái tháo đường của mình. Trên cơ sở các tri thức phát hiện được từ mô hình dự đoán, nhóm nghiên cứu đã xây dựng một giao tiếp trên nền web để người dùng có thể dễ dàng sử dụng các tri thức này vào việc dự đoán nguy cơ mắc bệnh của bản thân nhằm kịp thời phát hiện và có biện pháp chữa trị thích hợp.

Từ khóa - cây quyết định; dự đoán bệnh; đái tháo đường; khai phá dữ liệu; mô hình dự đoán.

1. Đặt vấn đề

Thông thường, để chuẩn đoán một người có mắc bệnh đái tháo đường hay không, phải thông qua các xét nghiệm đường máu, một biện pháp khá tốn kém và đòi hỏi phải có đủ trang thiết bị y tế hỗ trợ. Chính vì điều này nên đa số người dân không chủ động làm các kiểm tra định kỳ hay các xét nghiệm để chuẩn đoán sớm nguy cơ mắc bệnh đái tháo đường, mà chỉ thực hiện khi đã có các dấu hiệu phát bệnh, giai đoạn đã có các biến chứng nghiêm trọng [3]. Vì vậy, rất cần thiết phải có một công cụ để sử dụng và ít tốn kém cho phép người dùng có thể thường xuyên tự kiểm tra nguy cơ mắc bệnh đái tháo đường của mình, nhằm phát hiện và có biện pháp chữa trị kịp thời. Hệ thống này cho phép người dùng cung cấp các thông tin liên quan đến sức khỏe cá nhân như giới tính, chiều cao, cân nặng, huyết áp, tiền sử bệnh tật và các yếu tố di truyền thì có thể kiểm tra được nguy cơ mắc bệnh đái tháo đường của bản thân. Bài báo tập trung nghiên cứu về kỹ thuật cây quyết định trong khai phá dữ liệu để xây dựng mô hình dự đoán nguy cơ mắc bệnh đái tháo đường nhằm giúp người dùng phát hiện sớm nguy cơ mắc bệnh.

2. Giới thiệu về kỹ thuật phân lớp dữ liệu dựa vào cây quyết định

Trong lĩnh vực khai phá dữ liệu, cây quyết định (Decision Tree – DT) là một mô hình dự đoán thuộc lớp các bài toán phân lớp dùng để xác định lớp của các đối tượng cần dự đoán [1]. Bản chất cây quyết định dựa vào dãy các luật IF ... THEN để dự đoán lớp của đối tượng. Mỗi nút trong (internal node) của DT tương ứng với một biến, đường nối giữa một nút trong với nút con của nó thể hiện một giá trị cụ thể biến đó. Mỗi nút lá (leaf) đại diện

Abstract - Nowadays, diabetes mellitus has become a common condition in the world, including Vietnam. This disease causes a variety of serious complications if it is not diagnosed and treated in time. However, Vietnam has encountered many difficulties in early detection of diabetes, especially in remote areas where there are lack of standard medical equipments [3]. This paper focuses on analyzing the decision tree technique in data mining and a predictive model to help people find out the risks of getting diabetes mellitus by themselves. Basing on the knowledge learnt from the predictive model, the research group also built an interface on a web platform to help users learn the knowledge about the risk factors, predict how they are likely to develop this disease, to detect and decide appropriate treatments.

Key words - decision tree; disease prediction; diabetes mellitus; data mining; predictive model

cho giá trị dự đoán. Cây quyết định học để dự đoán giá trị của các biến phân loại bằng cách dựa vào tập dữ liệu huấn luyện (training data) để chọn ra nút gốc (root node) để phân tách cây bằng cách tính độ lợi thông tin (Information Gain - IG), quá trình phân tách cây được thực hiện một cách đệ qui cho đến khi không thể tiếp tục thực hiện việc phân tách cây được nữa [1], [4].

Cây quyết định được chia thành hai loại:

Cây hồi quy dùng để dự đoán giá trị của biến phân loại có kiểu dữ liệu giá trị như dự đoán doanh thu, lợi nhuận, giá thành sản phẩm... Thuật toán phổ biến dùng để xây dựng cây hồi quy là CART [5].

Cây phân lớp dùng để dự đoán giá trị của biến phân loại có kiểu dữ liệu phi giá trị như dự đoán khả năng mua hàng, khả năng bị bệnh, kết quả học tập của sinh viên (xuất sắc, giỏi, khá, trung bình, yếu)... Thuật toán phổ biến dùng để xây dựng cây phân lớp là ID3, J48, C4.5, C5.0 [5].

3. Xây dựng hệ thống dự đoán bệnh đái tháo đường

3.1. Mô tả hệ thống

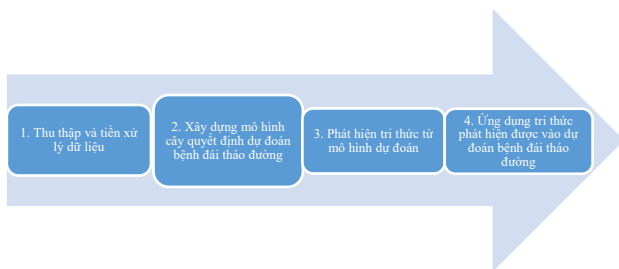
Mục đích: Ứng dụng cây quyết định xây dựng mô hình dự đoán nguy cơ mắc bệnh đái tháo đường.

Đầu vào: Gồm các đặc trưng liên quan như: giới tính, chiều cao, cân nặng, huyết áp, ...; thông tin về tiền sử bệnh; yếu tố di truyền.

Đầu ra: Dự đoán nguy cơ mắc bệnh: bình thường, tiền đái tháo đường, đái tháo đường.

3.2. Kịch bản triển khai hệ thống dự đoán bệnh

Hệ thống dự đoán bệnh đái tháo đường được tiến hành theo các bước chính như Hình 1.



Hình 1. Kịch bản triển khai hệ thống dự đoán bệnh

Bước 1. Thu thập và tiền xử lý dữ liệu

Dữ liệu thu thập được gần 8100 mẫu, là thông tin của

các bệnh nhân được Trung tâm Phòng chống Sốt rét - Bướu cổ tỉnh Quảng Nam, nơi đang thực hiện chương trình “Sàng lọc đái tháo đường” dưới sự chỉ đạo về chuyên môn, nghiệp vụ của Bệnh viện Nội tiết Trung ương và Ban điều hành dự án chống bệnh đái tháo đường quốc gia, thu thập qua 2 năm 2012-2013 [3].

Dữ liệu điều tra ban đầu gồm rất nhiều thuộc tính, sau quá trình tiền xử lý dữ liệu (sử dụng phương pháp trích chọn thuộc tính) để đánh giá mức độ ảnh hưởng của các thuộc tính đến nguy cơ mắc bệnh đái tháo đường, mô hình xác định được các thuộc tính có ảnh hưởng đến nguy cơ mắc bệnh đái tháo đường như Bảng 1.

Bảng 1. Mô tả dữ liệu

STT	Tên thuộc tính	Kiểu dữ liệu	Miền giá trị	Giải thích
1	ID	Nominal	000001-008084	Mã bệnh nhân, khóa
2	GioiTinh	Nominal	Nu, Nam	Giới tính của bệnh nhân
3	Tuoi	Interval	<50, 50-54, 55-59, 60-64, >64	Độ tuổi của bệnh nhân
4	TCCongViec	Nominal	HTTT, Nhe, TB, Nang	Tính chất công việc của bệnh nhân: Hoàn toàn tĩnh tại, nhẹ, trung bình hay nặng
5	TDHV	Nominal	TDHV1, TDHV2, TDHV3, TDHV4, TDHV5, TDHV6	Trình độ học vấn của bệnh nhân: Không biết đọc, không biết viết; Biết đọc, biết viết; Tốt nghiệp tiểu học; Tốt nghiệp trung học cơ sở; Tốt nghiệp phổ thông trung học; Tốt nghiệp trung học chuyên nghiệp, CD, ĐH hoặc cao hơn
6	MacBenhHA	Nominal	Co, Khong	Bệnh nhân đó có mắc bệnh huyết áp hay không
7	ChuaHA1	Nominal	Da, Chua, KhongMBHA	Bệnh nhân đã chữa bệnh tăng huyết áp bằng phương pháp ăn uống và luyện tập hay chưa
8	ChuaHA2	Nominal	Da, Chua, KhongMBHA	Bệnh nhân đã chữa bệnh tăng huyết áp bằng thuốc đông y, thuốc nam hay chưa
9	ChuaHA3	Nominal	Da, Chua, KhongMBHA	Bệnh nhân đã chữa bệnh tăng huyết áp bằng thuốc tây y hay chưa
10	BoMeDTD	Nominal	Co, Khong	Bố mẹ đẻ có bị mắc bệnh DTD hay không
11	ACEDTD	Nominal	Co, Khong	Anh, chị, em ruột có bị mắc bệnh DTD hay không
12	OngBaNoiDTD	Nominal	Co, Khong	Ông, bà nội ruột có bị mắc bệnh DTD hay không
13	ConDTD	Nominal	Co, Không	Con ruột có bị mắc bệnh DTD hay không
14	TS1	Nominal	Co, Khong	Đã có tiền sử về bệnh đột quỵ hay không
15	TS2	Nominal	Co, Khong	Đã có tiền sử về bệnh đau thắt ngực hay không
16	TS3	Nominal	Co, Khong	Đã có tiền sử về bệnh suy tim hay không
17	TS4	Nominal	Co, Khong	Đã có tiền sử về bệnh loét bàn chân hay không
18	TS5	Nominal	Co, Khong	Có bị cắt cụt chi hay không
19	RLMM	Nominal	Co, Khong	Đã bao giờ bị chuẩn đoán rối loạn mỡ máu hay chưa
20	CanNangNhat	Interval	<43, 43-54, 55-66, 67-78, 79-90, >90	Lúc nặng nhất là bao nhiêu kg? (ĐVT: Kg)
21	TuoiCanNangNhat	Interval	<27, 27-37, 38-48, 49-59, >59	Tuổi có cân nặng cao nhất?
22	DaMangThai	Nominal	Da, Chua, GTNam	Đã mang thai chưa?
23	ConNang	Interval	<1800, 1800-3600, >3600	Cân nặng nhất của con lúc sinh là bao nhiêu? (ĐVT: gam)
24	ConNhe	Interval	<1800, 1800-3600, >3600	Cân nhẹ nhất của con lúc sinh là bao nhiêu? (ĐVT: gam)
25	DTDMT	Nominal	Co, Khong, KhongBiet, GioiTinhNamHoacChuaMangThai	Đã bao giờ được chẩn đoán bị đái tháo đường khi mang thai không?
26	ChieuCao	Interval	<1.4, 1.4-1.59, 1.59-1.69, >1.69	Chiều cao của bệnh nhân (ĐVT: mét)
27	CanNang	Interval	<43, 43-54, 55-66, 67-78, 79-90, >90	Cân nặng của bệnh nhân (ĐVT: Kg)
28	BMI	Interval	<18, 18-22, 23-27, >27	Chỉ số BMI của bệnh nhân
29	BMINangNhat	Interval	<18, 18-22, 23-27, >27	Chỉ số BMI lúc cân nặng nhất của bệnh nhân
30	VE	Interval	<60, 60-70, 71-80, 81-90, >90	Vòng eo của bệnh nhân (cm)
31	VH	Interval	<80, 80-90, 91-100, >100	Vòng hông của bệnh nhân (cm)
32	HATTh	Interval	<100, 100-120, 121-140, 141-160, 161-180, >180	Huyết áp tâm thu của bệnh nhân (mmHg).
33	HATTr	Interval	<70, 70-89, 90-109, 110-129, >129	Huyết áp tâm trương của bệnh nhân (mmHg)
34	KQ	Nominal	BT, TienDTD, DTD	Kết quả, biến phân lớp.

Bước 2. Xây dựng mô hình dự đoán nguy cơ mắc bệnh

Mô hình cây quyết định dự đoán nguy cơ mắc bệnh đái tháo đường được xây dựng trên công cụ khai phá dữ liệu Business Intelligence Development Studio (BIDS) của Microsoft. BIDS là công cụ rất mạnh cho phép triển khai các mô hình khai phá dữ liệu, được sử dụng rộng rãi hiện nay bởi khả năng kết nối dễ dàng với nhiều nguồn dữ liệu, giao diện dễ sử dụng và nhất là khả năng biểu diễn tri thức phát hiện được rất trực quan dễ hiểu, dễ sử dụng. BIDS được tích hợp vào SQL SERVER 2005 trở về sau trong các phiên bản Enterprise hoặc Development [2], [6].

Sau khi thực hiện các thao tác tiền xử lý dữ liệu để phù hợp với mô hình khai phá dữ liệu, sử dụng Microsoft Decision Tree trong BIDS để xây dựng cây quyết định dự đoán nguy cơ mắc bệnh đái tháo đường. Kết quả cây quyết định dự đoán nguy cơ mắc bệnh đái tháo đường như Hình 4.

Bước 3. Phát hiện tri thức từ mô hình cây quyết định

Từ cây quyết định đã xây dựng, các tri thức phát hiện được cho dưới dạng các luật IF ... THEN như sau:

IF L_1 AND L_2 AND ... AND L_n THEN Kết quả = M

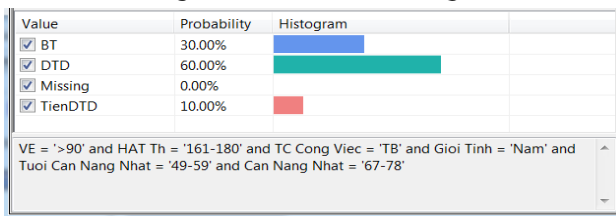
Trong đó:

$L_1, L_2, \dots, L_n$ là các biểu thức logic mà về trái là các thuộc tính và về phải là giá trị có thể có của thuộc tính đó.

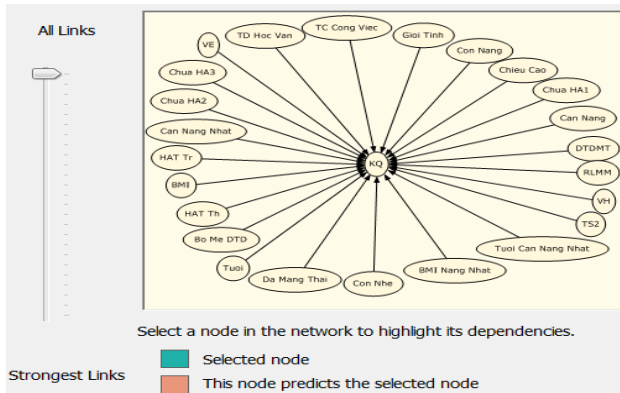
M là một trong các giá trị của biến kết quả cần dự đoán (bình thường, tiền đái tháo đường, đái tháo đường).

Chẳng hạn, một luật được trích ra từ cây quyết định dự đoán bệnh đái tháo đường đã xây dựng như sau:

IF VE = '>90' and HAT Th = '161-180' and TC Cong Viec = 'TB' and Gioi Tinh = 'Nam' and Tuoi Can Nang Nhat = '49-59' and Can Nang Nhat = '67-78' THEN KQ = "DTD"



(Nếu một người có vòng eo trên 90cm, có huyết áp tâm thu đo được nằm trong khoảng 161-180 mmHg, tính chất công việc trung bình, giới tính Nam, tuổi lúc cân nặng nằm trong khoảng 49-59 và trọng lượng nặng nhất trong khoảng 67-78kg thì người này có khả năng mắc bệnh đái tháo đường với xác suất 60%).



Hình 2. Mạng phụ thuộc của mô hình

Ngoài ra, mạng phụ thuộc của mô hình cho biết trọng số của các nhân tố ảnh hưởng đến nguy cơ mắc bệnh đái tháo đường như Hình 2.

Từ mạng phụ thuộc của mô hình, ta biết được mức độ ảnh hưởng của các nhân tố đến nguy cơ mắc bệnh đái tháo đường, được liệt kê theo mức độ ảnh hưởng giảm dần, như Bảng 2. **Bảng 2.** mức độ ảnh hưởng của các nhân tố đến nguy cơ mắc bệnh đái tháo đường

STT	Tên nhân tố	STT	Tên nhân tố
1	Huyết áp tâm thu	13	BMI
2	Tuổi	14	Giới tính
3	Vòng eo	15	Rối loạn mỡ máu
4	Cân nặng	16	BMI nặng nhất
5	Con nặng	17	Cân nặng nhất
6	Trình độ học vấn	18	Tuổi cân nặng nhất
7	Chiều cao	19	Chữa huyết áp 1
8	Vòng hông	20	Đã mang thai
9	Tính chất công việc	21	Bố mẹ đái tháo đường
10	Đái tháo đường mang thai	22	Tiền sử 2
11	Con nhẹ	23	Chữa huyết áp 3
12	Huyết áp tâm trương	24	Chữa huyết áp 2

Bước 4. Ứng dụng tri thức phát hiện được vào dự đoán bệnh đái tháo đường

Dựa vào các tri thức phát hiện được từ mô hình cây quyết định, một hệ thống giao tiếp được xây dựng trên nền web cho phép người dùng sử dụng dự đoán nguy cơ mắc bệnh đái tháo đường. Hệ thống sẽ cho ra kết quả về nguy cơ mắc bệnh của người dùng ứng với các thông tin người dùng đã cung cấp như Hình 3.

4. Kết luận và hướng phát triển

Cây quyết định là kỹ thuật khai phá dữ liệu được sử dụng phổ biến để giải quyết bài toán phân lớp dữ liệu bởi tính đơn giản, hiệu quả và nhất là nhất là khả năng biểu diễn tri thức phát hiện được rất trực quan, dễ hiểu, dễ sử dụng. Bài báo đã tìm hiểu về lý thuyết cây quyết định, từ đó nghiên cứu ứng dụng kỹ thuật này vào xây dựng mô hình dự đoán nguy cơ mắc bệnh đái tháo đường.

Dựa vào mô hình đã xây dựng, nhóm nghiên cứu đã xây dựng một ứng dụng trên nền web giúp cho người dùng dễ dàng sử dụng các tri thức phát hiện được từ mô hình trong việc kiểm tra nguy cơ mắc bệnh đái tháo đường của bản thân.

Hiện nay, có rất nhiều hệ thống (website) giới thiệu về bệnh đái tháo đường cũng như cách phòng chữa bệnh này. Tuy nhiên, các hệ thống này chỉ dừng lại ở việc cung cấp thông tin về bệnh, chưa giải quyết được vấn đề quan trọng là giúp người dùng có thể tự dự đoán được nguy cơ mắc bệnh đái tháo đường của bản thân, từ đó nâng cao ý thức trong việc chủ động phòng ngừa và phát hiện sớm nguy cơ mắc bệnh đái tháo đường nhằm giảm thiểu thiệt hại của bệnh đái tháo đường đối với xã hội.

Dự đoán

Tiền sử bệnh lý

Mắc bệnh huyết áp hay không?
Chữa huyết áp bằng phương pháp ăn uống, tập luyện hay chưa?
Chữa bệnh huyết áp bằng thuốc đông y, nam y hay chưa?
Chữa huyết áp bằng thuốc tây y hay chưa?
Đã có tiền sử về bệnh đái tháo đường hay không?
Đã có tiền sử về bệnh đau thắt ngực hay không?
Đã có tiền sử về bệnh suy tim hay không?
Đã có tiền sử về bệnh loét bàn chân hay không?
Có bị cắt cụt chi hay không?
Đã bao giờ bị chuẩn đoán rối loạn mỡ máu hay chưa?
Đã bao giờ được chẩn đoán bị đái tháo đường khi mang thai không?

Navigation
Đi Dự đoán
Đi Thông tin liên hệ

Chỉ số cơ thể

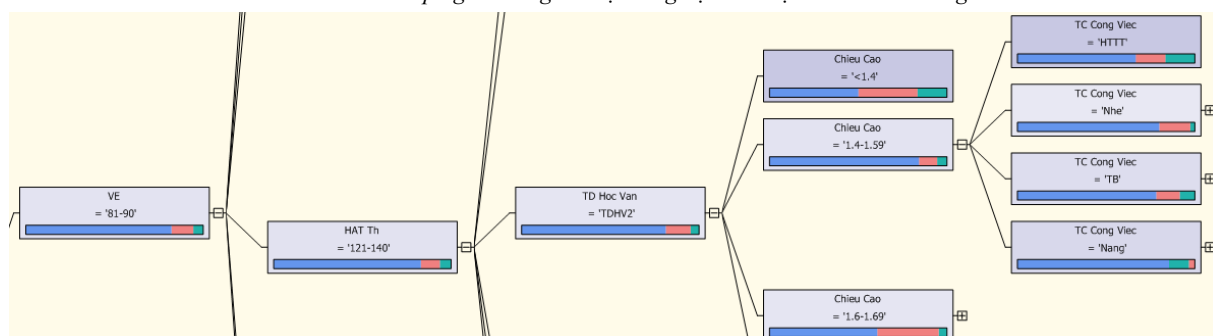
Lúc nặng nhất là bao nhiêu kg? (ĐVT: Kg)
Tuổi có cân nặng cao nhất?
Đã mang thai chưa?
Cân nặng nhất của con lúc sinh là bao nhiêu? (ĐVT: gam)
Cân nhẹ nhất của con lúc sinh là bao nhiêu? (ĐVT: gam)
Chiều cao của bệnh nhân (ĐVT: mét)
Cân nặng của bệnh nhân (ĐVT: Kg)
Chỉ số BMI của bệnh nhân
Chỉ số BMI lúc cân nặng nhất của bệnh nhân
Vòng eo của bệnh nhân
Vòng hông của bệnh nhân
Huyết áp tâm thu của bệnh nhân
Huyết áp tâm trương của bệnh nhân

Dự đoán

Kết quả dự đoán

Kết Quả :	Bệnh thường
Mức Độ Chính Xác	75 %

Hình 3. Giao tiếp người dùng với hệ thống dự đoán bệnh đái tháo đường



Hình 4. Một nhánh của cây quyết định dự đoán bệnh đái tháo đường

TÀI LIỆU THAM KHẢO

- [1] Rokach Lior; Maimon O., *Data mining with decision trees: theory and applications*, World Scientific Pub Co Inc. ISBN 978-9812771711, 2008.
- [2] JamieMacLennan, Z. T., Bogdan Crivat, *Data Mining with Microsoft SQL Server 2008*, Indianapolis, Indiana: Wiley Publishing, Inc, 2008.
- [3] Trung tâm Phòng chống Sốt rét-Bướu cổ tỉnh Quảng Nam, Tài liệu tập huấn về sàng lọc và quản lý tiền đái tháo đường và bệnh đái tháo đường tuýp 2 của Trung tâm Phòng chống Sốt rét - Bướu cổ tỉnh Quảng Nam, 2013.
- [4] Nguyễn Đức Thuần, 2013, *Nhập môn khai phá dữ liệu và quản trị tri thức*, NXB Thông tin và Truyền thông.
- [5] Nguyễn Văn Chức, Ứng dụng kỹ thuật cây quyết định trong khai phá dữ liệu xây dựng hệ thống tư vấn chọn ngành tuyển sinh đại học, *Tạp chí KH&CN ĐHDN*, số 1(74).2014, Quyển 2.
- [6] <http://msdn.microsoft.com/en-us/library/ms173767.aspx> (Introducing Business Intelligence Development Studio)

(BBT nhận bài: 27/08/2014, phản biện xong: 03/11/2014)