DATA MINING APPLICATION FOR BUILDING A SYSTEM TO SUPPORT THE SELECTION OF MAJORS AT DONG THAP UNIVERSITY

Huynh Le Uyen Minh
Dong Thap University

ARTICLE INFO

Received: 28/4/2022

Revised: 30/5/2022

Published: 31/5/2022

KEYWORDS

Admissions consulting
John Holland theory
Random forest
Data mining
Choosing a major

ABSTRACT

In this paper, we have approached data mining and John Holland's theory to build a recommendation model for major choosing at Dong Thap University. We collect data of students at the school and some students graduate on time, and then the pre-processing step is to transform the dataset into structured one, suited for the input of data mining algorithms used in the next step. The random forest model is learnt from the dataset to build the predictive model. In the system building step, we used the above model to build a suitable counseling function for candidates when they choose a major at Dong Thap University. In addition, we also built a function to help candidates in identifying the right career group for themselves. The experimental results can the improve the admissions consulting, contribute to improving the quality of studing and training, at the same time, candidates also know the admission information of the Dong Thap University more.

ỨNG DỤNG KHAI PHÁ DỮ LIỆU XÂY DỰNG MÔ HÌNH TƯ VẤN CHỌN NGÀNH HỌC TẠI TRƯỜNG ĐẠI HỌC ĐỒNG THÁP

Huỳnh Lê Uyên Minh
Trường Đại học Đồng Tháp

THÔNG TIN BÀI BÁO

Ngày nhận bài: 28/4/2022

Ngày hoàn thiện: 30/5/2022

Ngày đăng: 31/5/2022

TỪ KHÓA

Tư vấn tuyển sinh
Lý thuyết John Holland
Rừng ngẫu nhiên
Khai phá dữ liệu
Chọn ngành học

TÓM TẮT

Trong bài này, chúng tôi đã tiếp cận khai phá dữ liệu và lý thuyết mật mã John Holland để xây dựng mô hình tư vấn chọn ngành học tại Trường Đại học Đồng Tháp. Chúng tôi tiến hành thu thập dữ liệu của sinh viên đang học tại trường và một số sinh viên tốt nghiệp đúng tiến độ, sau đó thực hiện bước tiền xử lý dữ liệu, đưa dữ liệu về cấu trúc bảng. Tiếp đó, chúng tôi sử dụng giải thuật rừng ngẫu nhiên học từ dữ liệu để xây dựng mô hình dự báo. Trong bước xây dựng hệ thống, chúng tôi sử dụng mô hình thu được để xây dựng chức năng tư vấn sự phù hợp khi thí sinh chọn một ngành học tại trường Đại học Đồng Tháp. Bên cạnh đó, chúng tôi cũng xây dựng chức năng hỗ trợ thí sinh xác định nhóm ngành nghề phù hợp với cá nhân mình, dựa trên lý thuyết mật mã John Holland. Kết quả thu được có thể góp phần đẩy mạnh công tác tư vấn tuyển sinh, vừa hỗ trợ cho các thí sinh, vừa góp phần nâng cao chất lượng học tập và đào tạo, đồng thời cũng làm cầu nối để thí sinh biết đến các thông tin tuyển sinh của trường nhiều hơn.

DOI: https://doi.org/10.34238/tnu-jst.5915

Email: uyenminhdhdt@gmail.com

1. Giới thiệu

Trong những năm gần đây, công tác tuyển sinh được xem là một nhiệm vụ đặc biệt quan trọng, là yếu tố quyết định đến sự tồn tại và phát triển của mỗi trường đại học. Trong đó, vấn đề tư vấn tuyển sinh (TVTS) là một trong những yếu tố hàng đầu góp phần vào thành công trong công tác tuyển sinh, là kênh nối trực tiếp giữa nhà trường với thí sinh (TS) và phụ huynh. Nó có vai trò quan trọng, cần thiết trong việc quảng bá hình ảnh của nhà trường, đồng thời sẽ cung cấp đầy đủ các thông tin tuyển sinh đến học sinh, phụ huynh để giúp cho TS hiểu, tin tưởng khi lựa chọn học tại trường.

Việc lựa chọn ngành học sao cho phù hợp với năng lực bản thân luôn là vấn đề đầy băn khoăn của các bạn học sinh khi sắp rời ghế nhà trường, điều này sẽ ảnh hưởng trực tiếp đến chất lượng học tập của mỗi TS khi vào đại học và ảnh hưởng đến cơ hội việc làm sau khi ra trường. Đây cũng là vấn đề trăn trở không kém của các bậc phụ huynh và nhiều quan tâm của các cơ sở đào tạo.

Hoạt động TVTS của Trường Đại học Đồng Tháp (ĐHĐT) đã và đang được thực hiện dưới nhiều hình thức khác nhau như: tư vấn tuyển sinh tại các trường trung học phổ thông, tư vấn qua điện thoại, tư vấn trực tiếp và tư vấn qua mạng xã hội thông qua mục hỏi - đáp, đồng thời nhà trường cũng kết hợp đưa thông tin và tư vấn trên website http://tuyensinh.dthu.edu.vn/ với fanpage để cung cấp thông tin tuyển sinh. Tuy nhiên, các hình thức này chỉ nhằm hỗ trợ và cung cấp thông tin liên quan cho TS, mà chưa có chức năng tư vấn và định hướng ngành học phù hợp với năng lực, tính cách của mỗi TS.

Cùng với đó, hiện nay cơ sở dữ liệu về giáo dục đang được lưu trữ ở các trường đại học là rất lớn, là một trong những điều kiện thuận lợi cho việc nghiên cứu, ứng dụng khai phá dữ liệu (KPDL) để xây dựng các hệ thống gợi ý, tư vấn trong lĩnh vực này. Đây thực sự là vấn đề rất cần thiết cho các nhà quản lý giáo dục, giúp công tác quản lý, tư vấn trong giáo dục ngày càng hiệu quả. Đã có nhiều nghiên cứu liên quan đến việc ứng dụng KPDL trong giáo dục như sau:

Nhóm tác giả Lê Đức Thắng và cộng sự [1] đã đề xuất một phương pháp mới trong dự đoán kết quả học tập của sinh viên (SV) nhằm hỗ trợ SV lập kế hoạch học tập phù hợp. Tác giả đã thực nghiệm trên dữ liệu thực tế để xác định các SV có thuộc diện "cảnh báo" hay "không cảnh báo" đã cho thấy phương pháp này có khả năng dự đoán tốt hơn so với các phương pháp KPDL tiêu biểu.

Nghiên cứu của Nguyễn Thái Nghe [2] đã sử dụng các phương pháp dự đoán trong KPDL thông qua hai nghiên cứu, nghiên cứu thứ nhất liên quan đến việc sử dụng các giải thuật không cá nhân hóa như mạng Bayes và Cây quyết định, nghiên cứu thứ hai liên quan đến việc sử dụng giải thuật cá nhân hóa – lấy ý tưởng từ các kỹ thuật trong hệ thống gợi ý - như kỹ thuật phân rã ma trận thiên vị (Biased Matrix Factorization) nhằm dự đoán kết quả học tập cho từng cá nhân SV.

Nhóm tác giả Đỗ Thanh Nghị và cộng sự [3] đã giới thiệu tiếp cận khai mỏ dữ liệu để phát hiện môn học quan trọng ảnh hưởng đến kết quả học tập của SV ngành công nghệ thông tin (CNTT). Nhóm tác giả tiến hành sưu tập dữ liệu học tập của SV tốt nghiệp ngành CNTT tại Trường Đại học Cần Thơ, sau đó thực hiện bước tiền xử lý dữ liệu, đưa dữ liệu về cấu trúc bảng, sau đó đề xuất sử dụng giải thuật rừng ngẫu nhiên học từ dữ liệu để rút trích các môn học quan trọng trong chương trình đào tạo ngành CNTT.

Nghiên cứu của Lưu Hoài Sang và cộng sự [4] đề xuất một phương pháp dự báo kết quả học tập của SV bằng kỹ thuật học sâu nhằm khai thác cơ sở dữ liệu trong hệ thống quản lý SV tại các trường đại học. Dữ liệu sau khi thu thập được phân tích, tiền xử lý dữ liệu, thiết kế và huấn luyện mạng nơ-ron đa tầng. Kết quả thực nghiệm cho thấy mô hình đề xuất cho kết quả dự đoán khá chính xác và hoàn toàn khả thi để áp dụng vào thực tế.

Nghiên cứu của Nguyễn Đăng Nhượng [5] trình bày kết quả đã đạt được khi tiến hành áp dụng giải thuật gom cụm dữ liệu, kMeans [6] để khai thác thông tin từ điểm học sinh của trường Cao đẳng nghề Văn Lang Hà Nội. Tác giả tìm hiểu sự ảnh hưởng của vùng miền, của hoàn cảnh gia đình, dân tộc, đạo đức… đến kết quả học tập của học sinh, phân loại kết quả học tập để đánh giá một cách nhanh chóng nhận thức của người học.

Nhóm tác giả Nguyễn Thị Kim Sơn và cộng sự [7] nghiên cứu về việc xây dựng tập dữ liệu SV và kết quả ứng dụng kỹ thuật học máy để lập chương trình dự báo cho loại tốt nghiệp SV, dự báo các yếu tố trong tổ hợp tuyển sinh ảnh hưởng tới kết quả học tập của SV.

Có thể thấy được rằng, các nghiên cứu trên đây đều tập trung vào việc khai thác thông tin dữ liệu của người học, từ đó áp dụng các biện pháp tiền xử lý dữ liệu và giải thuật khai phá phù hợp để đưa ra những mô hình gợi ý, tư vấn cho người học các yếu tố liên quan đến việc học sao cho phù hợp, hiệu quả và kịp thời, giúp cho người học có được kết quả học tập tốt hơn. Tuy nhiên, chúng tôi chưa thấy có nghiên cứu nào liên quan đến việc KPDL trên người học để áp dụng trong vấn đề tư vấn chọn ngành học cho TS, đặc biệt là TS có nguyện vọng lựa chọn ngành học tại trường ĐHĐT.

Với mục đích tìm ra được mô hình trong tư vấn chọn ngành học sao cho phù hợp với năng lực, sở thích và tính cách của TS khi theo học tại trường ĐHĐT, chúng tôi đề xuất "*Xây dựng mô hình hỗ trợ ra quyết định trong tư vấn chọn ngành học tại trường Đại học Đồng Tháp*" dựa trên công nghệ khám phá tri thức và khai mỏ dữ liệu [8]. Qua đó có thể góp phần đẩy mạnh công tác TVTS, hỗ trợ cho các TS trong việc định hướng được nghề nghiệp phù hợp với năng lực của bản thân, góp phần nâng cao chất lượng đào tạo, chất lượng học tập, đồng thời qua kênh tư vấn ngành học cho TS, sẽ làm cầu nối để TS biết đến các thông tin tuyển sinh của trường nhiều hơn.

2. Phương pháp nghiên cứu

Các bước thực hiện nghiên cứu của chúng tôi bao gồm bước nghiên cứu cơ bản và bước vận dụng mô hình để xây dựng ứng dụng web.

Ở bước nghiên cứu cơ bản, chúng tôi tiến hành nghiên cứu tài liệu về kỹ thuật KPDL, lý thuyết trắc nghiệm chọn ngành nghề của tiến sỹ Tâm lý học John Holland, tìm hiểu thông tin tuyển sinh của trường ĐHĐT và quy định chung của Bộ Giáo dục và Đào tạo. Tiếp theo là thu thập thông tin của SV đang học tại trường ĐHĐT qua các khóa 2018, 2019, 2020 và của một số SV đã tốt nghiệp, dữ liệu tập trung ở 8 ngành học đông SV nhất. Sau đó thực hiện bước tiền xử lý dữ liệu để trích chọn, làm sạch và biến đổi dữ liệu về dạng cấu trúc bảng cho phù hợp, từ đó áp dụng giải thuật rừng ngẫu nhiên [12] để huấn luyện và đưa ra mô hình dự báo ngành học phù hợp cho TS, dựa trên những điểm nổi trội tiềm ẩn trong tính cách của mỗi TS (lý thuyết John Holand) và dựa trên những đặc thù chung của TS khi tham gia học tại trường ĐHĐT.

Ở bước vận dụng mô hình để xây dựng hệ thống, chúng tôi tập trung vào hai chức năng hỗ trợ cho TS: *Thứ nhất* là chức năng hỗ trợ TS tự đối chiếu sở thích, năng lực tự nhiên của cá nhân với yêu cầu của các nhóm ngành nghề, để từ đó TS có thể định hướng nghề nghiệp theo nhóm ngành phù hợp nhất dựa trên lý thuyết trắc nghiệm chọn ngành nghề của John Holland. *Thứ hai* là chức năng hỗ trợ ra quyết định trong tư vấn chọn ngành học tại trường ĐHĐT dựa trên công nghệ khám phá tri thức và KPDL, đưa ra dự báo sự phù hợp khi TS có nguyện vọng chọn học một ngành học nào đó tại trường ĐHĐT.

3. Cơ sở lý thuyết

3.1. Lý thuyết mật mã Holland

John Lewis Holland (1919 – 2008) là giáo sư xã hội học danh dự tại Trường Đại học Johns Hopkins và là một nhà tâm lý học Mỹ. Ông được biết đến như là tác giả học thuyết lựa chọn nghề nghiệp hay còn gọi là mã Holland (Holland Codes) và thường được viết tắt là RIASEC [9], [10].

Trên cơ sở lý thuyết này, John Holland đã xây dựng một bộ câu hỏi dành cho người muốn tự tìm hiểu mình. Qua nhiều năm phát triển, bộ trắc nghiệm này giúp cho người ta tự phát hiện được các kiểu người trội nhất đang tiềm ẩn trong con người mình để tự định hướng khi lựa chọn nghề.

Lý thuyết này dựa trên 8 luận điểm, trong đó 2 luận điểm đầu là: Hầu như ai cũng có thể được xếp vào 1 trong 6 kiểu người và có 6 môi trường hoạt động ứng đúng với 6 kiểu người kể trên [11], cụ thể:

Realistic (người thực tế, viết tắt là R): Thích làm với những vật cụ thể, máy móc, dụng cụ, cây cối, con vật hoặc các hoạt động ngoài trời.

Investigative (người thích nghiên cứu – I): Thích quan sát, tìm tòi, điều tra, phân tích, đánh giá hoặc giải quyết vấn đề.

Artistic (người có tính nghệ sĩ – A): Có khả năng nghệ thuật, sáng tác, trực giác và thích làm việc trong các tình huống không có kế hoạch trước như dùng trí tưởng tượng và sáng tạo.

Social (người có tính xã hội – S): Thích làm việc cung cấp hoặc làm sáng tỏ thông tin, thích giúp đỡ, huấn luyện, chữa trị hoặc chăm sóc sức khoẻ cho người khác, có khả năng về ngôn ngữ.

Enterprising (người dám nghĩ dám làm – E): Thích làm việc với những người khác, có khả năng tác động, thuyết phục, thể hiện, lãnh đạo hoặc quản lý các mục tiêu của tổ chức, các lợi ích kinh tế.

Conventional (người công chức – C): Thích làm việc với dữ liệu, con số, có khả năng làm việc văn phòng, thống kê, thực hiện các công việc đòi hỏi chi tiết, tỉ mỉ, cẩn thận hoặc làm theo hướng dẫn của người khác.

3.2. Rừng ngẫu nhiên cây quyết định

Tiếp cận rừng ngẫu nhiên do Breiman đưa ra là một trong những phương pháp tập hợp mô hình thành công nhất [12]. Giải thuật rừng ngẫu nhiên tạo ra một tập hợp các cây quyết định [13], [14] không cắt nhánh, mỗi cây được xây dựng trên tập mẫu bootstrap (lấy mẫu có hoàn lại từ tập học), tại mỗi nút phân hoạch tốt nhất được thực hiện từ việc chọn ngẫu nhiên một tập con các thuộc tính. Lỗi tổng quát của rừng phụ thuộc vào độ chính xác của từng cây thành viên trong rừng và sự phụ thuộc lẫn nhau giữa các cây thành viên. Giải thuật rừng ngẫu nhiên xây dựng cây không cắt nhánh nhằm giữ cho thành phần lỗi bias thấp (thành phần lỗi bias là thành phần lỗi của giải thuật học, nó độc lập với tập dữ liệu học) và dùng tính ngẫu nhiên để điều khiển tính tương quan thấp giữa các cây trong rừng. Tiếp cận rừng ngẫu nhiên cho độ chính xác cao khi so sánh với các thuật toán học có giám sát hiện nay. Như trình bày trong [12], rừng ngẫu nhiên học nhanh, chịu đựng nhiễu tốt và không bị tình trạng học vẹt. Giải thuật rừng ngẫu nhiên sinh ra mô hình có độ chính xác cao đáp ứng được yêu cầu thực tiễn cho vấn đề phân loại, hồi qui.

Giải thuật máy học rừng ngẫu nhiên có thể được trình bày ngắn gọn như sau:

– Từ tập dữ liệu học LS có m phần tử và n biến (thuộc tính), xây dựng T cây quyết định một cách độc lập nhau;

– Mô hình cây quyết định thứ t được xây dựng trên tập mẫu Bootstrap thứ t (lấy mẫu m phần tử có hoàn lại từ tập học LS);

– Tại nút trong, chọn ngẫu nhiên n’ biến (n’<< n) và tính toán phân hoạch tốt nhất dựa trên n’ biến này.

Cây được xây dựng đến độ sâu tối đa không cắt nhánh Kết thúc quá trình xây dựng T mô hình cơ sở, dùng chiến lược bình chọn số đông để phân lớp một phần tử mới đến X. Lặp lại tác vụ trên k lần để tạo ra một rừng gồm k cây thành viên, để thấy rõ quá trình xây dựng và áp dụng rừng ngẫu nhiên để phân lớp.

4. Kết quả thực nghiệm

4.1. Thu thập và tiền xử lý dữ liệu

Trong phần thực nghiệm, chúng tôi tiến hành thu thập dữ liệu kết quả học tập của SV các khóa 2018, 2019, 2020 trên 8 ngành học, do đây là những ngành học có số lượng SV tương đối đông, có thể sử dụng cho việc xây dựng mô hình, số lượng SV đã khảo sát theo mỗi ngành đào tạo được thể hiện trong bảng 1.

Dữ liệu thu thập được có dạng cấu trúc bảng, bao gồm hai tập tin: Một là, các thông tin của SV đang theo học tại trường và đã tốt nghiệp như: họ tên SV, ngành đang học, nơi ở, giới tính, ngày sinh, tên trường trung học phổ thông đã học, ngành học yêu thích, lý do chọn ngành đang

học, kết quả học trung học phổ thông, hình thức dự tuyển đại học, tên ngành đang học, nhóm khả năng sở thích phù hợp với bạn, sự phù hợp (đối với SV đã tốt nghiệp chúng tôi đánh giá cột này dựa trên việc ra trường đúng hạn), lý do phù hợp; Hai là, các thông tin liên quan đến việc lựa chọn nhóm khả năng sở thích phù hợp với SV dựa trên lý thuyết mật mã John Holland, bao gồm các kết quả về nhóm sở thích tương ứng với 6 nhóm.

Bảng 1. *Thống kê số lượng SV được khảo sát theo Khoa*

STT	Tên ngành	Số lượng
1	Công nghệ thông tin	105
2	Kế toán	93
3	Nông học	100
4	Nuôi trồng Thủy sản	75
5	Quản trị kinh doanh	99
6	Sư phạm Toán học	84
7	Tài chính ngân hàng	82
8	Nghệ thuật	90

Tiếp theo, chúng tôi tiến hành tiền xử lý dữ liệu: Dữ liệu thu thập được sẽ được tổng hợp và chuyển về một bảng duy nhất, mỗi cột (trường - field) của bảng biểu diễn thuộc tính đặc thù riêng của mỗi SV, mỗi dòng (record) mô tả thông tin đầy đủ của SV được khảo sát. Để làm được điều này chúng tôi tiến hành thực hiện ba bước:

Bước 1: Tổng hợp các dữ liệu thu được từ kết quả khảo sát SV dựa trên khả năng sở thích (theo mật mã John Holand), kết quả đưa về bảng dữ liệu gồm 8 cột, bao gồm mã số SV, họ tên, và 6 cột còn lại là tổng điểm ứng với mỗi nhóm sở thích, năng lực của mỗi SV.

Bước 2: Chuyển dữ liệu từ các kết quả về một bảng duy nhất, xóa bỏ các dữ liệu không hợp lệ, các thuộc tính không quan trọng. Ở bước này chúng ta thu được một bảng đã bỏ qua các thuộc tính không quan trọng, ví dụ như mã số, họ tên, ngày sinh,…

Bước 3: Dựa trên bảng dữ liệu vừa xây dựng ở bước 2, chúng tôi tiếp tục sử dụng các kỹ thuật trong tiền xử lý dữ liệu để loại bỏ nhiễu và các dữ liệu thiếu, không cần thiết, trích chọn những thuộc tính có giá trị cho việc phân lớp, chuyển đổi dữ liệu về dạng thích hợp để sử dụng một cách hiệu quả nhất, phù hợp cho quá trình xử lý.

4.2. Xây dựng mô hình dự báo

Chương trình xử lý của chúng tôi được thực hiện dựa trên gói dữ liệu đầu vào bao gồm 603 phần tử với 7 thuộc tính độc lập bao gồm: địa chỉ, giới tính, ngành học yêu thích, kết quả học tập ở trung học phổ thông, hình thức xét tuyển vào đại học, ngành học đang (đã) học, nhóm khả năng của SV – nhóm này được chúng tôi xử lý và tổng hợp từ kết quả khảo sát dựa trên lý thuyết mật mã John Holland và 1 thuộc tính phụ thuộc là sự phù hợp của SV với ngành học đang (đã) học. Quá trình thực nghiệm được tiến hành với ngôn ngữ lập trình Python trong môi trường Anaconda, gồm các bước sau:

Đầu tiên, chúng tôi sử dụng giao thức kiểm tra chéo (k-folds cross-validation) để phân chia tập dữ liệu ra làm hai gồm tập huấn luyện (training dataset) và tập kiểm tra (testing dataset). Chúng tôi đã sử dụng hàm train_test_split của sklearn.model_selection với thông số phân chia test_size = 0,3, random_state = 150 nhằm mang đến kết quả đánh giá tốt nhất cho mô hình.

Tiếp theo, nghiên cứu sử dụng gói chương trình rừng ngẫu nhiên RandomForestClassifier cung cấp sẵn trong sklearn.ensemble tiến hành xây dựng mô hình thực nghiệm rừng ngẫu nhiên với 100 cây quyết định và lưu mô hình bằng cách tuần tự hóa nó với hàm pickle.dump() với tên Randomforestmodel.pkl. Kết quả quá trình kiểm tra mang lại độ chính xác cao (96,69%) được thể hiện qua ma trận trực giao (confusion matrix) nhằm lưu trữ kết quả phân lớp - dự đoán ở giai đoạn kiểm tra.

Bên cạnh đó, chúng tôi cũng đánh giá mức độ quan trọng, độ ảnh hưởng của các thuộc tính độc lập đến kết quả của quá trình phân lớp, được trình bày như trong bảng 2.

Bảng 2. *Mức độ quan trọng của các thuộc tính*

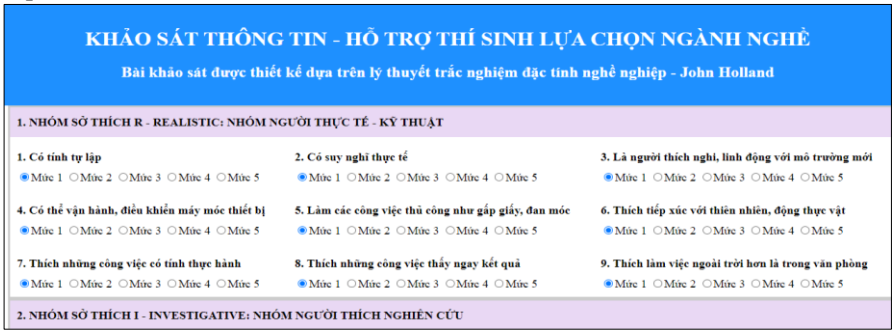
Tên thuộc tính	Ý nghĩa thuộc tính	Mức độ quan trọng (%)
Address	Địa chỉ	7,67
Sex	Giới tính	6,52
f_major	Ngành học yêu thích	17,87
Hs_results	Kết quả học tập ở trung học phổ thông	12,86
Form	Hình thức xét tuyển vào đại học	8,4
f_study	Ngành học đang (đã) học	21,2
Per_abilities	Nhóm khả năng của SV	25,48

4.3. Xây dựng ứng dụng

Trong phần này chúng tôi tiếp tục sử dụng mô hình Randomforestmodel.pkl được đánh giá và lưu trữ ở bước trên, vận dụng cho việc xây dựng ứng dụng web nhằm đáp ứng các chức năng trong việc tư vấn chọn ngành học cho TS, cụ thể như sau:

4.3.1. Chức năng: Khảo sát thông tin – hỗ trợ thí sinh lựa chọn ngành nghề

Ở giao diện này, người dùng sẽ thực hiện chọn vào các mục trắc nghiệm được xây dựng sẵn theo lý thuyết mật mã của John Holland dựa trên tính cách của mỗi cá nhân. Để tiết kiệm thời gian cho người dùng, chương trình sẽ mặc định mức độ thấp nhất trong mỗi câu trả lời là ở mức 1 (tương ứng 0 điểm), người sử dụng chỉ cần chọn vào những nhóm và những câu mà có khả năng khác, phù hợp với mức độ tính cách của mình. Giao diện khảo sát được minh họa như hình 1.



Hình 1. *Giao diện Khảo sát thông tin – hỗ trợ thí sinh lựa chọn ngành nghề*

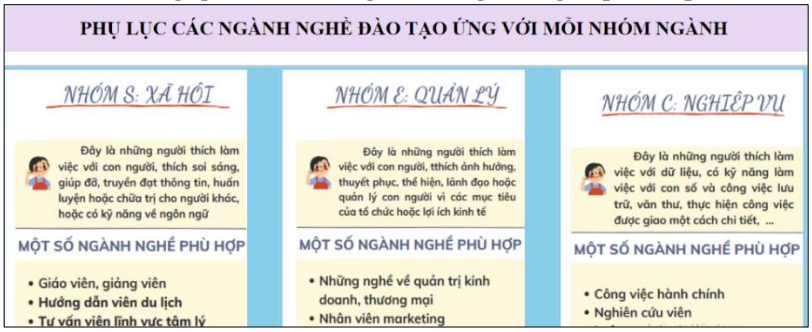
Sau khi thực hiện trắc nghiệm, nếu người dùng chọn vào chức năng "Kết quả khảo sát" sẽ mở ra giao diện thống kê số điểm tổng tương ứng với 6 nhóm ngành nghề đã được lựa chọn.

4.3.2. Chức năng: Thống kê kết quả khảo sát thông tin

Kết quả thống kê ở giao diện hình 2 sẽ cho biết người dùng thuộc nhóm tính cách, nhóm ngành nghề nào nhiều nhất (căn cứ vào thống kê nhóm có điểm số lớn nhất).

Hình 2. *Kết quả khảo sát thông tin – hỗ trợ thí sinh lựa chọn ngành nghề*

Trong giao diện này, chúng tôi cũng xây dựng thêm phụ lục tham khảo các ngành nghề ứng với từng nhóm tính cách, được thể hiện như trong hình 3, giúp người dùng có thể đối chiếu và tham khảo để có cái nhìn tổng quan với những nhóm ngành nghề phù hợp với cá nhân mình.



Hình 3. *Tham khảo các ngành nghề ứng với từng nhóm tính cách*

Nếu người dùng chọn vào chức năng "Tiếp theo", chương trình sẽ dẫn đến giao diện trang tư vấn chọn ngành học cho TS tại trường ĐHĐT, sẽ được trình bày trong phần dưới đây.

4.3.3. Chức năng: Tư vấn chọn ngành học cho thí sinh tại trường ĐHĐT

Ứng dụng này sẽ sử dụng mô hình đã đào tạo (Randomforestmodel.pkl) ở bước trên để dự đoán sự phù hợp khi TS chọn học một ngành tại trường ĐHĐT, dựa trên kết quả trắc nghiệm tính cách người dùng ở giao diện trước và kết hợp với các thông tin đặc điểm riêng của cá nhân TS khi tham gia học tại trường ĐHĐT. Để thực hiện, chúng tôi sử dụng các hàm flask, render_template, request, redirect, url_for từ Flask để nhúng mô hình học máy đã có vào ứng dụng web. Giao diện được trình bày như trong hình 4.

Hình 4. *Trang tư vấn chọn ngành học cho thí sinh tại trường ĐHĐT*

5. Kết luận

Chúng tôi vừa trình bày một tiếp cận khai mỏ dữ liệu và lý thuyết mật mã John Holland để tư vấn chọn ngành học cho TS tại trường ĐHĐT. Các bước thực hiện bao gồm thu thập dữ liệu của SV đang học và một số SV tốt nghiệp đúng tiến độ thuộc một số ngành học có đông SV, sau đó thực hiện bước tiền xử lý dữ liệu để có thể huấn luyện mô hình rừng ngẫu nhiên. Kết quả thu được của nghiên cứu có thể ứng dụng để giúp TS tham khảo trong việc lựa chọn ngành nghề phù hợp với tính cách sở thích của cá nhân, dự đoán sự phù hợp khi TS chọn học một ngành đào tạo tại trường ĐHĐT. Qua đó có thể góp phần đẩy mạnh công tác TVTS, hỗ trợ cho các TS trong việc định hướng được nghề nghiệp phù hợp với năng lực của bản thân, góp phần nâng cao chất lượng đào tạo, chất lượng học tập, đồng thời qua kênh tư vấn ngành học cho TS, sẽ làm cầu nối để TS biết đến các thông tin tuyển sinh của trường nhiều hơn. Tuy nhiên, nghiên cứu này cũng có những hạn chế, đó là chưa hỗ trợ tư vấn sự phù hợp cho tất cả các ngành học tại trường ĐHĐT.

Do đó, chúng tôi sẽ tiếp tục thu thập dữ liệu nhiều hơn nữa trong các nghiên cứu tiếp theo, nhằm xây dựng mô hình và ứng dụng một cách đầy đủ hơn, giúp cho TS có đầy đủ thông tin hơn khi cần tư vấn lựa chọn các ngành đào tạo tại trường ĐHĐT.

Lời cám ơn

Nghiên cứu này được hỗ trợ bởi đề tài mã số SPD2020.01.13

TÀI LIỆU THAM KHẢO/ REFERENCES

[1] D. T. Le, T. H. Truong, T. N. Nguyen, and X. H. Huynh, "Solutions to support students in making study plans based on rough set approach," In *Proc. of the IX National Scientific Conference on Basic and Applied Research in Information Technology (FAIR'9)*, '08, 2016, pp. 151-158.

[2] T. N. Nguyen, "Applying algorithms in data mining to support student learning planning," In *Information technology in decision support*, Can Tho: Can Tho University, 2015, pp. 18-34.

[3] T. N. Do, N. K. Pham, M. T. Nguyen, and T. H. Trinh, "Detection of the key courses affecting the learning outcomes of information technology students," *Can Tho University Journal of Science*, vol. 133, no. 1, pp. 49-57, 2014.

[4] H. S. Luu, T. D. Tran, T. H. Nguyen, and T. N. Nguyen, "Predicting student's performance through deep learning using a multi-layer perceptron," *Can Tho University Journal of Science*, vol. 56, no. 3A, pp. 20-28, 2020.

[5] D. N. Nguyen, "Data mining on learning outcomes of students at Van Lang Vocational College," M.S. thesis, University of Technology, VNU, Hanoi, 2012.

[6] J. MacQueen, "Some Methods for Classification and Analysis of Multivariate Observations," In *Proc. of 5th Berkeley Symposium on Mathematical Statistics and Probability*, '01, 1967, pp. 281-297.

[7] T. K. S. Nguyen, X. H. Nguyen, H. D. To, T. A. Pham, and T. T. T. Do, "Applying machine learning techniques in processing student data to assist university admission," *Hanoi Capital University Journal of Science*, vol. 52, pp. 121-133, 2021.

[8] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "From Data Mining to Knowledge Discovery in Databases," *AI Magazine*, vol. 17, no. 3, pp. 37-54, 1996.

[9] J. L. Holland, *Making Vocational Choices: A Theory of Carrers, Englewood Cliffs*. New Jersey: Prentice-Hall, 1973.

[10] J. L. Holland, *Making vocational choices, Editors and Publishers, 3rd ed.*, Odessa, FL: Psychological Assessment Resources, 1997.

[11] MU Career Center's Guide to Holland Code, *Part of the Career and Major Exploration Guide Series*, University of Missouri – MU Career Center, 2010.

[12] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5-32, 2001.

[13] L. Breiman, J. Friedman, R. Olshen, and C. Stone, *Classification and Regression Trees.* New York: Chapman & Hall, 1984.

[14] R. Quinlan, *C4.5: Programs for Machine Learning*. San Mateo: Morgan Kaufmann, 1993.