USING THE C4.5 ALGORITHM TO BUILD A DECISION TREE FOR THE CAREER CHOICE PROBLEM OF HIGH SCHOOL STUDENTS IN THAI NGUYEN PROVINCE

Bui Ngoc Tuan
TNU - University of Information and Communication Technology

ARTICLE INFO

Received: 22/3/2023
Revised: 27/4/2023
Published: 28/4/2023

KEYWORDS

High school students
Decision Tree
Algorithm C4.5
WEKA
J48

ABSTRACT

In this study, a decision tree was built and the career selection results of grade-12 students in Thai Nguyen province were predicted by algorithm C4.5. Research results have shown that this decision tree has high accuracy and is built based on data collection and preprocessing. However, it is found that the study can only be applied to Thai Nguyen province and needs to be optimized to give more accurate prediction results. Data was collected through a questionnaire of 900 grade-12 students at Ngo Quyen, Tran Quoc Tuan, and Dong Hy high schools in Thai Nguyen province. The author used primary factors to assess, including social needs, learning outcomes, family economic conditions, place of residence and gender. Research results showed that the social needs and learning ability were factors that have a great impact on students' intention to choose a career. At the same time, the classification based on algorithm C4.5 was confirmed to bring higher classification accuracy than the technique based on the Bayes algorithm. This research can contribute to support the university admissions process as well as to help guide the careers of high school students, mainly in grade 12.

SỬ DỤNG THUẬT TOÁN C4.5 XÂY DỰNG CÂY QUYẾT ĐỊNH CHO BÀI TOÁN LỰA CHỌN NGHỀ CHO HỌC SINH TRUNG HỌC PHỔ THÔNG TỈNH THÁI NGUYÊN

Bùi Ngọc Tuấn
Trường Đại học Công nghệ Thông tin và Truyền thông - ĐH Thái Nguyên

THÔNG TIN BÀI BÁO

Ngày nhận bài: 22/3/2023
Ngày hoàn thiện: 27/4/2023
Ngày đăng: 28/4/2023

TỪ KHÓA

Học sinh THPT
Cây quyết định
Thuật toán C4.5
Weka
J48

TÓM TẮT

Trong nghiên cứu này, cây quyết định được xây dựng và kết quả lựa chọn nghề của học sinh khối lớp 12 tại tỉnh Thái Nguyên được dự đoán bằng thuật toán C4.5. Kết quả nghiên cứu đã chỉ ra cây quyết định này có độ chính xác cao và được xây dựng dựa trên quá trình thu thập và tiền xử lý dữ liệu. Tuy nhiên, nhận thấy nghiên cứu chỉ có thể áp dụng cho tỉnh Thái Nguyên và cần được tối ưu hóa để đưa ra kết quả dự báo chính xác hơn. Dữ liệu được thu thập thông qua bảng hỏi 900 học sinh lớp 12 tại các trường trung học phổ thông Ngô Quyền, Trần Quốc Tuấn, Đồng Hỷ tỉnh Thái Nguyên. Tác giả sử dụng 5 nhóm nhân tố chính để đánh giá, bao gồm nhu cầu xã hội, kết quả học tập, điều kiện kinh tế gia đình, nơi ở và giới tính. Kết quả nghiên cứu cho thấy rằng, nhu cầu của xã hội và khả năng học tập là các yếu tố có tác động lớn đến ý định lựa chọn nghề của học sinh. Đồng thời, việc phân lớp dựa trên thuật toán C4.5 được khẳng định là mang lại độ chính xác phân lớp cao hơn so với kỹ thuật dựa trên thuật toán Bayes. Nghiên cứu này có thể đóng góp hỗ trợ cho quá trình tuyển sinh đại học cũng như giúp hướng nghiệp cho học sinh phổ thông chủ yếu là khối 12.

DOI: https://doi.org/10.34238/tnu-jst.7590

Email: bntuan@ictu.edu.vn

1. Mở đầu

Việc chọn cho mình nghề nghiệp đúng với khả năng và sở trường của học sinh trung học phổ thông là quan trọng, nó có ảnh hưởng đến tương lai và công việc của họ sau này. Vì vậy, cung cấp cho học sinh một công cụ hỗ trợ lựa chọn nghề nghiệp là cần thiết. Trong nghiên cứu này, tác giả sử dụng cây J48 trên thuật toán C4.5 để xây dựng cây quyết định trong việc hỗ trợ lựa chọn nghề cho học sinh trung học phổ thông tại tỉnh Thái Nguyên. Thông tin về năng lực, sở thích và tiềm năng của học sinh được thu thập thông qua khảo sát thực tế, tác giả đã áp dụng thuật toán C4.5 để xây dựng cây quyết định dựa trên thông tin thu thập được. Cây quyết định này có thể giúp học sinh có các thông tin cần thiết để chọn nghề học phù hợp với năng lực và sở trường của họ. Nghiên cứu này tiến hành đã đánh giá kết quả của cây quyết định bằng cách đối chiếu kết quả dự đoán của phương pháp với kết quả thực tế, và kết quả cho thấy cây J48 sử dụng trên C4.5 có kết quả cao trong việc dự báo chọn nghề của học sinh THPT tỉnh Thái Nguyên.

Đã có các nghiên cứu về sự lựa chọn nghề của học sinh như Bùi Thế Hồng [1] đã nghiên cứu "về thuật toán xây dựng cây quyết định và giảm tập luật", chỉ ra việc sử dụng thuật toán C4.5 để rút gọn các tập luật quyết định. Nghiên cứu của Nguyễn Thị Ngọc Ánh và cộng sự [2] trình bày ứng dụng phương pháp học máy - cây quyết định trong đánh giá diễn biến rừng ngập mặn xã Đất Mũi. Các tác giả đã sử dụng phương pháp học máy - cây quyết định để phân loại, hồi quy bằng cách xây dựng nhiều cây quyết định để so sánh chúng với nhau từ đó rút ra kết quả tối ưu.

D.W. Chapman [3] đã kết luận trong nghiên cứu "mô hình lựa chọn trường đại học của sinh viên", đưa ra 5 yếu tố bao gồm: Nỗ lực giao tiếp với sinh viên, chi phí, người quan trọng, khả năng và mức độ đam mê của học sinh để rút ra yếu tố quan trọng liên quan đến quyết định chọn trường của sinh viên. Cabera và La Nasa [4] đi sâu nghiên cứu 2 yếu tố là Cơ hội học tập trong tương lai và Cơ hội việc làm trong tương lai, và coi đây là các yếu tố quan trọng trong quyết định chọn trường đại học của sinh viên.

Nghiên cứu của Trần Văn Quý và cộng sự [5] đã trình bày "các yếu tố ảnh hưởng đến quyết định chọn trường đại học của học sinh THPT". Trong nghiên cứu này các tác giả sử dụng 5 nhân tố chính đó là (1) yếu tố cơ hội việc làm trong tương lai, (2) thông tin có sẵn về trường đại học, (3) yếu tố về bản thân cá nhân học sinh, (4) yếu tố về cá nhân có ảnh hưởng đến quyết định, (5) yếu tố đặc điểm cố định của trường đại học.

Lưu Ngọc Liêm [6] phân tích các yếu tố ảnh hưởng đến quyết định chọn trường của sinh viên Đại học Lạc Hồng. Tác giả cũng sử dụng 5 nhân tố bao gồm yếu tố về mức độ đa dạng và hấp dẫn ngành đào tạo; yếu tố về đặc điểm của trường đại học; yếu tố về khả năng đáp ứng sự mong đợi sau khi ra trường; yếu tố về những nỗ lực giao tiếp của trường đại học và yếu tố về danh tiếng của trường đại học.

Tác giả Nguyễn Mạnh Hà và cộng sự [7] khảo sát các yếu tố ảnh hưởng đến quyết định chọn trường Đại học Mở TP.HCM. Các tác giả đưa ra nhóm 7 nhân tố đó là nỗ lực của nhà trường đưa thông tin đến học sinh; chất lượng dạy và học; đặc điểm của bản thân sinh viên; công việc trong tương lai; khả năng đậu vào trường; người thân trong gia đình; người thân ngoài gia đình.

Tác giả Edmonds, Jill [8] đã nghiên cứu "những điều khác biệt trong việc lựa chọn ngành ở trường đại học". Trong nghiên cứu này tác giả sử dụng 4 nhân tố đó là đặc điểm của trường, sở thích, khả năng của học viên, cơ hội việc làm trong tương lai và giới tính.

Hiện tại, chưa có bất kỳ nghiên cứu nào sử dụng thuật toán C4.5 để phân tích những nhân tố ảnh hưởng đến việc lựa chọn nghề của học sinh. Do đó, mục đích của tác giả là tìm ra những nhân tố chủ yếu ảnh hưởng đến quyết định của học sinh lớp 12 trong việc lựa chọn nghề, tìm ra một phương pháp để nghiên cứu các yếu tố này đối với học sinh khối 12 tại tỉnh Thái Nguyên, và cung cấp thông tin và khuyến nghị cho các trường học, cha mẹ học sinh, và giáo viên để hỗ trợ học sinh trong quá trình lựa chọn nghề phù hợp với sở thích và năng lực của mình.

2. Phương pháp nghiên cứu

2.1. Các tập dữ liệu được sử dụng trong quá trình nghiên cứu

Dữ liệu sử dụng trong nghiên cứu được tác giả thu thập bằng bảng hỏi từ 900 học sinh khối 12 tại 3 trường THPT trên địa bàn tỉnh Thái Nguyên, bao gồm THPT Ngô Quyền, THPT Đồng Hỷ và Trường THPT Trần Quốc Tuấn. Từ đó, tác giả đã chọn ngẫu nhiên 150 phiếu để phân tích và rút ra các yếu tố ảnh hưởng đến việc chọn nghề của học sinh THPT tỉnh Thái Nguyên năm 2022.

2.2. Phương pháp nghiên cứu

Để tạo ra cây quyết định cho bài toán giúp học sinh khối 12 chọn nghề trên địa bàn tỉnh Thái Nguyên, tác giả đã thu thập dữ liệu về các nhân tố ảnh hưởng đến quyết định chọn nghề thông qua khảo sát học sinh. Các yếu tố này bao gồm học lực (HocL), nghề chọn (Nghe), nhu cầu của xã hội (Nhucau), nơi sống gia đình (Noi), điều kiện của gia đình (DieukienKT), và giới tính (GT).

Trước khi chạy mô hình, tác giả đã tiến hành tiền xử lý dữ liệu để chuẩn bị cho quá trình chạy dữ liệu. Các bước tiền xử lý dữ liệu bao gồm loại bỏ dữ liệu trùng lặp, xử lý các giá trị khuyết, và biến đổi dữ liệu sang dạng số.

Sau đó, tác giả đã sử dụng cây J48 để xây dựng cây quyết định dựa vào các thuộc tính của dữ liệu. Quá trình chạy kết quả của mô hình là chia dữ liệu thành các tập để huấn luyện và kiểm tra để đánh giá mô hình.

Để đánh giá hiệu quả của cây quyết định đã xây dựng, tác giả đã sử dụng các thang đo như độ chính xác và độ bao phủ. Kết quả được đánh giá trên cả tập dữ liệu huấn luyện và tập kiểm tra để đảm bảo khả năng sử dụng của mô hình.

Cuối cùng, tác giả đã thực hiện kiểm thử trên các dữ liệu mới để xem tính khả dụng của mô hình. Trong nghiên cứu, có ba thành phần cơ bản là các biến quyết định, các biến không thể kiểm soát và biến kết quả.

2.2.1. Khái niệm về cây quyết định

Cây quyết định được định nghĩa bởi Ross Quinlan [9], một nhà khoa học máy tính người New Zealand, vào năm 1986 chỉ ra như sau: "Cây quyết định là một mô hình học máy trong đó các quyết định được đưa ra dựa trên cây dạng phân cấp. Mỗi nút trên cây đại diện cho một quyết định, trong khi mỗi lá của cây đại diện cho một kết quả cuối cùng. Cây quyết định được xây dựng dựa trên các quy tắc học máy được học từ tập dữ liệu huấn luyện. Các ưu điểm của cây quyết định là xây dựng tương đối nhanh, đơn giản và dễ hiểu. Hơn nữa, các cây có thể dễ dàng được chuyển đổi sang các câu lệnh SQL để có thể được sử dụng để truy nhập cơ sở dữ liệu một cách hiệu quả. Cuối cùng, việc phân lớp dựa trên cây quyết định đạt được sự tương tự và đôi khi là chính xác hơn so với các phương pháp phân lớp khác".

2.2.2. Thuật toán C4.5

Cũng tác giả Ross Quinlan đưa ra thuật toán C4.5 [9] như sau: "C4.5 là một thuật toán học máy để xây dựng cây quyết định được phát triển bởi Ross Quinlan vào những năm 1993. C4.5 là phiên bản cải tiến của thuật toán ID3 trước đó, với khả năng xử lý các thuộc tính có giá trị liên tục và thiếu giá trị (missing values) trong dữ liệu huấn luyện. Thuật toán này sử dụng độ đo đóng góp thông tin (information gain) để chọn thuộc tính phân chia tốt nhất cho mỗi nút trên cây. Ngoài ra, C4.5 còn sử dụng kỹ thuật cắt tỉa cây (pruning) để giảm thiểu hiện tượng quá khớp và tăng khả năng tổng quát hóa của mô hình. Tư tưởng phát triển cây quyết định của C4.5 là phương pháp HUNT, với chiến lược phát triển theo độ sâu (depth-first strategy)".

2.2.3. Tổng quan về phần mềm Weka

Weka là một phần mềm mã nguồn mở do Ian H. Witten [10] phát triển để dùng cho học máy và khai thác dữ liệu. Weka được phát triển và viết bằng ngôn ngữ Java. Weka có thể được sử dụng trên các hệ điều hành khác nhau bao gồm Windows, MacOS và Linux.

Weka cung cấp cho người dùng một giao diện đồ họa thân thiện để xây dựng các mô hình học máy, khám phá dữ liệu, tiền xử lý dữ liệu và thực hiện các tác vụ khai thác dữ liệu khác. Weka cũng hỗ trợ nhiều thuật toán học máy phổ biến như cây quyết định, Naive Bayes, SVM, Neural Network, k-Means Clustering và nhiều thuật toán khác.

Weka cũng hỗ trợ các tính năng tiền xử lý dữ liệu như xử lý giá trị thiếu, loại bỏ nhiễu, chuẩn hóa dữ liệu, rút trích đặc trưng và biến đổi dữ liệu. Ngoài ra, Weka còn cho phép lưu các mô hình đã được huấn luyện để sử dụng lại trong tương lai.

Weka được sử dụng rộng rãi trong cộng đồng học máy và khai thác dữ liệu do tính linh hoạt và dễ sử dụng của nó. Các ứng dụng của Weka bao gồm phân loại văn bản, phát hiện gian lận thẻ tín dụng, phân tích tập trung khách hàng, phân loại hình ảnh và nhiều ứng dụng khác.

2.2.4. Dữ liệu đưa vào mô hình

Dữ liệu đưa vào mô hình Trong nghiên cứu, tập dữ liệu được sử dụng là file NGHE11.CSV gồm 150 bản ghi với 6 thuộc tính là: Trình độ học (HOL), Xã hội cần (nhucau), nghề lựa chọn (nghe), Nơi gia đình sinh sống (Noi), Điều kiện kinh tế gia đình (dieukiengd), và Giới tính (GT).

2.2.5. Phương pháp Cross-validation

Phương pháp Cross-validation được Sir Ronald A. Fisher [11] đưa ra như sau: Phương pháp Cross-validation là một kỹ thuật được sử dụng trong học máy và khai phá dữ liệu để đánh giá hiệu suất của mô hình trên dữ liệu huấn luyện. Kỹ thuật này thường được sử dụng để so sánh và đánh giá hiệu suất của các mô hình khác nhau trên cùng một tập dữ liệu. Weka cung cấp nhiều tùy chọn cho phương pháp cross-validation, bao gồm số lượng folds, cách chọn tập con (ngẫu nhiên hoặc theo thứ tự), và cách chia thành tập huấn luyện và tập kiểm tra. Để thực hiện phương pháp cross-validation trong Weka, người dùng có thể chọn tab Classify trên giao diện chính của Weka, sau đó chọn một mô hình và cấu hình các tham số trong phần More options. Kỹ thuật cross-validation sẽ chia dữ liệu thành các tập con (folds) và sử dụng một tập con để kiểm tra và các tập còn lại để huấn luyện mô hình. Điều này được lặp lại nhiều lần với các tập con khác nhau cho đến khi tất cả các tập con đã được sử dụng để kiểm tra.

3. Kết quả và bàn luận

3.1. Kết quả

Bảng 1. *Kết quả xác nhận phân lớp được thể hiện như sau*

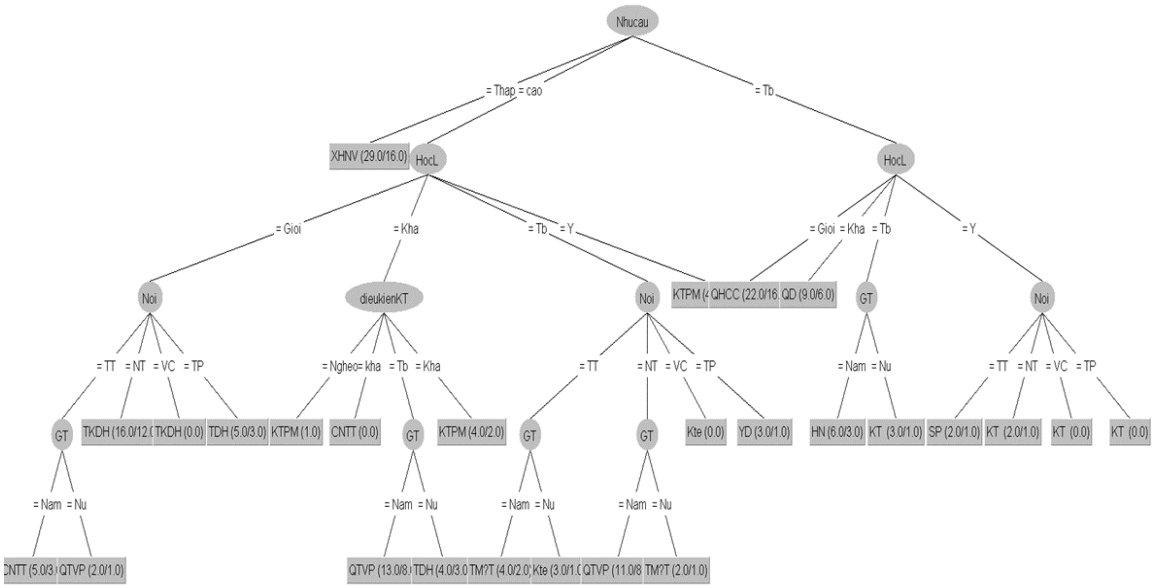
Các giá trị khi phân lớp	Số lượng	Tỷ lệ
Chính xác	135	90 %
Không chính xác	15	10%
Thống kê Kappa	0	
Sai số tuyệt đối trung bình	0,12	
Sai số chuẩn trung bình	0,25	
Sai số tuyệt đối tương đối		94,55 %
Sai số chuẩn tương đối		99,94 %
Tổng số dữ liệu của mô hình	150	

Kết quả của việc sử dụng cây J48 trên Weka để phân lớp dữ liệu được trình bày như sau:

Dữ liệu đầu vào: file NGHE11.CSV với 150 bản ghi và 6 thuộc tính: HOL, Nhucau, nghe, Noi, GT, Dieukiengd, Số bản ghi: 150, Số biến: 6 HOL, Nhucau, nghe, Noi, GT, Dieukiengd.

Qua bảng 1 ta nhận thấy về các giá trị phân lớp sau khi chạy mô hình, phân loại chính xác là 135/150 chiếm 90%; phân loại không chính xác là 15/150 chiếm 15%, sai số tuyệt đối trung bình là 0,12, sai số chuẩn trung bình là 0,25.

Kết quả hiển thị cây quyết định bằng Weka như Hình 1.



Hình 1. *Kết quả cây quyết định sử dụng thuật toán C4.5 bằng cây J48*

Sau khi sử dụng thuật toán C4.5 và cây J48 được kết quả như hình 1 ta xây dựng được tập các luật để tiến hành huấn luyện và sử dụng kết quả của các luật này vào kết luận và khuyến cáo học sinh, phụ huynh, giáo viên và các cơ sở giáo dục làm căn cứ để tư vấn cho học sinh lựa chọn nghề đúng với năng lực, sở thích và các điều kiện khác.

Tạo tập luật từ mô hình đã chạy:

Luat 1: If Nhucau= “Cao” and HoL=“Gioi” and Noi= “TP” then nghe = “KTPM” and “TDH”.

Luat 2: If Nhucau= “Cao” and HoL=“Gioi” and Noi= “TT” and GT=”Nam” then nghe = “CNTT”.

Luat 3: If Nhucau= “Cao” and HoL=“Gioi” and Noi= “TT” and GT=”Nu” then nghe = “QTVP”.

Luat 4: If Nhucau= “Cao” and HoL=“Gioi” and Noi=”NT” or Noi= “VC” then nghe = “TKDH”.

Luat 5: If Nhucau= “Cao” and HoL=“Gioi” and Dieukiengd= “kha” then nghe = “KTPM” and “CNTT”.

Luat 6: If Nhucau= “Cao” and HoL=“Gioi” and Dieukiengd= “ngheo” then nghe = “KTPM”.

Luat 7: If Nhucau= “Cao” and HoL=“Gioi” and Dieukiengd= “TB” and GT=”Nam” then nghe = “QTVP”.

Luat 8: If Nhucau= “Cao” and HoL=“Gioi” and Dieukiengd= “TB” and GT=”Nu” then nghe = “TDH”.

Luat 9: If Nhucau= “Cao” and HoL=“TB” and Noi= “TT” and GT=”Nam” then nghe = “TMDT”.

Luat 10: If Nhucau= “Cao” and HoL=“TB” and Noi= “TT” and GT=”Nu” then nghe = “Kte”.

Luat 11: If Nhucau= “Cao” and HoL=“TB” and Noi= “NT” and GT=”Nu” then nghe = “TMDT”.

Luat 12: If Nhucau= “Cao” and HoL=“TB” and Noi= “NT” and GT=”Nam” then nghe = “QTVP”.

Luat 13: If Nhucau= “Cao” and HoL=“TB” and Noi= “VC” then nghe = “Kte”.

Luat 14: If Nhucau= “Cao” and HoL=“TB” and Noi= “TP” then nghe = “Y”.

Luat 15: If Nhucau= “TB” and HoL=“Gioi” then nghe = “QHCC”.

Luat 16: If Nhucau= “TB” and HoL=“kha” then nghe = “QD”.

Luat 18: If Nhucau= “TB” and HoL=“ T B ” and GT=”Nu” then nghe = “KT”.

Luat 19: If Nhucau= “TB” and HoL=“ Y ” and Noi=”TT” then nghe = “SP”.

Luat 20: If Nhucau= “TB” and HoL=“ Y ” and Noi=”NT” and Noi=”VC” and Noi=”TP”then nghe = “KT”.

Luat 21: If Nhucau= “Thap” then nghe = “XHNV”.

Bảng 2. *Thuật toán Lớp được chọn trong WEKA - Tỷ lệ kết quả khi phân lớp*

Thuật toán	Lớp được chọn trong WEKA	Tỷ lệ kết quả khi phân lớp	
		Ban đầu	Kết quả sau
Bayes	BayesNet	66,66	90,00
	Naïve Bayes	75,00	86,00
	NaïveBayesMultiomialText	25,00	58,00
	NaïveBayesUpdatable	75,00	86,00
Tree	DecisionStump	75,00	88,00
	HoeffdingTree	75,00	86,34
	J48	87,67	92,00
	LMT	79,16	90,00
	RandomForest	92,67	91,31
	RandomTree	62,5	90,00
	REPTree	54,34	90,34

(Nguồn: Sau khi sử dụng WEKA trên cây J48 và các thuật toán C4.5)

Trong bảng 2, thuật toán Bayes được sử dụng để so sánh các lớp được chọn trong Weka giữa tỷ lệ ban đầu và kết quả sau phân lớp ví dụ với lớp BayesNet kết quả ban đầu chỉ có 66,66% sau khi trích chọn tăng lên 90%. Đối với thuật toán Tree sử dụng J48 kết quả ban đầu 87,67% sau trích chọn tăng lên 92%. Với thuật toán DecisionStump kết quả ban đầu là 75,00% sau trích chọn tăng lên 88%. Còn thuật toán HoeffdingTree cũng tăng từ 75% lên 86,34%. Hay thuật toán LMT tăng từ 79,16% lên 90%. Tuy nhiên ta nhận thấy J48 cho kết quả cao nhất.

3.2. Bàn luận

Nghiên cứu này tập trung vào việc sử dụng phương pháp học máy để giải quyết bài toán chọn nghề cho học sinh lớp 12 tại tỉnh Thái Nguyên bằng cách sử dụng Weka trên cây J48 và thuật toán C4.5. Tuy nhiên, để phương pháp này có tính khả thi và ứng dụng thực tế, yêu cầu thực hiện các bước tiền xử lý dữ liệu một cách chính xác và đầy đủ. Đồng thời, cần thực hiện các thí nghiệm và đánh giá để đưa ra kết quả chính xác và tin cậy hơn về độ chính xác của cây quyết định. Kết quả nghiên cứu cho thấy rằng phương pháp này là hiệu quả trong việc giúp các em học sinh khối 12 chọn nghề phù hợp.

4. Kết luận

Qua nghiên cứu cho thấy xây dựng cây quyết định từ dữ liệu khảo sát cung cấp độ chính xác cao trong việc phân tích sự lựa chọn nghề nghiệp của học sinh khối 12 tại tỉnh Thái Nguyên. Ngoài ra, kết quả phân loại dựa trên phương pháp Bayes cho kết quả độ chính xác cao nhất lên tới 90% cho thuật toán BayesNet sau khi trích chọn đặc trưng. Hơn nữa, kỹ thuật cây quyết định cũng cho kết quả đáng tin cậy với J48 đạt độ chính xác 92%. Tuy nhiên, tác giả lưu ý rằng kết quả có thể không áp dụng rộng rãi cho các nơi khác do sự khác biệt về điều kiện kinh tế, xã hội, văn hóa và giáo dục. Để đạt được kết quả dự báo chính xác hơn, cần tiếp tục tối ưu hóa mô hình và nâng cao chất lượng dữ liệu đầu vào.

Lời cảm ơn

Bài báo là sản phẩm của đề tài KH&CN cấp Đại học, mã số: ĐH2022-TN07-03, do Trường Đại học Công nghệ Thông tin và Truyền thông cấp kinh phí thực hiện.

TÀI LIỆU THAM KHẢO/ REFERENCES

[1] T. H. Bui, “On the algorithms for constructing decision trees and reducing rule sets,” (in Vietnamese), *Journal of Computer Science and Control*, vol.18, no. 4, pp. 323-332, 2002.

[2] T. N. A. Nguyen, D. H. Tran, P. H. Le, “Applying the method of machine learning - decision tree in assessing the changes of mangrove forest in Dat Mui commune,” (in Vietnamese), *Climate Change Science*, vol. 20, p. 33, 12/2021.

[3] D. W. Chapman, "A model of student college choice," *The Journal of Higher Education*, vol. 52, no. 5, pp. 490-505, 1981.

[4] A. Cabrera, "Understanding the college-choice process," *New directions for institutional research*, vol. 107, pp. 5-22, 2000.

[5] V. Q. Tran, “Factors influencing high school students' decision to choose a university,” (in Vietnamese), *Journal of Climate Change Science*, vol.15, pp. 89-91, 2009.

[6] N. L. Luu, “Identifying factors affecting the decision to choose a university among Lac Hong University students,” (in Vietnamese), Scientific research project, 2010.

[7] M. H. Nguyen, “A study of factors affecting students' decision to choose Ho Chi Minh City Open University,” (in Vietnamese), *Journal of Science, Ho Chi Minh City Open University*, vol. 6, no.2, pp.107-116. 2011.

[8] J. Edmonds, "Factors influencing choice of college major: what really makes a difference?" M.A. Thesis, Rowan University, 2012.

[9] J. R. Quinlan, *Programs for Machine Learning*, Morgan Kaufmann Publishers, Inc., 1993.

[10] I. H. Witten, “Data mining in bioinformatics using Weka,” *Bioinformatics*, vol. 20, no. 15, pp. 2479–2481, 2004.

[11] S. R. A. Fisher, "The Use of Multiple Measurements in Taxonomic Problems," *Annals of Eugenics*, vol.5, no.4, pp. 179-188, 1936.