HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

TRẦN XUÂN OANH

XÂY DỰNG HỆ THỐNG HỖ TRỢ RA QUYẾT ĐỊNH TRONG TƯ VẤN CHỌN NGÀNH NGHỀ CHO HỌC SINH TRUNG HOC PHỔ THÔNG

Chuyên ngành: KHOA HỌC MÁY TÍNH

Mã số: 8.48.01.01

TÓM TẮT LUẬN VĂN THẠC SĨ

HÀ NỘI - NĂM 2022

Luận văn được hoàn thành tại:

HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

Người hướng dẫn khoa học: PGS. TS. LÊ HỮU LẬP

Phản biện 1: PGS.TS. HOÀNG XUÂN DẬU

Phản biện 2: PGS.TS. PHẠM THANH GIANG

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: 14 giờ 00 ngày 20 tháng 12 năm 2022

Có thể tìm hiểu luận văn tại:

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

MỞ ĐẦU

1. Lý do chọn đề tài

Trong đời sống, đối với mỗi người, nghề nghiệp là điều có ý nghĩa vô cùng quan trọng. Do đó, trong thời điểm hiện tại, giáo dục hướng nghiệp ngày càng đóng vai trò to lớn trong việc giúp các học sinh có nhận thức đúng đắn về nghề nghiệp, qua đó, có được sự lựa chọn nghề nghiệp phù hợp với năng lực bản thân, đồng thời đáp ứng nhu cầu bức thiết của xã hội về nhân lực, góp phần sử dụng và phân luồng nguồn lao động hợp lý, giúp kinh tế, xã hội phát triển bền vững. Trong Văn kiện của Đảng có viết: "Coi trọng công tác hướng nghiệp và phân luồng học sinh trung học, chuẩn bị cho thanh niên, thiếu niên đi vào lao động nghề nghiệp phù hợp với sự chuyển dịch cơ cấu kinh tế trong cả nước và từng địa phương". Trong thời gian qua, hoạt động trong công tác giáo dục hướng nghiệp tại các trường trung học phổ thông còn tồn tại nhiều khiếm khuyết. Các chủ điểm nội dung trong giáo dục hướng nghiệp tại nhà trường vẫn còn thiếu sót: phiến diện, bản chất của các nghề chưa được làm rõ, những yêu cầu về năng lực, phẩm chất, của cá nhân chưa được xác định phù hợp với nghề được lựa chọn. Về mặt hình thức, cách truyền đạt còn thô cứng, nghèo nàn, mang tính hình thức, phổ cập, đại trà, các đối tượng học sinh thì chưa được phân hóa rõ ràng.

Hoàn cảnh khách quan đang trở nên ngày càng đa dạng và phức tạp, công nghệ thông tin cũng đang trên đà phát triển không ngừng. Trong bối cảnh đó, việc sử dụng các hệ thống trợ giúp, nhờ đó, sẽ làm thay đổi bộ mặt cũng như phương tiện giáo dục hướng nghiệp. Hệ trợ giúp quyết định - Decision Support System (DSS) do vậy, trở thành một công cụ hữu hiệu trong việc trợ giúp các em học sinh trung học phổ thông xác định rõ ràng nghề nghiệp của mình trong tương lai.

Chính bởi lẽ đó, là một người thầy đã có nhiều năm trực tiếp giảng dạy hàng ngày trong trường Trung học phổ thông (THPT), tôi quyết định chọn đề tài**" Xây dựng hệ thống hỗ trợ ra quyết định trong tư vấn chọn ngành nghề cho học sinh trung học phổ thông"** nhằm thử nghiệm công cụ hỗ trợ trong việc lựa chọn nghề nghiệp cho các học sinh ngay từ khi còn ngồi trên ghế nhà trường THPT.

2. Mục đích và nhiệm vụ nghiên cứu

Mục đích nghiên cứu: Sử dụng các công cụ trong khai phá dữ liệu để xây dựng hệ thống trợ giúp tư vấn hướng nghiệp cho học sinh trung học phổ thông. Áp dụng thử nghiệm cho một vài trường trung học phổ thông thuộc thành phố Hà Nội.

- **Nhiệm vụ nghiên cứu:**

Đưa ra một giải pháp từ việc phân loại dữ liệu trên các phiếu khảo sát thông tin lựa chọn ngành học, đến việc tiến hành khai thác xử lý chúng để đưa ra các tri thức cần thiết. Các tri thức này được tối ưu hóa và đem vào sử dụng một cách hiệu quả trong việc tư vấn chọn ngành học cho học sinh.

3. Đối tượng và phạm vi nghiên cứu

*** Đối tượng:** Nhu cầu tư vấn hướng nghiệp của học sinh trung học phổ thông.

*** Phạm vi nghiên cứu:** Đề tài tập trung nghiên cứu xây dựng hệ hỗ trợ giúp tư vấn hướng nghiệp cho học sinh trung học phổ thông dựa trên khai phá dữ liệu.

4. Phương pháp nghiên cứu

Luận văn sử dụng những phương pháp nghiên cứu sau đây:

- ➢ Phương pháp nghiên cứu tài liệu.
- ➢ Phương pháp điều tra và thu thập thông tin bằng bảng hỏi.

➢ Phương pháp thống kê toán học qua phiếu excel bảng điểm.

5. Cấu trúc của luận văn

Nội dung luận văn gồm 3 chương chính:

Chương 1: Hệ hỗ trợ giúp ra quyết định

Chương 2: Xây dựng hệ hỗ trợ tư vấn hướng nghiệp cho học sinh THPT.

Chương 3: Thiết lập hệ thống và thử nghiệm.

CHƯƠNG 1: CƠ SỞ LÝ THUYẾT

1.1 Tổng quan về hệ thống trợ giúp ra quyết định

Hệ thống hỗ trợ đưa ra quyết định gồm 5 thành phần chính:

- Người dùng.
- Giao diện tương tác với hệ thống.
- Mô hình cây quyết định.
- Cơ sở dữ liệu.
- Hệ thống điều phối.

Người dùng là người sẽ nhập các thông tin cần thiết và cần nhận được kết quả từ hệ thống hỗ trợ.

Giao diện tương tác với hệ thống là cửa sổ màn hình hiện lên cho phép người dùng nhấp chuột, nhập dữ liệu và thấy các thông tin cần thiết. Giao diện này bao gồm 2 vùng chính: vùng nhập dữ liệu và vùng hiển thị kết quả.

Mô hình cây quyết định là mô hình được sinh ra sau khi dữ liệu học máy được làm sạch và đưa vào Weka.

Cơ sở dữ liệu là cấu trúc và các bản ghi được lưu trữ để đưa vào xây dựng cây quyết định. Dữ liệu trong cơ sở dữ liệu này có thể được bổ sung thông qua các dữ liệu người dùng nhập vào để cải thiện mô hình cây quyết định.

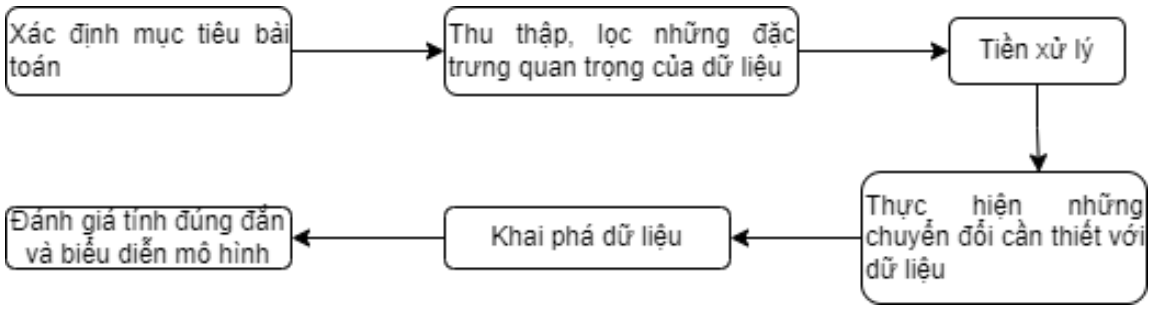
Hệ thống điều phối là hệ thống kết nối, điều phối để tương tác bổ sung bản ghi vào cơ sở dữ liệu, đưa dữ liệu vào để xây dựng mô hình cây quyết định, ghi nhận thông tin từ giao diện tương tác và áp dụng mô hình cây quyết định với thông tin ghi nhận được đó để cho ra kết quả sau cùng.

1.2 Khai phá dữ liệu

Dữ liệu thường chứa rất nhiều thông tin có giá trị, bổ ích đối với qui trình ra quyết định, tuy vậy với khối lượng dữ liệu rất lớn thì không thể phân tích bằng các phương pháp thủ công đồng thời cũng không thể dùng để truy vấn truyền thống (SQL) bởi vì thực ra còn nhiều kiểu truy vấn mà chúng ta quan tâm tới nó rất khó để miêu tả hay thực hiện miêu tả bằng ngôn ngữ vấn tin, ví dụ như: tìm tất cả các bản ghi nghi là gian lận, tìm tất cả các văn bản gần giống như văn bản A, không có quá nhiều thông tin trong các trường của CSDL…Do vậy, khai phá dữ liệu trở thành giải pháp hữu hiệu nhằm giải quyết vấn đề quá tải dữ liệu trong trong kỷ nguyên số hóa.

Có thể nói khai phá dữ liệu là quá trình trích xuất và khám phá các mẫu trong tập dữ liệu lớn liên quan đến các phương pháp kết hợp giữa học máy, thống kê và hệ thống cơ sở dữ liệu.

Hiện nay, khai phá dữ liệu và phát hiện tri thức được ứng dụng và triển khai trong thực tế, đem lại hiệu quả cao cho sản xuất kinh doanh và nghiên cứu khoa học.



Hình 1.1: Các bước trong quy trình khai phá dữ liệu

Quy trình khai phá dữ liệu bao gồm các bước sau:

- Xác định mục tiêu bài toán

- Thu thập, bổ sung các dữ liệu có liên quan đến mục tiêu bài toán thành một kho dữ liệu đầy đủ, với những thuộc tính quan trọng và cần thiết

- Thực hiện tiền xử lý dữ liệu trước khi đưa vào huấn luyện. Có hai cách thức tiền xử lý là rút gọn dữ liệu (khái quát hóa, tổng hợp, giảm số chiều dữ liệu, nén, rời rạc hóa hoặc giảm số lượng bản ghi đưa vào) và làm sạch dữ liệu (xử lý nhiễu, sai, thiếu dữ liệu)

- Thực hiện các chuyển đổi cần thiết như chuyển đổi kiểu, rời rạc hóa, phân giá trị thành nhóm, chuẩn hóa dữ liệu.

- Khai phá dữ liệu sử dụng chiến thuật, thuật toán phù hợp để đưa ra mô hình sau huấn luyện

- Đánh giá mô hình thông qua các kỹ thuật kiểm thử

Các kỹ thuật khai phá dữ liệu thường gặp là: phân lớp, phân cụm, dự đoán, hồi quy và mạng nơ-ron.

- Phân lớp

Thuật toán huấn luyện phân lớp sử dụng các mẫu được phân loại trước để xác định tập hợp các tham số cần thiết để phân loại thích hợp. Sau đó, thuật toán mã hóa các tham số này thành một mô hình được gọi là bộ phân loại (classifier).

- Phân cụm

Phân cụm có thể nói là xác định các lớp tương tự của các đối tượng. Bằng cách sử dụng các kỹ thuật phân cụm, ta có thể xác định thêm các vùng phân bố dày đặc hay thưa thớt trong không gian đối tượng và có thể khám phá mô hình phân phối tổng thể cũng như mối tương quan giữa các thuộc tính dữ liệu.

- Hồi quy

Kỹ thuật hồi quy thường được dùng để dự đoán. Phân tích hồi quy được sử dụng để thiết lập mô hình về mối quan hệ giữa một hoặc nhiều biến độc lập và biến phụ thuộc. Trong khai phá dữ liệu các biến độc lập là các thuộc tính đã biết và các biến phụ thuộc là những thuộc tính ta muốn dự đoán.

- **Luật kết hợp**

Trong các tập dữ liệu, mối liên hệ giữa các dữ liệu có thể được biểu diễn dưới dạng quy tắc, quan hệ nhân quả.

Các loại luật kết hợp: luật kết hợp đa cấp, luật kết hợp nhiều chiều, luật kết hợp định lượng.

- **Mạng nơron**

Mạng nơ-ron là một tập hợp các đơn vị đầu vào/đầu ra được kết nối và mỗi kết nối có trọng số đi kèm.

Trong giai đoạn huấn luyện, mạng học bằng cách điều chỉnh trọng số để có thể dự đoán đúng nhãn của các mẫu đầu vào.

Kết luận chương 1:

Khai phá dữ liệu là quá trình đi tìm tri thức được ẩn đằng sau các bộ dữ liệu, thường là dữ liệu lớn. Đặc biệt, áp dụng khai phá dữ liệu trong việc hỗ trợ quá trình định hướng nghề nghiệp và tuyển sinh đem lại lợi ích to lớn cho cả phía nhà trường và phụ huynh, học sinh. Trong chương I ta đã tìm hiểu các khái niệm cơ bản và các bước trong quá trình khai phá dữ liệu. Ta cũng đã xem xét các kỹ thuật khai phá dữ liệu phổ biến. Kỹ thuật phân lớp bằng mô hình dựng cây quyết định tỏ ra hiệu quả trong bài toán định hướng nghề nghiệp và tuyển sinh. Ta sẽ cùng phân tích kỹ hơn trong chương 2.

CHƯƠNG 2: XÂY DỰNG HỆ HỖ TRỢ TƯ VẤN HƯỚNG NGHIỆP CHO HỌC SINH THPT

Chương 2 sẽ tập trung vào phân tích cơ sở lý thuyết để xây dựng hệ thống hỗ trợ tư vấn hướng nghiệp bao gồm cơ sở lý luận Holland và thuật toán phân lớp bằng mô hình dựng cây quyết định. Ta cũng sẽ đi sâu vào thuật toán dựng cây quyết định Iterative Dichtomiser 3 (ID3).

2.1 Cơ sở lý luận John Holland

Lý thuyết mật mã Holland là thuộc lý thuyết về các đặc điểm cá nhân và nghề nghiệp do nhà tâm lý học John Holland (1919-2008) xây dựng. Ông được biết đến với công trình nghiên cứu về lý thuyết lựa chọn nghề nghiệp. Lý thuyết này được đánh giá là thực tế nhất, có nhiều cơ sở nghiên cứu nhất, được các nhà tư vấn nghề nghiệp ở Hoa Kỳ và nước ngoài sử dụng nhiều nhất.

Theo lý thuyết của Holland, có 6 môi trường làm việc tương ứng với 6 loại tính cách:

- Nhóm Kỹ thuật
- Nhóm Nghiên cứu khoa học
- Đoàn Nghệ thuật
- Nhóm Xã hội
- Nhóm Quản lý
- Nhóm Chuyên viên nghiệp vụ

Hình 2.1: 6 nhóm môi trường làm việc

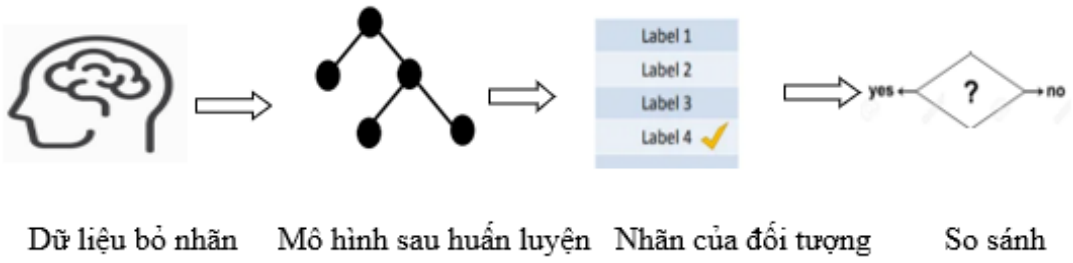
Hệ thống hỗ trợ tư vấn hướng nghiệp cho học sinh THPT áp dụng các kết quả từ lý thuyết mật mã Holland để gợi ý nhóm ngành phù hợp cho từng đối tượng giúp các em hiểu được điểm mạnh của mình và bớt bối rối khi đưa ra quyết định lựa chọn con đường đúng đắn. Nhờ đó, không phải cố gắng bằng mọi giá để vào được một trường cao đẳng hoặc đại học, bất kể chuyên ngành đó có phù hợp hay không. Đồng thời, giúp học sinh có cơ hội cao hơn trong các kỳ thi tuyển sinh.

2.2 Phân lớp dữ liệu với cây quyết định

Hình 2.2: Tạo mô hình huấn luyện

Việc phân lớp dữ liệu gồm có 2 bước:

Bước 1: Tạo mô hình từ dữ liệu huấn luyện. Kết quả của bước tạo mô hình là một mô hình toán học, cây quyết định hoặc tập hợp các luật để phân loại dữ



Hình 2.3 Ứng dụng mô hình phân lớp vào bài toán

liệu.

Bước 2: Ứng dụng mô hình huấn luyện vào bài toán. Hệ thống sẽ phân loại, gán nhãn cho dữ liệu mới dựa trên mô hình tạo ra ở bước 1. Tính đúng đắn của mô hình càng cao khi tỷ lệ dữ liệu phân lớp đúng càng cao.

2.3. Cây quyết định:

Cây quyết định (decision tree) là dạng cấu trúc biểu diễn tri thức dưới dạng cây nhằm mục đích phân chia đối tượng thành những lớp có nhãn.

Nếu sự lựa chọn các thuộc tính là hợp lý thì ta luôn tạo được cây quyết định phân loại đúng các đối tượng trong tập huấn luyện và thường tồn tại nhiều cây quyết định đúng. Tuy nhiên, điều quan trọng cây quyết định cần "đúng" không chỉ với các đối tượng trong tập huấn luyện mà còn đối với các đối tượng không nằm trong tập huấn luyện. Do đó, cây quyết định cần nắm bắt được những mối liên quan giữa các đối tượng trong một phân lớp và giá trị của chúng. Một cây quyết định đúng thường không quá phức tạp và mối liên hệ giữa nhãn với giá trị thuộc tính của đối tượng là có thể giải thích được.

```
Procedure build_tree (tập_mẫu, tập_thuộc_tính)
begin
    if  mọi mẫu trong tập_mẫu đều nằm trong cùng một phân lớp then
        return nút lá được gán nhãn là phân lớp đó
    else if tập_thuộc_tính rỗng then
        return nút lá được gán nhãn bởi tuyển chọn của tất cả các lớp
        trong tập_mẫu
    else
        begin
            chọn một thuộc tính T, lấy T làm nút gốc cho cây hiện tại;
            xóa nút T ra khỏi tập_thuộc_tính;
            với mỗi giá trị G của T;
            begin
                tạo nhánh mới cho cây gán nhãn G;
                Đặt vào phân vùng V các ví dụ trong tập_mẫu có
            giá trị G tại thuộc tính T;
```

2.4 Thuật toán Iterative Dichotomiser 3 (ID3)

Thuật toán ID3 được mô tả trong đoạn mã giả dưới đây.

```
                    Gọi build_tree (phân vùng V, tập_thuộc_tính), gắn
                    kết quả vào nhánh V
                end
            end
        end
```

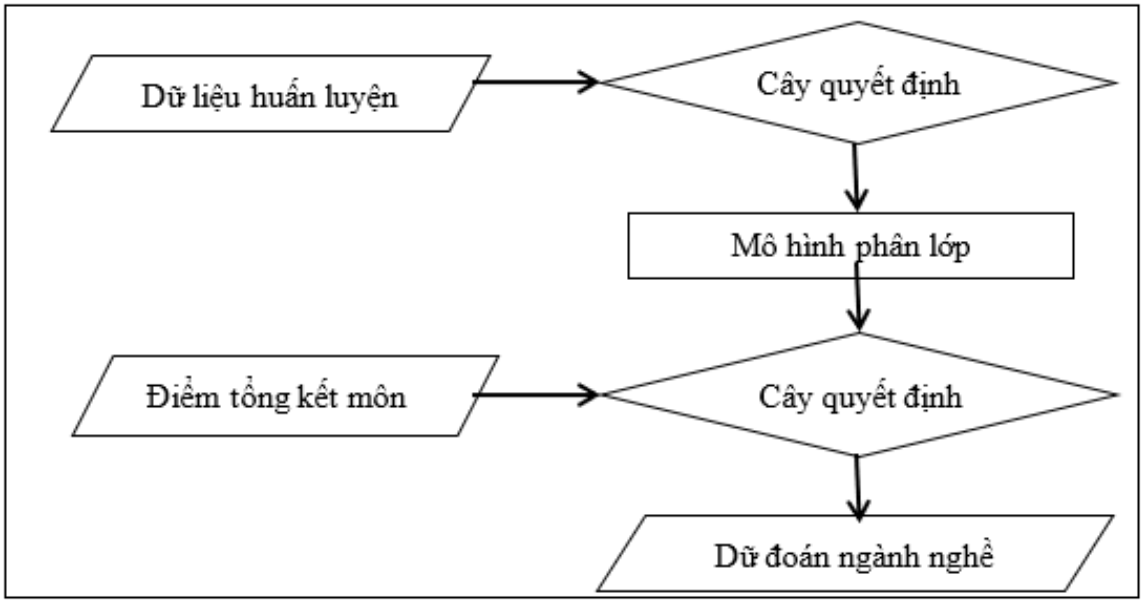
Hình 2.4 Mô tả thuật toán ID3

2.5 Xây dựng hệ thống hỗ trợ dựa trên cây quyết định

Về yêu cầu chức năng, hệ thống cần đáp ứng đầu vào, đầu ra như sau:

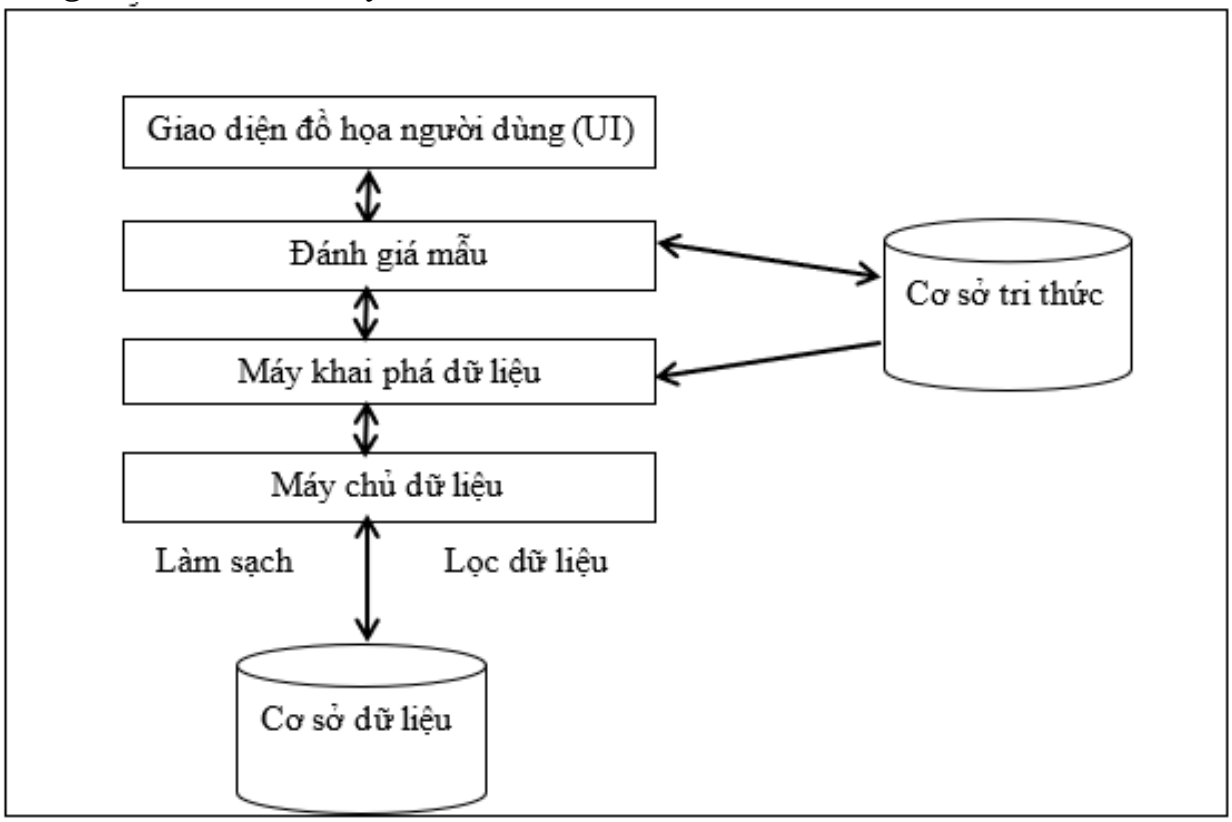
- ➢ Đầu vào: Kết quả học tập: điểm toán, văn, ngoại ngữ, lý, hóa, sinh, sử, địa.
- ➢ Đầu ra: Kết quả dự đoán 1 nhóm Holland và 3 ngành nghề thuộc nhóm và các trường cùng khối thi tương ứng học viên nên đăng ký xét tuyển.

Yêu cầu chức năng được mô tả chi tiết trong lưu đồ dưới đây:



Hình 2.5 Lưu đồ mô tả chức năng hệ thống hỗ trợ tư vấn hướng nghiệp

Về yêu cầu thiết kế, hệ thống cần đảm bảo có đầy đủ các thành phần trong kiến trúc dưới đây:



Hình 2.6 Yêu cầu kiến trúc hệ thống hỗ trợ tư vấn

Bộ công cụ Weka là một tập hợp các thuật toán học máy và các công cụ tiền xử lý dữ liệu bao gồm hầu như tất cả các thuật toán máy học cơ bản. Nó được thiết kế để ta có thể nhanh chóng thử các phương pháp hiện có trên tập dữ liệu mới theo những cách linh hoạt. Weka cung cấp hỗ trợ rộng rãi cho toàn bộ quá trình khai thác dữ liệu thử nghiệm, bao gồm chuẩn bị dữ liệu đầu vào, đánh giá sơ đồ học tập một cách thống kê,trực quan hóa dữ liệu đầu vào và kết quả học tập.

Kết luận chương II

Để xây dựng hệ thống hỗ trợ ra quyết định trong tư vấn chọn ngành nghề, lý thuyết John Holland về 6 nhóm ngành nghề cơ bản là lý thuyết điển hình để phân loại ngành nghề và kết quả tư vấn. Cây quyết định là cấu trúc cần xây dựng, đóng vai trò quyết định trong hệ thống phân loại ngành nghề. Trong bước xây dựng cây quyết định, thuật toán Iterative Dichotomiser 3 (ID3) được chọn để xây dựng. Hệ thống được đánh giá thông qua các chỉ số: Entropy, độ lợi thông tin (information gain) và tỷ suất lợi thông tin (information gain ratio).

CHƯƠNG III. THIẾT LẬP HỆ THỐNG TƯ VẤN VÀ THỬ NGHIỆM

3.1 Xác định mục tiêu của hệ thống và vấn đề cần giải quyết

Sử dụng phần mềm quản lý điểm sẽ hỗ trợ người dùng in ra các file excel bằng cách tổng hợp các bảng điểm dựa trên môn học, bảng tổng kết các môn học theo mỗi kỳ của học sinh các lớp như bảng sau:

STT	Họ tên	Toán	Vật lí	Hóa học	Sinh học	Tin học	Ngữ Văn	Lịch sử	Địa lí	Ngoại ngữ 1	Công nghệ	Năng khiếu	GDQP-AN	Thể dục	GDCD	TBM HK	Học lực	Hạnh kiểm
1	Nguyễn Thành An	8.9	9.5	8.6	8.3	7.6	6.5	7.8	8.8	6.8	8.5	8.2	8.4	Đ	8.9	8.2	G	T
2	Đinh Thị Mai Anh	9.3	9.5	9.3	7.8	8.2	6.9	7.7	9.1	7.1	8.4	8.3	7.4	Đ	9	8.3	G	T
3	Mai Thị Thục Anh	9.3	9.8	9.3	8.2	8.6	8.3	8	9.3	6.9	8.5	8.6	7.7	Đ	9.2	8.6	G	T
4	Kim Đức Dũng	8.8	9.7	7.6	8.2	7.9	6.7	8.1	8.8	8.2	8.7	8.2	6.9	Đ	8.9	8.2	G	T
5	Nguyễn Đình Dũng	9	8.2	9.5	9.3	7.9	6.6	7.8	8.9	6.8	8.1	8.2	7.2	Đ	9	8.2	G	T
6	Phùng An Duy	8.8	9.2	8.4	8.1	8	7.2	8.2	9	7.6	8.2	8.3	8	Đ	9.1	8.3	G	T
7	Trần Thị Thu Hà	8.9	9.7	9.5	8.6	8.4	7.4	7.7	9.1	7.2	8.5	8.5	7.3	Đ	9.1	8.5	G	T
8	Kim Thị Hồng Hạnh	8.7	9.2	8.9	7.5	8.5	7.1	7.7	8.9	6.8	8.7	8.2	7.2	Đ	8.7	8.2	G	T
9	Nguyễn Thị Thu Hằng	8.6	9.5	8.9	8.3	8.4	6.8	8	9.1	6.6	8.8	8.3	7.6	Đ	9	8.3	G	T
10	Đinh Thị Mai Hoa	8.7	8.1	7.8	7.5	8.1	8.1	7.8	8.9	9.2	8.2	8.2	7.1	Đ	8.5	8.2	G	T
11	Đinh Thị Ánh Hồng	8.8	9.4	9.6	8.5	8.3	7.5	8	8.9	6.9	8.6	8.4	7.8	Đ	9	8.4	G	T
12	Đặng Thị Mỹ Huệ	8.4	8.3	7.5	8.2	8	6.9	7.8	9	6.9	8.9	8	7.1	Đ	8.6	8	G	T
13	Nguyễn Ngọc Huyền	9.3	9.6	9.5	8.5	8.6	7.3	8	8.7	8.1	8.5	8.5	7.3	Đ	9.1	8.5	G	T
14	Nguyễn Tuyết Thương Huyền	9.3	9.7	9.1	8.5	8.2	7.4	8	8.9	7.5	8.3	8.4	7	Đ	9	8.4	G	T
15	Phạm Thị Ngọc Huyền	9.3	9.6	9.2	8.3	8.6	7.2	8	8.9	7.5	8.6	8.5	7.6	Đ	9.1	8.5	G	T
16	Nguyễn Thu Hương	8.9	9.2	9.3	7.9	8.2	6.9	7.7	9.2	6.8	8.5	8.2	7.1	Đ	8.9	8.2	G	T
17	Nguyễn Duy Khánh	9	9.3	9.3	7.9	8.3	6.4	7.7	8.8	6.2	8.8	8.1	7	Đ	8.8	8.1	K	T
18	Lê Thị Thùy Linh	8.9	9	7.9	8	8.1	6.6	8.1	9.4	7.3	8.1	8.2	7.5	Đ	8.9	8.2	G	T
19	Nguyễn Hải Linh	8.8	9.4	8.9	8.3	8.6	7.1	7.8	8.8	7.9	8.7	8.5	8.4	Đ	9.2	8.5	G	T
20	Nguyễn Hữu Linh	8.7	8.9	7.7	7.8	8.1	5.7	7.8	9.1	8	8.3	8	7	Đ	8.6	8	K	T

Bảng 3.1. Bảng điểm tổng kết

Dựa vào bảng điểm trên chúng ta sẽ rút ra được phương pháp chọn ngành nghề, trường học phù hợp với từng học sinh. Có 2 nguyên nhân ảnh hưởng đến việc chọn ngành nghề, trường học của học sinh như: nguyên nhân chủ quan và nguyên nhân khách quan. Trong đó nguyên nhân chủ quan của học sinh sẽ có ảnh hưởng phần lớn tới những quyết định then chốt này. Vì vậy mục tiêu của sự hỗ trợ tư vấn hướng nghiệp giúp tác giả đưa ra các phương pháp giải quyết vấn đề chọn ngành nghề, chọn trường phù hợp tất cả đều dựa trên nguyên nhân chủ quan của học sinh. Phương pháp này tập trung giải đáp từng vấn đề chính sau:

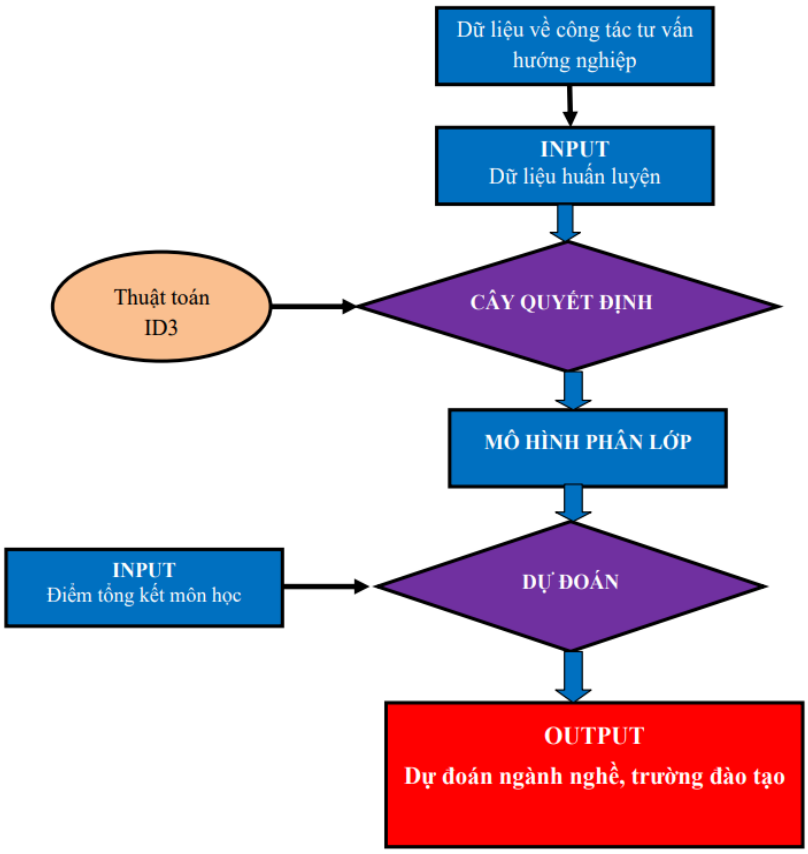
- Mối quan hệ giữa điểm trung bình môn học có ảnh hưởng đến quyết định cho sự lựa chọn của việc không tham gia kì thi xét tuyển đại học.

- Mối quan hệ giữa điểm trung bình môn học có ảnh hưởng đến quyết định cho sự lựa chọn khối thi, ngành thi, trường thi.

Mô tả hệ thống

Input: Thành tích học tập ở các môn: Điểm toán, lý, hóa, sinh, văn, sử, địa, ngoại ngữ.

Output: Hoạch định ngành nghề trong tương lai và 3-5 trường mà người tham gia sẽ đăng ký kỳ thi hoặc xét tuyển sau khi tốt nghiệp.



Hình 3.1 Mô hình hệ hỗ trợ tư vấn hướng nghiệp

3.2 Quy trình giải quyết bài toán

Có rất nhiều nguyên nhân đang ảnh hưởng đến việc chọn lọc ngành học của học sinh, tuy nhiên chúng ta chỉ chú trọng phân tích những yếu tố chính sau đây:

Thành tích học tập của học sinh: Điểm các môn Toán, Lý, Hóa, Sinh, Ngoại ngữ, Văn, Sử, Địa.

Ngành nghề: Danh sách các ngành nghề đào tạo trên toàn quốc.

Khối thi: Danh sách khối thi.

Trường: Danh sách các trường đại học, cao đẳng.

Trong dữ liệu ngay lúc đầu đang nắm giữ các thông tin hướng nghiệp như: lý lịch học sinh, kết quả học tập 3 năm của hơn 20.000 học sinh lớp 12 của các trường THPT trên địa bàn thành phố Hà Nội, danh mục các ngành nghề đào tạo của cả nước, Danh sách 481 trường Đại học, cao đẳng trên toàn quốc…, hệ thống trích lọc các dữ liệu cần thiết lập vào kho dữ liệu. Sau quá trình trích lọc dữ liệu thu được các kho dữ liệu sau:

Kết quả học tập (KQHT): Trích lọc các trường MaHS, điểm môn Toan, Ly, Hoa, Sinh, Van, Su, Đia, NgoaiNgu, HocLuc, ThiDH, KhoiThi, MaNganh, MaTruong.

Dữ liệu sau khi thu thập gồm 20550 bản ghi, xóa bỏ những bản ghi không chứa đủ thông tin của các thuộc tính để phân lớp dữ liệu còn lại 20303 bản ghi và 12 thuộc tính.

Có thể thấy rằng các thuộc tính điểm toán, lý, hóa, sinh, sử, địa, văn, ngoại ngữ, tbcm là kiểu dữ liệu số. Do đó trước khi thực hiện khai phá dữ liệu thì hệ thống cần thực hiện rời rạc hóa dữ liệu.

Đối với các thuộc tính trên được rời rạc hoá theo các nhóm: Nhóm Yếu: điểm < 5 Nhóm TB: 5<= điểm <= điểm = 8. Thông qua phần mềm Weka, các thuộc tính nói trên được chuyển đổi kiểu dữ liệu thành kiểu Nominal với 4 giá trị: Yeu (Yếu), TB (Trung bình), Kha (Khá), Gioi (Giỏi).

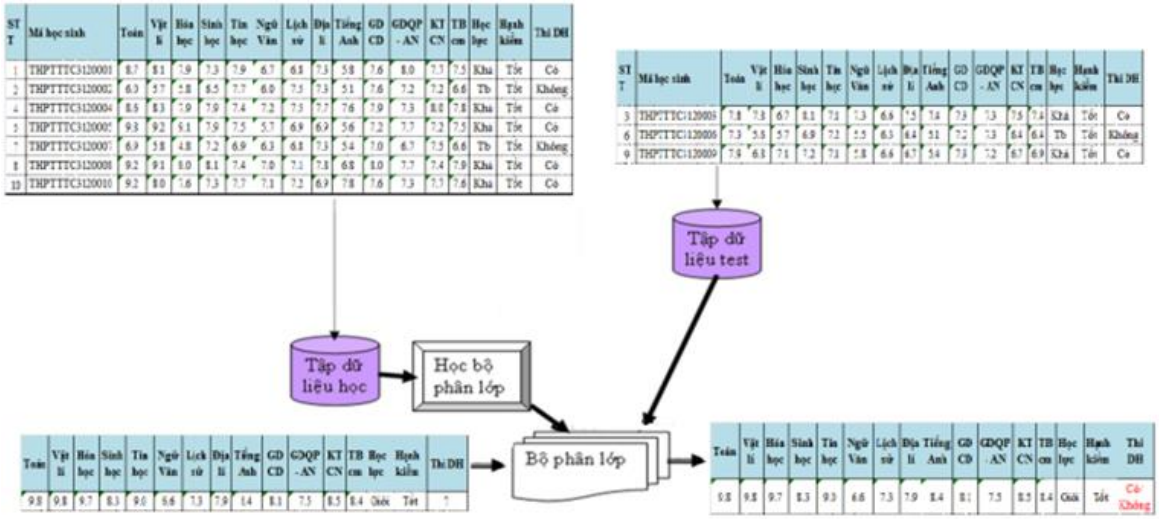
Từ các kho dữ liệu về tư vấn hướng nghiệp, thiết lập mô tơ khai phá dữ liệu bằng kỹ thuật phân lớp dữ liệu dựa trên cây quyết định để đưa ra tập luật cho hệ tư vấn hướng nghiệp.

Tiến hành khai phá dữ liệu bằng phần mềm Weka với tập dữ liệu gồm 14.203 bản ghi bao gồm các thuộc tính như toán, vật lý, hóa học, sinh, ngữ văn, lịch sử, địa lý, ngoại ngữ, thi đại học, khối thi, mã ngành. Để đạt được

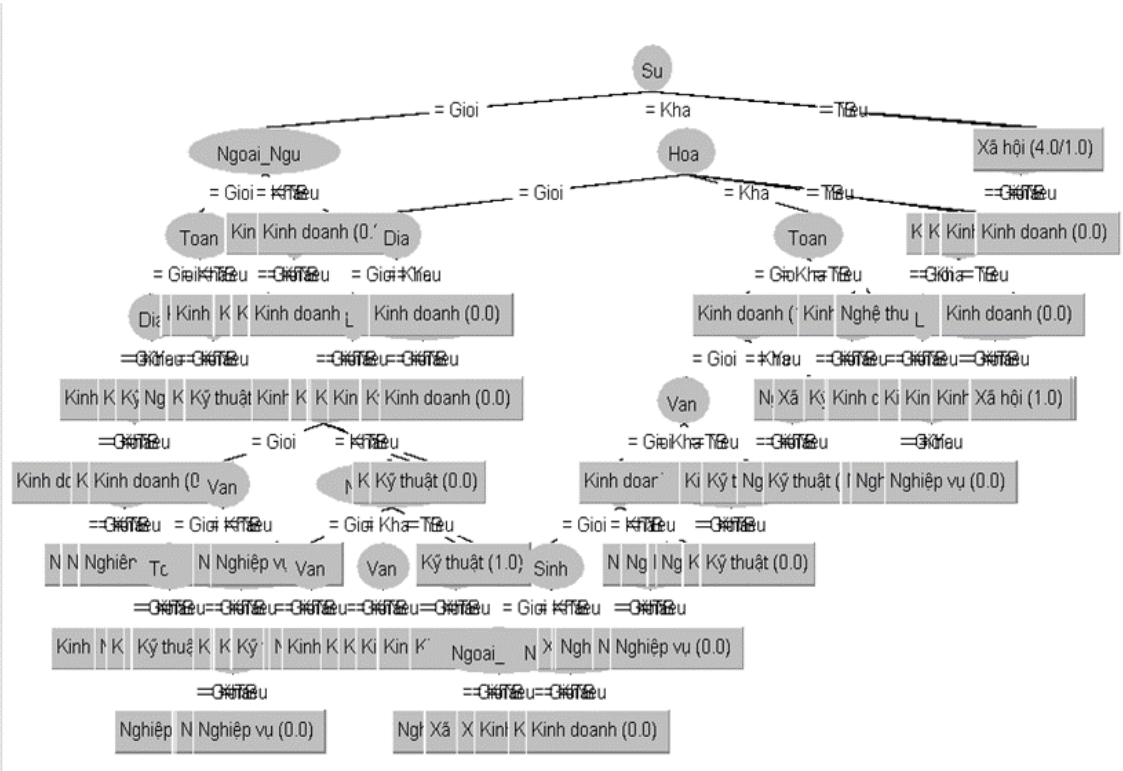
mục tiêu của bài toán, tác giả lựa chọn kỹ thuật phân lớp dữ liệu bằng cây quyết định.

Giai đoạn 1: Dự đoán học sinh có xét tuyển đại học không hay chỉ tham gia thi tốt nghiệp.

Giai đoạn này xây dựng bộ phân lớp bằng thuật toán cây quyết định ID3 trong phần mềm Weka, đầu vào là tập dữ liệu học



.Hình 3.2 Mô hình dự đoán thi đại học



Hình 3.3 Cây quyết định đầy đủ với thuộc tính Thidh

Giai đoạn 2: Dự đoán học sinh nên thi khối nào trong các khối cơ bản A, A1, B, C, D1 và học ngành nghề nào, trường nào.

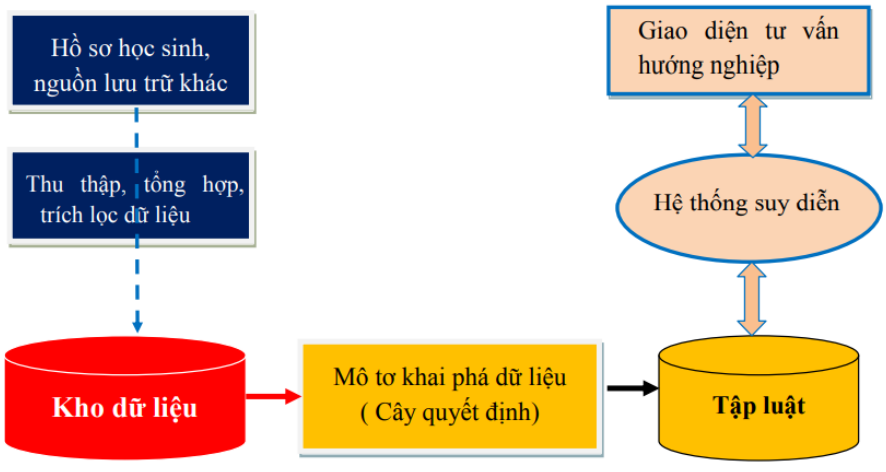
Quá trình dự đoán sử dụng bộ phân lớp dữ liệu đã được xây dựng ở Giai đoạn 1, áp dụng trên tập dữ liệu kiểm tra.

Kết quả đánh giá

H(x) = 2.340

Thuộc tính	Entropy	Gain	SplitInformation	Gain Ratio
Toan	2.311	0.029	1.326	0.023
Ly	2.325	0.015	1.239	0.012
Hoa	2.324	0.017	1.098	0.015
Sinh	2.324	0.017	1.098	0.015
Van	2.325	0.015	1.120	0.013
Su	2.313	0.027	0.786	0.035
Dia	2.321	0.019	0.965	0.020
Ngoai_Ngu	2.313	0.027	1.577	0.017

3.3 Cài đặt và thử nghiệm



Hình 3.4 Mô hình hệ hỗ trợ tư vấn hướng nghiệp

Chức năng chính của hệ thống là khi người dùng nhập vào các thông tin như điểm tổng kết các môn học thì hệ thống dự đoán xem học sinh đó có nên tham gia xét tuyển đại học hay không và dự đoán khối thi, nhóm ngành

nghề mà học sinh nên lựa chọn. Từ đó làm cơ sở tư vấn ngành nghề, trường để học sinh lựa chọn.

Dữ liệu dùng để thực nghiệm được thu thập tại 13 trường THPT trong khu vực nội thành Hà Nội. Dữ liệu thu thập là hồ sơ lưu trữ kết quả học tập và các báo cáo tình hình sau khi tốt nghiệp của các năm 2019, 2020, 2021. Dữ liệu được tập hợp trong 1 file; qua trích lọc và làm sạch, bao gồm 20.430 bản ghi và 12 thuộc tính.

Chia dữ liệu nguồn thành 2 tập dữ liệu: dữ liệu huấn luyện (70%) và dữ liệu kiểm tra (30%).

Ngoài ra thu thập thêm các dữ liệu liên quan đến tư vấn hướng nghiệp và tiến hành lưu trữ dữ liệu trong các bảng sau:

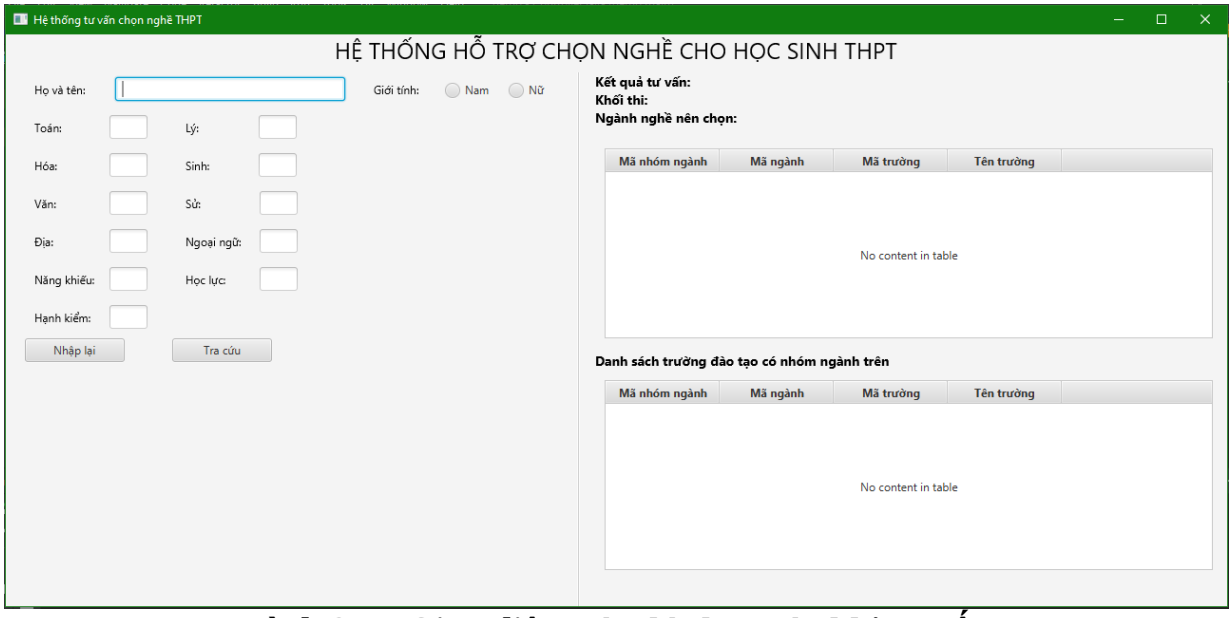
- Bảng KetquaHT: Lưu trữ thông tin về kết quả học tập của học sinh.
- Bảng NganhNghe: Lưu trữ thông tin mã ngành, tên nghành.
- Bảng Khoi: Lưu trữ thông tin mã khối, tên khối, tên môn thi.
- Bảng Trường: Lưu trữ thông tin mã trường, tên trường.

Hệ hỗ trợ tư vấn hướng nghiệp được xây dựng trên cơ sở công nghệ sau:

- Phần mềm mã nguồn mở Weka 3.9.6 để rời rạc dữ liệu và tạo bộ luật.
- Microsoft Excel 2022 để lưu trữ kho dữ liệu.
- Java làm ngôn ngữ lập trình để xây dựng hệ thống suy diễn từ kho luật và JavaFX thiết kế giao diện tương tác với người dùng.

Phần giao diện hệ hỗ trợ tư vấn hướng nghiệp bao gồm:

Giao diện tư vấn hướng nghiệp: Trên giao diện này người dùng cần nhập điểm tổng kết các môn học sau đó sẽ được hệ thống đưa ra các tư vấn như nên chọn khối thi nào, ngành nào, và liệt kê một số trường có đào tạo ngành nghề vừa nêu.

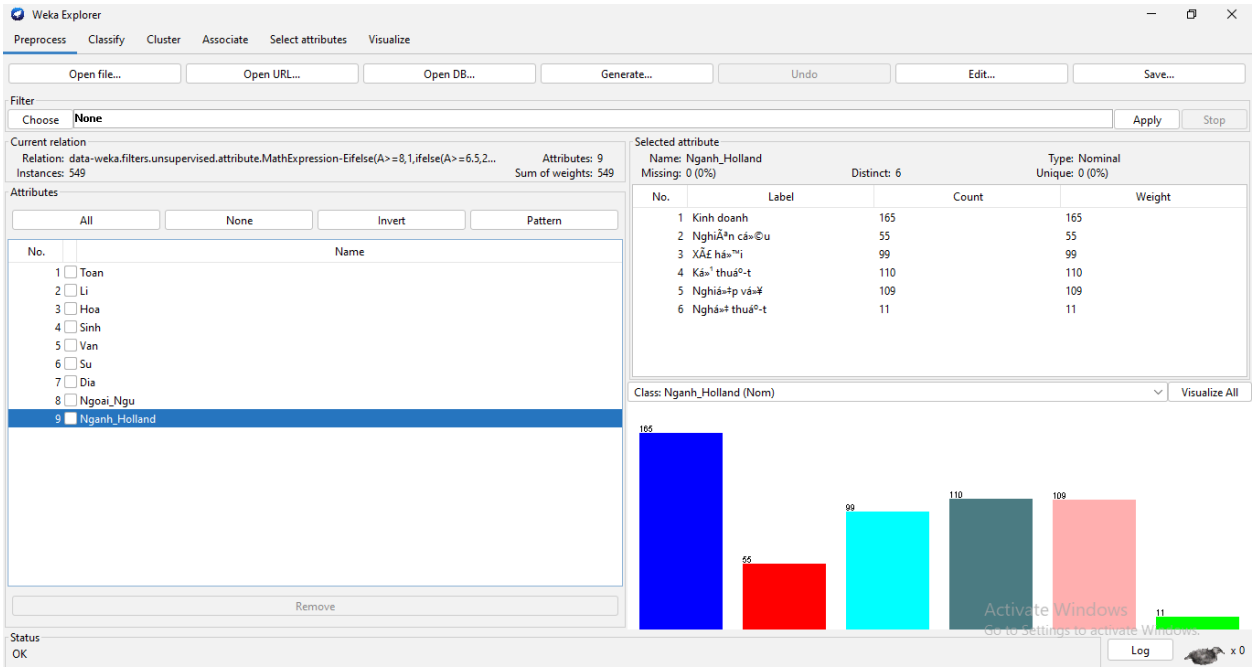


Hình 3.11 Giao diện màn hình trước khi tư vấn

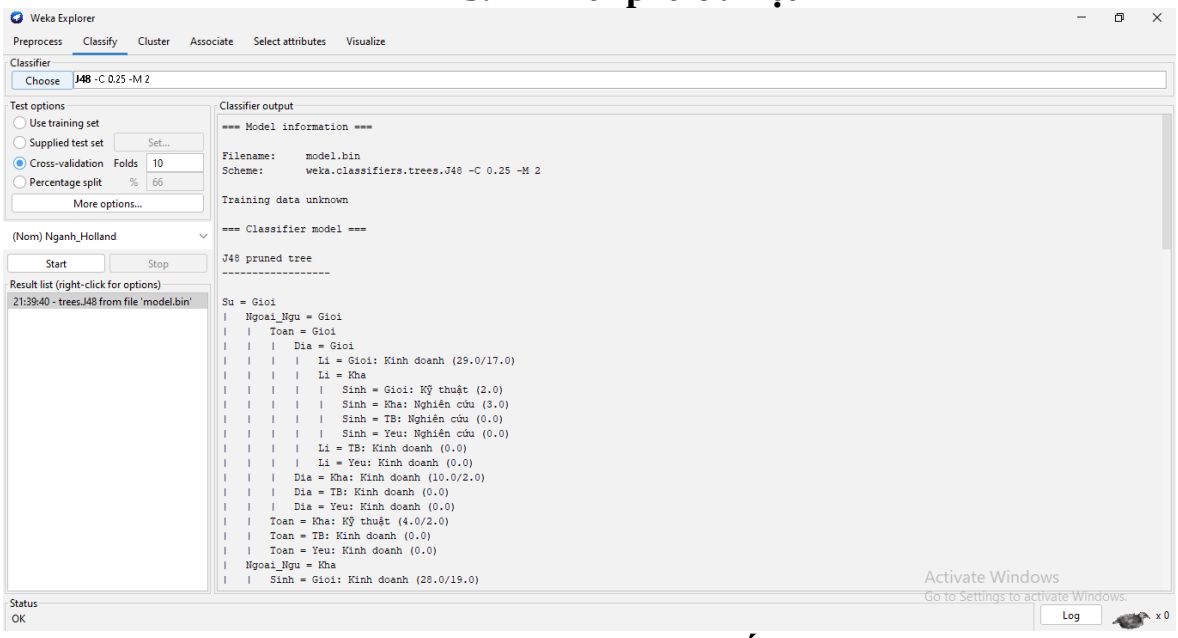
Giao diện test dữ liệu: Trong phần này người xây dựng hệ thống có thể test đoợc bộ dữ liệu kiểm tra độ chính xác của hệ thống.

Toán	Lý	Hóa	Sinh	Văn	Lịch sử	Địa lý	Ngoại ngữ	Khối thi	Mã nghề học	Kết quả khối	Kết quả ngành nghề	Te: OK
7.9	5.9	5.5	8.5	6.9	5.8	2.1	2.9	A1	D4802	A1	D3401	OK
7.3	10	5.9	8.2	6	5.3	8.6	7.9	A1	D4401	A1	D8501	OK
7.8	8.7	5.2	7.8	6.5	6.7	6.1	4.1	A1	D0122	A1	D6203	OK
8.3	7.1	0.1	5	9.3	8.6	8.4	3.9	NGH	504202	NGH	5085	OK
6.6	7.1	10	5.6	5.1	1.7	5.3	3	A1	D2102	A1	D8102	OK
9.2	5.8	0.5	1.1	7.7	9.5	8.4	2.8	A1	D8105	A1	D6203	OK
8.5	7.3	7	5.5	8.4	9.5	5.4	2.9	A1	D3404	A1	D5106	OK
5.3	0.5	3.1	3.3	6	5.1	6.6	7.7	NGH	503103	NGH		NO
7.3	1.7	8.7	0.5	8.2	8.3	6.1	5.8	NGH	504802	NGH	503403	OK
5.3	1.7	5.4	1.9	5.9	10	8.3	9.3	A1	D6203	A1	D2103	OK
5.4	4.3	8.1	2.1	5.9	3	8.1	6	NGH	5081	NGH		NO
6.5	4.1	5.9	4.2	5.6	0.7	7.9	4.4	NGH	5072	NGH		NO
8.4	5.9	5.7	7.6	8	8.6	9.7	6.3	A1	D0604	A1	D0122	OK
7.6	9.5	9.2	8.6	5.1	9.3	9.4	8.9	A1	D3105	A1	D6203	OK
7.8	5	1.8	6.1	8	6.8	9.2	6.8	A1	D4802	A1	D5104	OK
9.3	1.4	7	9	6	0.1	8.8	0.4	A1	D0114	A1		NO
5.5	8.1	8.2	7	6.6	3.6	8.9	7.8	A1	D3101	A1		NO
7.8	7.1	9.6	8.9	8.5	7.4	5.9	7.6	A1	D0401	A1	D0604	OK
9.5	8.7	7.7	6	9.1	6.7	7.8	2.5	A1	D0101	A1	D3104	OK
7.3	9.7	7.1	3	6.3	6.6	7	8.7	A1	D5102	A1	D4802	OK
8.3	8.8	6.7	9.3	5.6	6	8.3	8.4	A1	D7207	A1	D0122	OK
8.7	9	7.7	5.1	9.4	6.8	8.9	1	NGH	502202	NGH	503204	OK
6.6	7.5	9	7.4	5.7	9.9	1.2	6.2	A1	D5102	A1	D5403	OK

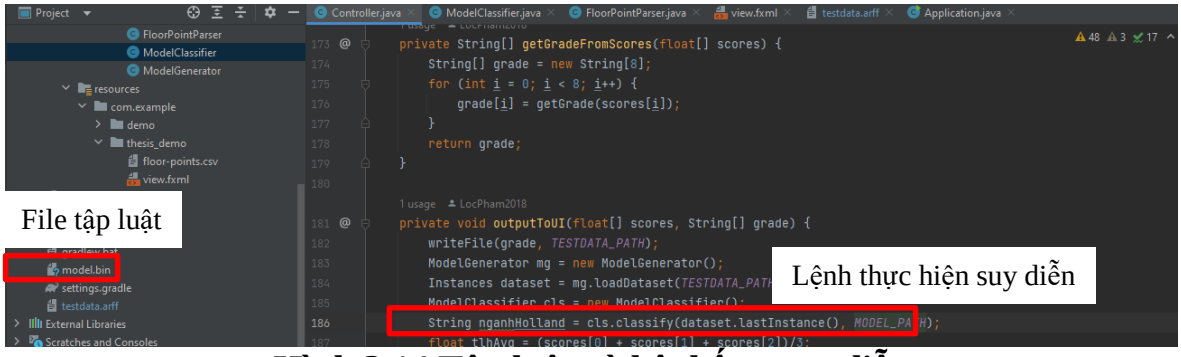
Hình 3.12 Giao diện màn hình test dữ liệu



Hình 3.12 Khai phá dữ liệu



Hình 3.13 Dữ liệu cây quyết định



Hình 3.14 Tập luật và hệ thống suy diễn

Ưu điểm của hệ thống bao gồm:

- Chương trình xây dựng trên nền tảng Microsoft Excel và C#, áp dụng giải thuật ID3 xây dựng cây quyết định, có dung lượng nhỏ, không cần cài thêm môi trường hỗ trợ.

- Chương trình đã xây dựng hoàn chỉnh một mô hình khai phá dữ liệu, có đánh giá kết quả trong quá trình chạy thuật toán.

- Giao diện dễ hiểu, trực quan, người dùng không chuyên cũng có thể dễ dàng sử dụng.

- Phù hợp với yêu cầu, quy mô của bài toán đề ra.

Bên cạnh đó hệ thống vẫn tồn tại những nhược điểm:

- Để có cây quyết định tối ưu và tập luật tối ưu cần phải qua quá trình tinh chỉnh, cắt tỉa cây, tuy nhiên nội dung này của hệ thống còn chưa được chú trọng nghiên cứu.

- Phương pháp lưu trữ kết quả của hệ thống đơn giản và thiếu bảo mật.

Sau khi xây dựng mô hình, tiến hành thử nghiệm với tập dữ liệu kiểm thử để kiểm tra xem độ chính xác của mô hình. Kết quả thu được như sau:

Số lượng mẫu huấn luyện: 22.430 mẫu.

Số lượng mẫu kiểm thử: 6.729 mẫu.

Số lượng mẫu đúng: 5.013 mẫu, chiếm 75%

Số lượng mẫu sai: 1.716 mẫu, chiếm 25%

KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

1/ Về mặt lý thuyết, luận văn đã tiến hành phân tích, nghiên cứu, tìm hiểu được các bước, quy trình của công tác tư vấn hướng nghiệp.

Nắm được các phương pháp và mô hình khai phá dữ liệu, áp dụng để giải quyết yêu cầu bài toán đặt ra. Cụ thể là nghiên cứu và vận dụng thuật toán ID3 xây dựng cây quyết định để khai phá dữ liệu giáo dục, rút ra các tập luật dự đoán lực học của học sinh để tư vấn hướng nghiệp.

2/ Về mặt thực tiễn, luận văn đã nêu được giải pháp kỹ thuật để vận dụng và xây dựng hệ thống trợ giúp ra quyết định trong công tác tư vấn hướng nghiệp của các trường THPT, các trường ĐH, CĐ và các trung tâm hướng nghiệp. Có thể thấy rằng việc kết hợp lý thuyết về mô hình khai phá dữ liệu và thuật toán xây dựng cây quyết định là rất cần thiết, nó giúp giảm thiểu đáng kể thời gian trong việc tìm kiếm, xác định thông tin để phục vụ cho công tác tư vấn hướng nghiệp.

3/ Đặc thù của công tác tư vấn hướng nghiệp là mang nặng tính chất định tính, có nhiều yếu tố ảnh hưởng đến sự lựa chọn ngành học, trường học của các em học sinh, vì vậy hệ thống được xây dựng chỉ mang tính hỗ trợ là chính.

4/ Hiện tại, hệ thống chỉ chạy thử nghiệm trên cơ sở dữ liệu đã trích xuất ra tập tin Excel, chưa chạy trực tiếp trên hệ quản trị cơ sở dữ liệu SQL Server. Do đó chưa có sự kết nối với chương trình quản lý điểm của các trường THPT. Đây cũng là một yêu cầu cần thực hiện trong thời gian sau.

Trên cơ sở nghiên cứu luận văn, hướng phát triển đề xuất của tác giả là tiếp tục.

Quá trình nghiên cứu không tránh khỏi những sai sót nhất định, tác giả mong muốn nhận được góp ý từ các Thầy Cô để luận văn được hoàn thiện hơn.

DANH MỤC CÁC TÀI LIỆU THAM KHẢO

Tiếng Việt

[1] Trần Văn Hải, (2015), Ứ"ng dụng thuật toán học máy SVM trong tư vấn hướng nghiệp cho học sinh trung học phổ thông".

[2] Hoàn Kiếm, Đỗ Phúc (2005), "Giáo trình khai phá dữ liệu", Trung tâm nghiên cứu phát triển công nghệ thông tin, Đại học Quốc gia thành phố Hồ Chí Minh, TP. Hồ Chí Minh, Việt Nam, 204.

Tiếng nước ngoài

[3] D. Hand, H. Mannila, and P. Smyth (2001), Principles of Data Mining, The MIT Press, London, England, 241.

[4] D. Hand, H. Mannila, and P. Smyth (2001), Principles of Data Mining, The MIT Press, London, England, 230.

[5] T. K. Leung, C. Victoria, P. Chen, W. Jiang, and Y. A. Aslandogan (2001), Data Minning Methods and applications.

[6] U.Fayyad, G. Piatetsky-Shapiro, P.Smyth (1996). From Data Mining to Knowledge Discovery in Databases. AI Magazine.

[7]. M. Fayyad, G. P. Shapiro, P. Smyth, and R. Uthurusamy (1996). Advances in Knowledge Discovery and Data Mining. AAAI Press, Menlo Park, CA.

[8] D. Hand, H. Mannila, and P. Smyth (2001). Principles of Data Mining. The MIT Press, London, England.

[9] T. K. Leung, C. Victoria, P. Chen, W. Jiang, and Y. A. Aslandogan (2001). Data Minning Methods and applications.

[10] M. Kantardzic (2003). Data Mining: Concepts, Models, Method, and Algorithms. John Wiley & Sons, New York, NY.

[11] P. Gray, H. J. Watson (1998), Decision Support in Data Warehouse, Prentice Hall.

[12] T. Mitchell (1999). Machine learning and data mining. Communications of the ACM, 42(11): pp. 30-36.

[13] J. Han and M. Kamber (2006). Data Mining: Concepts and Techniques. Universityoflllinois, Morgan Kaufmann Publishers.

[14] L. Zhao, S. Lee and S.P Jeong (2021), Decision Tree Application to Classification Problems with Boosting Algorithm.

[15] Bharati M. Ramageri (2006), Data mining techniques and applications.

[16] K. Caudle, L. Pyeatt, A. Morast, C. Karlssoon, R. C. Hoover, Building a Better Decision Tree by Delaying the Split Decision.

[17] J. S. Deogun (1987), A conceptual approach to decision support systerm models.

[18] J.R Quinlan (1986), Induction of Decision Trees.

[10] M. Kantardzic (2003). Data Mining: Concepts, Models, Method, and Algorithms. John Wiley & Sons, New York, NY.

[11] P. Gray, H. J. Watson (1998), Decision Support in Data Warehouse, Prentice Hall.

[12] T. Mitchell (1999). Machine learning and data mining. Communications of the ACM, 42(11): pp. 30-36.

[13] J. Han and M. Kamber (2006). Data Mining: Concepts and Techniques. Universityoflllinois, Morgan Kaufmann Publishers.

[14] L. Zhao, S. Lee and S.P Jeong (2021), Decision Tree Application to Classification Problems with Boosting Algorithm.

[15] Bharati M. Ramageri (2006), Data mining techniques and applications.

[16] K. Caudle, L. Pyeatt, A. Morast, C. Karlssoon, R. C. Hoover, Building a Better Decision Tree by Delaying the Split Decision.

[17] J. S. Deogun (1987), A conceptual approach to decision support systerm models.

[18] J.R Quinlan (1986), Induction of Decision Trees.