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 HỌC PHỔ THÔNG

LUẬN VĂN THẠC SỸ KỸ THUẬT

(Theo định hướng ứng dụng)

HÀ NỘI - 2022

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 HỌC PHỔ THÔNG

CHUYÊN NGÀNH : KHOA HỌC MÁY TÍNH

Mã số: 8.48.01.01

LUẬN VĂN THẠC SỸ KỸ THUẬT

NGƯỜI HƯỚNG DẪN KHOA HỌC: PGS.TS. LÊ HỮU LẬP

HÀ NỘI – 2022

LỜI CAM ĐOAN

Tôi xin cam đoan đây là công trình nghiên cứu và tìm hiểu của riêng tôi.

Các số liệu, kết quả nêu trong luận văn là trung thực và chưa từng được ai công bố trong bất kỳ công trình nào khác.

Tác giả luận văn

Trần Xuân Oanh

LỜI CẢM ƠN

Để thực hiện và hoàn thành đề tài nghiên cứu khoa học này, em đã nhận được rất nhiều sự hỗ trợ, giúp đỡ. Nghiên cứu khoa học cũng được hoàn thành dựa trên sự tham khảo, học tập kinh nghiệm từ các kết quả nghiên cứu liên quan. Đặc biệt hơn nữa là sự hợp tác của cán bộ, thầy cô và học sinh của trường trung học phổ thông Mỹ Đức B thành phố Hà Nội.

Trước tiên, em xin gửi lời cảm ơn sâu sắc đến Thầy PGS. TS Lê Hữu Lập, người trực tiếp hướng dẫn khoa học đã luôn dành nhiều thời gian, công sức hướng dẫn em trong suốt quá trình thực hiện nghiên cứu và hoàn thành đề tài nghiên cứu khoa học.

Em xin trân trọng cảm ơn ban giám hiệu nhà trường. Khoa sau đại học và quan hệ Quốc tế cùng toàn thể các thầy cô khoa Công nghệ thông tin, trường Học Viện Công Nghệ Bưu Chính Viễn Thông cơ sở I- Hà Nội đã tận tình truyền đạt những kiến thức quý báu, giúp đỡ em trong quá trình học tập và nghiên cứu.

Tuy có nhiều cố gắng, nhưng trong đề tài nghiên cứu khoa học này không tránh khỏi những thiếu sót. Em kính mong Quý thầy cô, các chuyên gia, đồng nghiệp và bạn bè những người quan tâm đến đề tài, tiếp tục có những ý kiến đóng góp, giúp đỡ để đề tài được hoàn thiện hơn.

Trân trọng cảm ơn!

Tác giả

Trần Xuân Oanh

MỤC LỤC

BẢNG KÝ HIỆU VIẾT TẮT

STT	KÝ HIỆU	GIẢI NGHĨA
1	CSDL	Cơ sở dữ liệu
2	DSS	Decision Support System (Hệ thống hỗ trợ ra quyết định)
3	EI	Environment Information (Thông tin môi trường)
4	EM	Expectation - Maximization (Tối ưu hóa kỳ vọng)
5	GT	Goal Tree (Cây mục tiêu)
6	ID3	Iterative Dichotomizer 3
7	PAM	Partition Around Medoids (Phân vùng quanh medoid)
8	SQL	Structured Query Language (Ngôn ngữ truy vấn có cấu trúc)
9	THPT	Trung học phổ thông
10	UI	User Interface (Giao diện người dùng)

DANH MỤC HÌNH VẼ

DANH MỤC BẢNG

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.

Trong trường trung học phổ thông có nhiều phương pháp để giáo dục hướng nghiệp cho học sinh: qua những hoạt động dạy và học các bộ môn khoa học cơ bản, môn công nghệ. Ngoài ra, có thể thông qua các hoạt động ngoại khóa, hoặc thông qua những hoạt động giáo dục hướng nghiệp chính quy, những buổi sinh hoạt hướng nghiệp. Tuy nhiên, các biện pháp mang tính tuyên truyền bộc lộ nhiều điểm yếu:

Chưa cá nhân hóa theo đặc điểm về giới tính, gia cảnh, tôn giáo, vùng miền... của học sinh.

Chưa thu thập nhận xét của thầy cô chủ nhiệm với học viên.

Chưa dựa trên điểm số, kết quả học tập để minh chứng cho lực học làm cơ sở.

Nhìn chung, các phương pháp trên mới chỉ nhắm tới mục tiêu cung cấp kiến thức mà chưa đáp ứng được tiêu chí nâng cao năng lực nhận thức bản thân, qua đó,

phát triển năng lực chọn nghề cho các bạn học sinh và đặc biệt các phương pháp trên không thể giúp các em giải quyết được những bối rối, băn khoăn trong quá trình chọn ngành, chọn nghề.

Song song với sự phát triển như vũ bão của khoa học kĩ thuật, nền kinh tế tri thức cũng là tương lai, đường hướng phát triển của kinh tế thế giới cùng với sự hỗ trợ, đồng hành của những ngành phát triển công nghệ cao. Việt Nam cũng không nằm ngoài xu hướng phát triển đó. Hòa vào tình hình chung của đất nước và thế giới, xã hội hóa giáo dục trở thành một trong những hướng đi thiết yếu. Trong thực tế có rất nhiều các bạn sinh viên ra trường thất nghiệp, hoặc phải làm trái ngành, trái nghề, không đúng với chuyên môn các bạn được đào tạo sau khi tốt nghiệp, dẫn tới năng suất lao động không cao, tỷ lệ bỏ việc nhiều. Bởi thế, vấn đề lựa chọn ngành nghề sao cho đúng đắn, nghề nghiệp được định hướng tốt ngay từ trên ghế nhà trường là nhu cầu vô cùng bức thiết. 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.

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.

Mặc dù có nhiều cố gắng nhưng do thời gian và trình độ còn có hạn chế, luận

2. Tổng quan

a/ Tổng quan về vấn đề nghiên cứu

Nghề nghiệp là một lĩnh vực hoạt động lao động mà trong đó nhờ được đào tạo, con người có những kiến thức, những kỹ năng chuyên môn để làm ra các sản phẩm vật chất hay tinh thần nào đó đáp ứng được nhu cầu của xã hội.

Nghề nghiệp trong xã hội không phải là một cái gì cố định, cứng nhắc. Mới nghe qua chúng ta sẽ dễ nhầm tưởng với công việc nhưng nó không phải là bỏ sức lao động ra làm việc để nhận lại tiền công để trang trải cuộc sống.

Định hướng lựa chọn nghề nghiệp là sự tác động của gia đình, nhà trường, bạn bè , xã hội và môi trường sống xung quanh vào các bạn trẻ. Từ đó các em có thể căn cứ trên năng lực học tập, sở thích cá nhân và đam mê để đưa ra quyết định lựa chọn nghề nghiệp trong tương lai một cách đúng đắn .

Tư vấn hướng nghiệp là một hình thức tư vấn và hỗ trợ của các cán bộ tư vấn nghề nghiệp cho các em về nhu cầu lao động của xã hội, khuynh hướng thị trường lao động cùng với năng lực học tập của các em nhằm giúp các em học sinh có cái nhìn đúng và từ đó đưa ra quyết định lựa chọn phù hợp nhất cho bản thân.

Có 2 loại hình tư vấn hướng nghiệp:

- Tư vấn hướng nghiệp theo nhóm: là loại hình tư vấn hướng nghiệp mà các học sinh trong cùng một nhóm (khối, lớp) được tư vấn cùng một lúc, trong cùng một không gian. Loại hình này thường tiết kiệm chi phí và có thể lồng ghép với nhiều hoạt động phong phú.
- Tư vấn hướng nghiệp cho từng cá nhân: là loại hình tư vấn hướng nghiệp mà mỗi cá nhân học viên được tư vấn riêng biệt, phù hợp với đặc điểm của từng học sinh. Loại hình này là tư vấn sâu hơn, tốn kém nhưng lại đi vào chi tiết, đôi khi có thể dự đoán kết quả cho từng em.

Nhóm lý thuyết cá nhân là lý thuyết liệt kê ra những năng lực nhận biết và đặc điểm phát triển của mọi người để có thể tìm kiếm công việc thích hợp. Lý thuyết mật mã John Holland là một lý thuyết tiêu biểu cho nhóm lý thuyết cá nhân.

Tư vấn tuyển sinh là một bước trong quá trình tư vấn hướng nghiệp mà mọi học sinh đều được cung cấp thông tin cần thiết về các cơ sở đào tạo bậc cao như: trường trung cấp nghề, cao đẳng, đại học để các bạn có được thông tin và quyết định đúng đắn.

Tư vấn viên là người có đảm nhiệm việc tư vấn hướng nghiệp cho từng cá nhân trong trường THPT. Hiện nay còn ít trường có tư vấn viên mà thường là các thầy cô, cán bộ trong nhà trường đảm nhiệm.

Hệ thống hỗ trợ tư vấn hướng nghiệp là hệ thống hỗ trợ các tư vấn viên trong quá trình tư vấn tuyển sinh để tư vấn được chính xác, hiệu quả cho từng cá nhân và tối ưu hóa chi phí cho bài toán này. Phần dưới đây sẽ tìm hiểu rõ hơn bài toán hỗ trợ tư vấn hướng nghiệp.

b/ Giới thiệu về bài toán hệ hỗ trợ tư vấn hướng nghiệp

Hướng nghiệp ở Việt Nam hiện nay đang là vấn đề đang được đông đảo các bộ phận quan tâm. Mà đặc biệt là công tác tư vấn hướng nghiệp trong tuyển sinh đại học. Mục đích chính của công tác tư vấn tuyển sinh này là làm thế nào để giúp các em học sinh trung học phổ thông chọn được ngành học phù hợp với năng lực của mình.

Trong luận văn này chúng ta sẽ tập trung vào nghiên cứu,phân tích các vấn đề liên quan đến tư vấn hướng nghiệp của các nhóm nghề dựa trên cơ sở lý luận của tiến sĩ John Holland nhằm giúp các em thí sinh có thể lựa chọn được ngành học phù hợp với bản thân.

Nội dung luận văn sẽ đi sâu vào thuật toán ID3 (Iterative Dichotomizer 3), cách thức khai phá dữ liệu từ tập dữ liệu có sẵn trong trường học về kết quả học tập, thông tin cá nhân… của học sinh. Từ tập dữ liệu huấn luyện sử dụng phần mềm Weka để tạo cây quyết định bằng thuật toán ID3, sau đó rút ra tập luật từ cây quyết định này.

Để xây dựng "hệ thống hỗ trợ ra quyết định trong tư vấn chọn ngành nghề cho học sinh trung học phổ thông", ta sẽ thu thập các dữ liệu liên quan nằm trong phạm vi nghiên cứu đề tài như: Tổng điểm trung bình theo từng môn học (Toán, Vật lý,

Hóa học, Sinh học, Văn, Lịch sử, Địa lý, Ngoại ngữ) của lớp 10, 11, 12; thông tin cá nhân; phiếu khảo sát học sinh; phiếu nhận xét giáo viên chủ nhiệm và tập dữ liệu huấn luyện.

Sau khi dữ liệu được thu thập, làm sạch, hệ thống sẽ thực hiện chức năng phân tích kho dữ liệu đã có sẵn và tìm ra quy luật nhờ mô hình đã được xây dựng để tư vấn cho các thí sinh sau khi tốt nghiệp THPT. Ngoài ra dữ liệu thu thập được còn có thể được dùng để đánh giá, dự báo nhu cầu và nguồn lao động của từng ngành học.

Để tìm hiểu về quá trình khai phá dữ liệu và phát hiện tri thức thực hiện như thế nào và bằng những kĩ thuật gì, chúng ta sẽ phân tích kĩ ở phần sau.

3. Mục đích nghiên cứu

a/ Mục tiêu của luận văn

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.

b/ Kết quả cần đạt

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.

4. Đối tượng và phạm vi nghiên cứu

a/ Giới hạn nghiên cứu

- Về khách thể: Học sinh lớp 12 tại trường trung học phổ thông Mỹ Đức B thành phố Hà Nội và dữ liệu được chọn từ các môn học của 3 năm học thuộc cấp 3 (năm học 2018-2019, 2019-2020, 2020-2021)

- Về đối tượng: Nhu cầu tư vấn hướng nghiệp của học sinh trung học phổ thông.

b/ 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.

c/ 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.
- Sử dụng ngôn ngữ lập trình Java để viết phần mềm ứng dụng.

CHƯƠNG I. HỆ THỐNG TRỢ GIÚP RA QUYẾT ĐỊNH

1.1 Tổng quan về hệ thống trợ giúp ra quyết định

1.1.1 Khái niệm

Hệ hỗ trợ đưa ra quyết định là hệ thống các máy tính có khả năng tương tác để giúp con người đưa ra quyết định. Hệ thống này sử dụng dữ liệu và mô hình để giải quyết những vấn đề phi cấu trúc.

Các hệ thống hỗ trợ đưa ra quyết định thường có các tính chất sau:

- Là phần mềm máy tính.
- Có chức năng trợ giúp con người đưa ra quyết định.
- Giải quyết những bài toán phi cấu trúc.
- Có khả năng tương tác được với người dùng.
- Áp dụng nhiều mô hình phân tích và mô hình dữ liệu trong tính toán.

1.1.2 Các thành phần của hệ thống hỗ trợ đưa 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 nhìn 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.

1.1.3 Phương pháp xây dựng

Mục đích của hệ thống là hỗ trợ con người đưa ra quyết định trong bối cảnh hoạt động và ngữ cảnh của tổ chức. Để đưa ra quyết định hiệu quả, người ra quyết định phải tuân theo một quy trình được xác định rõ ràng. Quá trình ra quyết định là một quá trình nhận biết tình huống, tạo ra và phân tích các hướng hành động thay thế, lựa chọn một giải pháp thay thế và thực hiện quyết định dựa trên các mục tiêu nhất định. Hệ thống hỗ trợ đưa ra quyết định có 5 chức năng [13]:

- Xử lý dữ liệu.
- Xây dựng mô hình.
- Phân tích mục tiêu.
- Nhận dạng và phân tích vấn đề .
- Quy trình giải pháp.

Mô hình hệ thống hỗ trợ quyết định (DSS) là một mô hình tích hợp cho các hệ thống hỗ trợ quyết định dựa trên năm khía cạnh này.

Ta sẽ đi tìm hiểu cách xây dựng hệ thống DSS theo 5 chức năng này.

1.1.3.1 Xử lý dữ liệu

Quá trình ra quyết định về cơ bản là một quá trình chuẩn bị và trình bày thông tin. Do đó, việc xử lý dữ liệu hiệu quả và tạo ra thông tin ảnh hưởng đáng kể đến quá trình ra quyết định. Một tập hợp các cặp sự kiện có thứ tự và xác suất xảy ra của chúng trong tương lai được gọi là: Thông tin môi trường (EI). Việc ra quyết định

thường được phân thành ba loại dựa trên EI: ra quyết định chắc chắn, với rủi ro hoặc không chắc chắn. [5]

- Ra quyết định một cách chắc chắn xảy ra khi EI hoàn toàn được nắm rõ bởi người ra quyết định.
- Ra quyết định với rủi ro là ra quyết định có cấu trúc bán phần, xảy ra khi có yếu tố xác suất trong EI.
- Việc ra quyết định trong điều kiện không chắc chắn xảy ra ngay cả khi người ra quyết định không có kiến thức về các xác suất trong EI.

DSS có thể hỗ trợ các kiểu ra quyết định này với sự trợ giúp của lý thuyết tập hợp mở và các quy tắc quyết định, nhưng trực giác và khả năng phán đoán của người ra quyết định đóng một vai trò lớn.

1.1.3.2 Xây dựng mô hình

Xây dựng mô hình là sự chuyển đổi từ mô tả cấu trúc con người nhận biết được sang mô hình DSS có thể nhận biết được. Để làm được điều này cấu trúc cần phải được biểu diễn dưới dạng cây hoặc dạng lưu trữ khác mà trí tuệ nhân tạo có thể được lưu trữ trong cơ sở tri thức của DSS.

1.1.3.3 Phân tích mục tiêu

Trong quá trình ra quyết định, các mục tiêu chỉ ra kết quả mà ta cần hướng đến. Trên thực tế, mục tiêu đóng vai trò là cơ sở để đo lường hiệu quả của các lựa chọn thay thế. Do đó, mục tiêu xác định thứ tự ưu tiên so với các lựa chọn thay thế.

Mục tiêu cho biết hướng thay đổi mong muốn, trong đó một thuộc tính là thông số hiệu suất, đặc tính, yếu tố hoặc tài sản. Ví dụ: mục tiêu “lợi nhuận” có thể được chia thành mục tiêu “giảm chi phí” và “tăng doanh thu” (mục tiêu phụ), và sau đó, mục tiêu “tăng doanh thu” có thể được thể hiện dưới dạng “giá cả” và thuộc tính "số lượng".

1.1.3.4 Nhận biết và phân tích vấn đề

Quá trình ra quyết định có liên quan chặt chẽ đến việc giải quyết vấn đề bằng nhận thức. Do đó, một vấn đề quyết định có thể được giải quyết bằng cách giảm bớt và tổng hợp các mục tiêu và mục tiêu phụ [4]. Quá trình rút gọn mục tiêu và các mục

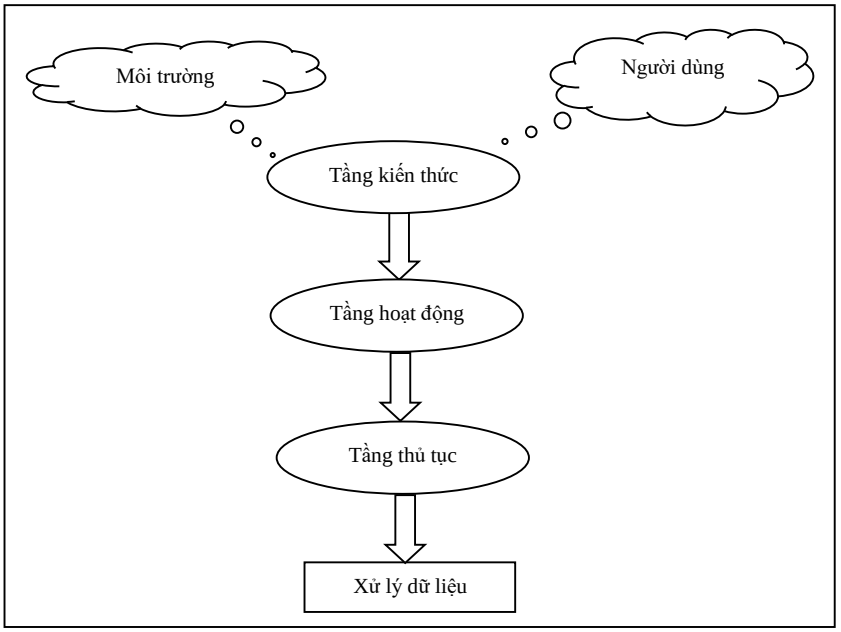
tiêu con có thể được biểu diễn trong một cấu trúc cây được gọi là: Cây mục tiêu (GT). Mỗi nút của GT đại diện cho một nhiệm vụ được xử lý bởi người ra quyết định, chẳng hạn như phân tích cuối cùng của việc ra quyết định, hoặc bởi một mô hình, chẳng hạn như một chức năng tiện ích đại diện cho sự đánh đổi giá trị giữa các mục tiêu hoặc thuộc tính thay thế. Các lá của Cây mục tiêu đại diện cho các thuộc tính được sử dụng để đánh giá một chức năng tiện ích hoặc một mô hình ở cấp cao hơn tiếp theo của cây. Một thuộc tính, như đã được đề cập trước đó, được xem như một tham số. Để nhận biết và đánh giá một tham số như vậy, nó cần được coi như một bài toán có thể được rút gọn thành các bài toán con.

1.1.3.5 Quy trình giải pháp

Mục tiêu chính của DSS là hỗ trợ người ra quyết định trong quá trình ra quyết định bằng cách tạo ra một hệ thống quyết định. Chức năng chính của hệ thống là quá trình tìm giải pháp. Quy trình tìm giải pháp hỗ trợ người ra quyết định giải quyết vấn đề bằng cách cung cấp một môi trường để tạo và đánh giá một tập hợp các giải pháp thay thế. Quy trình giải pháp của DSS có hai giai đoạn:(1) tạo giải pháp và (2) phân tích giải pháp. Tạo ra sự giải pháp là quá trình quét môi trường bên trong và bên ngoài để hình thành thông tin giải pháp thay thế phù hợp với mục tiêu [1]

Phân tích giải pháp là quá trình đánh giá và phân tích hậu quả của mỗi phương án dựa trên thông tin, mục tiêu sẵn có cũng như trực giác và phán đoán của người ra quyết định. Quy trình này là quy trình giải quyết vấn đề

Tồn tại nhiều mức độ trừu tượng giữa hệ thống xử lý dữ liệu thô (hay chính xác hơn là các bit) và người dùng cuối cùng xử lý các vấn đề trừu tượng, chẳng hạn như giải quyết vấn đề không có cấu trúc và phân tích mục tiêu. Để giảm mức độ trừu tượng, ta đưa ra hệ thống phân cấp liên quan đến ba tầng trừu tượng: tầng kiến thức, tầng hoạt động và tầng thủ tục.



Hình 1.1: Các mức trừu tượng của DSS

Tầng kiến thức là một phần trừu tượng của thế giới thực liên quan đến một người ra quyết định. Ở tầng kiến thức, DSS hiểu được vấn đề của người dùng, được nêu một cách trừu tượng, bằng cách truy xuất khối kiến thức và xử lý vấn đề. DSS phân tích kiến thức liên quan đến cấu trúc và mục tiêu của cũng như các phương pháp tổng hợp và giảm thiểu rủi ro.

Ở tầng hoạt động, DSS vận hành với chức năng liên kết các hoạt động ở tầng kiến thức với các hoạt động ở tầng thủ tục nhằm cung cấp môi trường phân tích quyết định để hỗ trợ người ra quyết định. Nói cách khác, DSS ở tầng này quản lý tất cả các hoạt động của hệ thống từ tầng kiến thức đến tầng thủ tục. Kiến trúc DSS này cung cấp một Hệ thống điều khiển (CS) để quản lý, điều phối và kiểm soát các hoạt động chặt chẽ.

Ở tầng thủ tục, DSS có liên quan đến thao tác và sửa đổi.

Như vậy, với kiến trúc trên, DSS thỏa mãn các mục tiêu đề ra là hỗ trợ người dùng đưa ra quyết định dựa trên các phân tích đúng đắn với tính toán và bằng chứng rõ ràng về tri thức và bối cảnh.

1.2 Khai phá dữ liệu

1.2.1 Tổng quan về khai phá dữ liệu

Trong thời đại công nghệ bùng nổ lượng thông tin trên các cơ sở dữ liệu tăng lên đến mức chóng mặt. Sau khoảng hai năm người ta ước tính số lượng của các thông tin trên các cơ sở dữ liệu trên toàn cầu tăng gấp đôi cả về số lượng hồ sơ dữ liệu và số lượng các thuộc tính, các trường. Nguồn tài nguyên khổng lồ này có giá trị rất lớn khi nó được khai phá và phát hiện ra được kiến thức tiềm ẩn. Lượng kiến thức này là rất bé so với nguồn dữ liệu khổng lồ. Vì thế việc tìm ra chúng là việc khó khăn vì những kiến thức tìm ẩn này thường là rất nhỏ so với lượng dữ liệu khổng lồ. 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 như vậy 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.

- Theo tiến sĩ U.M.Fayyad: “Khai phá dữ liệu, thường được xem là việc khám phá tri thức trong các cơ sở dữ liệu, là một quá trình trích xuất những thông tin ẩn, trước đây chưa biết và có khả năng hữu ích, dưới dạng các qui luật, ràng buộc, qui tắc trong cơ sở dữ liệu” [8]

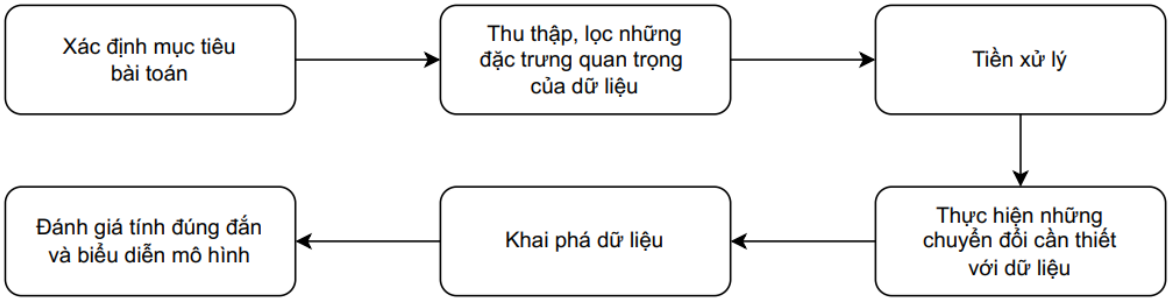
- Tiến sĩ Aleksi Kallio [10] có viết: "Khai phá dữ liệu là quá trình ứng dụng các phương pháp tính toán trên một lượng lớn dữ liệu để tìm thấy những thông tin mới có liên quan và không dễ dàng nhận thấy."

Vì thế chúng ta có thể hiểu quá trình khám phá tri thức tiềm ẩn trong cơ sở dữ liệu chính là khai thác dữ liệu. Nói một cách rõ hơn, nó là quá trình lọc, nhằm tạo ra tri thức hoặc mẫu mới nhưng có ích từ cơ sở dữ liệu lớn.

Như vậy 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. Chẳng hạn như hệ thống SKICAT được sử dụng vào việc phân tích ảnh vệ tinh, phân loại và sắp xếp nhóm các vật thể không gian từ các ảnh quan sát vũ trụ: hệ thống xử lý sự cố CASSIOPEE được dùng để phát hiện và tiên đoán những sự cố của máy bay Boeing, hệ thống TASA dùng để phân tích các lỗi báo động trên đường truyền trong lĩnh vực viễn thông.

1.2.2 Quy trình khai phá tri thức trong CSDL



Hình 1.2 Các bước trong quy trình khai phá dữ liệu

- **Bước 1: Xác định mục tiêu bài toán**

Trong mọi dự án, việc tìm hiểu về bài toán cần giải quyết là nhiệm vụ tiên quyết. Định nghĩa bài toán sẽ quyết định cách thu thập, trích xuất dữ liệu, cách lựa chọn thuật toán trong tất cả những bước sau này. Bởi vậy, để tạo tiền đề thuận lợi cho quá trình khai phá dữ liệu, tránh những sai sót không đáng có, việc mô tả chính xác bài toán là vô cùng quan trọng.

- **Bước 2: Thu thập, trích xuất những đặc trưng quan trọng**

Dựa trên kết quả của bước 1, khi mục tiêu đã được xác định, các dữ liệu có liên quan cần được thu thập và bổ sung, tập hợp thành kho dữ liệu đầy đủ, sẵn sàng. Những dữ liệu này phải đầy đủ, bao gồm những thuộc tính quan trọng cần thiết. Đây là bước rất quan trọng, bởi dữ liệu không liên quan có thể gây nhiễu, dữ liệu thiếu đầy đủ sẽ làm mô hình bị sai, dữ liệu quá nhiều và không đủ đặc trưng sẽ làm mô

hình bị quá khớp. Do đó, cần đặc biệt lưu ý trong quá trình lựa chọn tiêu chí trích xuất và thu thập dữ liệu.

● **Bước 3: Tiền xử lý**

Dữ liệu thu thập được trong bước 2 còn ở dạng thô, có thể thiếu giá trị, thiếu thuộc tính, trùng lặp, không hợp lệ... Nếu đưa trực tiếp dữ liệu thô này vào huấn luyện có thể khiến đưa ra mô hình sai lệch rất nhiều so với những gì dữ liệu phản ánh. Vì vậy ta cần thực hiện bước tiền xử lý trước khi đưa dữ liệu vào khai phá, huấn luyện cho mô hình.

Có hai dạng tiền xử lý thường gặp là rút gọn và làm sạch dữ liệu.

Rút gọn dữ liệu là việc khái quát hóa, tổng hợp, giảm số chiều dữ liệu, nén, rời rạc hóa hoặc giảm số lượng bản ghi đưa vào.

- Để khái quát hóa và tổng hợp dữ liệu ta có thể gộp hai hay nhiều thuộc tính làm một, đưa các dữ liệu ở mức thấp, chi tiết gom thành dữ liệu khái quát.

- Để giảm số chiều dữ liệu, ta cần loại bỏ những thuộc tính thừa, không liên quan bằng những thuật toán như Heuristic, cây quyết định, vét cạn. Đối với cây quyết định, ta rời rạc hóa các giá trị liên tục để giảm chiều dữ liệu.

- Để nén dữ liệu, biến đổi wavelet là phương án thường dùng.

Quá trình rút gọn dữ liệu cần sự khéo léo và kỹ thuật hợp lý để việc rút gọn không làm mất đặc trưng của tập dữ liệu.

Làm sạch dữ liệu là việc xử lý dữ liệu trong trường hợp bị nhiễu hoặc dữ liệu bị sai, thiếu, không nhất quán...

- Khi dữ liệu bị thiếu, không nhất quán phương án xử lý thông thường là bỏ qua mẫu bị sai hoặc thiếu nếu số lượng mẫu nhiều hơn mức tối thiểu cần thiết và quá nhiều trường dữ liệu bị thiếu. Tuy nhiên, khi số lượng mẫu dữ liệu không đủ nhiều thì cần cân nhắc điền bằng tay những trường bị thiếu hoặc dùng các phép toán học để bổ sung.

- Với dữ liệu bị nhiễu thì có nhiều phương án xử lý hơn:

✓ Chia giỏ dữ liệu theo chiều rộng bằng cách chia miền giá trị thành N khoảng có cùng kích thước, hoặc theo chiều sau bằng cách chia miền giá trị thành N khoảng

có số mẫu tương đương. Sau đó khử nhiễu bằng các phương pháp giá trị trung bình, biên của giỏ...

✓ Dùng phương pháp hồi quy tuyến tính để tìm được quan hệ giữa các biến hoặc thuộc tính, sau đó suy ra thuộc tính từ giá trị của thuộc tính khác.

✓ Tổ chức các giá trị tương tự nhau thành các cụm và xem xét các giá trị ở ngoài cụm để làm mịn dữ liệu.

● **Bước 4: Thực hiện những chuyển đổi cần thiết**

Kiểu dữ liệu cần có để đưa vào mỗi thuật toán là khác nhau, do vậy, ta cần chuyển đổi dữ liệu sang dạng cần thiết trước khi đưa vào khai phá, huấn luyện. Các dạng chuyển đổi thông dụng được liệt kê sau đây:

Chuyển đổi kiểu dữ liệu từ dạng logic nhị phân (true-false) sang dữ liệu số nguyên hoặc theo chiều ngược lại.

Rời rạc hóa: Đưa dữ liệu từ miền giá trị có tính liên tục thành các nhãn rời rạc thay cho giá trị thực. Đây cũng là kiểu chuyển đổi cơ bản sẽ được áp dụng trong luận văn này để thực hiện phân lớp dùng cây quyết định.

Phân giá trị trong một cột thành nhóm rồi chuyển đổi giá trị thành tên nhóm giúp thu gọn phạm vi giá trị của mẫu.

Chuẩn hóa các giá trị theo tỷ lệ để đưa về một dải giá trị nhất định (thường là các khoảng 0.0 đến 1.0, -1.0 đến 1.0). Kiểu chuyển đổi dữ liệu này thường được tìm thấy trong các bài toán liên quan đến độ sai lệch hoặc dùng mạng nơ-ron để gán nhãn dữ liệu.

● **Bước 5: Khai phá dữ liệu**

Đây là bước cốt lõi trong toàn bộ quá trình khai phá dữ liệu. Tại bước này, cần áp dụng những chiến thuật khai phá dữ liệu cùng thuật toán phù hợp để tìm ra thông tin từ dữ liệu đã được chuẩn bị kỹ càng trong 4 bước trước đó. Kết quả của bước này chính là mô hình sau huấn luyện. Mô hình sẽ khám phá ra những kiểu mẫu, quy luật của dữ liệu để đưa ra xu hướng dự đoán. Có nhiều kỹ thuật có thể được kể đến như: phân nhóm (clustering), luật kết hợp (Association rules), hồi quy (regression), phân lớp (classification). Ta sẽ nói rõ hơn về các kỹ thuật này trong phần ngay sau đây.

- **Bước 6: Đánh giá**

Để thuận tiện hơn cho quá trình xem xét kết quả, đối chiếu mẫu, mô hình trong bước này cũng có thể được minh họa, biểu diễn bằng bảng biểu, sơ đồ. Sau đó, mô hình sẽ được kiểm thử, đánh giá tính đúng đắn, độ chính xác bằng những kỹ thuật kiểm thử mô hình. Kỹ thuật phổ biến trong việc đánh giá độ chính xác của mô hình là k-fold với việc chia dữ liệu ra thành k nhóm (fold), lấy ra 1 nhóm, loại bỏ nhãn của nhóm đó rồi đưa vào mô hình được huấn luyện từ (k-1) nhóm còn lại và đối chiếu với nhãn ban đầu.

1.2.3 Các kỹ thuật khai phá dữ liệu

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

- **Phân lớp**

Phân lớp là kỹ thuật khai thác dữ liệu được áp dụng phổ biến nhất, sử dụng một tập hợp các mẫu để tạo ra mô hình có thể phân loại tổng thể các bản ghi. Các ứng dụng phát hiện gian lận và tín dụng đặc biệt phù hợp với loại kỹ thuật này.

Cách tiếp cận này thường sử dụng cây quyết định hoặc các thuật toán phân loại dựa trên mạng nơron. Quá trình phân lớp dữ liệu bao gồm huấn luyện và phân lớp. Trong bước huấn luyện, dữ liệu huấn luyện được phân tích bằng thuật toán phân lớp. Tập hợp dữ liệu được sử dụng để ước tính độ chính xác của các quy tắc phân lớp. Nếu độ chính xác là chấp nhận được thì mô hình sẽ được áp dụng cho các bộ dữ liệu mới.

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).

Các loại mô hình phân loại: phân loại theo cây quyết định, phân loại Bayes, mạng nơ-ron.

- **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. Việc phân loại là phương pháp hiệu quả để phân biệt các nhóm hoặc lớp đối tượng nhưng việc định nghĩa các nhãn trước là rất tốn thời gian và công sức. Vì vậy, đôi khi ta dùng phân cụm làm bước tiền xử lý để lựa chọn và phân loại tập hợp con các thuộc tính. Ví dụ: tạo nhóm khách hàng dựa trên sản phẩm họ mua, phân loại gen có chức năng tương tự.

Các thuật toán phân cụm: phân vùng xung quang medoids (PAM), tối đa hóa kỳ vọng (EM), K-means...

- **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.

Trong thực tế, việc dự đoán là không đơn giản. Ví dụ, khối lượng bán hàng, lượng hàng tồn kho, giá cả và tỷ lệ hỏng hóc của sản phẩm đều rất khó dự đoán vì chúng có thể phụ thuộc vào tương tác của nhiều biến độc lập. Do đó, các kỹ thuật phức tạp hơn (ví dụ: hồi quy logistic, cây quyết định, hoặc lưới thần kinh) có thể cần được áp dụng để đưa ra dự đoán. Hồi quy và phân loại đều được dùng để dự đoán nhưng dữ liệu của hồi quy là liên tục còn phân loại thì là rời rạc.

Các loại phương pháp hồi quy: hồi quy tuyến tính, hồi quy tuyến tính đa biến, hồi quy phi tuyến, hồi quy phi tuyến tính đa biến.

- **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. Ví dụ, thông tin về thống kê mua hàng ở siêu thị: "80% người mua váy là phụ nữ, 70% người mua váy sẽ mua thêm giày". Quan hệ có thể kết hợp từ nhiều luật để đi đến kết luận.

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. Mạng nơ-ron có khả năng đáng chú ý về việc rút ra ý nghĩa từ những dữ liệu không chính xác và có thể được sử dụng để trích xuất các mẫu và phát hiện các xu hướng quá phức tạp đối với con người hoặc các kỹ thuật máy tính khác. Chúng rất phù hợp cho các đầu vào và đầu ra có giá trị liên tục. Mạng nơron được dùng hiệu quả nhất trong việc xác định các mẫu hoặc xu hướng dữ liệu và rất thích hợp để dự đoán, dự báo nhu cầu.

Các loại mạng nơ-ron: mạng truyền ngược (back propagation).

Kết luận chương I

Khai phá dữ liệu là quá trình đi tìm tri thức được ẩn đằng các 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 II. 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.

Các luận điểm của lý thuyết mật mã Holland có những luận điểm chính sau đây:

• Nếu một người chọn một công việc phù hợp với tính cách của mình, anh ta sẽ dễ dàng thành công trong nghề đó và thành công hơn. Nói cách khác, những người làm việc trong một môi trường giống với tính cách của họ thường dễ thành công và hài lòng với công việc của họ.

• Hầu như tất cả mọi người đều có thể được xếp vào một trong sáu loại tính cách và có 6 môi trường làm việc tương ứng với sáu loại tính cách:

✓ Nhóm kỹ thuật: có sở thích và khả năng tìm tòi, khám phá, có thể dùng máy móc, làm tốt những việc đòi hỏi thao tác khéo léo của cơ thể trong các lĩnh vực: ô tô, điện, điện lạnh, cơ khí, điện tử, tin học hoặc các lĩnh vực đòi hỏi sự tinh xảo, tỉ mỉ như bonsai, nấu ăn, chế tác đồ thủ công mỹ nghệ, …

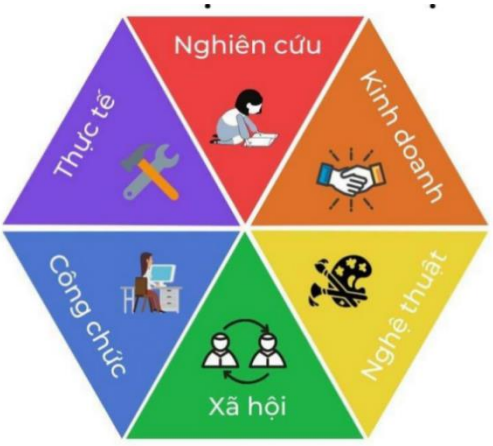
✓ Nhóm nghiên cứu khoa học: có tính tập trung, có niềm say mê lĩnh vực chuyên sâu và khả năng làm việc độc lập với hệ thống khái niệm, có thể tìm ra quy luật, biểu diễn tư duy trừu tượng qua hệ thống ký hiệu, có khả năng thiết kế sáng tạo.

✓ Đoàn nghệ thuật: có năng khiếu về các lĩnh vực nghệ thuật như hội họa, âm nhạc, múa... Có óc sáng tạo và khả năng tư duy về không gian, cảm nhận nghệ thuật.

✓ Nhóm xã hội: có khả năng giao tiếp lịch thiệp, hoạt ngôn, thích tiếp xúc với mọi người, muốn được nói, biết lắng nghe.

✓ Nhóm quản lý: có khả năng lãnh đạo, ra lệnh, sắp xếp công việc, có trí nhớ tốt, tập trung cao, cảm xúc ổn định, có tư duy hệ thống.

✓ Nhóm chuyên viên nghiệp vụ: thích hợp với công việc bàn giấy, tỉ mỉ, thận trọng, có hiểu biết về lĩnh vực chuyên sâu của mình và hiểu biết rộng các lĩnh vực lân cận.



Hình 2.1: 6 nhóm môi trường làm việc

Lý thuyết mật mã của Holland được áp dụng rộng rãi cho những người mới bắt đầu khám phá sở thích và nghề nghiệp. Trong thực tế công tác tư vấn hướng nghiệp ở nước ta, nếu được sử dụng đúng cách thì lý thuyết mật mã Holland sẽ đem lại nhiều lợi ích.

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

2.2.1 Mô tả bài toán

Bài toán phân lớp dữ liệu giải quyết vấn đề gán nhãn cho các mẫu mới với độ chính xác cao nhất nhằm phân loại mẫu mới vào phân lớp thích hợp.

Dữ liệu đầu vào: Tập hợp các mẫu (dữ liệu huấn luyện) và một nhãn phân lớp tương ứng với mỗi mẫu dữ liệu.

Output: Mô hình dự đoán, tức là cây quyết định dùng để phân lớp dữ liệu cho mẫu mới.

2.2.2 Quá trình phân lớp dữ liệu

Việc phân lớp dữ liệu gồm có 2 bước:

Bước 1: Tạo mô hình từ dữ liệu huấn luyện.

Tập hợp dữ liệu huấn luyện: là thông tin đầy đủ về mẫu huấn luyện để hệ thống tham chiếu và xây dựng mô hình.

Mẫu huấn luyện là thuộc tính được định nghĩa các giá trị đặc trưng và gán nhãn đúng. 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 liệu.

Hình 2.2: Tạo mô hình huấn luyện

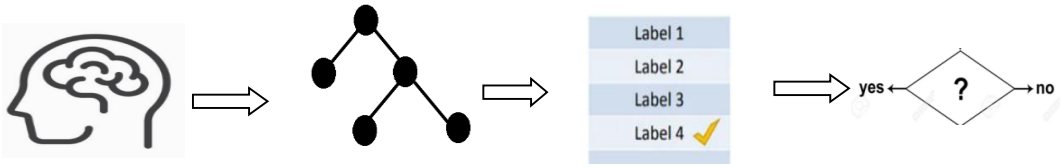
Bước 2: Ứng dụng mô hình huấn luyện vào bài toán.

Khi có dữ liệu mới được nhập vào, hệ thống sẽ phân loại, gán nhãn cho dữ liệu dựa trên mô hình huấn luyện được tạo ra ở bước 1.

Để đánh giá tính đúng đắn của mô hình ta sử dụng k-folds. Dữ liệu đã có nhãn được chia thành k nhóm, trong đó, k - 1 nhóm được dùng để huấn luyện, nhóm còn

lại được dùng để kiểm chứng nhãn phân loại. Nếu nhãn hệ thống gắn trùng với nhãn của đối tượng thì kết quả là đúng, nếu không là sai.

Tính đúng đắn của mô hình càng cao khi tỉ lệ dữ liệu phân lớp đúng càng cao.



Dữ liệu bỏ nhãn Mô hình sau huấn luyện Nhãn của đối tượng So sánh

Hình 2.3 Ứng dụng mô hình phân lớp vào bài toán

2.3 Cây quyết định

2.3.1 Khái niệm

Tập hợp các thuộc tính với giá trị đi kèm với nó biểu diễn một mẫu dữ liệu cụ thể, ta gọi nó là đối tượng. Mỗi thuộc tính là đặc tính của mẫu dữ liệu đó. Giá trị của thuộc tính là rời rạc.

Mỗi đối tượng có nhãn, là tên phân lớp của của đối tượng trong bài toán phân loại.

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.

Cây quyết định được cấu tạo bởi các nút và các nhánh có phân chia cấp độ trong đó:

- Nhánh (branch): là những khoảng giá trị rời rạc mà thuộc tính có thể mang. Trên hình vẽ 2.4 được biểu diễn dưới dạng đường nối 2 nút (node).

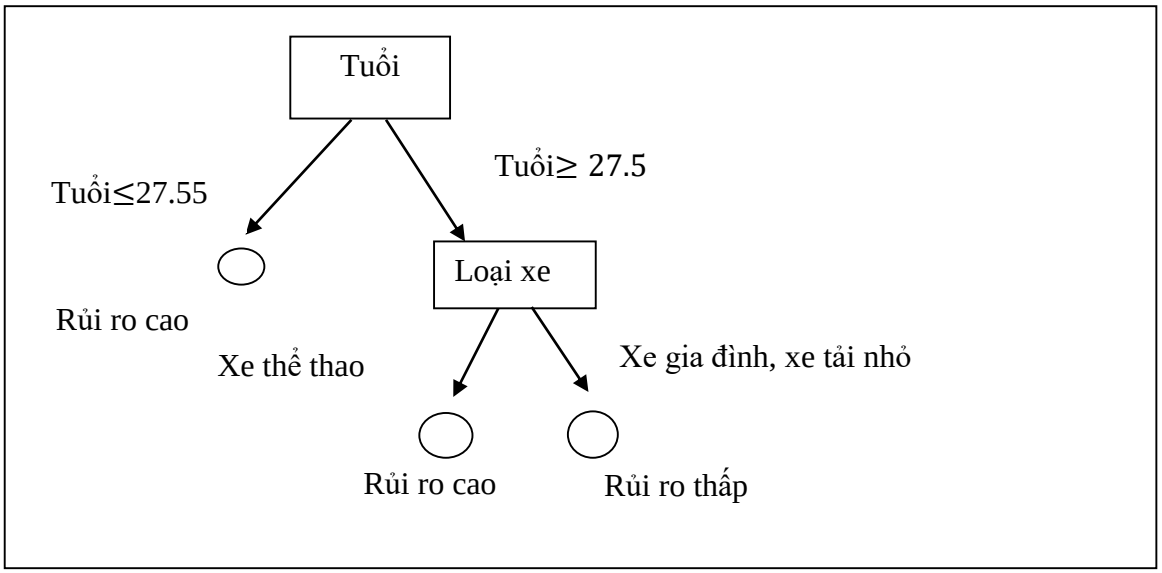
- Nút (node): Giống như các nút của của cấu trúc dữ liệu cây thông thường.

- Nút gốc (root node) là nút bắt đầu của cây, nút gốc không có nút cha.

- Nút lá (leaf node) là nút không có nút con, là kết quả phân lớp của cây quyết định.

Trên hình vẽ 2.4 được biểu diễn bằng hình tròn. Nút con là kết quả.

- Nút trong (internal node) là các nút được biểu diễn bằng hình chữ nhật trong hình 2.4, có cả nút cha và có ít nhất một nút con. Nút trong và nút cha là tên thuộc tính.



Hình 2.4 Cây quyết định

Để xác định một đối tượng mang nhãn nào, ta đi từ gốc của cây quyết định, đánh giá các giá trị từng thuộc tính của đối tượng, đi theo nhánh thích hợp. Quá trình rẽ nhánh dừng khi bắt gặp một nút lá. Sau cùng, nhãn của nút lá là nhãn của đối tượng.

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.

2.3.2 Các bước dựng cây quyết định

Để phân loại được đối tượng theo cây quyết định thì việc đầu tiên cần làm là xây dựng cây quyết định. Để xây dựng được cây quyết định, ta tuân theo hai bước: tạo cây và tỉa cây.

- Tạo cây quyết định.

Bởi cây quyết định là cấu trúc phân tầng từ trên xuống dưới, việc tạo cây cũng cần tuân theo cấu trúc phân tầng này. Cây quyết định được tạo từ trên xuống, bắt đầu từ nút gốc, chứa thuộc tính phân loại đầu tiên. Toàn bộ dữ liệu trong tập huấn luyện cần thỏa mãn thuộc tính ở nút gốc. Sau đó, ta tiến hành chọn các thuộc tính phân hoạch. Quá trình phân chia lặp đi lặp lại cho tới khi thỏa mãn các điều kiện sau:

- Mọi đối tượng thuộc về một nút đều nằm trong một lớp.

- Không còn thuộc tính phân hoạch nào để tiếp tục phân chia dữ liệu.

- Không còn phần tử nào thuộc nút để tiếp tục phân chia dữ liệu.

- Tỉa cây quyết định.

Sau khi cây được tạo, sẽ có những nhánh chứa phần tử không thuộc lớp nào hoặc các phần tử mang thuộc tính hỗn loạn. Ở bước tỉa cây, ta sẽ loại bỏ các nhánh và phần tử nhiễu này.

2.4 Thuật toán Iterative Dichotomiser 3 (ID3)

2.4.1 Tổng quan

ID3 được Quinlan công bố vào cuối thập niên 70, thế kỷ XX. Năm 1986, ID3 được giới thiệu và mô tả chi tiết trong mục Induction on decision trees, machine learning. ID3 dựng cây quyết định theo hướng từ trên xuống (top-down) bằng cách lựa chọn thuộc tính tốt nhất để triển khai mỗi bước.

ID3 là một trong những thuật toán khai phá dữ liệu đơn giản nhưng lại vô cùng hiệu quả. ID3 có cách biểu diễn tri thức học được dễ hiểu và trong sáng, heuristic đơn giản, có hiệu quả tốt trong xử lý dữ liệu nhiễu. Bởi vậy, cho tới nay, dù có nhiều thuật

toán dựng cây quyết định mới được tạo ra, ID3 vẫn có tính ứng dụng thực tiễn cao và phổ biến.

Đầu vào giải thuật: Tập dữ liệu huấn luyện gồm các mẫu huấn luyện. Mỗi mẫu là một đối tượng gồm thuộc tính mang giá trị và phân lớp (còn gọi là "nhãn") của đối tượng.

Đầu ra: Cây quyết định có khả năng phân lớp (hoặc "gán nhãn") đối tượng. Cây có khả năng gán nhãn đúng cho đối tượng mới, không nằm trong tập huấn luyện.

2.4.2 Mô tả giải thuật

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

```
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;
```

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

end

end

end

Hình 2.5: Mô tả thuật toán ID3

Với mỗi thuộc tính bất kỳ của tập hợp thuộc tính, dữ liệu huấn luyện đều có thể được phân lớp thành những tập con mang giá trị chung của thuộc tính. ID3 chọn nút gốc để bắt đầu và sử dụng đệ quy, coi nút hiện tại là nút gốc của phân vùng tập hợp mẫu. Quá trình đệ quy kết thúc khi mọi phân vùng nằm trong một phân lớp và phân lớp đó chính là lá của cây quyết định.

2.4.3 Cách chọn thuộc tính phân hoạch

Với cùng một tập mẫu, ta có thể xây dựng nhiều cây quyết định với độ sâu-rộng và độ phức tạp khác nhau, phụ thuộc vào thứ tự lựa chọn thuộc tính để triển khai trên cây quyết định. Bởi vậy, cách chọn thuộc tính phân hoạch là yếu tố quyết định độ phức tạp của cây.

Để chọn được thứ tự và thuộc tính tốt nhất tạo cây quyết định trong mỗi bước, thuật toán ID3 đánh giá các đại lượng: độ lợi thông tin (information gain), độ pha trộn Entropy và tỷ suất độ lợi thông tin (information gain ratio). Ta sẽ xem xét từng tiêu chí này để phân tích phương án chọn thuộc tính.

2.4.4 Độ pha trộn Entropy

Entropy trong lý thuyết thông tin là khái niệm mở rộng từ entropy trong cơ và nhiệt động lực học. Về khái niệm, entropy trong một tập hợp S được định nghĩa là số lượng các bit cần thiết để mã hóa thông tin của một phần tử lấy ngẫu nhiên trong S. Hay nói cách khác, entropy đặc tả sự hỗn loạn của tín hiệu trong một sự kiện ngẫu

nhiên. Thông tin được định nghĩa là thành phần không hỗn loạn ngẫu nhiên của tín hiệu. Như vậy, entropy trở thành thang đo độ thuần nhất của thông tin.

Gọi tập hợp S là tập hợp các mẫu huấn luyện, lấy ngẫu nhiên một mẫu x trong S, ta rút ra các đặc tính sau của Entropy H(x):

- $0 \leq H(x) \leq 1$
- H(x) = 0 khi và chỉ khi S là thuần nhất, nghĩa là ta luôn chắc chắn rằng mẫu x thuộc phân lớp L, đồng nghĩa với việc tất cả các mẫu thuộc S đều thuộc phân lớp L.
- H(x) = 1 khi và chỉ khi S có độ hỗn loạn tối đa, với mỗi mẫu thuộc một phân lớp và không có quy luật hay sự trùng lặp.
- $0 < H(x) < 1$ đồng nghĩa với việc tập mẫu S có số lượng mẫu thuộc các loại không bằng nhau.
- Ta có công thức tổng quát để tính Entropy của mẫu ngẫu nhiên rời rạc x trong tập S, với x có thể nhận n giá trị như sau:

$$H(x) = - \sum_{i=0}^{n} p(i) \log_2 p(i)$$

Với: p(i) là xác suất mẫu được gán nhãn (i).

p(i) được tính bằng công thức:

$$p(i) = \frac{C_i}{D}$$

Với: C là số lượng mẫu được gán nhãn (i), D là số lượng mẫu của tập S.

Nhìn vào công thức có thể thấy entropy phản ánh đúng khái niệm là giá trị kỳ vọng của độ ngạc nhiên của các giá trị mà mẫu x có thể mang.

2.4.5 Độ lợi thông tin (information gain)

Độ lợi thông tin là thang đo độ hiệu quả của thuộc tính được lựa chọn để phân loại. Đại lượng này phụ thuộc vào hai đại lượng: thông tin và entropy.

Độ lợi thông tin của thuộc tính A trong tập hợp mẫu S, G (S, A) được tính bằng:

$$G(S, A) = H(x) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} H(y)$$

Với: - Sv là tập hợp con của S sao cho mọi mẫu thuộc Sv đều có giá trị thuộc tính A bằng v.

- H(y) là entropy của mẫu ngẫu nhiên rời rạc y trong tập Sv.

- Values(A) là tập hợp các giá trị có thể có của thuộc tính A.

- |S| là số lượng mẫu của tập hợp S.

- |Sv| là số lượng mẫu của tập hợp Sv.

Ý nghĩa của Gain (S, A): Số lượng bit có thể giảm trong việc mã hóa phần tử ngẫu nhiên trong tập mẫu huấn luyện S, khi biết giá trị thuộc tính A.

Trong quá trình xây dựng cây quyết định, tại mỗi bước triển khai cây của thuật toán ID3, thuộc tính triển khai được chọn là thuộc tính có giá trị Gain lớn nhất.

2.4.6 Tỷ suất độ lợi thông tin (Information Gain Ratio)

Cây quyết định tốt là cây có chiều rộng hợp lý, không quá sâu, bởi nếu cây đi sâu sẽ dễ dẫn đến hiện tượng quá khớp (overfitting). Việc tăng độ rộng của cây quyết định là một trong những mục tiêu chính của quá trình chọn thuộc tính phân hoạch.

Để đạt được điều này ta cần tính tỷ số của tổng lượng thông tin thu được trên số lượng nhánh. Độ đo tỷ suất độ lợi thông tin được tính theo công thức:

$$GainRation(S, A) = \frac{Gain(A)}{SplitInfo(S,A)}$$

Với:

$$SplitInfo(S, A) = \sum_{i=1}^{c} \frac{|S_i|}{S} \log_2 \frac{|S_i|}{|S|}$$

Trong đó:

SplitInfo (S, A) là thông tin phân tách của S trên cơ sở giá trị thuộc tính phân loại A

Si là tập con trong số c tập con của S.

2.4.7 Ví dụ tính toán

STT	toan	ly	hoa	sinh	su	dia	van	eng	thidh
1	kha	gioi	TB	kha	yeu	gioi	yeu	kha	Do
2	gioi	TB	kha	gioi	kha	yeu	yeu	yeu	Do
3	TB	yeu	TB	TB	gioi	kha	TB	gioi	Do
4	gioi	TB	gioi	kha	yeu	gioi	yeu	kha	Truot
5	yeu	kha	gioi	TB	gioi	TB	kha	gioi	Do
6	gioi	gioi	yeu	TB	kha	gioi	gioi	TB	Truot
7	kha	gioi	kha	kha	yeu	TB	yeu	gioi	Truot
8	TB	kha	gioi	gioi	yeu	kha	gioi	gioi	Do
9	yeu	TB	kha	TB	TB	gioi	gioi	TB	Truot
10	yeu	gioi	yeu	yeu	kha	kha	gioi	yeu	Truot
11	yeu	kha	TB	gioi	yeu	TB	TB	kha	Do
12	gioi	yeu	TB	kha	gioi	TB	yeu	gioi	Truot
13	kha	yeu	kha	gioi	yeu	yeu	kha	yeu	Truot
14	TB	gioi	kha	TB	kha	yeu	gioi	yeu	Do
15	gioi	yeu	gioi	yeu	kha	kha	TB	yeu	Do

Bảng 2.1 Dữ liệu xếp loại học sinh

Lớp P: thidh= ”Do”

Lớp N: thidh=”Truot”

$$\text{Info}(p,n) = \text{Info}(8,7) = -\frac{8}{15}\log_2\frac{8}{15} - \frac{7}{15}\log_2\frac{7}{15} = 0.997$$

Đối với môn Toán

toan	Số lượng	thidh		Info
		Do	Truot	
gioi	5	2	3	0.971
kha	3	1	2	0.918
TB	3	3	0	0.0
yeu	4	2	2	1.0

Bảng 2.2 Thông tin thuộc tính “toan”

$$\text{Entropy(toan)} = \frac{5}{15}(0.971) + \frac{3}{15}(0.918) + \frac{3}{15}(0.0) + \frac{4}{15}(1.0) = 0.774$$
$$\text{Gain(toan)} = 0.223$$

Đối với môn Lý

ly	Số lượng	thidh		Info
		Do	Truot	
gioi	5	2	3	0.971
kha	3	3	0	0.0
TB	3	1	2	0.918
yeu	4	2	2	1.0

Bảng 2.3 Thông tin thuộc tính "ly"

$$\text{Entropy(ly)} = 0.774$$

$$\text{Gain(ly)} = 0.223$$

Đối với môn Hóa

hoa	Số lượng	thidh		Info
		Do	Truot	
gioi	4	3	1	0.811
kha	5	2	3	0.971
TB	4	3	1	0.811
yeu	2	0	2	0.0

Bảng 2.4 Thông tin thuộc tính "hoa"

$$\text{Entropy(hoa)} = 0.756$$

$$\text{Gain(hoa)} = 0.243$$

Đối với môn Sinh học

sinh	Số lượng	thidh		Info
		Do	Truot	
gioi	4	3	1	0.811

kha	4	1	3	0.811
TB	5	3	2	0.971
yeu	2	1	1	1.0

Bảng 2.5 Thông tin thuộc tính "sinh"

$$\text{Entropy(sinh)} = 0.890 \text{Gain(sinh)} = 0.107$$

Đối với môn Sử học

su	Số lượng	thidh		Info
		Do	Truot	
gioi	3	2	1	0.918
kha	5	3	2	0.971
TB	1	0	1	0.0
yeu	6	3	3	1.0

Bảng 2.6 Thông tin thuộc tính "su"

$$\text{Entropy(su)} = 0.907$$

$$\text{Gain(su)} = 0.090$$

Đối với môn Địa lý

dia	Số lượng	thidh		Info
		Do	Truot	
gioi	4	1	3	0.811
kha	4	3	1	0.811
TB	4	2	2	1.0
yeu	3	2	1	0.918

Bảng 2.7 Thông tin thuộc tính "dia"

$$\text{Entropy(dia)} = 0.883$$

$$\text{Gain(dia)} = 0.11$$

Đối với môn Ngữ văn

van	Số lượng	thidh		Info
		Do	Truot	
gioi	5	2	3	0.971
kha	2	1	1	1.0
TB	3	3	0	0.0
yeu	5	2	3	0.971

Bảng 2.8 Thông tin thuộc tính "nguvan"

$$\text{Entropy}(\text{van}) = 0.781$$

$$\text{Gain}(\text{van}) = 0.21$$

Đối với môn Ngoại ngữ

eng	Số lượng	thidh		Info
		Do	Truot	
gioi	5	3	2	0.971
kha	3	2	1	0.918
TB	2	0	2	0.0
yeu	5	3	2	0.971

Bảng 2.9 Thông tin thuộc tính "ngoaingu"

$$\text{Entropy}(\text{eng}) = 0.831$$

$$\text{Gain}(\text{eng}) = 0.166$$

Độ lợi thông tin của môn Hóa là lớn nhất (0.243) nên ta chọn phân tích thuộc tính này.

Phân tích Gain Ratio

$$\text{SplitInformation}(S, \text{toan}) = -\frac{5}{15}\log_2\frac{5}{15} - \frac{3}{15}\log_2\frac{3}{15} - \frac{3}{15}\log_2\frac{3}{15} - \frac{4}{15}\log_2\frac{4}{15}$$

$$= 1.966$$

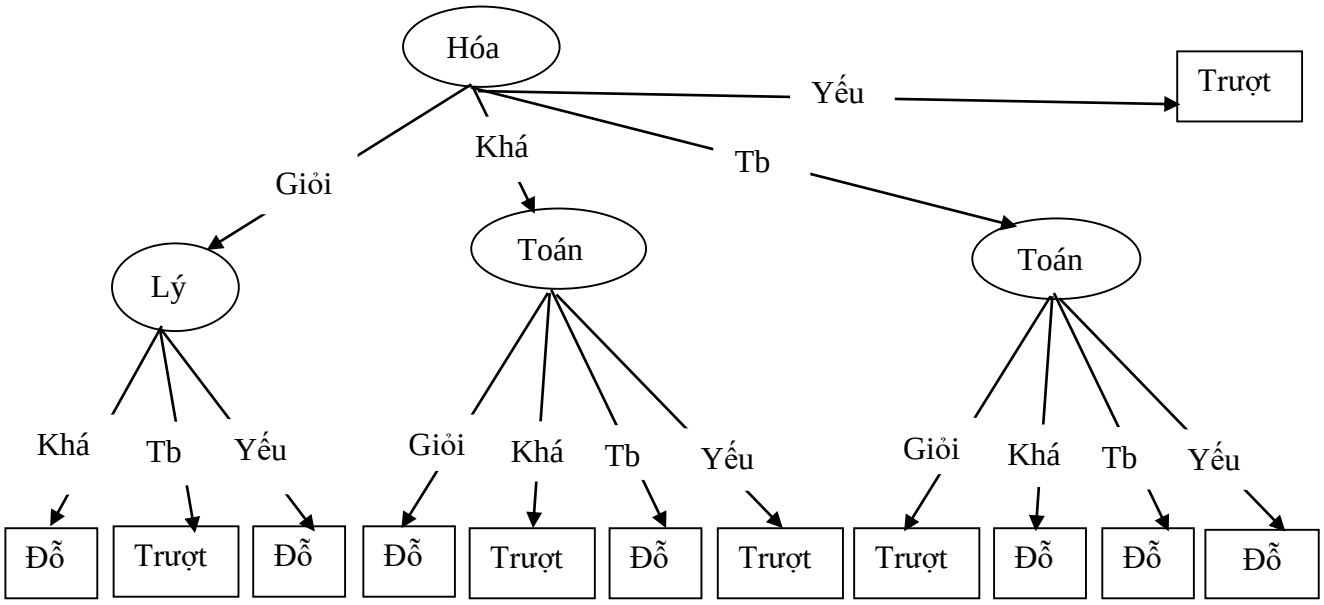
$$\text{GainRatio}(S, \text{toan}) = \frac{0.223}{1.966} = 0.119$$

Tính tương tự cho các thuộc tính còn lại ta có bảng dưới đây.

Thuộc tính	Entropy	Gain	SplitInformation	Gain Ratio
toan	0.774	0.223	1.966	0.119
ly	0.774	0.223	1.966	0.119
hoa	0.756	0.243	1.933	0.122
sinh	0.890	0.107	1.933	0.054
su	0.907	0.090	1.782	0.051
dia	0.883	0.114	1.990	0.057
van	0.781	0.216	1.909	0.113
eng	0.831	0.116	1.909	0.061

Bảng 2.10 Bảng thống kê các môn học của học sinh

Tiếp tục quá trình như trên ta thu được cây quyết định như hình dưới:



Hình 2.6 Cây quyết định

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

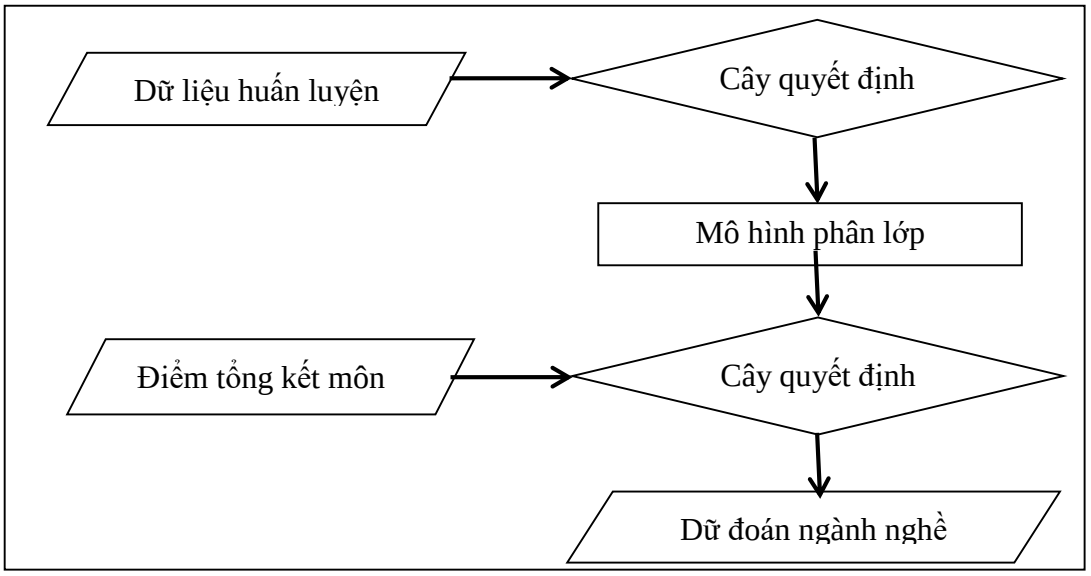
2.5.1 Yêu cầu cơ bản của hệ thống

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.

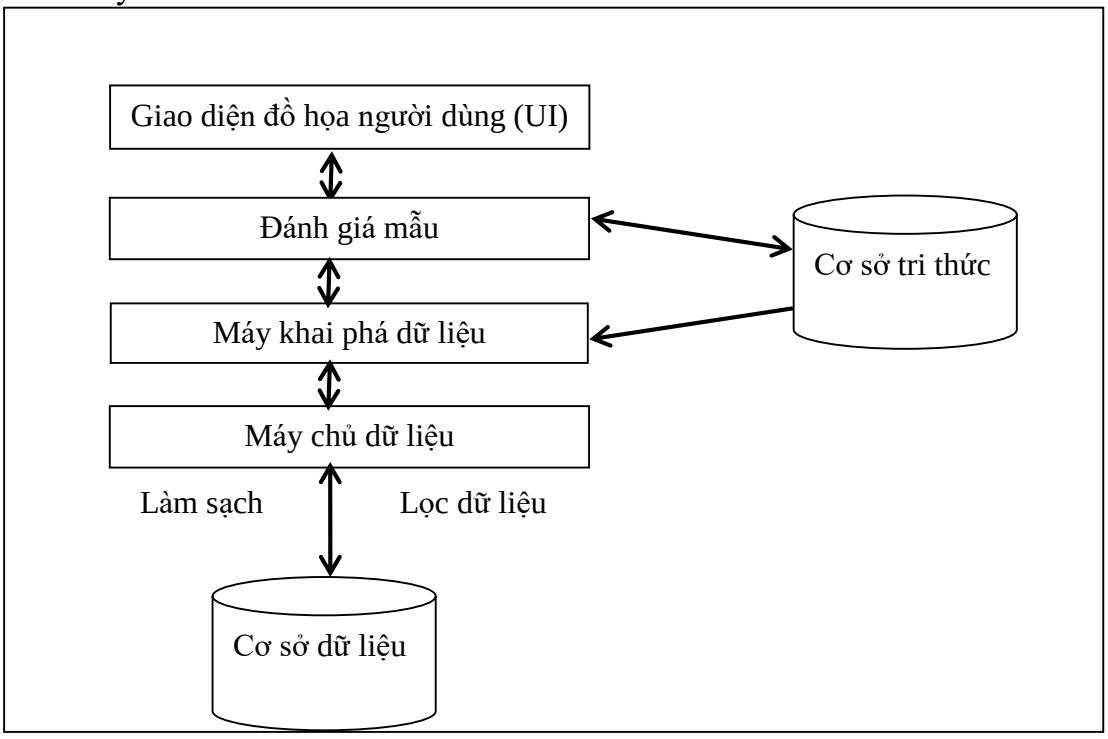
- Đầu ra: Kết quả dự đoán 1 nhóm Holland và 3 ngành nghề thuộc nhóm và các trường cùng khối thi tương ứng học viên nên đăng ký xét tuyển.

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



Hình 2.7 Lưu đồ mô tả chức năng hệ thống hỗ trợ tư vấn hướng nghiệp

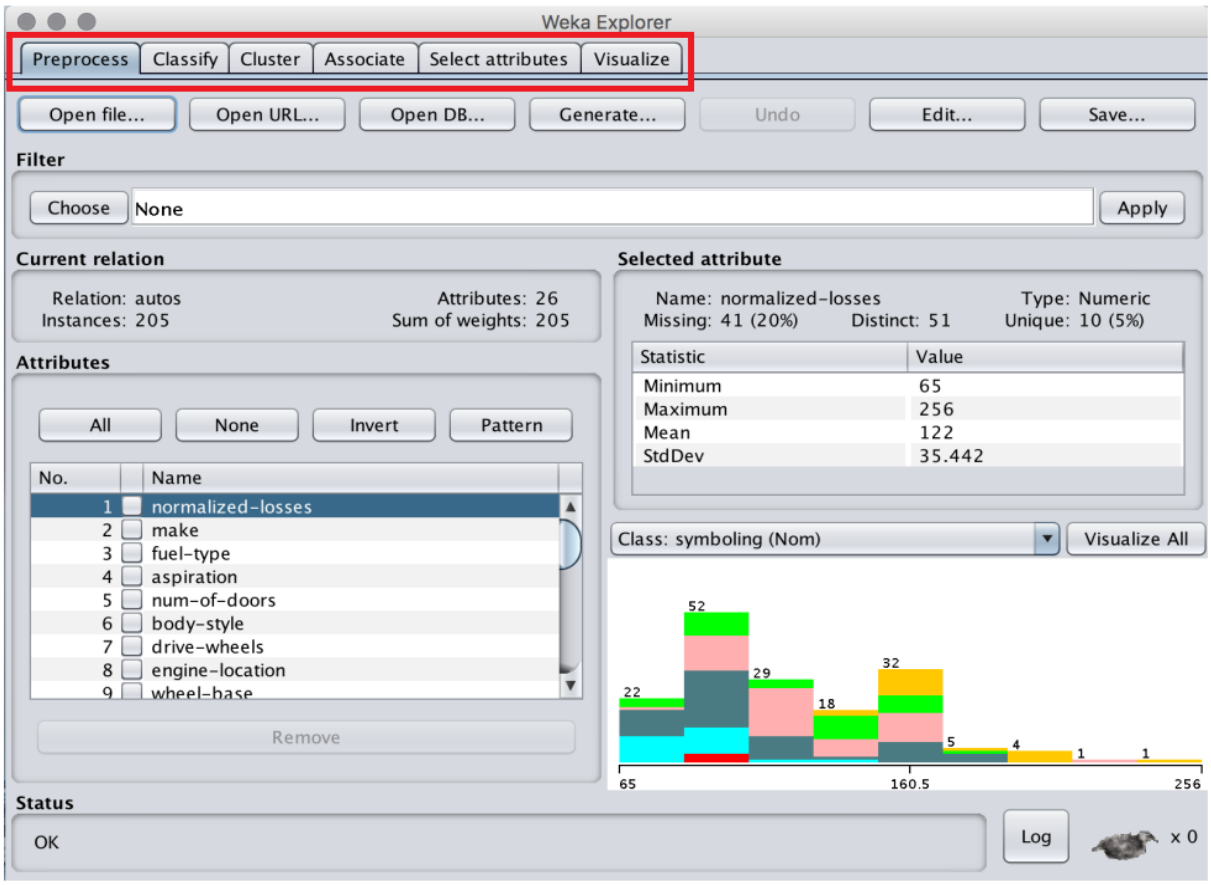
Về yêu cầu thiết kế, hệ thống cần đảm bảo có đầy đủ các thành phần trong kiến trúc dưới đây:



Hình 2.8 Yêu cầu kiến trúc hệ thống hỗ trợ tư vấn

2.5.2 Phần mềm Weka Explorer

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. Cũng như hàng loạt các thuật toán học tập khác, nó bao gồm một loạt các công cụ tiền xử lý. Bộ công cụ đa dạng và toàn diện này được truy cập thông qua một giao diện chung để người dùng có thể so sánh các phương pháp khác nhau và xác định những phương pháp phù hợp nhất cho vấn đề đang giải quyết gọi là Weka Explorer. WEKA được phát triển tại đại học Waikato ở New Zealand; tên viết tắt của Waikato Environment for Knowledge Analysis. Bên ngoài trường đại học, WEKA, được phát âm cùng vần với Mecca, là một loài chim không biết bay với bản tính tò mò chỉ được tìm thấy trên các hòn đảo của New Zealand. Hệ thống được viết bằng Java và được phân phối theo các điều khoản của giấy phép công cộng GNU. Nó chạy trên hầu hết mọi nền tảng và đã được thử nghiệm trên các hệ điều hành Linux, Windows và Macintosh.



Hình 2.9 Giao diện phần mềm Weka

WEKA Explorer có 6 chức năng chính:

- **Preprocessing (Tiền xử lý):** Các công cụ tiền xử lý trong WEKA được gọi là: "Bộ lọc". Tiền xử lý truy xuất dữ liệu từ một tệp, cơ sở dữ liệu SQL hoặc URL (đối với các tập dữ liệu rất lớn, có thể cần lấy mẫu phụ vì tất cả dữ liệu được lưu trữ trong bộ nhớ chính). Dữ liệu có thể được xử lý trước bằng một trong các công cụ tiền xử lý của Weka. Tab tiền xử lý hiển thị biểu đồ với thống kê cho thuộc tính hiện được chọn. Biểu đồ cho tất cả các thuộc tính có thể được xem đồng thời trong một cửa sổ riêng biệt. Một số bộ lọc hoạt động khác nhau, tùy thuộc vào việc một lớp thuộc tính đã được thiết lập hay chưa. Hộp bộ lọc được sử dụng để thiết lập bộ lọc cần thiết. WEKA chứa các bộ lọc cho tiết chế, chuẩn hóa, lấy mẫu lại, lựa chọn thuộc tính, kết hợp thuộc tính.

- **Classify (Phân loại):** Các công cụ phân loại có thể được sử dụng để thực hiện phân tích sâu hơn về dữ liệu được xử lý trước. Nếu dữ liệu yêu cầu một vấn đề

phân loại hoặc hồi quy, nó có thể được xử lý bằng cách sử dụng tab phân loại. Một mô hình phân loại được sản xuất trên dữ liệu được đào tạo đầy đủ. WEKA bao gồm tất cả các kỹ thuật học tập chính để phân loại và hồi quy: Bayesian bộ phân loại, cây quyết định, bộ quy tắc, máy vectơ hỗ trợ, các perceptron hậu cần và nhiều lớp, tuyến tính hồi quy và các phương pháp lân cận gần nhất. Nó cũng chứa "những người học meta" như đóng gói, xếp chồng, tăng cường và các chương trình thực hiện điều chỉnh tham số tự động bằng cách sử dụng xác thực chéo, phân loại nhạy cảm với chi phí, v.v.Các thuật toán học tập có thể được đánh giá bằng cách sử dụng xác thực chéo hoặc tập hợp tạm dừng và Weka cung cấp tiêu chuẩn các phép đo hiệu suất bằng số (ví dụ: độ chính xác, lỗi bình phương trung bình gốc), cũng như các phương tiện đồ họa cho trực quan hóa hiệu suất của bộ phân loại (ví dụ: đường cong ROC và đường cong thu hồi độ chính xác). Có thể hình dung dự đoán của một mô hình phân loại hoặc hồi quy, cho phép xác định các giá trị ngoại lệ, tải và lưu các mô hình đã được tạo.

- **Cluster (Cụm):** WEKA chứa các "cụm" để tìm các nhóm cá thể trong tập dữ liệu. Công cụ cụm cung cấp quyền truy cập vào các thuật toán phân cụm của Weka, chẳng hạn như k-means, một lược đồ phân nhóm phân cấp tăng dần theo phương pháp heuristic. Cluster các bài tập có thể được trực quan hóa và so sánh với các cụm thực tế, được xác định bởi một trong các thuộc tính trong dữ liệu.
- **Associate (Liên kết):** Các công cụ liên kết có các thuật toán tạo quy tắc kết hợp. Nó có thể được sử dụng để xác định mối quan hệ giữa các nhóm thuộc tính trong dữ liệu.
- **Select attributes (Chọn thuộc tính):** Thú vị hơn trong bối cảnh tin sinh học là tab thứ năm, cung cấp các phương pháp để xác định các tập hợp con của các thuộc tính mang tính dự đoán của thuộc tính mục tiêu trong dữ liệu. Weka chứa một số các phương pháp tìm kiếm thông qua không gian của các tập con thuộc tính, các thước đo đánh giá cho các thuộc tính và thuộc tính tập hợp con. Các phương pháp tìm kiếm như tìm kiếm ưu tiên nhất, thuật toán di truyền, lựa chọn chuyển tiếp và xếp hạng thuộc tính. Cả hai phương pháp tìm kiếm và phương pháp đánh giá khác nhau có thể được kết hợp, làm cho hệ thống trở nên rất linh hoạt.

- **Visualize (Trực quan hoá):** Các công cụ trực quan hóa hiển thị một ma trận gồm các biểu đồ phân tán. Thực tế hình dung rất hữu ích giúp xác định những khó khăn trong vấn đề học tập. WEKA hình dung kích thước đơn (1D) cho đơn thuộc tính và thứ nguyên kép (2D) cho các cặp thuộc tính. Nó là để hình dung mối quan hệ hiện tại trong các lô 2D. Không tí nào phần tử ma trận có thể được chọn và phóng to trong một cửa sổ riêng biệt, nơi người ta có thể phóng to các tập hợp con của dữ liệu và truy xuất thông tin về các điểm dữ liệu riêng lẻ. Tùy chọn "Jitter" để xử lý các thuộc tính danh nghĩa cho để lộ các điểm dữ liệu bị che khuất cũng được cung cấp.

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

Ngày nay, các công ty phần mềm đang triển khai phần mềm quản lý điểm đang thí điểm ở các trung tâm, cơ sở đào tạo và các trường THPT trên khắp cả nước.

Sau khi sử dụng phần mềm này 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:

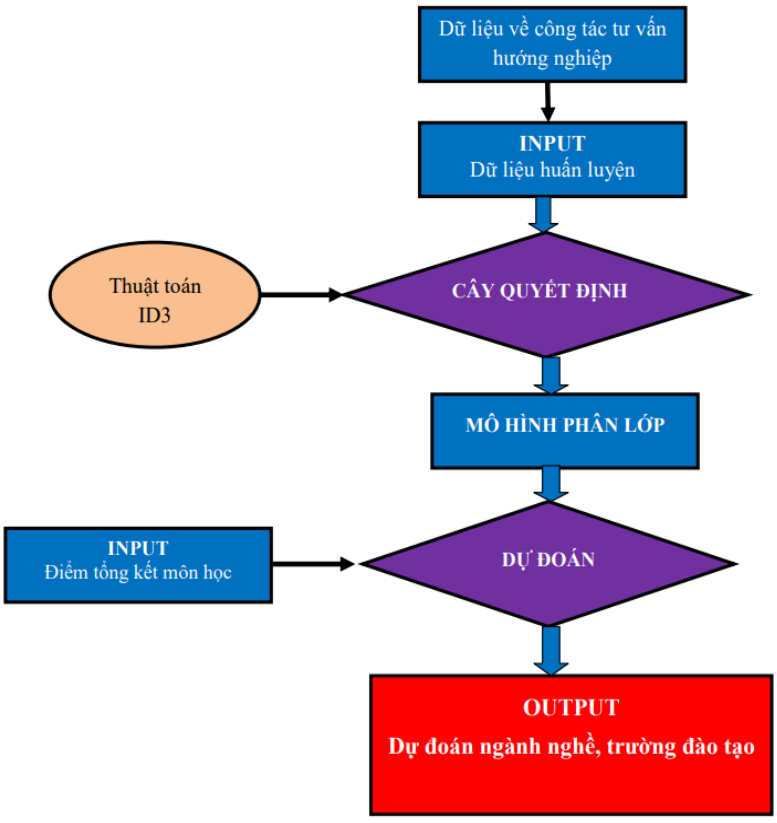
- Mối quan hệ giữa điểm trung bình môn học có ảnh hưởng đến quyết định cho sự lựa chọn của việc không tham gia kì thi xét tuyển đại học.

- 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

3.2.1 Thu thập, trích lọc dữ liệu

Theo dữ liệu nguồn công bố việc khai thác để tìm ra thông tin hữu ích cho việc tư vấn hướng nghiệp nhằm thu thập từ các luồng thông tin khác nhau. Dữ liệu sẽ được

tập hợp từ các sơ yếu lý lịch, thành tích đạt được trong quá trình học tập của học sinh THPT, các ngành học, …

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:

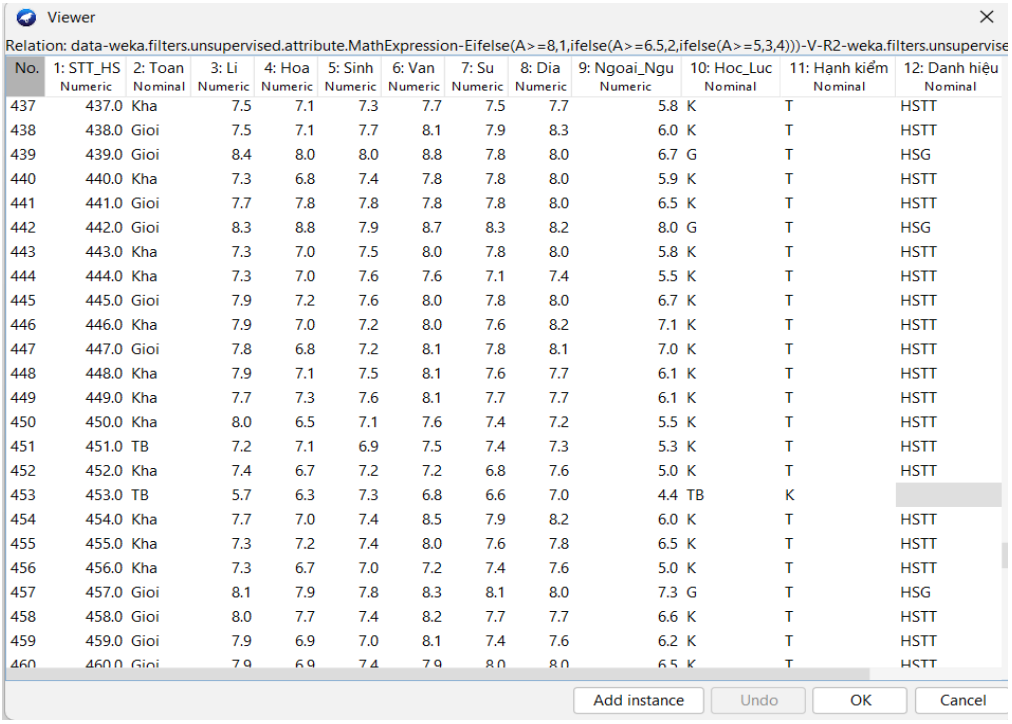
Thành tích học tập của học sinh: Điểm các môn Toán, Lý, Hóa, Sinh, Ngoại ngữ, Văn, Sử, Địa.

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.

Ví dụ dữ liệu ban đầu được lưu trong các bảng như sau:



No.	1: STT_HS Numeric	2: Toan Nominal	3: Li Numeric	4: Hoa Numeric	5: Sinh Numeric	6: Van Numeric	7: Su Numeric	8: Dia Numeric	9: Ngoai_Ngu Numeric	10: Hoc_Luc Nominal	11: Hạnh kiểm Nominal	12: Danh hiệu Nominal
437	437.0	Kha	7.5	7.1	7.3	7.7	7.5	7.7	5.8	K	T	HSTT
438	438.0	Gioi	7.5	7.1	7.7	8.1	7.9	8.3	6.0	K	T	HSTT
439	439.0	Gioi	8.4	8.0	8.0	8.8	7.8	8.0	6.7	G	T	HSG
440	440.0	Kha	7.3	6.8	7.4	7.8	7.8	8.0	5.9	K	T	HSTT
441	441.0	Gioi	7.7	7.8	7.8	7.8	7.8	8.0	6.5	K	T	HSTT
442	442.0	Gioi	8.3	8.8	7.9	8.7	8.3	8.2	8.0	G	T	HSG
443	443.0	Kha	7.3	7.0	7.5	8.0	7.8	8.0	5.8	K	T	HSTT
444	444.0	Kha	7.3	7.0	7.6	7.6	7.1	7.4	5.5	K	T	HSTT
445	445.0	Gioi	7.9	7.2	7.6	8.0	7.8	8.0	6.7	K	T	HSTT
446	446.0	Kha	7.9	7.0	7.2	8.0	7.6	8.2	7.1	K	T	HSTT
447	447.0	Gioi	7.8	6.8	7.2	8.1	7.8	8.1	7.0	K	T	HSTT
448	448.0	Kha	7.9	7.1	7.5	8.1	7.6	7.7	6.1	K	T	HSTT
449	449.0	Kha	7.7	7.3	7.6	8.1	7.7	7.7	6.1	K	T	HSTT
450	450.0	Kha	8.0	6.5	7.1	7.6	7.4	7.2	5.5	K	T	HSTT
451	451.0	TB	7.2	7.1	6.9	7.5	7.4	7.3	5.3	K	T	HSTT
452	452.0	Kha	7.4	6.7	7.2	7.2	6.8	7.6	5.0	K	T	HSTT
453	453.0	TB	5.7	6.3	7.3	6.8	6.6	7.0	4.4	TB	K	
454	454.0	Kha	7.7	7.0	7.4	8.5	7.9	8.2	6.0	K	T	HSTT
455	455.0	Kha	7.3	7.2	7.4	8.0	7.6	7.8	6.5	K	T	HSTT
456	456.0	Kha	7.3	6.7	7.0	7.2	7.4	7.6	5.0	K	T	HSTT
457	457.0	Gioi	8.1	7.9	7.8	8.3	8.1	8.0	7.3	G	T	HSG
458	458.0	Gioi	8.0	7.7	7.4	8.2	7.7	7.7	6.6	K	T	HSTT
459	459.0	Gioi	7.9	6.9	7.0	8.1	7.4	7.6	6.2	K	T	HSTT
460	460.0	Gioi	7.9	6.9	7.4	7.9	8.0	8.0	6.5	K	T	HSTT

Bảng 3.2 Bảng dữ liệu Kết quả học tập của học sinh

Mã khối	Tên khối	Môn thi
A	Khối A	Toán, Lý, Hóa
A1	Khối A1	Toán, Lý (đề thi khối A), Tiếng Anh
B	Khối B	Sinh, Toán, Hóa
C	Khối C	Văn, Sử, Địa
D1	Khối D1	Văn, Toán, Tiếng Anh
D2	Khối D2	Văn, Toán, Tiếng Nga
D3	Khối D3	Văn, Toán, Tiếng Pháp
D4	Khối D4	Văn, Toán, Tiếng Trung
D5	Khối D5	Văn, Toán, Tiếng Đức
D6	Khối D6	Văn, Toán, Tiếng Nhật
V	Khối V	Toán, Lý (đề thi khối A), Vẽ mỹ thuật
V1	Khối V1	Toán, Ngữ văn (đề thi khối D), Vẽ mỹ thuật
T	Khối T	Sinh, Toán (đề thi khối B), Năng khiếu TDTT
M	Khối M	Văn, Toán (đề thi khối D), Năng khiếu
N	Khối N	Văn (đề thi khối C), 2 môn năng khiếu Nhạc
H	Khối H	Văn (đề thi khối C), Năng khiếu - Mỹ thuật
H1	Khối H1	Toán, Ngữ văn (thi đề khối D), Vẽ Trang trí màu
R	Khối R	Văn, Sử (đề thi khối C), Năng khiếu báo chí
S	Khối S	Văn (đề thi khối C), 2 môn năng khiếu Điện ảnh
K	Khối K	Toán, Lý, môn kỹ thuật nghề

Bảng 3.3 Khối thi-môn thi

Mã Ngành	Tên Ngành
D15	Khoa học giáo dục và đào tạo giáo viên
D1501	Khoa học giáo dục
D150101	Giáo dục học
D150115	Quản lý giáo dục
D1502	Đào tạo giáo viên
D150201	Giáo dục mầm non
D150202	Giáo dục Tiểu học
D150203	Giáo dục Đặc biệt
D150204	Giáo dục Công dân
D150205	Giáo dục Chính trị
D150206	Giáo dục Thể chất
D150207	Huấn luyện thể thao*
D150208	Giáo dục Quốc phòng - An ninh

D150209	Sư phạm Toán học
D150210	Sư phạm Tin học
D150211	Sư phạm Vật lý
D150212	Sư phạm Hoá học
D150213	Sư phạm Sinh học
D150215	Sư phạm Kỹ thuật công nghiệp
D150215	Sư phạm Kỹ thuật nông nghiệp
D150217	Sư phạm Ngữ văn
D150218	Sư phạm Lịch sử
D150219	Sọ phạm Địa lý
D150221	Sư phạm Âm nhạc
D150222	Sư phạm Mỹ thuật
D150223	Sư phạm Tiếng Bana
50	Trình độ cao đẳng nghiệp
50150201	Sư phạm dạy nghề
50210101	Kỹ thuật điêu khắc gỗ
D150227	Sư phạm Tiếng H'mong
D150228	Sư phạm Tiếng Chăm
D150229	Sư phạm Tiếng M'nông
D150230	Sư phạm Tiếng Xêđăng
D150231	Sư phạm Tiếng Anh
D150232	Sư phạm Tiếng Nga
D150233	Sư phạm Tiếng Pháp
…….	…….
……..	……

Bảng 3.4 Dữ liệu ngành nghề

STT	Mã trường	Tên trường
1	A	ĐẠI HỌC QUỐC GIA HÀ NỘI
2	QHI	TRƯỜNG ĐẠI HỌC CÔNG NGHỆ
3	QHT	TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN
4	QHX	TRƯỜNG ĐẠI HỌC KHOA HỌC XÃ HỘI VÀ NHÂN VĂN
5	QHF	TRƯỜNG ĐẠI HỌC NGOẠI NGỮ
6	QHE	TRƯỜNG ĐẠI HỌC KINH TẾ

…	…	…
…	…	…
…	…	…
478	CYA	TRƯỜNG CAO ĐẲNG Y TẾ ĐỒNG THÁP
479	CYG	TRƯỜNG CAO ĐẲNG Y TẾ KIÊN GIANG
480	CYV	TRƯỜNG CAO ĐẲNG Y TẾ TIỀN GIANG
481	YTV	TRƯỜNG CAO ĐẲNG Y TẾ TRÀ VINH

Bảng 3.5. Dữ liệu trường đại học, cao đẳng trên cả nước

STT	Mã trường	Tên cơ sở đào tạo	Địa chỉ	Cơ quản chủ quản
		Khối địa phương		
1	HNIVC	CĐN Công Nghiệp	Q.Đống Đa	UBND TP
2	CDT0124	CĐN Kỹ Thuật Nghiệp Vụ	Q. Cầu Giấy	UBND TP
3	HACTECH	CĐN Bách Khoa	Q. Hai Bà Trưng	UBND TP
4	CDT0104	CĐN Kỹ Thuật Công Nghệ	H. Đông Anh	UBND TP
5	CDD0126	CĐN Việt Nam – Hàn Quốc	H. Đông Anh	UBND TP
6	CĐT0105	CĐ Cơ Điện	Q. Cầu Giấy	UBND TP

Bảng 3.6. Dữ liệu trường cao đẳng nghề tại Hà Nội

3.2.2. Tạo kho dữ liệu tư vấn hướng nghiệp

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 như bảng 3.2 được chuyển về dạng dữ liệu rời rạc.

3.2.3 Tạo kho dữ liệu tư vấn hướng nghiệp

Làm sạch dữ liệu.

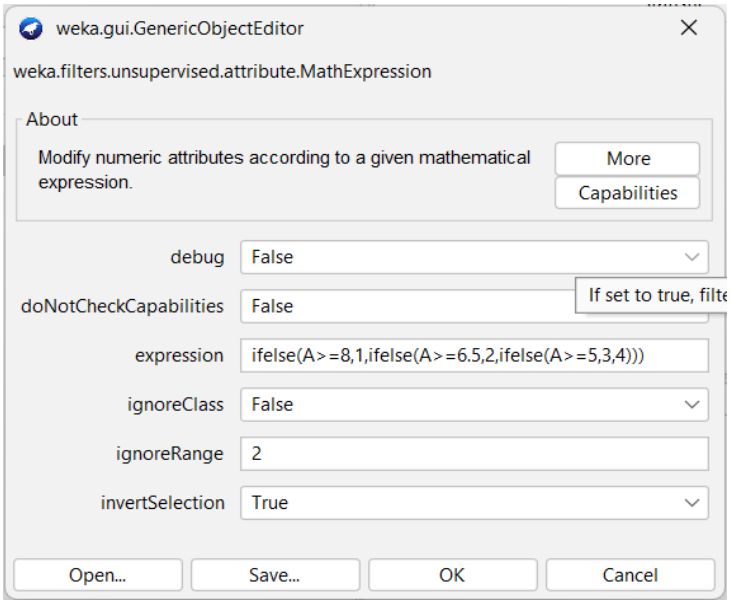
Dữ liệu sau khi thu thập gồm 20550 bản ghi, xóa bỏ những bản ghi không chứa đủ thông tin của các thuộc tính để phân lớp dữ liệu còn lại 20303 bản ghi và 12 thuộc tính.

Rời rạc hóa dữ liệu.

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. Ví dụ thực hiện rời rạc hóa dữ liệu cho thuộc tính Toan (điểm toán) với phần mềm Weka Chuyển đổi kiểu dữ liệu của thuộc tính Toán thành kiểu Nominal với 4 giá trị: Yeu (Yếu), TB (Trung bình), Kha (Khá), Gioi (Giỏi).

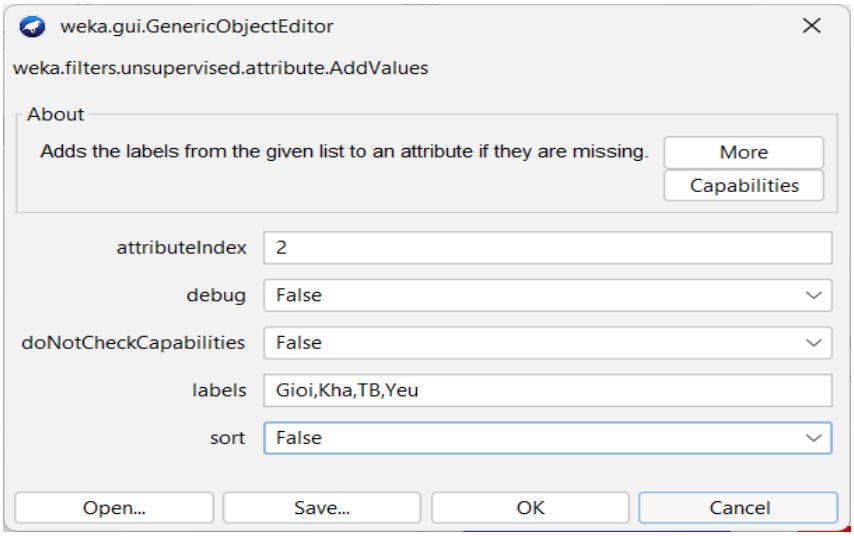
Bước 1: Thiết lập thông số cho các giá trị mới Chọn mục MathExpression trong cây thư mục rồi thiết lập các thông số:



Hình 3.2 Thiết lập thông số cho giá trị mới trong weka

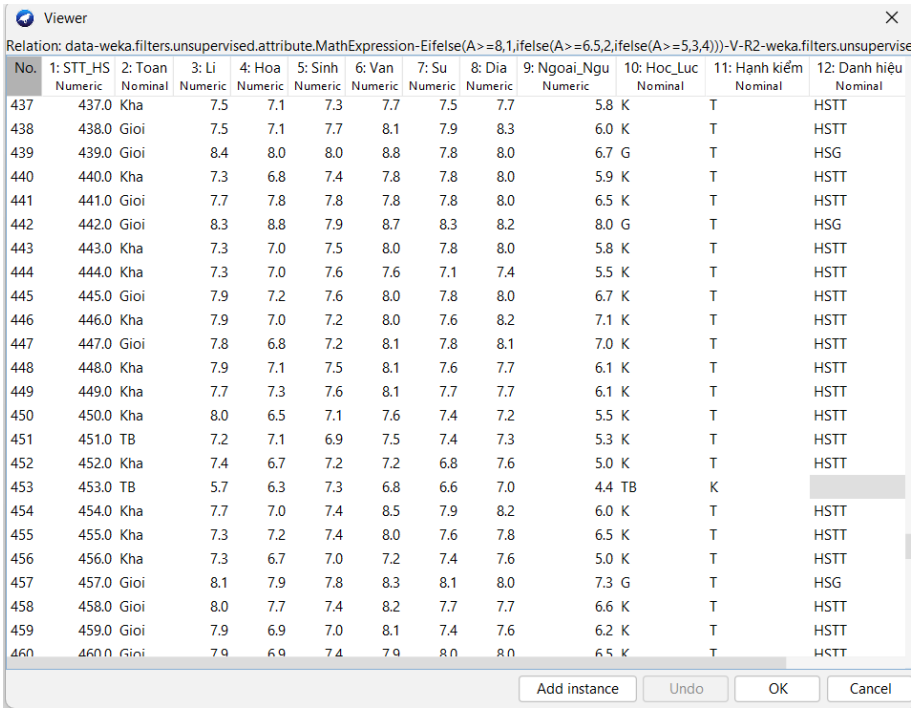
Bước 2: Chuyển đổi kiểu dữ liệu sang kiểu Nominal.

Bước 3: Thêm các giá trị mới.



Hình 3.3 Thêm giá trị mới cho thuộc tính rời rạc

Dữ liệu sau khi rời rạc hóa thuộc tính Toan như sau:



Hình 3.4 Thuộc tính "toan" sau khi rời rạc

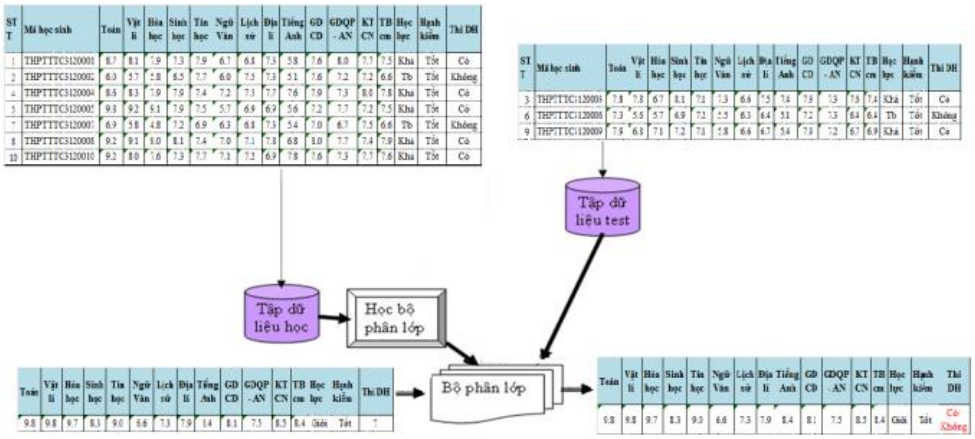
Với các thuộc tính Ly, Hoa, Sinh, Nguvan, Lichsu, Dia, Ngoaingu, TBcm tiến hành rời rạc tương tự như thuộc tính Toan.

3.2.4 Khai phá dữ liệu phát hiện tri thức

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.



Hình 3.5 Mô hình dự đoán thi đại học

Thực hiện phân lớp dữ liệu bằng cây quyết định với thuật toán ID3 trong phần mềm Weka thu được cây quyết định và tập luật sau:

=== Run information ===

Scheme: weka.classifiers.trees.Id3

Relation: Holland

Instances: 11203

Attributes: 9

Toan

Li

Hoa

Sinh

Van

Su

Dia

Ngoai_Ngu

Nganh_Holland

Test mode: 10-fold cross-validation

=== Classifier model (full training set) ===

Id3

Toan = Gioi

| Dia = Gioi

| | Ngoai_Ngu = Gioi

| | | Su = Gioi

| | | | Li = Gioi

| | | | | Van = Gioi

| | | | | | Hoa = Gioi

| | | | | | | Sinh = Gioi: Kinh doanh

| | | | | | | Sinh = Kha: Kinh doanh

| | | | | | | Sinh = TB: null

| | | | | | | Sinh = Yeu: null

| | | | | | Hoa = Kha

| | | | | | | Sinh = Gioi: Kinh doanh

| | | | | | | Sinh = Kha: Kinh doanh

| | | | | | | Sinh = TB: null

| | | | | | | Sinh = Yeu: null

| | | | | | Hoa = TB: null

| | | | | | Hoa = Yeu: null

| | | | | Van = Kha

| | | | | | Sinh = Gioi

| | | | | | | Hoa = Gioi: Kinh doanh

| | | | | | | Hoa = Kha: Kinh doanh

| | | | | | | Hoa = TB: null

| | | | | | | Hoa = Yeu: null

| | | | | | Sinh = Kha

| | | | | | | Hoa = Gioi: Kinh doanh

| | | | | | | Van = Gioi: Kinh doanh

| | | | | | | Van = Kha: Kinh doanh

| | | | | | | Van = TB: null

| | | | | | | Van = Yeu: null

| | | | | | Li = TB: null

| | | | | | Li = Yeu: null

| | | | | Hoa = TB: null

| | | | | Hoa = Yeu: null

| | | | Sinh = TB: null

| | | | Sinh = Yeu: null

| | | Su = Kha

| | | | Hoa = Gioi

| | | | | Sinh = Gioi: Kinh doanh

| | | | | Sinh = Kha: Kinh doanh

| | | | | Sinh = TB: null

| | | | | Sinh = Yeu: null

| | | | Hoa = Kha

| | | | | Sinh = Gioi

| | | | | | Van = Gioi: Kinh doanh

| | | | | | Van = Kha: Kỹ thuật

| | | | | | Van = TB: null

| | | | | | Van = Yeu: null

| | | | | Sinh = Kha

| | | | | | Van = Gioi: Kỹ thuật

| | | | | | Van = Kha: Kỹ thuật

| | | | | | Van = TB: null

| | | | | | Van = Yeu: null

| | | | | Sinh = TB: null

| | | | | Sinh = Yeu: null

| | | | Hoa = TB: null

| | | | Hoa = Yeu: null

| | | Su = TB: null

| | | Su = Yeu: null

| | Ngoai_Ngu = Kha

| | | Hoa = Gioi

| | | | Sinh = Gioi: Kinh doanh

| | | | Sinh = Kha

| | | | | Li = Gioi

| | | | | | Van = Gioi: null

| | | | | | Van = Kha: Kinh doanh

| | | | | | Van = TB: Kinh doanh

| | | | | | Van = Yeu: null

| | | | | Li = Kha: Kinh doanh

| | | | | Li = TB: null

| | | | | Li = Yeu: null

| | | | Sinh = TB: null

| | | | Sinh = Yeu: null

| | | Hoa = Kha

| | | | Su = Gioi

| | | | | Li = Gioi: Kinh doanh

| | | | | Li = Kha: Nghiệp vụ

| | | | | Li = TB: null

| | | | | Li = Yeu: null

| | | | Su = Kha

| | | | | Sinh = Gioi: Xã hội

| | | | | Sinh = Kha

| | | | | | Van = Gioi

| | | | | | | Li = Gioi: Kỹ thuật

| | | | | | | Li = Kha: Xã hội

| | | | | | | Li = TB: null

| | | | | | | Li = Yeu: null

| | | | | | Van = Kha

| | | | | | | Li = Gioi: Xã hội

| | | | | | | Li = Kha: Nghiên cứu

| | | | | | | Li = TB: null

| | | | | | | Li = Yeu: null

| | | | | | Van = TB: Kỹ thuật

| | | | | | Van = Yeu: null

| | | | | Sinh = TB: Nghiệp vụ

| | | | | Sinh = Yeu: null

| | | | Su = TB: null

| | | | Su = Yeu: null

| | | Hoa = TB: null

| | | Hoa = Yeu: null

| | Ngoai_Ngu = TB

| | | Li = Gioi

| | | | Sinh = Gioi: Nghiệp vụ

| | | | Sinh = Kha: Kỹ thuật

| | | | Sinh = TB: null

| | | | Sinh = Yeu: null

| | | Li = Kha

| | | | Sinh = Gioi

| | | | | Hoa = Gioi

| | | | | | Su = Gioi: Kinh doanh

| | | | | | Su = Kha: Kỹ thuật

| | | | | | Su = TB: null

| | | | | | Su = Yeu: null

| | | | | Hoa = Kha: Kinh doanh

| | | | | Hoa = TB: null

| | | | | Hoa = Yeu: null

| | | | Sinh = Kha

| | | | | Van = Gioi: Kinh doanh

| | | | | Van = Kha: Xã hội

| | | | | Van = TB: null

| | | | | Van = Yeu: null

| | | | Sinh = TB: null

| | | | Sinh = Yeu: null

| | | Li = TB: Kỹ thuật

| | | Li = Yeu: null

| | Ngoai_Ngu = Yeu: null

| Dia = Yeu: Kỹ thuật

Toan = Kha

| Ngoai_Ngu = Gioi

| | Li = Gioi

| | | Su = Gioi

| | | | Van = Gioi: Xã hội

| | | | Van = Kha: Kỹ thuật

| | | | Van = TB: null

| | | | Van = Yeu: null

| | | Su = Kha

| | | | Sinh = Gioi: Kinh doanh

| | | | Sinh = Kha

| | | | | Van = Gioi: Kinh doanh

| | | | | Van = Kha: Xã hội

| | | | | Van = TB: null

| | | | | Van = Yeu: null

| | | | Sinh = TB: null

| | | | Sinh = Yeu: null

| | | Su = TB: null

| | | Su = Yeu: null

| | Li = Kha

| | | Van = Gioi

| | | | Sinh = Gioi: Nghiên cứu

| | | | Sinh = Kha

| | | | | Su = Gioi: Nghiệp vụ

| | | | | Su = Kha: Kinh doanh

| | | | | Su = TB: null

| | | | | Su = Yeu: null

| | | | Sinh = TB: null

| | | | Sinh = Yeu: null

| | | Van = Kha

| | | | Dia = Gioi

| | | | | Hoa = Gioi: null

| | | | | Hoa = Kha: Xã hội

| | | | | Hoa = TB: Kinh doanh

| | | | | Hoa = Yeu: null

| | | | Dia = Kha: Nghiệp vụ

| | | | Dia = Yeu: null

| | | Van = TB: Nghiệp vụ

| | | Van = Yeu: null

| | Li = TB: Kinh doanh

| | Li = Yeu: null

| Ngoai_Ngu = Kha

| | Van = Gioi

| | | Hoa = Gioi: Nghiệp vụ

| | | Hoa = Kha

| | | | Li = Gioi: Nghiên cứu

| | | | Li = Kha

| | | | | Dia = Gioi: Kinh doanh

| | | | | Dia = Kha: Kinh doanh

| | | | | Dia = Yeu: null

| | | | Li = TB: null

| | | | Li = Yeu: null

| | | Hoa = TB: null

| | | Hoa = Yeu: null

| | Van = Kha

| | | Li = Gioi

| | | | Dia = Gioi

| | | | | Sinh = Gioi

| | | | | | Su = Gioi: Kinh doanh

| | | | | | Su = Kha

| | | | | | | Hoa = Gioi: Xã hội

| | | | | Li = Gioi: null

| | | | | Li = Kha: Kinh doanh

| | | | | Li = TB

| | | | | | Hoa = Gioi: null

| | | | | | Hoa = Kha: Kinh doanh

| | | | | | Hoa = TB: Kinh doanh

| | | | | | Hoa = Yeu: null

| | | | | Li = Yeu: null

| | | | Sinh = TB: null

| | | | Sinh = Yeu: null

| | | Van = TB: Xã hội

| | | Van = Yeu: null

| | Dia = Yeu: null

| Ngoai_Ngu = TB

| | Sinh = Gioi: null

| | Sinh = Kha: Xã hội

| | Sinh = TB: Kỹ thuật

| | Sinh = Yeu: null

| Ngoai_Ngu = Yeu: Xã hội

Toan = Yeu

| Li = Gioi: null

| Li = Kha: Xã hội

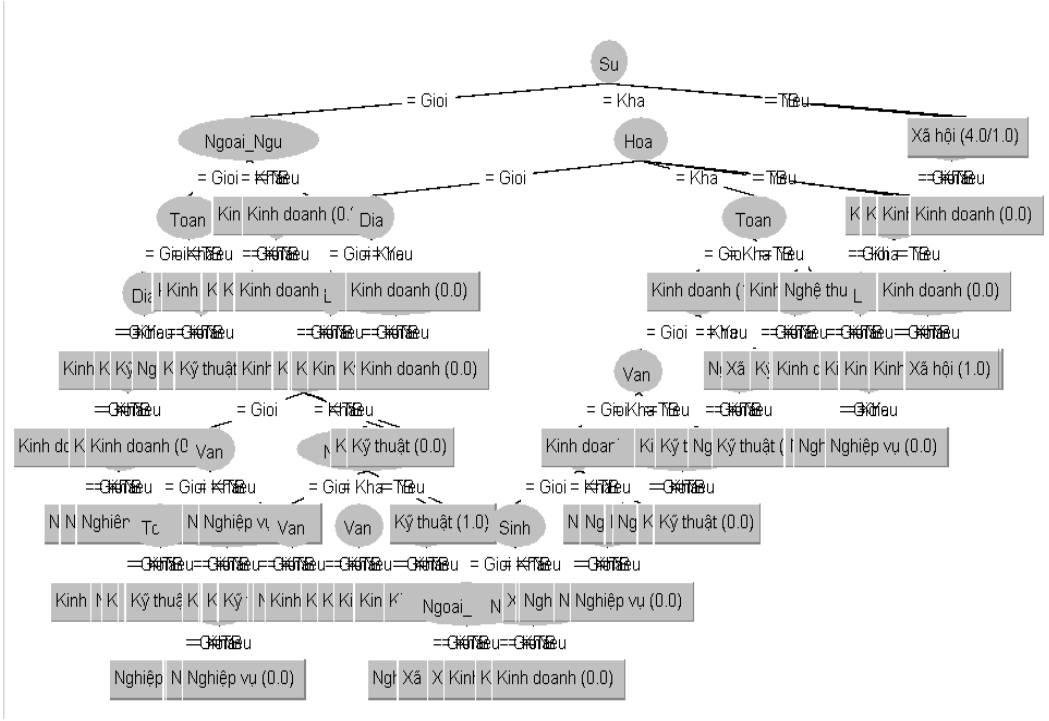
| Li = TB: null

| Li = Yeu: Xã hội

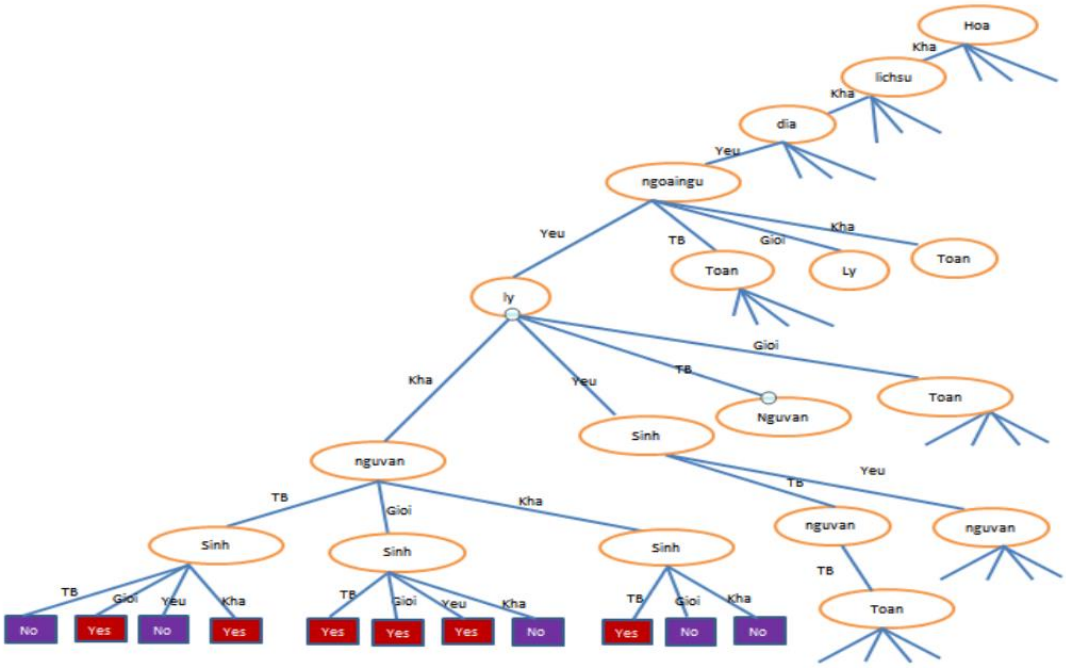
Time taken to build model: 0 seconds

=== Stratified cross-validation ===

=== Summary ===



Hình 3.6 Cây quyết định đầy đủ với thuộc tính Thidh



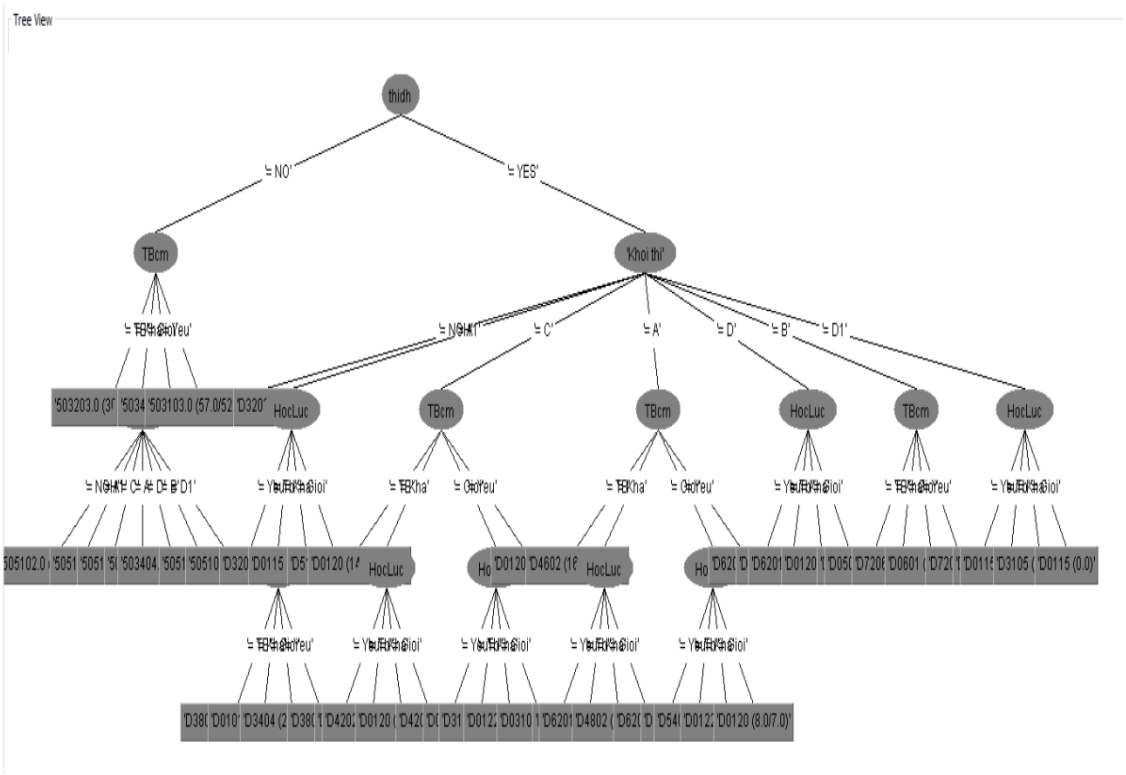
Hình 3.7 Nhánh trái cây quyết định Thidh

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

Phân lớp dữ liệu với các thuộc tính toán, vật lý, hóa học, sinh học, ngữ văn, lịch sử, địa lý, tiếng anh, ThiDH, khối thi, ngành nghề:

STT_HS	Toan	Li	Hoa	Sinh	Van	Su	Dia	Ngoai_Ngu	Hoc_Luc	Nganh_Holland
1	8.9	9.5	8.6	8.3	6.5	7.8	8.8	6.8	G	Kỹ thuật
2	9.3	9.5	9.3	7.8	6.9	7.7	9.1	7.1	G	Kinh doanh
3	9.3	9.8	9.3	8.2	8.3	8.0	9.3	6.9	G	Kinh doanh
4	8.8	9.7	7.6	8.2	6.7	8.1	8.8	8.2	G	Nghệ thuật
5	9.0	8.2	9.5	9.3	6.6	7.8	8.9	6.8	G	Kỹ thuật
6	8.8	9.2	8.4	8.1	7.2	8.2	9.0	7.6	G	Nghiệp vụ
7	8.9	9.7	9.5	8.6	7.4	7.7	9.1	7.2	G	Kinh doanh
8	8.7	9.2	8.9	7.5	7.1	7.7	8.9	6.8	G	Nghiên cứu
9	8.6	9.5	8.9	8.3	6.8	8.0	9.1	6.6	G	Kỹ thuật
10	8.7	8.1	7.8	7.5	8.1	7.8	8.9	9.2	G	Nghiệp vụ
11	8.8	9.4	9.6	8.5	7.5	8.0	8.9	6.9	G	Nghiệp vụ
12	8.4	8.3	7.5	8.2	6.9	7.8	9.0	6.9	G	Xã Hội

Hình 3.8 Mô hình dự đoán khối thi, ngành nghề



Hình 3.9 Cây quyết định đầy đủ

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

Thuộc tính Toan có độ lợi thông tin cao nhất nên ta chọn thuộc tính này

Xét 4 trường hợp:

Toan = Gioi: H(x) = 2.317

Thuộc tính	Entropy	Gain	SplitInformation	Gain Ratio
Ly	2.298	0.019	0.837	0.023
Hoa	2.286	0.031	1.029	0.030
Sinh	2.307	0.010	1.047	0.009
Van	2.297	0.019	1.083	0.018
Su	2.290	0.027	0.860	0.032
Dia	2.276	0.041	0.766	0.053
Ngoai_Ngu	2.290	0.027	1.417	0.019

Thuộc tính Dia có độ lợi thông tin cao nhất (0.041) nên ta chọn thuộc tính này, phân tích tiếp ta thu được sơ đồ cây:

Dia = Gioi

| Ngoai_Ngu = Gioi

| | Su = Gioi

| | | Li = Gioi

| | | | Van = Gioi

| | | | | Hoa = Gioi

| | | | | | Sinh = Gioi: Kinh doanh

| | | | | | Sinh = Kha: Kinh doanh

| | | | | Hoa = Kha

| | | | | | Sinh = Gioi: Kinh doanh

| | | | | | Sinh = Kha: Kinh doanh

| | | | Van = Kha

| | | | | Sinh = Gioi

| | | | | | Hoa = Gioi: Kinh doanh

| | | | | | Hoa = Kha: Kinh doanh

| | | | | Sinh = Kha

| | | | | | Hoa = Gioi: Kinh doanh

| | Su = Kha

| | | Hoa = Gioi

| | | | Sinh = Gioi: Kinh doanh

| | | | Sinh = Kha: Kinh doanh

| | | Hoa = Kha

| | | | Sinh = Gioi

| | | | | Van = Gioi: Kinh doanh

| | | | | Van = Kha: Kỹ thuật

| | | | Sinh = Kha

| | | | | Van = Gioi: Kỹ thuật

| | | | | Van = Kha: Kỹ thuật

| Ngoai_Ngu = Kha

| | Hoa = Gioi

| | | Sinh = Gioi: Kinh doanh

| | | Sinh = Kha

| | | | Li = Gioi

| | | | | Van = Kha: Kinh doanh

| | | | | Van = TB: Kinh doanh

| | | | Li = Kha: Kinh doanh

| | Hoa = Kha

| | | Su = Gioi

| | | | Li = Gioi: Kinh doanh

| | | | Li = Kha: Nghiệp vụ

| | | Su = Kha

| | | | Sinh = Gioi: Xã hội

| | | | Sinh = Kha

| | | | | Van = Gioi

| | | | | | Li = Gioi: Kỹ thuật

| | | | | | Li = Kha: Xã hội

| | | | | Van = Kha

| | | | | | Li = Gioi: Xã hội

| | | | | | Li = Kha: Nghiên cứu

| | | | | Van = TB: Kỹ thuật

| | | | Sinh = TB: Nghiệp vụ

| Ngoai_Ngu = TB

| | Li = Gioi

| | | Sinh = Gioi: Nghiệp vụ

| | | Sinh = Kha: Kỹ thuật

| | Li = Kha

| | | Sinh = Gioi

| | | | Hoa = Gioi

| | | | | Su = Gioi: Kinh doanh

| | | | | Su = Kha: Kỹ thuật

| | | | Hoa = Kha: Kinh doanh

| | | Sinh = Kha

| | | | Van = Gioi: Kinh doanh

| | | | Van = Kha: Xã hội

| | Li = TB: Kỹ thuật

Dia = Yeu: Kỹ thuật

Toan = Kha: H(x) = 2.341

Thuộc tính	Entropy	Gain	SplitInformation	Gain Ratio
Ly	2.322	0.019	0.920	0.021
Hoa	2.302	0.039	0.596	0.066
Sinh	2.335	0.016	0.673	0.023
Van	2.314	0.028	0.964	0.029
Su	2.312	0.029	0.428	0.068
Dia	2.332	0.010	0.985	0.010
Ngoai_Ngu	2.300	0.042	1.386	0.030

Thuộc tính Ngoai_Ngu có độ lợi thông tin cao nhất nên ta chọn thuộc tính này, phân tích tiếp ta thu được sơ đồ cây như sau:

Ngoai_Ngu = Gioi

| Li = Gioi

| | Su = Gioi

| | | Van = Gioi: Xã hội

| | | Van = Kha: Kỹ thuật

| | Su = Kha

| | | Sinh = Gioi: Kinh doanh

| | | Sinh = Kha

| | | | Van = Gioi: Kinh doanh

| | | | Van = Kha: Xã hội

| Li = Kha

| | Van = Gioi

| | | Sinh = Gioi: Nghiên cứu

| | | Sinh = Kha

| | | | Su = Gioi: Nghiệp vụ

| | | | Su = Kha: Kinh doanh

| | Van = Kha

| | | Dia = Gioi

| | | | Hoa = Kha: Xã hội

| | | | Hoa = TB: Kinh doanh

| | | Dia = Kha: Nghiệp vụ

| | Van = TB: Nghiệp vụ

| Li = TB: Kinh doanh

Ngoai_Ngu = Kha

| Van = Gioi

| | Hoa = Gioi: Nghiệp vụ

| | Hoa = Kha

| | | Li = Gioi: Nghiên cứu

| | | Li = Kha

| | | | Dia = Gioi: Kinh doanh

| | | | Dia = Kha: Kinh doanh

| Van = Kha

| | Li = Gioi

| | | Dia = Gioi

| | | | Sinh = Gioi

| | | | | Su = Gioi: Kinh doanh

| | | | | Su = Kha

| | | | | | Hoa = Gioi: Xã hội

| | | | Li = Kha: Kinh doanh

| | | | Li = TB

| | | | | Hoa = Kha: Kinh doanh

| | | | | Hoa = TB: Kinh doanh

| | Van = TB: Xã hội

Ngoai_Ngu = TB

| Sinh = Kha: Xã hội

| Sinh = TB: Kỹ thuật

Ngoai_Ngu = Yeu: Xã hội

Toan = TB: H(x) = 2.195

Thuộc tính	Entropy	Gain	SplitInformation	Gain Ratio
Ly	1.997	0.198	0.918	0.216
Hoa	2.057	0.138	0.792	0.174
Sinh	1.837	0.358	0.549	0.653
Van	1.982	0.214	0.702	0.304
Su	2.135	0.061	0.276	0.219
Dia	1.897	0.298	0.959	0.311
Ngoai_Ngu	1.744	0.451	1.221	0.369

Thuộc tính Ngoai_Ngu có độ lợi thông tin cao nhất nên ta chọn thuộc tính này. Xảy ra 4 trường hợp:

Toan = TB, Ngoai_Ngu = Gioi: Nganh_Holland = Nghiệp vụ

Toan = TB, Ngoai_Ngu = Kha: H(x) = 2.225

Thuộc tính	Entropy	Gain	SplitInformation	Gain Ratio
Ly	2.035	0.190	0.971	0.196
Hoa	2.068	0.157	0.722	0.218
Sinh	2.006	0.220	0.353	0.623
Van	1.983	0.243	0.567	0.429
Su	2.113	0.113	0.353	0.319
Dia	1.870	0.355	0.837	0.425

Thuộc tính Dia có độ lợi thông tin cao nhất nên ta chọn thuộc tính này. Có hai trường hợp xảy ra:

Toan = TB, Ngoai_Ngu = Kha, Dia = “Gioi. Phân tích tiếp ta sẽ có:

- Ly = Kha: Nganh_Holland = Nghiệp vụ

- Ly = TB: Nganh_Holland = Xã hội

Toan = TB, Ngoai_Ngu = Kha, Dia = Kha. Phân tích tiếp ta sẽ có sơ đồ cây như sau:

Van = Kha

| Sinh = Gioi: Nghiên cứu

| Sinh = Kha

| | Ly = Kha: Kỹ thuật

| | Ly = TB

| | | Hoa = Kha: Kinh doanh

| | | Hoa = TB: Kinh doanh

Van = TB: Xã hội

Toan = TB, Ngoai_Ngu = TB: H(x) = 0.811

Thuộc tính	Entropy	Gain	SplitInformation	Gain Ratio
Ly	0.811	0	0	-
Hoa	0.811	0	0	-
Sinh	0	0.811	0.811	1
Van	0.500	0.311	1	0.311
Su	0.811	0	0	-
Dia	0.689	0.122	0.811	0.151

Thuộc tính Sinh có độ lợi thông tin cao nhất nên ta chọn thuộc tính này. Xảy ra hai trường hợp:

- Sinh = Kha: Nganh_Holland = Xã hội

- Sinh = TB: Nganh_Holland = Kỹ thuật

Toan = TB, Ngoai_Ngu = Yeu: Nganh_Holland = Xã hội

Toan = Yeu: H(x) = 0.811

Thuộc tính	Entropy	Gain	SplitInformation	Gain Ratio
Ly	0.689	0.122	0.811	0.151
Hoa	0.811	0	0	-
Sinh	0.811	0	0	-
Van	0.689	0.122	0.811	0.151
Su	0.811	0	0	-
Dia	0.811	0	0	-
Ngoai_Ngu	0.811	0	0	-

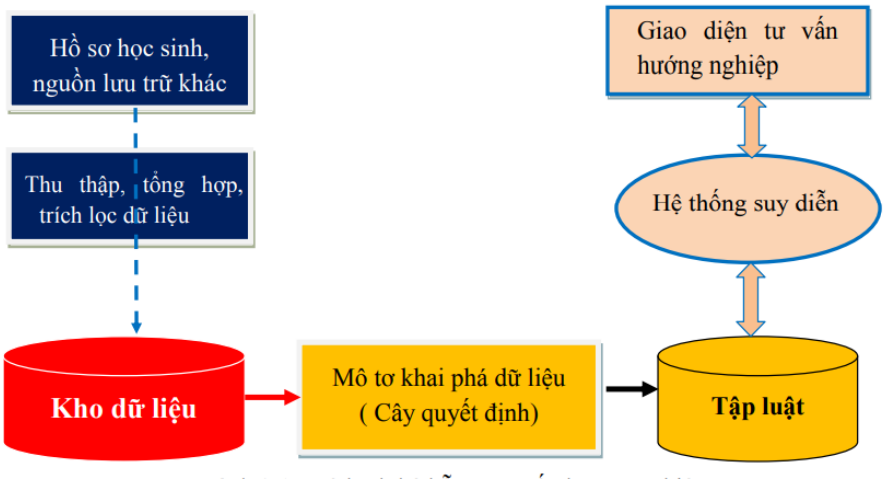
Thuộc tính Ly và Van đều có độ lợi thông tin cao nhất. Chọn thuộc tính Ly, xảy ra hai trường hợp:

- Ly = Kha: Nganh_Holland = Xã hội

- Ly = Yeu: Nganh_Holland = Xã hội

3.3 Cài đặt và thử nghiệm

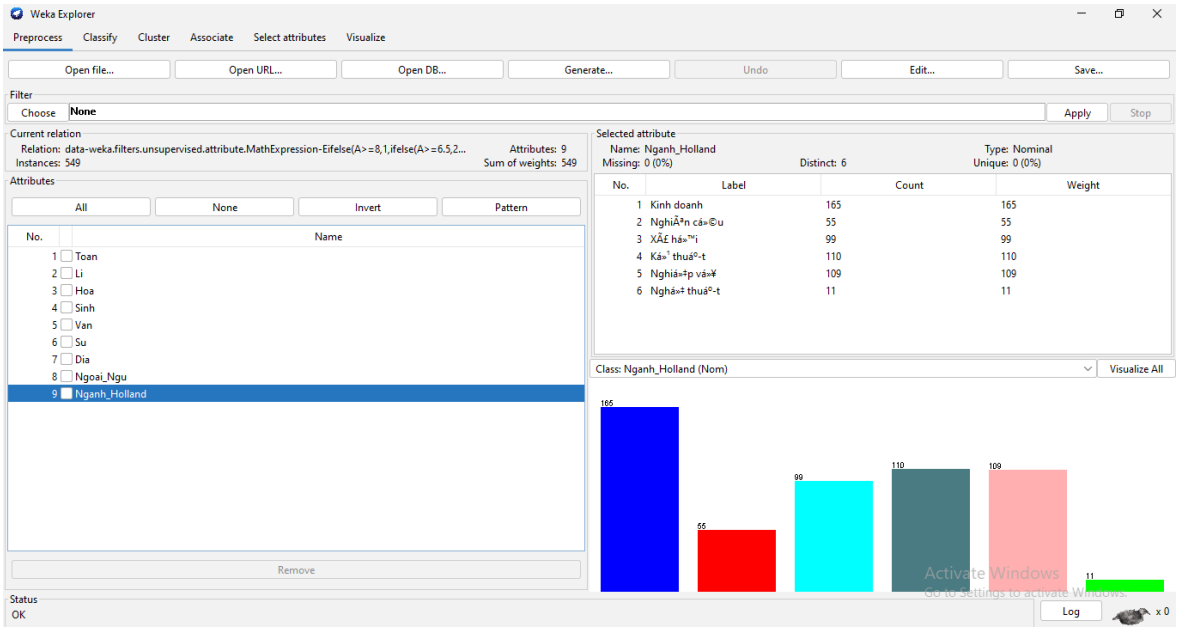
3.3.1 Mô hình hệ hỗ trợ tư vấn hướng nghiệp



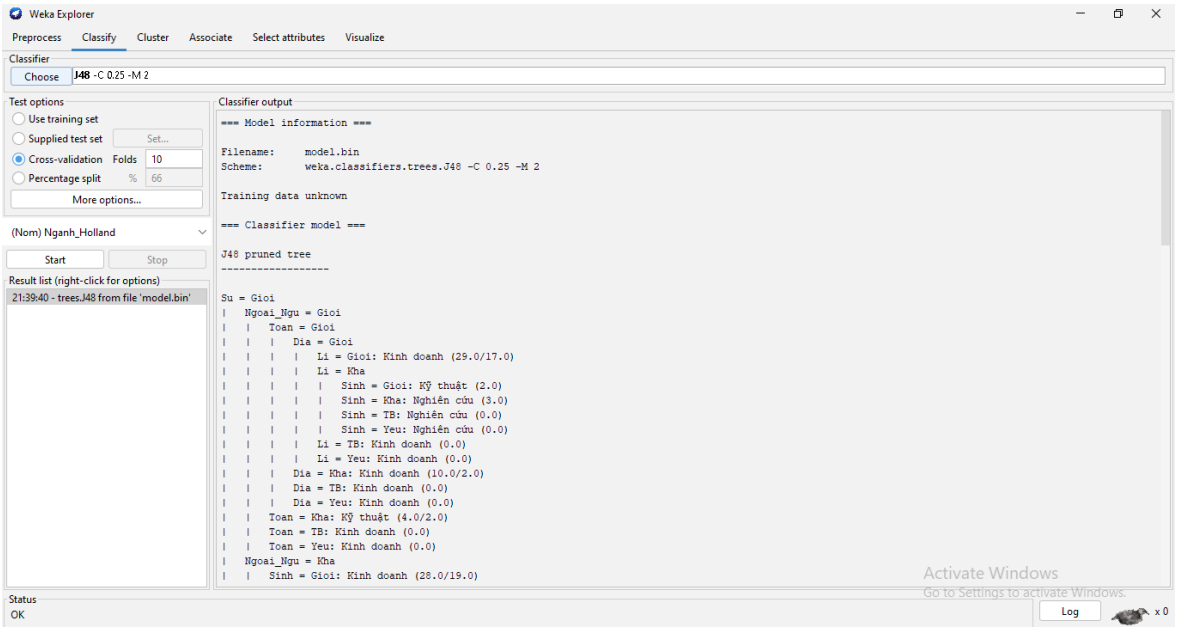
Hình 3.10 Mô hình hệ hỗ trợ tư vấn hướng nghiệp

Họ tên	Toán	Vật lí	Hóa học	Sinh học	Ngữ Văn	Lịch sử	Địa lí	Ngoại ngữ 1	Năng khiếu	Học lực	Hạnh kiểm
Nguyễn Thành An	G	G	G	G	K	K	G	K	G	G	T
Đinh Thị Mai Anh	G	G	G	K	K	K	G	K	G	G	T
Mai Thị Thục Anh	G	G	G	G	G	G	G	K	G	G	T
Kim Đức Dũng	G	G	K	G	K	G	G	G	G	G	T
Nguyễn Đình Dũng	G	G	G	G	K	K	G	K	G	G	T
Phùng An Duy	G	G	G	G	K	G	G	K	G	G	T
Trần Thị Thu Hà	G	G	G	G	K	K	G	K	G	G	T
Kim Thị Hồng Hạnh	G	G	G	K	K	K	G	K	G	G	T
Nguyễn Thị Thu Hằng	G	G	G	G	K	G	G	K	G	G	T
Đinh Thị Mai Hoa	G	G	K	K	G	K	G	G	G	G	T
Đinh Thị Ánh Hồng	G	G	G	G	K	G	G	K	G	G	T
Đặng Thị Mỹ Huệ	G	G	K	G	K	K	G	K	G	G	T
Nguyễn Ngọc Huyền	G	G	G	G	K	G	G	G	G	G	T
Nguyễn Tuyết Thương Huyền	G	G	G	G	K	G	G	K	G	G	T
Phạm Thị Ngọc Huyền	G	G	G	G	K	G	G	K	G	G	T
Nguyễn Thu Hương	G	G	G	K	K	K	G	K	G	G	T
Nguyễn Duy Khánh	G	G	G	K	TB	K	G	TB	G	K	T
Lê Thị Thùy Linh	G	G	K	G	K	G	G	K	G	G	T
Nguyễn Hải Linh	G	G	G	G	K	K	G	K	G	G	T

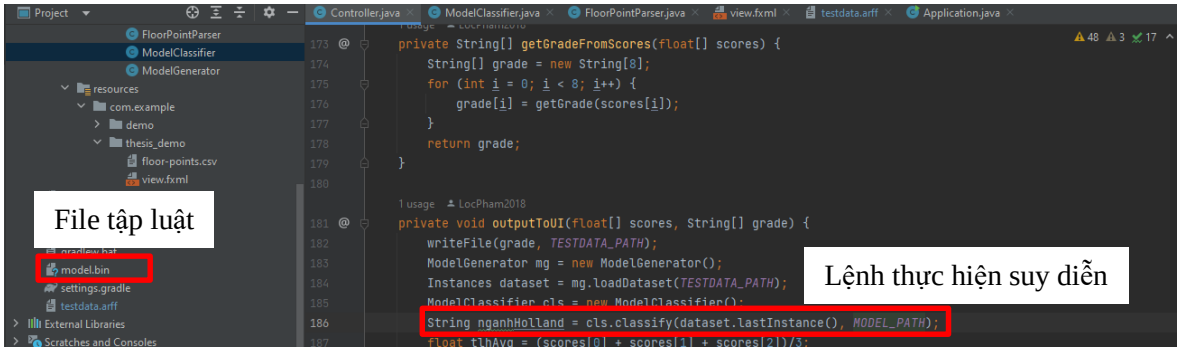
Hình 3.11 Kho dữ liệu



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

3.3.2 Chức năng của 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.

3.3.3 Chuẩn bị và thiết kế CSDL

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 gồm 20.550 bản ghi và 30 thuộc tính. Sau khi trích lọc và làm sạch, dữ liệu còn lại gồm 20.430 bản ghi và 12 thuộc tính.

Chia dữ liệu nguồn thành 2 tập dữ liệu:

- Dữ liệu huấn luyện: 15.700 bản ghi chiếm 70% dữ liệu ban đầu.

- Dữ liệu kiểm tra: 6.730 bản ghi chiếm 30% dữ liệu ban đầu.

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 bao gồm Mã học sinh, toán, vật lý, hóa học, sinh học, ngữ văn, lịch sử, địa lý, tiếng anh,

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.

3.3.4 Công nghệ sử dụng

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.

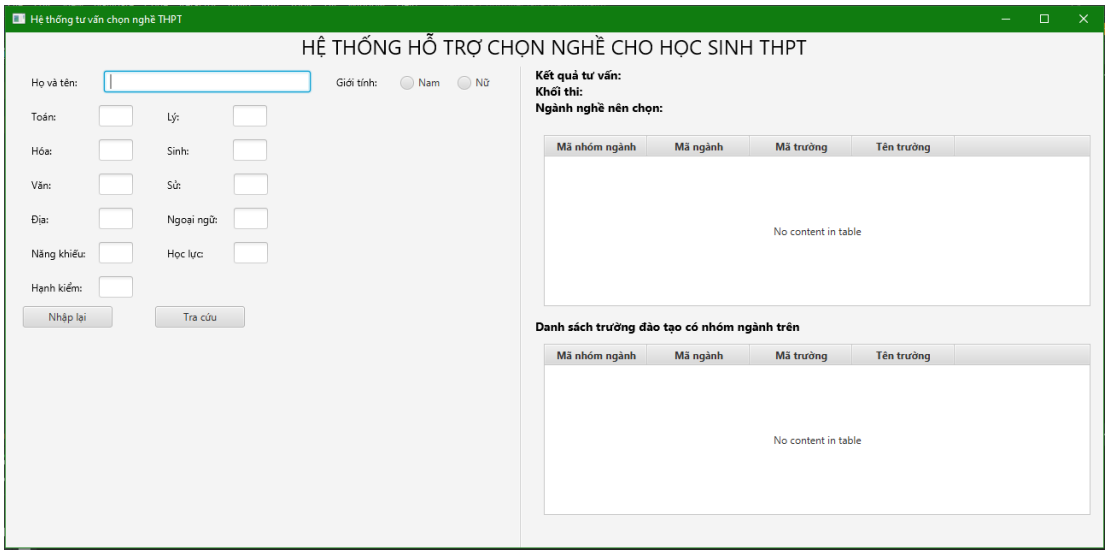
- Microsoft Excel 2022 để lưu trữ kho dữ liệu.

- 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.

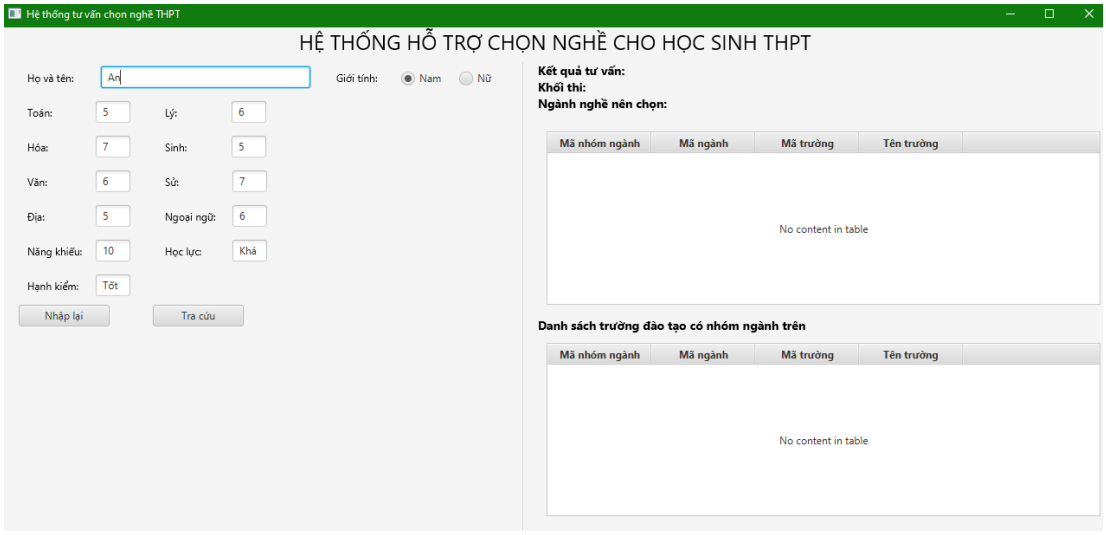
3.3.5 Giao diện hệ hỗ trợ tư vấn hướng nghiệp

3.3.5.1 Phần giao diện tư vấn hướng nghiệp

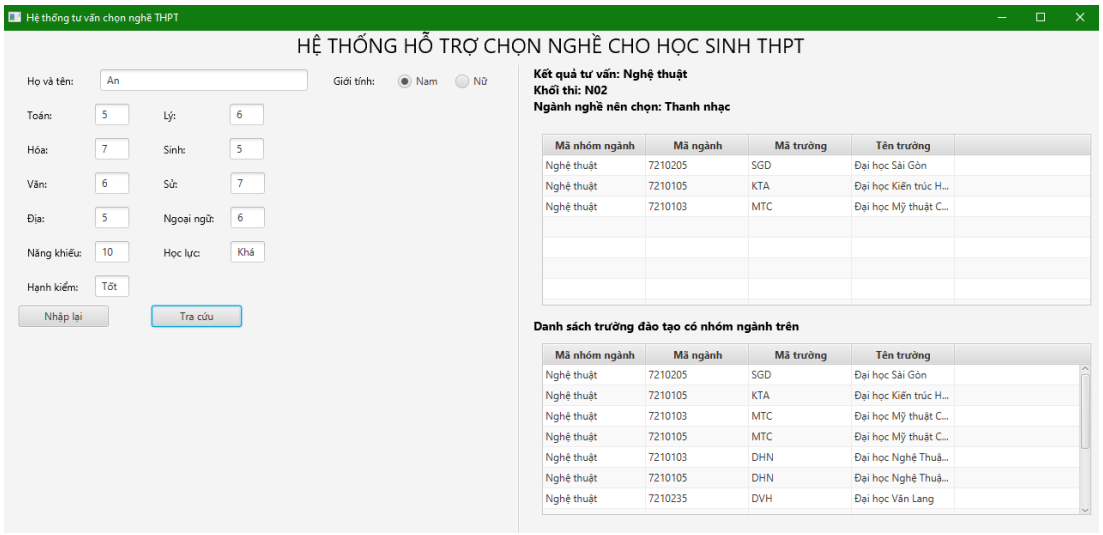
Tương tác với hệ hỗ trợ tư vấn hướng nghiệp qua giao diện này người dùng cần nhập điểm tổng kết các môn học sau đó sẽ được hệ thống đưa ra các tư vấn như nên chọn khối thi nào, ngành nào, và liệt kê một số trường có đào tạo ngành nghề vừa nêu.



Hình 3.15 Giao diện màn hình trước khi tư vấn



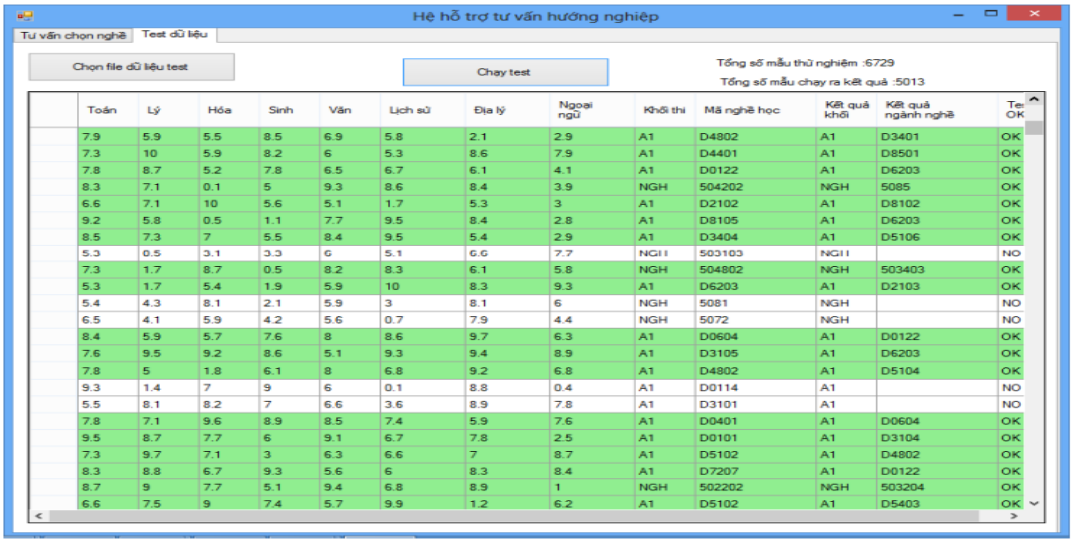
Hình 3.16 Giao diện màn hình nhập dữ liệu



Hình 3.17 Giao diện màn hình sau khi trả về kết quả

3.3.5.2. Phần giao diện test dữ liệu

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



Hình 3.18 Giao diện màn hình test dữ liệu

3.3.6 Đánh giá ưu, nhược điểm của hệ thống

3.3.6.1 Ưu điểm

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.

3.3.6.2 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.

3.3.7 Đánh giá kết quả thử nghiệm

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), "Ứng 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. Universityofllllinois, 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.

BẢN CAM ĐOAN

Tôi cam đoan đã thực hiện việc kiểm tra mức độ tương đồng nội dung luận văn qua phần mềm kiemtratailieu một cách trung thực và đạt kết quả mức độ tương đồng .....% toàn bộ nội dung luận văn. Bản luận văn kiểm tra qua phần mềm là bản cứng luận văn đã nộp để bảo vệ trước hội đồng. Nếu sai tôi xin chịu các hình thức kỷ luật theo quy định hiện hành của Học viện.

Ngày..... tháng ..... năm .....

HỌC VIÊN CAO HỌC

(Ký và ghi rõ học tên)

Trần Xuân Oanh