ỨNG DỤNG CÁC THUẬT TOÁN HỌC MÁY ĐỂ ĐÁNH GIÁ BỘ CƠ SỞ DỮ LIỆU TRONG PHÂN LOẠI RỐI LOẠN PHỔ TỰ KỶ

Phạm Quang Thuận[a*], Nguyễn Đình Thuân[b]

[a]Trung tâm Thông tin-Thư viện, Trường Cao đẳng Sư phạm Trung ương-Nha Trang, Khánh Hòa, Việt Nam
[b]Khoa Hệ thống thông tin, Trường Đại học Công nghệ Thông tin, Đại học Quốc gia TP.Hồ Chí Minh, Việt Nam
[*]Tác giả liên hệ: Email: thuanpq@sptwnt.edu.vn

Lịch sử bài báo
Nhận ngày 04 tháng 02 năm 2020
Chỉnh sửa lần 1 ngày 08 tháng 3 năm 2020 | Chỉnh sửa lần 2 ngày 10 tháng 5 năm 2020
Chấp nhận đăng ngày 23 tháng 9 năm 2020

Tóm tắt

Bài báo này, chúng tôi trình bày kết quả đánh giá bộ cơ sở dữ liệu trong phân loại rối loạn phổ tự kỷ (ASD) trẻ em trên kho dữ liệu UCI. Chúng tôi tiến hành đánh giá bộ dữ liệu với các thuật toán SVM và Random Forest, đồng thời khảo sát thêm các thuật toán Decision Trees, Logistic Regression, K-Nearest-Neighbors, Naïve Bayes, và mạng nơ-ron Multi Layer Perceptron (MLP). Kết quả thử nghiệm trên bảy thuật toán cho kết quả phân loại cao phù hợp với các nghiên cứu trước đó. Chúng tôi kết luận bộ dữ liệu phân loại rối loạn phổ tự kỷ trẻ em trên kho dữ liệu UCI là đáng tin cậy.

Từ khóa: Rối loạn phổ tự kỷ; Sàng lọc rối loạn phổ tự kỷ; Thuật toán học máy.

DOI: http://dx.doi.org/10.37569/DalatUniversity.10.3.649(2020)

Loại bài báo: Bài báo nghiên cứu gốc có bình duyệt

APPLICATION OF MACHINE LEARNING ALGORITHMS TO EVALUATE THE UCI DATABASE IN THE CLASSIFICATION OF AUTISM SPECTRUM DISORDERS

Pham Quang Thuan[a*], Nguyen Dinh Thuan[b]

[a]The Library-Information Center, Nha Trang National College of Pedagogy, Khanhhoa, Vietnam
[b]The Faculty of Information Systems, University of Information Technology, Vietnam National University Ho Chi Minh City, Hochiminh City, Vietnam
[*]Corresponding author: Email: thuanpq@sptwnt.edu.vn

Article history
Received: February 4th, 2020
Received in revised form (1st): March 8th, 2020 | Received in revised form (2nd): May 10th, 2020
Accepted: September 23rd, 2020

Abstract

In this article, we present the results of an evaluation of the autism spectrum disorder classification (ASD) of children in the UCI database. We evaluated the data set with the SVM and Random Forest algorithms and also investigated the Decision Tree, Logistic Regression, K-Nearest-Neighbors, Naïve Bayes, and Multi-Layer Perceptron (MLP) algorithms. All algorithms give high classification results consistent with previous studies. We conclude that the data set for classifying children's autism spectrum disorders in the UCI database is reliable.

Keywords: Autism spectrum disorder; Machine learning algorithms; Screening autism spectrum disorder.

DOI: http://dx.doi.org/10.37569/DalatUniversity.10.3.649(2020)

Article type: (peer-reviewed) Full-length research article

1. ĐẶT VẤN ĐỀ

"Rối loạn phổ tự kỷ (ASD) là một dạng khuyết tật phát triển tồn tại trong cuộc đời, thường xuất hiện trong ba năm đầu đời. ASD là do rối loạn thần kinh gây ảnh hưởng đến chức năng hoạt động của não bộ. ASD có thể xảy ra ở bất cứ cá nhân nào không phân biệt giới tính, chủng tộc hoặc điều kiện kinh tế-xã hội. Đặc điểm của ASD là những khiếm khuyết về tương tác xã hội, giao tiếp ngôn ngữ và phi ngôn ngữ, có hành vi, sở thích và hoạt động mang tính hạn hẹp và lặp đi lặp lại" (The United Nations, n.d). Ở Việt Nam chưa có số liệu chính xác, theo ước tính do Cục Bảo trợ xã hội-Bộ Lao động Thương binh và Xã hội (LĐTB&XH) hiện có khoảng hơn 200,000 người bị ASD. Song theo cách tính của Tổ chức WHO, con số này tầm khoảng 500,000 và thực tế số lượng trẻ được chẩn đoán và điều trị ngày càng tăng từ năm 2000 đến nay. Báo cáo của Viện Khoa học Giáo dục Việt Nam cho biết, nghiên cứu mô hình tàn tật ở trẻ em của khoa Phục hồi Chức năng, Bệnh viện Nhi Trung ương giai đoạn 2000-2007 đã thống kê số lượng trẻ mắc chứng ASD đến khám năm 2007 tăng gấp 50 lần so với thời điểm bảy năm trước đó, xu thế mắc cũng tăng nhanh từ 122% đến 268% trong giai đoạn 2004-2007 so với năm 2000. Trên thế giới, tỷ lệ trẻ được phát hiện và chẩn đoán ASD tăng một cách đáng kể. Điển hình như ở Mỹ, trước đây tỷ lệ này là 1/1,000 thì nay đã tăng lên 1/68 từ năm 2012 (Doanh, 2018).

Việc chẩn đoán ASD chủ yếu được thực hiện thông qua các biểu hiện lâm sàng bằng quan sát trực tiếp hoặc phỏng vấn người chăm sóc. Quy trình chẩn đoán ASD thường rất phức tạp, chủ quan và nhiều thách thức. Theo tiêu chuẩn của WHO, chẩn đoán cho các rối loạn phát triển của một trẻ cần năm chuyên gia, theo tiêu chuẩn của Mỹ là sáu chuyên gia, cùng theo dõi trẻ trong tối thiểu một tháng ở ba môi trường khác nhau (phòng khám hoặc trung tâm, gia đình, và cộng đồng) (Nguyễn, 2012). Thực tế để tiến hành các chẩn đoán ASD thường mất nhiều thời gian và phụ thuộc trình độ của các chuyên gia lâm sàng nên các nhà khoa học đã nghĩ đến một quy trình chẩn đoán mới để tăng độ chính xác và tiết kiệm thời gian.

Quá trình chẩn đoán ASD là một vấn đề phân loại điển hình trong đó bác sĩ lâm sàng đang cố gắng xây dựng một mô hình tự động (phân loại) bằng cách sử dụng học máy để đoán xem một trường hợp có phải là ASD hay không. Trình phân loại này thường được xây dựng từ bộ dữ liệu đầu vào (các trường hợp trước đây có và không bị ASD được phân loại bởi một công cụ chẩn đoán), sau đó đánh giá trên các trường hợp thử nghiệm độc lập (trường hợp mới) để đo lường hiệu quả của nó trong việc dự đoán ASD. Nhìn chung, quá trình chẩn đoán trong nghiên cứu ASD là một nhiệm vụ phân loại.

Nhiều thuật toán học máy đã được các nhà khoa học đã áp dụng trên các bộ cơ sở dữ liệu khác nhau và thu được những kết quả nghiên cứu khả quan (Bảng 1). Trong số các thuật toán cho kết quả dự đoán cao là SVM và Random forest. Tuy nhiên, phần lớn các bộ dữ liệu có sẵn về ASD liên quan đến di truyền. Một vài ví dụ trong số này là bộ dữ liệu AGRE (Geschwind & ctg., 2001), cơ sở dữ liệu quốc gia Hoa Kỳ về nghiên cứu ASD NDAR (Hall, Huerta, McAuliffe, & Farber, 2012) và AC (Fischbach & Lord, 2010). Các bộ dữ liệu này đều không được công khai nên khó tiếp cận trong quá trình nghiên cứu. Có rất ít bộ dữ liệu ASD liên quan đến hành vi dùng để sàng lọc ASD. Hiện

nay chỉ có bộ dữ liệu sàng lọc ASD (trẻ em (Thabtah, 2017c), trẻ vị thành niên (Thabtah, 2017b) và người lớn (Thabtah, 2017a)) được giáo sư Fadi Fayez Thabtab, Đại học Auckland, New Zealand công bố trên kho dữ liệu UCI vào tháng 12 năm 2017 là bộ dữ liệu hành vi ASD. Để đánh giá độ tin cậy của các bộ dữ liệu này các nhà khoa học cần thêm nhiều nghiên cứu.

Trong bài báo này, chúng tôi tiến hành đánh giá bộ dữ liệu sàng lọc ASD trẻ em. Lý do chọn bộ dữ liệu này là để tương thích với bộ dữ liệu dùng để kiểm nghiệm mà chúng tôi thu thập được tại Trung tâm Tư vấn và Hỗ trợ Giáo dục Đặc biệt thuộc Trường Cao đẳng Sư phạm Trung ương-Nha Trang–đơn vị có chức năng tổ chức công tác tư vấn và hỗ trợ giáo dục trẻ có nhu cầu đặc biệt. Các bước chúng tôi thực hiện đánh giá bộ dữ liệu như sau: (1) Sử dụng các thuật toán Decision Trees, Logistic Regression, K-Nearest-Neighbors, Naïve Bayes, MLP để xây dựng mô hình trên bộ dữ liệu UCI; (2) Kiểm nghiệm các mô hình trên bộ dữ liệu thực tế; và (3) Kết luận.

2. NỘI DUNG NGHIÊN CỨU

2.1. ASD và các phương pháp sàng lọc ASD

2.1.1. Giới thiệu về ASD

ASD là một rối loạn phát triển lan tỏa gây cản trở các kỹ năng cá nhân trong xã hội, tạo ra các hành vi lặp đi lặp lại và tác động đến giao tiếp bằng lời nói hoặc biểu hiện sự gián đoạn từ trung bình đến nặng (Pennington, Cullinan, & Southern, 2014). Các triệu chứng ASD dễ thấy hơn và dễ nhận biết ở trẻ từ hai đến ba tuổi. Theo Towle và Patrich (2016), cứ 68 trẻ thì có một trẻ bị chứng ASD. Do đó, các phương pháp sàng lọc khác nhau đã được phát triển bởi các chuyên gia y tế và bác sĩ tâm thần hàng đầu trên thế giới nhằm tìm cách xác định các đặc điểm của ASD ở giai đoạn nguyên thủy để sẵn sàng cung cấp các hình thức can thiệp cần thiết (Robins, Fein, Barton, & Green, 2001).

Chẩn đoán ASD thường được tiến hành bởi các bác sĩ chuyên khoa trong môi trường lâm sàng sử dụng một quy trình chẩn đoán lâm sàng (*Clindical Judgment*) và dựa vào các chỉ số hành vi có thể quan sát, đo lường được. Các mô hình hiện thường dựa trên ý tưởng càng nhiều chỉ số đánh giá thì độ chính xác phân loại càng cao. Các phương pháp sàng lọc ASD thường dựa trên các phương pháp chẩn đoán lâm sàng do đó chúng thường mất thời gian do bộ sàng lọc có quá nhiều tiêu chí. Điều này đòi hỏi cần một phương pháp mới để khắc phục tình trạng này.

Các công cụ sàng lọc ASD thường được sử dụng các quy tắc do các chuyên gia khoa học tâm thần và hành vi xây dựng. Chất lượng của kết quả phân loại phụ thuộc đáng kể vào sự đóng góp chủ quan của các chuyên gia này và trình độ diễn giải của các nhân viên lâm sàng khi thực hiện đánh giá. Vì vậy, chẩn đoán ASD có thể được trao cho học máy–nơi các quyết định được đưa ra tự động dựa trên các thuật toán thông minh. Sử dụng học máy kết quả sẽ không bị ảnh hưởng của con người trong quá trình phân loại. Tuy nhiên các mô hình học máy sẽ không thay thế bác sĩ lâm sàng mà sẽ là công cụ hỗ trợ để cải thiện việc ra quyết định chẩn đoán.

2.1.2. Các phương pháp sàng lọc ASD

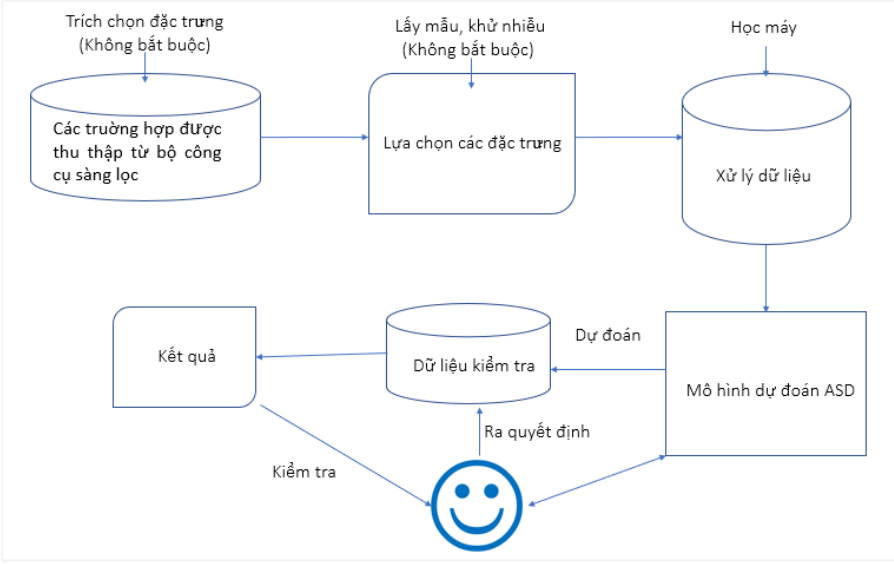
Quy trình chẩn đoán ASD rất khác nhau tuỳ theo cách tiếp cận, với mỗi công cụ chẩn đoán khác nhau lại có thể có quy trình khác nhau. Thông thường, quá trình chẩn đoán sẽ đến sau bước sàng lọc. Trẻ sẽ được sàng lọc để xác định nguy cơ ASD trước khi tiến hành một chẩn đoán chuyên sâu gồm các bước: (1) Mô tả lí do và mục đích chẩn đoán; (2) Phân tích tiền sử phát triển; (3) Nghiên cứu chẩn đoán tâm lý (sử dụng các công cụ chẩn đoán); và (4) Kết luận và đưa ra lời khuyên (Nguyễn, 2012).

Các công cụ sàng lọc và hỗ trợ chẩn đoán ASD phổ biến hiện nay là: Bảng kiểm sàng lọc tự kỷ ở trẻ nhỏ (*Check-list for Autism in Toddlers–CHAT*), Bảng kiểm sàng lọc tự kỷ ở trẻ nhỏ có sửa đổi (*Modifier Check-list Autism in Toddlers–M-CHAT 23*), Thang chẩn đoán tự kỷ tuổi ấu thơ (*Childhood Autism Rating Scale–CARS*), Bảng phỏng vấn chẩn đoán tự kỷ có điều chỉnh (*The Autism Diagnostic Interview-Revised–ADI-R*),Bảng quan sát chẩn đoán tự kỷ (*The Autism Diagnostic Observation Schedule– ADOS*), Thang đánh giá tự kỷ Gilliam (*Gilliam Autism Rating Scale–GARS),* và AQ (Nguyễn, 2012).

2.2. Ứng dụng học máy trong phân loại ASD

2.2.1. Mô hình học máy trong phân loại ASD

Mô hình học máy trong phân loại ASD được đề xuất bởi Thabtah (2018) thể hiện trong Hình 1.



Hình 1. Mô hình phân loại ASD sử dụng học máy

Nguồn: Thabtah (2018).

Các yêu cầu cần thiết ứng dụng học máy trong phân loại ASD là:

- *Đầu vào*: Tập dữ liệu, thông thường được thu thập bởi các công cụ sàng lọc như ADOS, AQ…

- *Xử lý*: Các thuật toán học máy bao gồm các thuật toán trích chọn đặc trưng và xử lý sẽ được áp dụng trên tập dữ liệu đã được thu thập.

- *Kết quả*: Là một mô hình dự đoán dùng để phân loại cho các trường hợp thử nghiệm.

- *Đánh giá của chuyên gia lâm sàng*: Đây là người sẽ đánh giá kết quả của mô hình dự đoán học máy để đưa ra kết quả quyết định cuối cùng. Kết quả của mô hình học máy sẽ giúp các chuyên gia lâm sàng rút ngắn được thời gian và nâng cao hiệu quả chẩn đoán ASD.

2.2.2. Các công trình ứng dụng học máy trong phân loại ASD

Để tiến hành các chẩn đoán ASD thường tốn nhiều thời gian và phụ thuộc trình độ của các chuyên gia lâm sàng nên các nhà khoa học đã nghĩ đến một quy trình chẩn đoán mới để tăng độ chính xác và tiết kiệm thời gian. Với sự phát triển của công nghệ thông tin, nhiều nhà khoa học đã áp dụng phương pháp học máy để hỗ trợ việc chẩn đoán. Các công trình tiêu biểu được thể hiện trong Bảng 1.

Bảng 1. Các nghiên cứu ứng dụng học máy trong phân loại ASD

Tác giả	Phương pháp	Bộ dữ liệu	Kết quả phân loại
Wall, Kosmicki, DeLuca, Harstad, và Fusaro (2012)	ADTree và Random Tree	AGRE và AC	Gần đạt 100.00%. ADTree (100.00%)
Mythili và Shanavas (2014)	ADTree	AGRE	87.00%
Bone, Goodwin, Black, Lee, Audhkhasi, và Narayanan (2014)	Random Forest	Georgia ADD Network 2008, Georgia ADD Network 2010	86.50%
Ramani và Sivaselvi (2017)	Naïve Bayes, SVM, Random Tree, C4.5, CS-CRT	CART (UCLA's Center for Autism Research and Treatment)	Random Tree (88.46%)
Stevens, Atchison, Stevens, Hong, Granpeesheh, Dixon, và Linstead (2017)	K-means	SKILIS	
Gök (2019)	Bayes network	InceRNA	
Demirhan (2018)	SVM, KNN, Random forest	Tự xây dựng	95.00%, 89.00%,100.00%
Basu (2018)	Autism Screening Adult Data Set	Decision Tree, Random Forest, Support Vector Machinees, KNN, Naïve Bayes, Logistcis Regression, Linear Discrimiant	SVM (100.00%)
McNamara, Lora, Yang, Flores, và Daly (2018)	Autism Screening Adult Data Set	Decision Tree, Random Forest	61.00%,79.00%

Từ Bảng 1 chúng ta có thể thấy, các nhà khoa học đã áp dụng nhiều thuật toán học máy trên các bộ dữ liệu khác nhau và cho kết quả rất khả quan. Tuy nhiên hầu hết các nghiên cứu đề sử dụng các bộ dữ liệu riêng và liên quan đến di truyền như của các tác giả Mythili & Shanavas (2014), Ramani & Sivaselvi (2017), Wall và ctg. (2012)… Các bộ dữ liệu thử nghiệm không được công khai nên gây khó khăn trong việc đánh giá so sánh kết quả.

Hiện nay chỉ có bộ dữ liệu sàng lọc ASD (trẻ em, trẻ vị thành niên, và người lớn) được giáo sư Fadi Fayez Thabtab, Đại học Auckland, New Zealand công bố trên kho dữ liệu UCI vào tháng 12 năm 2017. Trên bộ dữ liệu sàng lọc ASD người lớn (*Autism Screening Adult Data Set*) đã có các nghiên cứu của tiến sĩ Kanad Basu khảo sát các thuật toán học máy Decision Trees, Random Forest, Support Vector Machines (SVM), k-Nearest Neighbors(kNN), Naive Bayes Classification, Logistic Regression, Linear Discriminant Analysis (LDA), và Multi Layer Perception (MLP) (Basu, 2018) giải quyết bài toán phân loại ASD với người lớn. Nghiên cứu của tiến sĩ Basu Kanad chỉ ra rằng giải thuật SVM và Random Forest là hai giải thuật tốt nhất để phân loại ASD. Một nghiên cứu khác của Brian McNamara và cộng sự khi khảo sát hai giải thuật Decision Trees, Random Forest cũng chỉ ra sự hiệu quả của giải thuật Random Forest trong phân loại ASD (McNamara & ctg., 2018). Tuy nhiên các nghiên cứu của các giả Kanad Basu và Brian McNamara mới chỉ khảo sát các thuật toán học máy trên bộ dữ liệu sàng lọc ASD người lớn đồng thời không có dữ liệu thực tế để kiểm nghiệm từ đánh giá hiệu quả của mô hình học máy.

Từ các nghiên cứu trên chỉ ra rằng, các thuật toán học máy nổi bật là SVM, Random forest, ADTree... có hiệu quả trong xây dựng mô hình học máy để hỗ trợ quá trình phân loại ASD.

2.3. Đánh giá bộ cơ sở dữ liệu phân loại ASD

2.3.1. Bộ dữ liệu

Bộ dữ liệu huấn luyện: Chúng tôi sử dụng bộ dữ liệu sàng lọc ASD trẻ em (*Autistic Spectrum Disorder Screening Data for Children Data Set*) được công bố trên bộ dữ liệu UCI. Bộ dữ liệu dùng cho các nghiên cứu sàng lọc, phân loại, dự đoán chứng ASD ở trẻ em. Bộ dữ liệu có 292 trường hợp với 21 đặc trưng, trong đó có 141 trường hợp được phân lớp là ASD và 151 trường hợp không được phân lớp bị ASD (Hình 2).

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	age	gender	ethnicity	jundice	austim	contry_of_res	used_app_before	result	age_desc	relation	Class/ASD
2	1	1	0	0	1	1	0	1	0	0	6	m	Others	no	no	Jordan	no	5	'4-11 years'	Parent	NO
3	1	1	0	0	1	1	0	1	0	0	6	m	'Middle Ea	no	no	Jordan	no	5	'4-11 years'	Parent	NO
4	1	1	0	0	0	1	1	1	0	0	6	m	?	no	no	Jordan	yes	5	'4-11 years'	?	NO
5	0	1	0	0	1	1	0	0	0	1	5	f	?	yes	no	Jordan	no	4	'4-11 years'	?	NO
6	1	1	1	1	1	1	1	1	1	1	5	m	Others	yes	no	'United States'	no	10	'4-11 years'	Parent	YES
7	0	0	1	0	1	1	0	1	0	1	4	m	?	no	yes	Egypt	no	5	'4-11 years'	?	NO
8	1	0	1	1	1	1	0	1	0	1	5	m	White-Eur	no	no	'United Kingdc	no	7	'4-11 years'	Parent	YES
9	1	1	1	1	1	1	1	1	0	0	5	f	'Middle Ea	no	no	Bahrain	no	8	'4-11 years'	Parent	YES
10	1	1	1	1	1	1	1	0	0	0	11	f	'Middle Ea	no	no	Bahrain	no	7	'4-11 years'	Parent	YES

Hình 2. Hình ảnh 10 trường hợp đầu tiên của bộ dữ liệu sàng lọc ASD trẻ em

Bộ dữ liệu kiểm nghiệm: Bộ dữ liệu kiểm định được chúng tôi xây dựng với sự hỗ trợ của các chuyên gia của Trung tâm Tư vấn và Hỗ trợ giáo dục đặc biệt, Trường Cao đẳng Sư phạm Trung ương-Nha Trang. Các bước xây dựng bộ dữ liệu này như sau: (1) Các chuyên gia sử dụng bộ câu hỏi AQ-10 cho trẻ em trên ứng dụng ASD Test để đánh giá các trường hợp mắc ASD tại trung tâm; và (2) Tiến hành mã hóa dữ liệu. Kết quả chúng tôi thu được 18 trường hợp trong đó có 10 trường hợp đã được chẩn đoán lâm sàng mắc ASD và 8 trường hợp không bị ASD (Hình 3).

A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	age	Gender	Jundice	Austim	Relation	Class
1	1	1	1	1	1	1	1	1	1	5	F	NO	NO	Health care professional	YES
0	0	1	1	1	1	1	1	1	1	5	F	NO	NO	Health care professional	YES
1	0	1	1	1	1	1	1	0	0	4	F	NO	NO	Health care professional	YES
1	1	1	1	1	1	1	1	1	1	5	F	NO	NO	Health care professional	YES
1	1	1	1	1	1	1	1	1	1	8	F	NO	NO	Health care professional	YES
0	1	1	1	1	1	1	1	1	1	5	F	NO	NO	Health care professional	YES
1	1	1	1	1	1	1	1	1	1	5	F	NO	NO	Health care professional	YES
1	0	1	1	1	1	0	1	1	1	5	M	NO	NO	Health care professional	YES
1	0	1	1	1	1	1	0	1	1	4	F	NO	NO	Health care professional	YES
1	1	1	1	1	1	1	1	1	1	4	F	NO	NO	Health care professional	YES
1	0	0	1	0	1	0	1	1	0	11	M	NO	NO	Health care professional	NO
1	1	0	1	0	0	0	0	0	0	4	M	NO	NO	Health care professional	NO
1	1	0	0	0	0	0	0	0	0	7	M	NO	NO	Health care professional	NO
0	1	0	0	0	1	1	1	1	1	6	F	NO	NO	Health care professional	NO
0	1	1	1	0	1	0	0	0	1	10	F	NO	NO	Health care professional	NO
0	1	1	0	1	1	0	0	0	1	5	M	NO	NO	Health care professional	NO
0	1	0	0	0	1	1	1	0	0	11	M	NO	NO	Health care professional	NO
0	0	1	0	1	0	0	0	0	1	7	M	NO	NO	Health care professional	NO

Hình 3. Bộ dữ liệu thực tế trẻ mắc chứng ASD được xây dựng bởi các chuyên gia của Trung tâm Tư vấn và Hỗ trợ giáo dục đặc biệt, Trường Cao đẳng Sư phạm Trung ương-Nha Trang

2.3.2. Xây dựng mô hình dự đoán

Các bước xây dựng mô hình dự đoán như sau:

Bước 1: Làm sạch dữ liệu. Ở bước này chúng tôi tiến hành xóa các trường hợp có dữ liệu bị thiếu. Trong 292 trường hợp của bộ dữ liệu có 44 trường hợp có dữ liệu bị thiếu (NA) ở các thuộc tính *ethnicity* (tôn giáo) và *relation* (người thực hiện kiểm tra). Các giá trị còn thiếu chủ yếu là kiểu dữ liệu phân loại. Điều này gây khó khăn cho việc tạo các giá trị thay thế vì chúng ta không thể thay thế giá trị trung bình cho các biến không phải là kiểu dữ liệu số. Sau khi xóa các trường hợp này, tập dữ liệu còn 248 trường hợp. Trong đó có 126 trường hợp được phân loại ASD và 122 trường hợp không bị ASD.

Bước 2: Lựa chọn đặc trưng: Chúng tôi sử dụng phương pháp Chi Square (CHI) (Bahassine, Madani, Al-Sarem, & Kissi, 2018; Thabtah, 2018) để đánh giá giá độ liên quan của các đặc trưng tới kết quả phân lớp.

CHI-SQ (Công thức 1) tính toán mối tương quan giữa các biến thuộc tính (*variable–v*) và biến mục tiêu (*class–l*) sử dụng xác suất mong đợi và kết quả quan sát của chúng trong tập dữ liệu huấn luyện (*T*).

$$CHI - Square(v,l) = \frac{S \times (AD - BC)^2}{(A+C) \times (B+D) \times (A+B) \times (C+D)} \quad (1)$$

Trong đó: A là tần số cặp (v,l) trong T; B là tần số của biến v không có lớp l trong T; C là tần số của lớp l không có biến v trong T; D là tần số của các trường hợp không có cả (v,l) trong T; S là kích thước của T.

Kết quả, với phương pháp CHI-SQ chúng tôi thu được 10 đặc trưng đầu tiên A1-A10 của bộ dữ liệu có mối tương quan nhiều nhất đến biến phân lớp. Chúng tôi chọn 10 đặc trưng này để xây dựng mô hình.

Bước 3: Xây dựng mô hình: Với 10 đặc trưng thu được từ quá trình lựa chọn đặc trưng, chúng tôi xây dựng mô hình dự đoán dựa trên các thuật toán học máy SVM và Random Forest. Tuy nhiên để chọn được mô hình tốt nhất, chúng tôi tiến hành khảo sát thêm các thuật toán: Decision Trees, Logistic Regression, K-Nearest-Neighbors, Naïve Bayes, và mạng nơ ron Multi Layer Perceptron. Bộ dữ liệu sẽ được chia làm hai phần: 80% (198 trường hợp) dùng để huấn luyện mô hình và 20% (50 trường hợp) dùng để xác thực, kiểm thử.

Để nâng cao chất lượng mô hình chúng tôi sử dụng kỹ thuật xác thực chéo (*Cross-validation*) với k = 10. Vì số lượng dữ liệu hạn chế, nếu lấy quá nhiều dữ liệu trong tập huấn luyện ra làm dữ liệu xác thực, phần dữ liệu còn lại không đủ để xây dựng mô hình. Lúc này, tập xác thực phải thật nhỏ để giữ được lượng dữ liệu huấn luyện đủ lớn. Xác thực chéo là một cải tiến của xác thực với lượng dữ liệu trong tập xác thực là nhỏ nhưng chất lượng mô hình được đánh giá trên nhiều tập xác thực khác nhau. Đây là một phương pháp kiểm tra được sử dụng để đánh giá hiệu suất của của mô hình dự đoán (Kohavi, 1995). Để cài đặt các thuật toán học máy, chúng tôi sử dụng máy tính Intel ®, Core i5-5200U CPU 2.2 GHz, Ram 8GB và sử dụng các gói thư viện sklearn và keras của Python trên môi trường lập trình PyScripter. Kết quả xây dựng mô hình thể hiện thông qua các thang đánh giá Accuracy, Sensitivity (Recall), Specificity, Precision, F-1, cross_val_score (Bảng 2).

Bảng 2. Kết quả xây dựng mô hình

STT	Thuật toán học máy	Thang đánh giá					
		Accuracy	Sensitivity (Recall)	Specificity	Precision	F-1	cross_val_score
1	Decision Trees	0.92	0.96	0.86	0.90	0.91	0.93
2	Random Forest	0.94	1.00	0.86	0.90	0.92	0.93
3	SVM	1.00	1.00	1.00	1.00	1.00	1.00
4	Logistic Regression	1.00	1.00	1.00	1.00	1.00	0.99
5	K-Nearest-Neighbors	0.92	1.00	0.82	0.88	0.89	0.90
6	Naïve Bayes	0.72	0.64	0.82	0.82	0.78	0.64
7	Multi Layer Perceptron	0.96	0.96	0.95	0.96	0.96	1.00

Từ kết quả thử nghiệm ở Bảng 2, dựa vào thang đo độ chính xác phân loại thì các giải thuật SVM, Logistic Regression Multilayer Perceptron, K-Nearest-Neighbors, và Random Forest cho kết quả phân loại ASD cao. Các kết quả này phù hợp với các nghiên cứu trước đó được thể hiện ở Bảng 1.

2.3.3. Thử nghiệm trên bộ cơ sở dữ liệu thực tế

Chúng tôi tiến hành thử nghiệm mô hình của bảy thuật toán trên bộ dữ liệu thực tế. Kết quả dự đoán được thể hiện trên Bảng 3.

Bảng 3. Kết quả thực nghiệm bảy thuật toán trên bộ dữ liệu thực tế

Thuật toán	DecisionTrees	RandomForest	SVM	LogisticRegression	KNN	NaiveBayes	MLP
Số lượng trường hợp dự đoán đúng	17	18	18	18	18	15	18
Tỷ lệ (%)	94%	100%	100%	100%	100%	83%	100%

Từ Bảng 3 chúng tôi rút ra nhận xét, các thuật toán RandomForest, SVM, LogisticRegression, KNN, và MLP cho kết quả dự đoán tốt trên bộ dữ liệu thực tế. Điều này có thể giải thích bởi bộ dữ liệu được xây dựng trên bộ câu hỏi được các chuyên gia tâm lý phát triển và thử nghiệm nên các đặc trưng đã thể hiện được độ tin cậy. Mặt khác các đặc trưng ứng với các trường hợp bị ASD và không bị ASD trong bộ dữ liệu khá rõ ràng.

Căn cứ vào kết quả xây dựng mô hình, kết quả thực nghiệm trong nghiên cứu của chúng tôi và kết quả nghiên cứu của Thabtah (2018) thì mô hình thuật toán SVM là tin cậy. Nó có thể dùng để phát triển ứng dụng sàng lọc ASD trẻ em.

3. KẾT LUẬN VÀ KIẾN NGHỊ

3.1. Kết luận

Chúng tôi đã tiến hành đánh giá bộ dữ liệu sàng lọc ASD trẻ em với các thuật toán SVM và Random Forest, đồng thời khảo sát thêm các thuật toán Decision Trees, Logistic Regression, K-Nearest-Neighbors, Naïve Bayes, và MLP. Kết quả thử nghiệm trên các bảy thuật toán cho kết quả phân loại cao phù hợp với các nghiên cứu trước đó. Chúng tôi đề xuất sử dụng mô hình thuật toán SVM để sử dụng phát triển ứng dụng sàng lọc ASD trẻ em

Như vậy, có thể khẳng định rằng bộ dữ liệu dùng để xây dựng các mô hình phân loại ASD trẻ em là đáng tin cậy. Bộ dữ liệu này có thể sử dụng để xây dựng các mô hình hỗ trợ sàng lọc ASD. Đây là một hướng nghiên cứu khả quan có thể áp dụng vào thực tiễn trong tương lai.

3.2. Kiến nghị

Trên cơ sở các kết quả đã thu được, hướng phát triển tiếp theo của chúng tôi là: (1) Tiếp tục kết hợp với các chuyên gia giáo dục đặc biệt thu thập xây dựng bộ dữ liệu sàng lọc ASD trẻ em ở Việt Nam; và (2) Phát triển ứng dụng sàng lọc ASD cho trẻ em Việt Nam. Với tỷ lệ trẻ em ở Việt Nam mắc chứng ASD ngày càng tăng, ứng dụng sàng lọc sẽ giúp cha mẹ và người chăm sóc có thể sàng lọc ASD sớm từ đó có những biện pháp can thiệp kịp thời góp phần giảm gánh nặng cho gia đình và xã hội.

TÀI LIỆU THAM KHẢO

Bahassine, S., Madani, A., Al-Sarem, M., & Kissi, M. (2018). Feature selection using an improved Chi-square for Arabic text classification. *Journal of King Saud University-Computer and Information Sciences, 32*(2), 225-231. https://doi.org/10.1016/j.jksuci.2018.05.010.

Basu, K. (2018). *Autism screening adult data set : A machine learning approach*. Retrieved from https://github.com/kbasu2016/Autism-Detection-in-Adults/blob/master/proposal.pdf.

Bone, D., Goodwin, M. S., Black, M. P., Lee, C. C., Audhkhasi, K., & Narayanan, S. (2014). Applying machine learning to facilitate autism diagnostics: Pitfalls and promises. *Journal of Autism and Developmental Disorders, 45*(5), 1121-1136. https://doi.org/10.1007/s10803-014-2268-6.

Demirhan, A. (2018). Performance of machine learning methods in determining the autism spectrum disorder cases. *Mugla Journal of Science and Technology, 4*(1), 79-84. https://doi.org/10.22531/muglajsci.422546.

Doanh, Đ. (2018). *Cần sớm hoàn thiện và đưa tài liệu về hỗ trợ trẻ em tự kỷ vào cuộc sống*. Được truy lục từ http://laodongxahoi.net/can-som-hoan-thien-va-dua-tai-lieu-ve-ho-tro-tre-em-tu-ky-vao-cuoc-song-1310672.html.

Fischbach, G. D., & Lord, C. (2010). The simons simplex collection: A resource for identification of autism genetic risk factors. *Neuron, 68*(2), 192-195. https://doi.org/10.1016/j.neuron.2010.10.006.

Geschwind, D. H., Sowinski, J., Lord, C., Iversen, P., Shestack, J., Jones, … Spence, S. J. (2001). The autism genetic resource exchange: A resource for the study of autism and related neuropsychiatric conditions. *The American Journal of Human Genetics, 69*(2), 463-466. https://doi.org/10.1086/321292.

Gök, M. (2019). A novel machine learning model to predict autism spectrum disorders risk gene. *Neural Computing and Applications, 31*(10), 6711-6717. https://doi.org/10.1007/s00521-018-3502-5.

Hall, D., Huerta, M. F., McAuliffe, M. J., & Farber, G. K. (2012). Sharing heterogeneous data: The national database for autism research. *Neuroinformatics, 10*(4), 331-339. https://doi.org/10.1007/s12021-012-9151-4.

Kohavi, R. (1995). *A study of cross-validation and bootstrap for accuracy estimation and model selection*. Paper presented at The 14th International Joint Conference on Artificial Intelligence, Quebec, Canada.

Mythili, M. S., & Shanavas, A. R. M. (2014). A study on autism spectrum disorders using classification techniques. *International Journal of Soft Computing and Engineering*, *4*(5), 88-91.

McNamara, B., Lora, C., Yang, D., Flores, F., & Daly, P. (2018). *Machine learning classification of adults with autism spectrum disorder*. Retrieved from http://rstudio-pubs-static.s3.amazonaws.com/383049_1faa93345b324da6a1081506f371a8dd.html.

Nguyễn, N. T. A. (2012). Một số vấn đề cơ bản trong chuẩn đoán rối loạn phổ tự kỷ. *Tạp Chí Khoa Học ĐHQGHN, Khoa Học Xã Hội và Nhân Văn*, *28*, 143-147.

Pennington, M. L., Cullinan, D., & Southern, L. B. (2014). Defining autism: Variability in state education agency definitions of and evaluations for autism spectrum disorders. *Autism Research and Treatment*, *2014*, 1-8. https://doi.org/10.1155/2014/327271.

Ramani, R. G., & Sivaselvi, K. (2017). Autism spectrum disorder identification using data mining techniques. *International Journal of Pure and Applied Mathematics*, *117*(16), 427-436.

Robins, D. L., Fein, D., Barton, M. L., & Green, J. A. (2001). The modified checklist for autism in toddlers: An initial study investigating the early detection of autism and pervasive developmental disorders. *Journal of Autism and Developmental Disorders*, *31*(2), 131-144. https://doi.org/10.1023/A:1010738829569.

Stevens, E., Atchison, A., Stevens, L., Hong, E., Granpeesheh, D., Dixon, D., & Linstead, E. (2017). *A cluster analysis of challenging behaviors in autism spectrum disorder*. Paper presented at The 16th IEEE International Conference on Machine Learning and Applications, Cancun, Mexico. https://doi.org/10.1109/ICMLA.2017.00-85.

Thabtah, F. A. (2017a). *Autism screening adult data set*. Retrieved from https://archive.ics.uci.edu/ml/datasets/Autism+Screening+Adult.

Thabtah, F. A. (2017b). *Autistic spectrum disorder screening data for adolescent data set*. Retrieved from https://archive.ics.uci.edu/ml/datasets/Autistic+Spectrum+Disorder+Screening+Data+for+Adolescent+++.

Thabtah, F. A. (2017c). *Autistic spectrum disorder screening data for children data set*. Retrieved from https://archive.ics.uci.edu/ml/datasets/Autistic+Spectrum+Disorder+Screening+Data+for+Children++.

Thabtah, F. A. (2018). *Detecting autistic traits using computational intelligence & machine learning techniques*. Retrived from http://eprints.hud.ac.uk/id/eprint/34844/.

The United Nations. (n.d). *World autism awareness day 2 April*. Retrieved from https://www.un.org/en/observances/autism-day/background.

Towle, P. O., & Patrick, P. A. (2016). Autism spectrum disorder screening instruments for very young children: A systematic review. *Autism Research and Treatment*, *2016*, 1-29. https://doi.org/10.1155/2016/4624829.

Wall, D. P., Kosmicki, J., DeLuca, T. F., Harstad, E., & Fusaro, V. A. (2012). Use of machine learning to shorten observation-based screening and diagnosis of autism. *Translational Psychiatry*, *2*(4), 1-8. https://doi.org/10.1038/tp.2012.10.