BỘ GIÁO DỤC VÀ ĐÀO TẠO
ĐẠI HỌC ĐÀ NẴNG

HOÀNG THỊ THANH HIỀN

ỨNG DỤNG KHAI PHÁ DỮ LIỆU ĐỂ XÂY DỰNG HỆ THỐNG CHẨN ĐOÁN BỆNH TRẦM CẢM CHO HỌC SINH PHỔ THÔNG

Chuyên ngành: Hệ thống thông tin
Mã số : 60.48.01.04

TÓM TẮT LUẬN VĂN THẠC SĨ HỆ THỐNG THÔNG TIN

Đà Nẵng - Năm 2016

Công trình được hoàn thành tại
ĐẠI HỌC ĐÀ NẴNG

Người hướng dẫn khoa học: PGS.TS. NGUYỄN THANH BÌNH

Phản biện 1: TS. Phạm Anh Phương

Phản biện 2: PGS.TS. Lê Văn Sơn

Luận văn đã được bảo vệ trước hội đồng chấm Luận văn tốt nghiệp thạc sĩ Hệ thống thông tin họp tại Đại học Đà Nẵng vào ngày 31 tháng 7 năm 2016.

Có thể tìm hiểu luận văn tại:

- Trung tâm Học liệu, Đại học Đà Nẵng
- Thư viện Trường Đại học Sư Phạm, Đại Học Đà Nẵng

MỞ ĐẦU

1. Tính cấp thiết của đề tài

Hiện nay, trầm cảm là một bệnh diễn ra khá phổ biến và có tác động phức tạp đến đời sống xã hội, nhất là đối với lứa tuổi thanh thiếu niên. Nguyên nhân chủ yếu dẫn đến hiện tượng này chính là vấn đề về sức khoẻ tâm thần.

Sức khỏe tâm thần là một cấu phần quan trọng trong sức khỏe tổng thể của trẻ. Vấn đề sức khỏe tâm thần ở thời kỳ vị thành niên có mối liên quan chặt chẽ với rất nhiều hành vi như: uống rượu, hút thuốc, sử dụng ma túy, nguy cơ tự tử,… sẽ ảnh hưởng đến sức khỏe của trẻ khi trưởng thành.

Ở tuổi vị thành niên, rối loạn trầm cảm thường biểu hiện là những thay đổi về cảm xúc như: cảm thấy buồn, khóc, vô vọng; không quan tâm đến những hoạt động vui chơi, giải trí hay suy giảm các hoạt động học tập; ăn không ngon miệng; hay thay đổi về giấc ngủ; hay có những khó chịu trong cơ thể một cách mơ hồ; ngoài ra trẻ còn nghĩ rằng không thể làm được việc gì đúng, cảm thấy cuộc sống không có ý nghĩa hoặc vô vọng [6].

Trầm cảm ảnh hưởng rất lớn đến năng lực học tập, giao tiếp; sự hình thành phát triển các mối quan hệ xã hội, hoàn thiện thể chất, tinh thần và tính cách của trẻ. Nếu rối loạn trầm cảm không được quan tâm phòng ngừa và can thiệp phù hợp sẽ tăng gánh nặng cho gia đình và xã hội. Do đó, yêu cầu cấp thiết cần phải có hệ thống chẩn đoán sàng lọc lâm sàng để phát hiện sớm các biểu hiện rối loạn trầm cảm ở học sinh phổ thông nhằm đưa ra các giải pháp can thiệp kịp thời trong việc phát triển sức khỏe.

Hiện nay, việc ứng dụng Công nghệ thông tin (CNTT) vào lĩnh vực y tế còn hạn chế, nhất là việc hỗ trợ tìm kiếm, khai thác thông tin nhằm chẩn đoán các biểu hiện lâm sàng. Trong đó, khai phá dữ liệu

là một kỹ thuật thường được áp dụng để hỗ trợ đưa ra các quyết định khá chính xác.

Xuất phát từ những thực tế trên, tôi đã chọn đề tài "**Ứng dụng khai phá dữ liệu để xây dựng hệ thống chẩn đoán bệnh trầm cảm cho học sinh phổ thông"** để nghiên cứu luận văn thạc sĩ của mình.

2. Mục tiêu nghiên cứu và nhiệm vụ nghiên cứu

❖ **Mục tiêu nghiên cứu**

Nghiên cứu kỹ thuật khai phá dữ liệu và ứng dụng để xây dựng hệ thống hỗ trợ chẩn đoán bệnh rối loạn trầm cảm.

❖ **Nhiệm vụ nghiên cứu:**

- Tìm hiểu về bệnh RLTC, tiến hành điều tra thu thập dữ liệu.
- Nghiên cứu lý thuyết về kỹ thuật phân lớp bằng thuật toán cây quyết định và thuật toán phân cụm.
- Xây dựng mô hình để chẩn đoán bệnh RLTC cho học sinh dựa vào kỹ thuật cây quyết định.
- Ứng dụng công cụ hỗ trợ khai phá Business Intelligence để xây dựng và kiểm tra các mô hình.
- Đánh giá kết quả dự đoán của mô hình và lựa chọn mô hình tốt nhất để chẩn đoán bệnh RLTC.
- Ứng dụng kỹ thuật cây quyết định xây dựng hệ thống chẩn đoán bệnh trầm cảm của học sinh.
- Ứng dụng kỹ thuật phân cụm để phân tích các đặc trưng của bệnh RLTC.

3. Đối tượng và phạm vi nghiên cứu

❖ **Đối tượng nghiên cứu**

- Dữ liệu nghiên cứu bao gồm các đặc điểm cá nhân và một số yếu tố liên quan đến biểu hiện rối loạn trầm cảm của học sinh.
- Các kỹ thuật khai phá dữ liệu, công cụ khai phá dữ liệu và mô-đun lập trình trong khai phá dữ liệu.

❖ **Phạm vi nghiên cứu**

- Số liệu thu thập gồm các hồ sơ bệnh án thuộc đối tượng trẻ vị thành niên từ 12 -18 tuổi.

- Các kỹ thuật: phân lớp bằng cây quyết định và kỹ thuật phân cụm.

- Công cụ hỗ trợ khai phá BI và các môđun hỗ trợ.

- Xây dựng hệ thống chẩn đoán bệnh và tìm ra những đặc trưng của bệnh rối loạn trầm cảm.

4. Phương pháp nghiên cứu

❖ **Phương pháp nghiên cứu lý luận**

- Thu thập, đọc hiểu thông tin từ các tài liệu, giáo trình liên quan đến khai phá dữ liệu.

- Nghiên cứu các kỹ thuật phân lớp dữ liệu dựa vào cây quyết định, ứng dụng các kỹ thuật đó để chuẩn đoán bệnh rối loạn trầm cảm dựa vào các thông tin đầu vào.

- Tìm hiểu các kỹ thuật phân cụm để phân tích những đặc trưng của bệnh RLTC.

❖ **Phương pháp nghiên cứu thực tiễn**

- Sử dụng kiến thức khai phá dữ liệu cộng với tri thức chuyên gia bác sĩ, y học chứng cớ và y học thực chứng trong quá trình khai phá dữ liệu y khoa.

- Tiến hành so sánh kết quả của các kỹ thuật KPDL để lựa chọn kỹ thuật cho kết quả chính xác nhất.

- Xây dựng hệ thống nhằm hỗ trợ bác sĩ trong việc chẩn đoán và điều trị bệnh.

5. Bố cục của luận văn

Ngoài các phần như mở đầu, kết luận và hướng phát triển, đề tài gồm 3 chương:

Chương 1: Tổng quan về khai phá dữ liệu. Chương này tìm hiểu và trình bày các nội dung: nghiên cứu tổng quan về KPDL; các kỹ

thuật khai phá dữ liệu bằng cây quyết định và kỹ thuật phân cụm.

Chương 2: Nghiên cứu và xử lý dữ liệu về bệnh rối loạn trầm cảm. Chương này trình bày nội dung sau: tìm hiểu đặc điểm tâm lý của tuổi vị thành niên, khái niệm về bệnh rối loạn trầm cảm, đặc điểm lâm sàng, các yếu tố liên quan đến bệnh rối loại trầm cảm trên cơ sở đó thu thập và xử lý dữ liệu nghiên cứu bệnh RLTC. Ngoài ra còn trình bày các công cụ xây dựng mô hình khai phá dữ liệu.

Chương 3: Xây dựng hệ thống chẩn đoán bệnh RLTC dựa trên khai phá dữ liệu. Chương này trình bày dữ liệu yêu cầu cho việc xây dựng mô hình, những tham số hỗ trợ cho các thuật toán từ đó ứng dụng kỹ thuật cây quyết định để chẩn đoán bệnh và sử dụng kỹ thuật phân cụm để phân tích các đặc trưng của bệnh rối loại trầm cảm cho học sinh. Đồng thời chương này xây dựng chương trình cài đặt thử nghiệm việc chẩn đoán bệnh bằng kỹ thuật quyết định và kỹ thuật phân cụm trên cơ sở dữ liệu bệnh RLTC.

CHƯƠNG 1
TỔNG QUAN VỀ KHAI PHÁ DỮ LIỆU

1.1. KHAI PHÁ DỮ LIỆU

1.1.1. Định nghĩa

Khai phá dữ liệu (KPDL) là một khái niệm ra đời vào những năm cuối của thập kỷ 80. Nó bao hàm một loạt các kỹ thuật nhằm phát hiện ra các thông tin có giá trị tiềm ẩn trong các tập dữ liệu lớn. Về bản chất, KPDL liên quan đến việc phân tích các dữ liệu và sử dụng các kỹ thuật để tìm ra các mẫu có tính chính quy trong tập dữ liệu. Ở mức độ tổng quát, ta có thể coi mục đích chính của quá trình KPDL là mô tả và dự đoán [5].

Nói tóm lại: KPDL là một quá trình phát hiện những tri thức mới từ những dữ liệu đã thu thập được.

1.1.2. Quy trình KPDL

Quy trình KPDL bao gồm các bước như trong hình sau:

Bước 1. Làm sạch dữ liệu (Data cleaning)

Bước 2. Tích hợp dữ liệu (Data integration)

Bước 3. Trích chọn dữ liệu (Data selection)

Bước 4. Chuyển đổi dữ liệu (Data transformation)

Bước 5. Khai phá dữ liệu (Data mining).

Bước 6. Ước lượng mẫu (Knowledge evaluation)

Bước 7. Biểu diễn tri thức (Knowledge presentation)

1.1.3. Các kỹ thuật KPDL

- *Cây quyết định (Decision Tree)*

- *Phân lớp dữ liệu (Data Classification)*
- *Phân cụm dữ liệu (Data Clustering)*
- *Khai phá luật kết hợp (Association Rule)*
- *Hồi quy (Regression)*
- *Giải thuật di truyền (Genetic Algorithm)*
- *Mạng Nơron (Neural Network)*

1.1.4. Những ứng dụng của KPDL

1.2. KPDL BẰNG KỸ THUẬT CÂY QUYẾT ĐỊNH

1.2.1. Giới thiệu

Trong lĩnh vực KPDL, 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. DT dựa vào dãy các luật để dự đoán lớp các đối tượng. DT có cấu trúc biểu diễn dạng cây.

1.2.2. Cấu trúc DT

1.2.3. Các bước xây dựng DT

1.2.4. Ưu điểm của DT

1.2.5. Thuật toán ID3

a. Giới thiệu

b. Lựa chọn thuộc tính để kiểm tra

c. Giải thuật ID3

d. Ví dụ

e. Đánh giá thuật toán ID3

1.2.6. Thuật *toán* C4.5

a. Giới thiệu

Thuật toán C4.5 được phát hiện và công bố bởi Quinlan vào năm

1996. Thuật toán C4.5 là thuật toán được cải tiến từ thuật toán ID3, C4.5 giải quyết được hầu hết hạn chế của ID3. C4.5 sử dụng độ đo là Gain Ratio thực hiện phân lớp tập mẫu dữ liệu theo chiến lược ưu tiên theo chiều sâu và được dùng rộng rãi trong các ứng dụng phân lớp với lượng dữ liệu cỡ vài trăm nghìn bản ghi.

b. Thuật toán C4.5

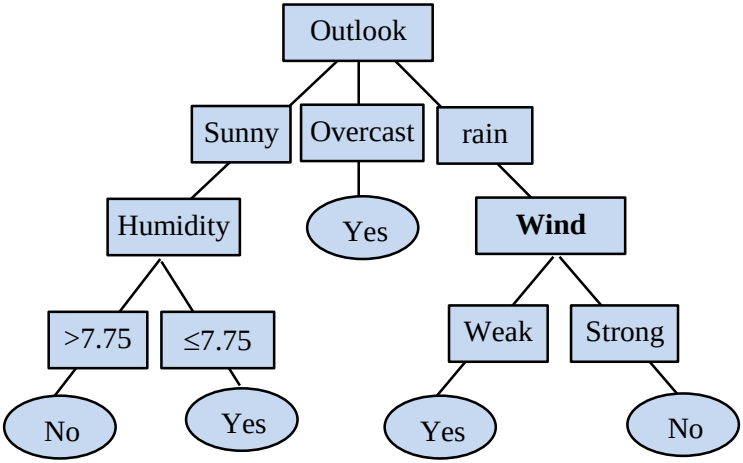
c. Một số cải tiến so với ID3

d. Ví dụ thuật toán C4.5

Bảng 1.2. Ví dụ thuật toán C4.5 dữ liệu chơi golf

Day	**Outlook**	**Temp**	**Humidity**	**Wind**	**Play?**
1	Sunny	Hot	85	Weak	No
2	Sunny	Hot	90	Strong	No
3	Overcast	Hot	78	Weak	Yes
4	Rain	Mild	96	Weak	Yes
5	Rain	Cool	80	Weak	Yes
6	Rain	Cool	70	Strong	No
7	Overcast	Cool	65	Weak	Yes
8	Sunny	Mild	95	Weak	No
9	Sunny	Cold	70	Weak	Yes
10	Rain	Mild	80	Strong	Yes
11	Sunny	Mild	70	Strong	Yes
12	Overcast	Mild	90	Strong	Yes
13	Overcast	Hot	75	Weak	Yes
14	Rain	Mild	80	Strong	No

Cây quyết định được sinh ra từ dữ liệu trên



***H**ình 1.4. Cây quyết định chơi golf từ thuật toán C4.5*

Luật rút ra từ cây quyết định:

Luật 1: if (Outlook = Sunny) and (Humidity >77.5) then Play= No

Luật 2: if (Outlook = Sunny) and (Humidity ≤77.5) then Play= Yes

Luật 3: if (Outlook = Overcast) then Play= Yes

Luật 4: if (Outlook = Rain) and (Wind= Weak) then Play= Yes

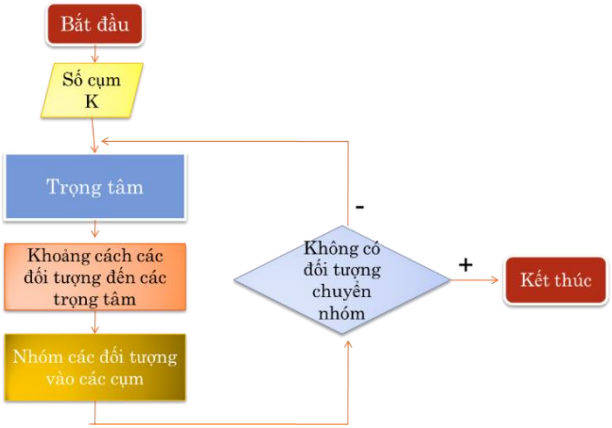
Luật 5: if (Outlook = Rain) and (Wind= Strong) then Play= No

1.3. KPDL BẰNG KỸ THUẬT PHÂN CỤM

1.3.1. Giới thiệu

1.3.2. Thuật toán K-Means

Thuật toán K-Means thực hiện qua các bước như *Hình 1.5.*



Hình 1.5. Sơ đồ thuật toán K-Means

KẾT LUẬN CHƯƠNG I

Chương này đã tập trung nghiên cứu tổng quan về KPDL; khái niệm KPDL; quá trình KPDL và các kỹ thuật KPDL. Trong đó đặc biệt là KPDL bằng kỹ thuật cây quyết định, nêu được giải thuật ID3 và giải thuật C4.5. Thuật toán C4.5 được cải tiến nhiều so với thuật toán ID3, C4.5 giải quyết hầu hết các bài toán mà ID3 chưa thể giải quyết được. C4.5 sử dụng RatioGain để xác định điểm chia tốt nhất. Ngoài ra, phần này còn trình bày thêm kỹ thuật phân cụm dữ liệu bằng thuật toán K-Means dựa trên khoảng cách của các đối tượng.

CHƯƠNG 2
NGHIÊN CỨU VÀ XỬ LÝ DỮ LIỆU VỀ BỆNH RỐI LOẠN TRẦM CẢM

Trong chương này, luận văn sẽ trình bày về đặc điểm tâm sinh lý của tuổi vị thành niên, khái niệm về bệnh RLTC, thực trạng bệnh RLTC hiện nay và đưa ra các đặc điểm lâm sàng chung cũng như đặc điểm lâm sàng của bệnh RLTC ở tuổi vị thành niên nói riêng và liệt kê các triệu chứng của bệnh RLTC. Trên cơ sở đó để thu thập xử lý số liệu thực tế về bệnh RLTC tại tỉnh Quảng Trị. Ngoài ra chương này cũng trình bày thêm các công cụ khai phá dữ liệu để từ đó xây dựng mô hình chẩn đoán bệnh RLTC.

2.1. ĐẶC ĐIỂM TÂM SINH LÝ CỦA TUỔI VỊ THÀNH NIÊN

2.2. KHÁI NIỆM VỀ BỆNH RỐI LOẠN TRẦM CẢM

Hiện nay, khái niệm RLTC của Tổ chức Y tế thế giới đang được áp dụng rộng rãi trong thực hành lâm sàng tâm thần học ở hầu hết các quốc gia trên thế giới. "Trầm cảm là một hội chứng bệnh lý biểu hiện đặc trưng bởi khí sắc trầm, mất mọi quan tâm thích thú, giảm năng lượng dẫn đến sự mệt mỏi và giảm hoạt động kèm theo một số triệu chứng phổ biến về rối loạn hành vi nhận thức, sự tập trung chú ý, tình dục, giấc ngủ và ăn uống; Các triệu chứng này phải kéo dài ít nhất 1 tuần [10].

2.3. THỰC TRẠNG BỆNH RỐI LOẠN TRẦM CẢM

2.4. ĐẶC ĐIỂM LÂM SÀNG CỦA BỆNH RLTC

2.4.1. Đặc điểm lâm sàng chung

2.4.1. Đặc điểm rối loạn trầm cảm ở trẻ vị thành niên

- Khí sắc trầm cảm: Trẻ có cảm giác buồn chán mơ hồ, không giải thích được nguyên cớ, hay cáu kỉnh.
- Giảm hứng thú trong học tập, công việc được giao phó.
- Tư duy: Khó tập trung chú ý, khó tiếp thu trong học tập.

- Các hoạt động xã hội: Trẻ thu mình cô lập không muốn giao tiếp hoặc tham gia các hoạt động đoàn thể, phàn nàn không có bạn thân hoặc khó chia sẻ với bạn.

- Rối loạn ăn: Thường nổi bật là cảm giác chán ăn, không có hứng thú trong ăn uống, mất cảm giác ngon miệng, hậu quả là trẻ bị giảm cân.

- Rối loạn giấc ngủ, trẻ ngủ nhiều hơn bình thường hoặc ngủ ít, trong rất nhiều trường hợp trẻ thường xuyên có ác mộng.

- Tự sát là triệu chứng rất quan trọng cần được quan tâm trong bệnh lý trầm cảm ở trẻ vị thành niên.

2.5. CHẨN ĐOÁN BỆNH RỐI LOẠN TRẦM CẢM

Chẩn đoán giai đoạn trầm cảm dựa vào những triệu chứng sau: [14].

- 3 triệu chứng đặc trưng:

+ Khí sắc trầm

+ Mất mọi quan tâm và thích thú

+ Giảm năng lượng dẫn đến mệt mỏi và giảm hoạt động

- 7 triệu chứng phổ biến:

+ Giảm sút sự tập trung và sự chú ý

+ Giảm sút tính tự trọng và lòng tin

+ Những ý tưởng bị tội và không xứng đáng

+ Nhìn vào tương lai ảm đạm và bi quan

+ Ý tưởng và hành vi tự huỷ hoại hoặc tự sát

+ Rối loạn giấc ngủ bệnh nhân thường mất ngủ vào cuối giấc.

+ Ăn ít ngon miệng.

2.6. CÁC YẾU TỐ LIÊN QUAN ĐẾN BỆNH RLTC

2.6.1. Yếu tố gia đình

2.6.2. Yếu tố học đường

2.6.3. Yếu tố xã hội

2.7. THU THẬP VÀ XỬ LÝ DỮ LIỆU NGHIÊN CỨU BỆNH RLTC

2.7.1. Thu thập dữ liệu

Việc thu thập dữ liệu tuỳ theo lĩnh vực ngành nghề. Mỗi ngành, dữ liệu lưu trữ có nguyên tắc riêng. Việc lấy dữ liệu cần thực hiện đúng theo quy định hay quy chế của cơ quan, đơn vị. Dữ liệu khai phá trong luận văn là hồ sơ bệnh án, do đó khi mượn hồ sơ bệnh án để nghiên cứu cũng phải có đơn xin mượn hồ sơ bệnh án và giấy giới thiệu của cơ quan quản lý.

Số liệu thu thập được 4000 hồ sơ bệnh án trong thời gian 3 năm (1/2013-1/2016) gồm các hồ sơ bệnh nhân thuộc đối tượng trẻ vị thành niên từ 12-18 tuổi, đã đến thăm khám tại Khoa tâm thần - bệnh viện đa khoa tỉnh Quảng Trị, một số phòng khám tư nhân và Phòng Y tế các trường THPT, THCS trên địa bàn tỉnh Quảng Trị. Các thông tin dữ liệu phân bổ rời rạc, có khoảng 1500 bệnh án điện tử còn lại là hồ sơ bệnh án lưu ở sổ sách. Do đó, phải được thu thập, gộp lại một cách thủ công vào file excel.

Dữ liệu được thu thập dựa trên bảng câu hỏi thu thập thông tin về bệnh rối loạn trầm cảm (xem phần *Phụ lục*), dưới sự tư vấn về chuyên môn, nghiệp vụ của bác sĩ tại khoa tâm thần bệnh viên Đa khoa tỉnh Quảng Trị nên dữ liệu có tính trung thực khách quan cao.

2.7.2. Xử lý dữ liệu

Bước 1. Làm đầy hoặc loại bỏ các bản ghi có trường dữ liệu bị thiếu và loại bỏ các dữ liệu bị trùng lặp.

- Sử dụng lệnh Data filter trong Excel để tìm ra các trường dữ liệu rỗng và tiến hành làm đầy dữ liệu như trường *Giới tính*

- Xoá những bản ghi có giá trị rỗng mà không thể làm đầy được.

- Sử dụng lệnh Remove Duplicates trong excel để xoá những bản ghi trùng lặp.

Bước 2. Tiến hành mã hóa tên các trường dữ liệu, giá trị của dữ liệu đảm bảo tính nhất quán.

- Tiến hành bỏ dấu
- Loại bỏ thông tin về họ tên bệnh nhân
- Năm sinh được chuyển thành Tuổi
- Trường kết quả: Yes = Mắc bệnh, No = Không mắc bệnh.

Sau khi thu thập và xử lý dữ liệu, luận văn tiến hành nghiên cứu các công cụ xây dựng mô hình KPDL.

2.8. CÁC CÔNG CỤ XÂY DỰNG MÔ HÌNH KPDL

2.8.1. Hệ quản trị CSDL SQL Server 2014

2.8.2. Công cụ xây dựng mô hình KPDL Business Intelligence

2.9. KPDL VỚI MSSQL SERVER 2014 ANALYSIS SERVICES

2.9.1. Môi trường phát triển ứng dụng

2.9.2. Các thuật toán KPDL trong MSSQL Server 2014

KẾT LUẬN CHƯƠNG 2

Chương 2 đã trình bày được đặc điểm tâm sinh lý của tuổi vị thành niên; nêu được khái niệm; đặc điểm lâm sàng của bệnh rối loạn trầm cảm, trình bày được các triệu chứng của bệnh RLTC, nêu được các yếu tố ảnh hưởng đến bệnh rối loạn trầm cảm và trình bày được các cộng cụ hỗ trợ xây dựng mô hình khai phá dữ liệu. Chương này đã giới thiệu cách thu thập và xử lý dữ liệu về bệnh trầm cảm của học sinh phổ thông. Đề tài đã thu thập và xử lý 4000 dòng dữ liệu từ Khoa tâm thần - bệnh viện đa khoa tỉnh Quảng Trị và một số phòng khám tư nhân, Phòng Y tế các trường THPT, THCS trên địa bàn tỉnh Quảng Trị để xây dựng cơ sở dữ liệu làm tiền giải quyết bài toán chẩn đoán bệnh RLTC.

CHƯƠNG 3
XÂY DỰNG HỆ THỐNG CHẨN ĐOÁN BỆNH RỐI LOẠN TRẦM CẢM DỰA TRÊN KHAI PHÁ DỮ LIỆU

Sau khi đã xây dựng CSDL trong chương 2, chương này trình bày kỹ thuật cây quyết định để KPDL và xây dựng mô hình chuẩn đoán bệnh RLTC. Sau đó, chương này sử dụng kỹ thuật phân cụm để đưa ra các đặc trưng cho từng cụm bệnh nhân của bệnh này. Cuối chương là phần cài đặt hệ thống thử nghiệm các tập luật đã khai phá được. Hệ thống này cho phép người dùng nhập dữ liệu về đặc điểm cá nhân cũng như các thông tin liên quan đến bệnh RLTC và đưa ra các dự báo về khả năng mắc chứng bệnh này.

3.1. XÂY DỰNG CSDL TRONG SQL SERVER

3.1.1. Mô tả dữ liệu

3.1.2. Bài toán chẩn đoán bệnh RLTC

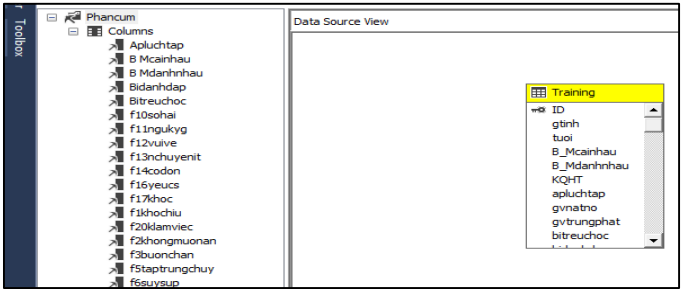
Để biết được một người có mắc phải bệnh RLTC hay không thì bệnh nhân phải đến bệnh viện hoặc các phòng khám tư để thăm khám lâm sàng, làm các xét nghiệm máu hoặc kiểm tra tuyến giáp và điền vào bảng câu hỏi để đánh giá tâm lý tại các cơ sở y tế có đầy đủ các trang thiết bị hỗ trợ. Tuy nhiên, vấn đề đặt ra là ta có thể chẩn đoán được kết quả tương tự như vậy nhưng không cần làm các xét nghiệm y khoa tốn kém thời gian và mất nhiều chi phí. Thay vào đó ta có thể dự đoán được bệnh thông qua các dữ liệu từ thông tin cá nhân, yếu tố gia đình, yếu tố trường học, việc học tập, mối quan hệ bạn bè, biểu hiện về sức khoẻ trong 7 ngày qua như rối loạn giấc ngủ, mệt mỏi, ăn không ngon miệng, cảm thấy buồn, khó chịu, không tập trung công việc, suy sụp tinh thần, ít trò chuyện….. được hay không?

3.1.3. Lựa chọn giải thuật giải quyết bài toán.

Ở đề tài này, chúng tôi không có tham vọng sẽ tìm hiểu hết tất cả các thuật toán nhưng sẽ nêu ra các thuật toán thông dụng để từ đó áp dụng vào đề tài. Kết quả mong muốn là xây dựng 2 mô hình dựa vào thuật toán cây quyết định DT, thuật toán phân cụm để có được các tập luật nhằm đưa ra quyết định chẩn đoán cho bài toán.

3.1.4. Xây dựng mô hình

Dữ liệu nguồn được dùng để xây dựng mô hình là 4000 bệnh nhân. Sau khi kết nối DataSource trong SQL Server Data Tools để đến SQL Server, thì Data Source View sẽ có khung nhìn các bảng CSDL với trường *Khóa là ID* như *Hình 3.1*.



Hình 3.1. Khung nhìn các bảng dữ liệu

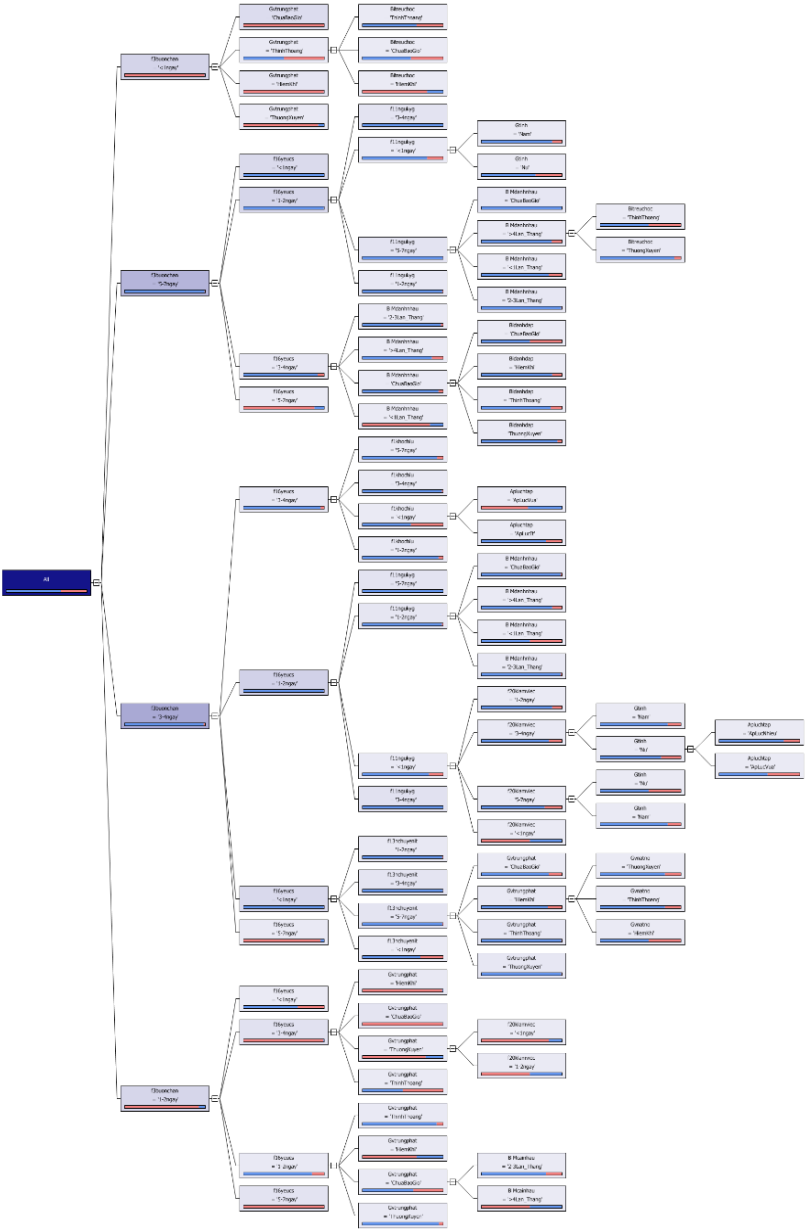
3.2. ỨNG DỤNG KỸ THUẬT DT ĐỂ XÂY DỰNG MÔ HÌNH CHẨN ĐOÁN BỆNH RLTC

3.2.1. Dữ liệu yêu cầu cho việc xây dựng mô hình DT

3.2.2. Những tham số được hỗ trợ trong thuật toán DT

3.2.3. Xây dựng mô hình KPDL và kết quả đạt được

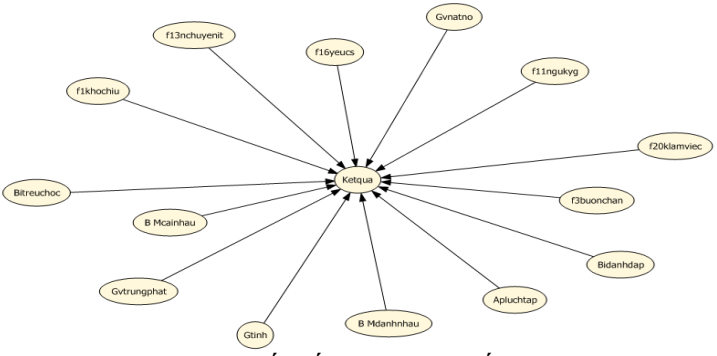
Luận văn này chọn thuật toán Decision Trees, cho kết quả mô hình như *Hình 3.4*.



Hình 3.4. Cây quyết định chuẩn đoán bệnh RLTC

3.2.4. Mức độ phụ thuộc của các dấu hiệu chẩn đoán bệnh RLTC

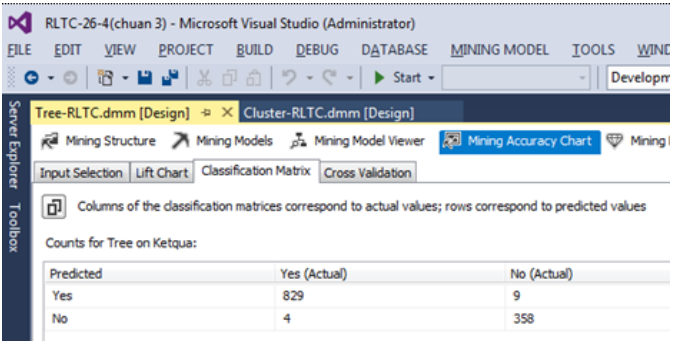
Từ kết quả của KPDL theo mô hình DT, hệ thống cũng cho thấy các yếu tố ảnh hưởng đến kết quả bênh RLTC được thể hiện trong *Hình 3.5.*



Hình 3.5. Các yếu tố ảnh hưởng đến bệnh RLTC

3.2.5. Đánh giá mô hình dự đoán

Với dữ liệu huấn luyện là 70% và dữ liệu test là 30%, mô hình đưa ra kết quả chẩn đoán với độ chính xác lên đến: 98.91%.



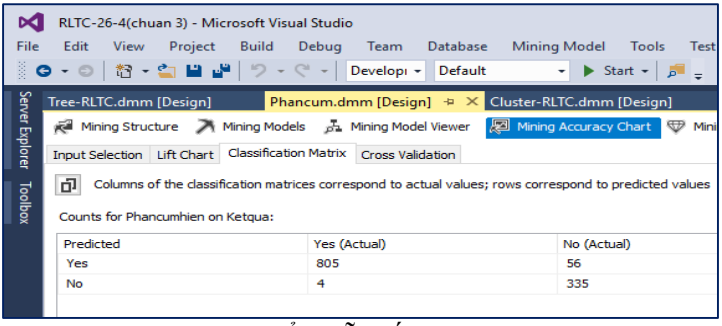
Hình 3.6. Ma trận biểu diễn khả năng chẩn đoán mô hình DT

3.3. ỨNG DỤNG KỸ THUẬT PHÂN CỤM ĐỂ PHÂN TÍCH CÁC ĐẶC TRƯNG CỦA BỆNH RLTC

3.3.1. Dữ liệu yêu cầu cho việc xây dựng mô hình phân cụm

3.3.2. Xây dựng mô hình phân cụm

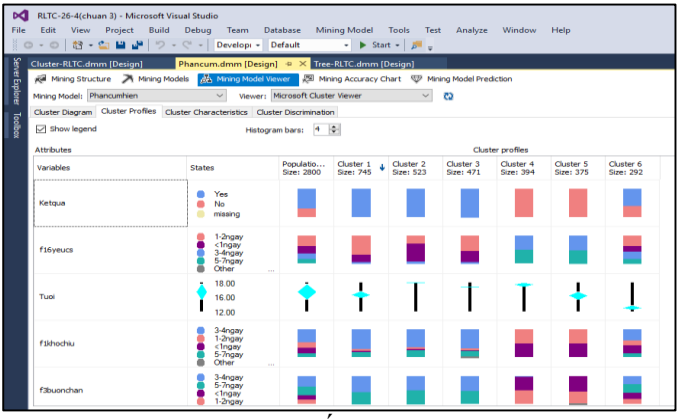
Đề tài sử dụng thuật toán K-mean để phân cụm, Kết quả cho thấy mô hình có độ chính xác đến 95%.



Hình 3.8. Ma trận biểu diễn kết quả mô hình phân cụm

3.3.3. Kết quả mô hình phân cụm của bệnh RLTC.

Mô hình cho ra kết quả như *Hình 3.9.*



Hình 3.9. Những đặc trưng chi tiết của bệnh RLTC trong từng cụm

Kết quả 6 cụm dữ liệu với các đặc trưng khác nhau như sau:

Cụm 1: Cụm này đa số là những BN có giới tính là Nữ, ở tuổi từ 15-16, trong một tuần BN có mắc phải các dấu hiệu sau: cảm thấy buồn, khó chịu, quá mệt mỏi không thể làm gì được từ 3-4 ngày, cảm thấy vui vẻ/hạnh phúc, yêu thích cuộc sống chỉ có 1-2 ngày,thường xuyên bị giáo viên trừng phạt, chứng kiến bố mẹ/ người thân cải nhau từ 3-4 lần/tháng. Cụm này đa số BN bị mắc bệnh RLTC.

Cụm 2. Cụm này đa số là những BN có giới tính là Nữ, ở tuổi từ 18, trong một tuần BN có mắc phải dấu hiệu sau: cảm thấy buồn, ít nói chuyện hơn bình thường, quá mệt mỏi không thể làm gì được từ 5-7 ngày, cảm thấy vui vẻ/hạnh phúc, yêu thích cuộc sống chỉ có <1ngày, khó chịu, không muốn ăn từ 3-4 ngày, thường xuyên bị giáo viên trừng phạt và chứng kiến bố mẹ hoặc người thân cải nhau từ >4 lần/tháng. Cụm này đa số BN bị mắc bệnh RLTC.

Cụm 3. Cụm này đa số là những BN có giới tính là Nữ, ở tuổi 17, trong một tuần BN có các dấu hiệu sau: cảm thấy buồn, khó chịu, hay khóc, quá mệt mỏi không thể bắt đầu những việc bình thường từ 3-4 ngày, cảm thấy vui vẻ/hạnh phúc, yêu thích cuộc sống từ 1-2 ngày, ngủ không yên giấc từ 5-7 ngày, thường xuyên bị bạn trêu chọc, chứng kiến bố mẹ hoặc người thân cải nhau từ 3-4 lần/tháng. Cụm này đa số BN bị mắc bệnh RLTC.

Cụm 4. Cụm này đa số là những người có giới tính là Nam, độ tuổi 18, trong một tuần BN có các dấu hiệu sau: thấy khó chịu, không thể bắt đầu những việc bình thường chỉ từ 1-2 ngày, cảm thấy vui vẻ/hạnh phúc, yêu thích cuộc sống từ 3-4 ngày, chưa bao giờ bị giáo viên trừng phạt, chứng kiến bố mẹ hoặc người thân cải nhau <1 lần/tháng. Cụm này đa số BN không bị mắc bệnh RLTC.

Cụm 5. Cụm này đa số là những người có giới tính là Nam, độ tuổi 15, trong một tuần BN có các dấu hiệu sau: thấy khó chịu, cảm thấy buồn chán, không muốn ăn hoặc ăn không ngon miệng chỉ có <

1 ngày, cảm thấy vui vẻ/hạnh phúc từ 5-7 ngày, yêu thích cuộc sống từ 3-4 ngày, không thể bắt đầu những việc bình thường từ 1-2 ngày, hiếm khi bị bạn trêu chọc, chưa bao giờ chứng kiến bố mẹ hoặc người thân cải nhau. Cụm này đa số BN không bị mắc bệnh RLTC.

Cụm 6. Cụm này đa số là những BN có giới tính là Nam, ở tuổi 12-15, trong một tuần BN có các dấu hiệu sau: thấy khó từ 3-4 ngày, cảm thấy vui vẻ/hạnh phúc từ 1-2 ngày, cảm thấy tràn đầy hy vọng vào tương lai từ 3-4 ngày, gặp khó khăn để tập trung chú ý vào những việc mình đang làm từ 1-2 ngày, áp lực học tập vừa phải, chưa bao giờ bị giáo viên trừng phạt, chứng kiến bố mẹ hoặc người thân cải nhau từ >4 lần/tháng. Cụm này đa số BN bị mắc bệnh RLTC.

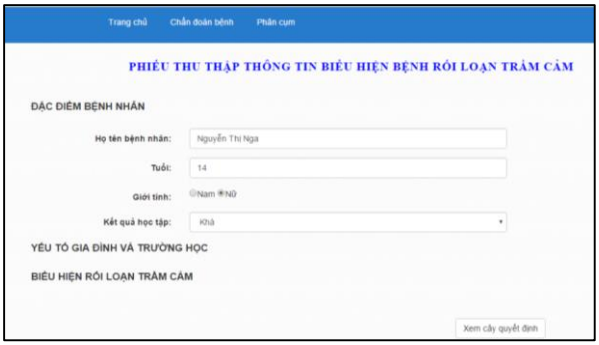
2.10. XÂY DỰNG HỆ THỐNG CHẨN ĐOÁN BỆNH RLTC.

Hệ thống được xây dựng bằng ngôn ngữ C# trên nền tảng .Net Framework. Bước đầu, hệ thống kết nối với CSDL và sử dụng các tập luật đã xây dựng ở chương 3 để đưa ra các dự đoán về khả năng mắc bệnh của BN. Người dùng có thể nhập thông tin của BN vào hệ thống và hệ thống sẽ đưa ra các kết quả chuẩn đoán kèm theo độ chính xác của dự đoán.

Giao diện sử dụng mô hình được xây dựng như sau:

Hình 3.10. Giao diện người dùng với hệ thống chẩn đoán bệnh RLTC

Trang nhập thông tin bệnh nhân



Hình 3.11. Giao diện thu thập thông tin người bệnh

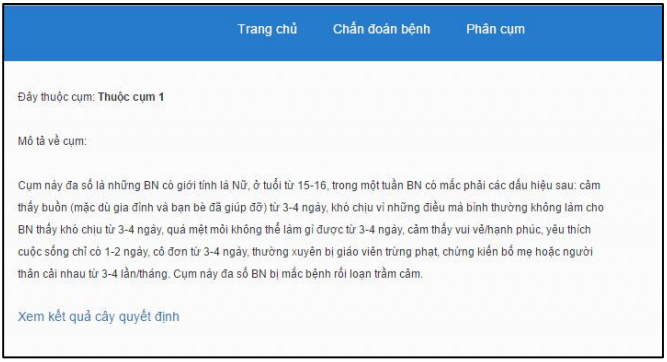
Hệ thống cho phép người dùng nhập vào những thông tin cần thiết cho việc chẩn đoán như: thông tin cá nhân, yếu tố gia đình, trường học và những biểu hiện rối loạn trầm cảm trong một tuần qua, sau đó hệ thống thực hiện chức năng chẩn đoán và cho ra kết quả chẩn đoán bệnh cho người dùng.

Hệ thống chẩn đoán bệnh rối loạn trầm cảm được xây dựng dựa vào mô hình cây quyết định cho kết quả như *Hình 3.12.*

Hình 3.12. Kết quả chẩn đoán bệnh từ mô hình cây quyết định

Với dữ liệu người dùng đã nhập ở *Hình 3.11* hệ thống chẩn đoán bệnh cho kết quả là Yes ứng với bệnh nhân bị bệnh rối loạn trầm cảm với độ chính xác là: 97.23%.

Ứng với thông tin thu thập như *Hình 3.11* hệ thống cũng đưa ra các đặc trưng cơ bản cho cụm này.



Hình 3.13. Kết quả mô hình phân cụm

KẾT LUẬN CHƯƠNG 3

Chương này đã sử dụng thuật toán Decision Trees để xây dựng mô hình chẩn đoán bệnh RLTC với kết quả dự đoán của mô hình là: 98.91% và đưa ra được mức độ phụ thuộc của các dấu hiệu chẩn đoán bệnh RLTC. Ngoài ra, còn ứng dụng được kỹ thuật phân cụm với thuật toán K-Means để xác định các đặc trưng cơ bản của bệnh RLTC với độ chính xác của mô hình là 95% và đã xây dựng được Demo cho phép người dùng nhập dữ liệu về đặc điểm cá nhân cũng như các thông tin liên quan đến bệnh RLTC và đưa ra các dự báo về khả năng mắc chứng bệnh này. Demo được sử dụng ngôn ngữ C# trên nền tảng .Net Framework với giao diện thân thiện dễ sử dụng.

KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN CỦA ĐỀ TÀI

Kết quả đạt được:

Qua quá trình nghiên cứu và tìm hiểu về các vấn đề liên quan tới khai phá dữ liệu, luận văn đã hoàn thành và đạt được một số kết quả như sau:

Luận văn đã trình bày được cơ sở lý thuyết liên quan đến hai kỹ thuật cây quyết định và kỹ thuật phân cụm. Nghiên cứu quy trình khai phá dữ liệu. Ngoài ra, luận văn cũng tìm hiểu các công cụ Microsoft SQL Server 2014 Analysis services và Business Intelligence để xây dựng mô hình KPDL.

Đối với bài toán chẩn đoán bệnh trầm cảm cho học sinh, luận văn đã xây dựng mô hình chẩn đoán bệnh dựa trên thuật toán cây quyết định. Với dữ liệu huấn luyện ban đầu, mô hình cho phép phân tích các yếu tố ảnh hưởng đến kết quả chẩn đoán bệnh, mức độ phụ thuộc của từng triệu chứng đối với bệnh RLTC. Bên cạnh đó luận văn ứng dụng kỹ thuật phân cụm để tìm ra những đặc trưng của bệnh RLTC.

Trên cơ sở tri thức phát hiện được từ mô hình chẩn đoán, luận văn xây dựng được hệ thống hiển thị mô hình cây quyết định, mô hình phân cụm đến người dùng cuối.

Với việc triển khai hệ thống thử nghiệm cho thấy có khả năng ứng dụng kết quả này trong việc chẩn đoán bệnh RLTC cho học sinh. Hệ thống chẩn đoán bệnh này sẽ góp phần hỗ trợ cho nhà quản lý, bác sĩ và người dân tự theo dõi, phát hiện sớm bệnh RLTC để điều trị kịp thời.

Bên cạnh những kết quả đạt được, để đưa mô hình dự đoán vào ứng dụng một cách hiệu quả hơn thì cần tiếp tục đầu tư thu thập dữ liệu BN nhiều hơn nữa. Triển khai dự đoán, kiểm chứng thực tế và

đánh giá kết quả một cách thường xuyên. Bản thân tôi nhận thấy đây là hướng tiếp cận đúng đắn và có tính thực tiễn cao

Hướng phát triển

Hạn chế của đề tài là mẫu dữ liệu sử dụng cho mô hình chưa nhiều, dữ liệu đầu vào cho mô hình mới chỉ được thu thập từ các hồ sơ bệnh án ở đối tượng trẻ vị thành niên thuộc tỉnh Quảng Trị. Vì vậy, mô hình chỉ chẩn đoán bệnh rối loại trầm cảm cho học sinh từ 12-18 tuổi. Trong thời gian tới, sẽ tiếp tục thu thập thêm dữ liệu để hoàn thiện mô hình, đồng thời nghiên cứu phát triển mô hình để xây dựng hệ thống chẩn đoán cho nhiều đối tượng khác nhau và tiếp tục nghiên cứu kỹ thuật cây quyết định để xây dựng mô hình chẩn đoán thêm một số bệnh khác…