Journal of Science and Transport Technology

University of Transport Technology

Application of data science approach to predicting the cultivation ages of ginseng and analyzing affecting variables

Do Quang Hung[1], Ngo Thi Thu Tinh[1*], Nguyen Phuong Linh[2]

[1] University of Transport Technology, Hanoi 100000, Vietnam

[2] Hanoi - Amsterdam Highschool for the Gifted, Hanoi 100000, Vietnam

Article info

Type of article:
Original research paper

***Corresponding author:**
E-mail address:
tinhntt@utt.edu.vn

Received: 21/7/2022
Accepted: 22/8/2022
Published: 24/8/2022

Abstract: The cultivation ages of ginseng are important factors that influence the quality and price of ginseng. Recent advances in data science have created great benefits for various practical applications. In data science, machine learning plays a vital role to discover the insights from data. This study develops and assesses the performance of three machine learning models, including Extreme Gradient Boosting (XGB), Light Gradient Boosting (LGB), and Gradient Boosting (GB), in predicting the cultivation age of ginseng (CAG). The models are developed based on 106 data samples with nine input parameters and one output parameter. The K-fold cross-validation technique is used to improve the models' generalizability and predictive performance. Importantly, the XGB model is optimized to find the hyperparameters. The predictive performance of the optimal XGB model is compared to the performance of the LG and GB models. The results show that the XGB is the best model with very high predictive performance (R^2=0.964, RMSE=0.148 years, MAE=0.107 years). The sensitivity analysis using the feature importance is performed to evaluate the influence of input variables on the predicted CAG.

Keywords: Cultivation age of ginseng (CAG), Machine learning (ML), Extreme Gradient Boosting (XGB), Data science.

Tạp chí điện tử
Khoa học và Công nghệ Giao thông
Trường Đại học Công nghệ GTVT

Ứng dụng phương pháp khoa học dữ liệu để dự báo tuổi phát triển của sâm và phân tích các yếu tố ảnh hưởng

Đỗ Quang Hưng[1], Ngô Thị Thu Tình[1], Nguyễn Phương Linh[2]
[1]Khoa Công nghệ thông tin, Trường Đại học Công nghệ Giao thông Vận tải, Hà Nội, Việt Nam
[2]Trường THPT Chuyên Hà Nội - Amsterdam, Hà Nội, Việt Nam

Thông tin bài viết
Tác giả liên hệ:
Địa chỉ E-mail:
tinhntt@utt.edu.vn

Ngày nộp bài: 21/7/2022
Ngày chấp nhận:22/8/2022
Ngày đăng bài:24/8/2022

Tóm tắt: Tuổi phát triển của sâm (Cultivation ages of ginseng – CAG) là yếu tố quan trọng ảnh hưởng đến chất lượng và giá thành của sâm. Những tiến bộ gần đây trong khoa học dữ liệu đã tạo ra những lợi ích to lớn cho đa dạng các ứng dụng thực tế. Trong lĩnh vực khoa học dữ liệu, học máy đóng một vai trò quan trọng để khám phá thông tin chi tiết từ dữ liệu. Nghiên cứu này dựa trên cơ sở dữ liệu thực nghiệm thu thập được nhằm xây dựng và đánh giá hiệu suất của 3 mô hình máy học: Tăng cường độ dốc cực cao - Extreme Gradient Boosting (XGB), Tăng cường độ dốc nhẹ - Light Gradient Boosting (LGB) và Tăng cường độ dốc - Gradient Boosting (GB) trong việc dự đoán CAG. Các mô hình được phát triển dựa trên 106 mẫu dữ liệu với chín tham số đầu vào và một tham số đầu ra. Kỹ thuật xác thực chéo K-lần được sử dụng để nâng cao khả năng tổng quát hóa và hiệu suất dự báo của mô hình. Quan trọng hơn, trong nghiên cứu này các mô hình máy học được tối ưu hóa để lựa chọn các siêu tham số. Hiệu suất dự báo của 3 mô hình XGB, LGB và GB sau khi tối ưu hóa tham số được so sánh để chọn ra mô hình máy học tốt nhất nhằm dự báo CAG. Kết quả cho thấy XGB là mô hình tốt nhất với hiệu suất dự đoán rất cao (R^2=0,964; RMSE=0,148 năm, MAE=0,107 năm). Ngoài ra, kỹ thuật tầm quan trọng của tính năng (Feature importance) được thực hiện để đánh giá ảnh hưởng của các biến đầu vào đối với CAG dự đoán.
Từ khóa: Tuổi phát triển của sâm (CAG), mô hình máy học (ML), mô hình tăng cường độ dốc (XGB), Khoa học dữ liệu.

1. Đặt vấn đề

Từ hàng nghìn năm trước, nhân sâm đã là một vị thuốc quý trong lĩnh vực y học cổ truyền. Cho đến nay, nhân sâm vẫn được ưa chuộng trên toàn thế giới [1]. Sâm có nhiều tác dụng rất tốt trong hỗ trợ điều trị và chăm sóc sức khỏe, cụ thể là tăng cường khả năng miễn dịch, phòng chống ung thư [2], chống ô xi hóa [3], cải thiện chức năng nhận thức thần kinh [4], điều trị rối loạn lipid máu và các lợi ích khác mà không có tác dụng phụ [5]. Với nhiều tác dụng như vậy, chất lượng của sâm có liên quan trực tiếp đến sức khỏe và lợi ích của người sử dụng. Chất lượng của sâm chịu ảnh hưởng của nhiều yếu tố như loại sâm, xuất xứ, tuổi phát triển của sâm, phương thức trồng trọt và công nghệ sản xuất [6]. Trong đó, tuổi phát triển của sâm (CAG) là một yếu tố quan trọng, quyết định đến chất lượng và giá cả của sâm vì nó ảnh hưởng phần lớn đến việc tích lũy các hợp chất hoạt tính sinh học của sâm [7]. Nhìn chung, sâm có tuổi phát triển càng lâu năm càng có giá trị. Cùng một loại sâm nhưng tuổi phát triển khác nhau có giá bán

trên thị trường rất khác nhau. Lợi dụng điều này, sâm non tuổi có thể bị pha tạp chất hoặc tráo đổi thành sâm lâu đời để bán với giá cao. Do đó, điều quan trọng là phải phát triển một phương pháp đáng tin cậy để xác định tuổi phát triển của sâm nhằm chống lại việc làm sai lệch tuổi sâm.

Phương pháp truyền thống để xác định tuổi phát triển sâm là quan sát các đặc điểm hình thái và vi thể của sâm như số lượng ngạnh, số vết tích trên thân sâm trước khi thu hoạch [7]. Tuy nhiên, phương pháp này đòi hỏi người mua phải có kỹ năng phân biệt tốt, và nó không mang tính khách quan [8]. Hơn nữa, phương pháp quan sát này cũng không thể áp dụng đối với các sản phẩm sâm trên thị trường đã mất đi các đặc điểm hình thái này.

Cho đến nay, một số phương pháp mới và hiệu quả hơn trong xác định tuổi phát triển của sâm đã được nghiên cứu như: phương pháp cộng hưởng từ hạt nhân (NMR), phương pháp sắc ký lớp mỏng (TLC), phương pháp sắc ký lỏng hiệu năng cao (HPLC), phương pháp sắc ký lỏng hiệu suất cực cao ghép đầu dò khối phổ (UPLC/Q-TOF-MS) [9], [10], [7]. Các phương pháp này sử dụng kỹ thuật phân tích hiện đại, cho phép phân tích định tính và định lượng các thành phần hoạt tính trong dược liệu, từ đó xây dựng mô hình dự báo tuổi phát triển của sâm. Tuy nhiên, các kỹ thuật xử lý khá phức tạp, tốn nhiều thuốc thử, tốn nhiều thời gian và chi phí. Hơn nữa, một số nghiên cứu có bản chất tuyến tính, không thể mô tả đầy đủ các mối quan hệ nội tại giữa các cấu hình hóa lý và năm tăng trưởng của sâm. Kết quả dự đoán sẽ dễ dàng thất bại đối với các bài toán phức tạp, đặc biệt là khi mối quan hệ giữa các biến đầu vào và đầu ra không rõ ràng [11]. Vì vậy, một phương pháp hiện đại để dự đoán tuổi phát triển của sâm cần được phát triển nhằm giảm chi phí, thời gian, cho kết quả dự báo tin cậy.

Khoa học dữ liệu là ngành khoa học nhằm rút ra những hiểu biết sâu sắc từ dữ liệu bao gồm cả dữ liệu thô và dữ liệu không có cấu trúc. Trong những năm gần đây, máy học (ML) – một lĩnh vực của khoa học dữ liệu đang dần trở nên phổ biến và được ứng dụng trong nhiều lĩnh vực khoa học như kỹ thuật dân dụng [12], khoa học thực vật [13], y tế và chăm sóc sức khỏe [11], [14]. Ưu điểm của ML là dựa trên cơ sở dữ liệu sẵn có, nó có thể học các hành vi cơ bản của một hệ thống phức tạp mà không cần biết trước mối quan hệ giữa các biến đầu vào và đầu ra, từ đó dự báo được tham số đầu ra. ML sử dụng các thuật toán cho phép máy tính học từ các dữ liệu sẵn có nhằm thực thiện các công việc thay vì phải trình một cách rõ ràng. Vì vậy, trong nghiên cứu này, dựa trên một bộ dữ liệu gồm 106 mẫu, ba mô hình học máy XGB, LGB, GB được xây dựng để dự báo CAG. Việc quan trọng nhất khi xây dựng các mô hình học máy chính là lựa chọn các siêu tham số để mô hình đạt hiệu quả dự báo tốt nhất. Ở đây, các mô hình máy học sử dụng được tối ưu hóa nhằm lựa chọn ra các siêu tham số. Tiếp theo, hiệu suất dự báo của ba mô hình XGB, LGB, và GB tối ưu được so sánh. Các tiêu chí đánh giá hiệu suất dự báo sử dụng là Hệ số xác định (R^2), Sai số tuyệt đối trung bình (MAE), Sai số toàn phương trung bình (RMSE). Kết quả là mô hình XGB cho hiệu suất dự báo tốt nhất, được lựa chọn để dự báo tuổi phát triển của sâm. Hơn thế nữa, trong nghiên cứu này mức độ ảnh hưởng của các tham số đầu vào đến tuổi phát triển của sâm được đánh giá bằng kỹ thuật “Tầm quan trọng của các tính năng”.

2. Cơ sở dữ liệu

Hiệu suất dự đoán của mô hình học máy phụ thuộc vào nhiều yếu tố, chẳng hạn như tính đầy đủ của dữ liệu đào tạo, số lượng dữ liệu, mối quan hệ giữa dữ liệu đầu vào và đầu ra. Trong nghiên cứu này, một bộ dữ liệu bao gồm 106 dữ liệu thực nghiệm được thu thập từ bài báo đã được đăng trên các tạp chí uy tín trên thế giới [11].

Bộ dữ liệu sử dụng trong nghiên cứu này bao gồm 9 thông số đầu vào là: Chiều dài mẫu, cm (X_1), Trọng lượng mẫu, g (X_2), hàm lượng chất hòa tan trong cồn, % (X_3), hàm lượng chất hòa tan trong nước, % (X_4); Rg1 ,% (X_5); Rd ,% (X_6); Re, % (X_7); Rb1, % (X_8); F11, % (X_9). Chỉ có duy nhất một biến đầu ra là tuổi phát triển của sâm, năm (Y). Trong

đó Rg1, Rd, Re, Rb1, F11 là các hoạt chất chính của sâm. Bảng 1 trình bày chi tiết ký hiệu, đơn vị, số lượng và phân tích thống kê, bao gồm: giá trị trung bình, độ lệch chuẩn (Std), giá trị nhỏ nhất, giá trị lớn nhất (max), giá trị ở góc 10%, 20%... của các tham số đầu vào cũng như tham số đầu ra.

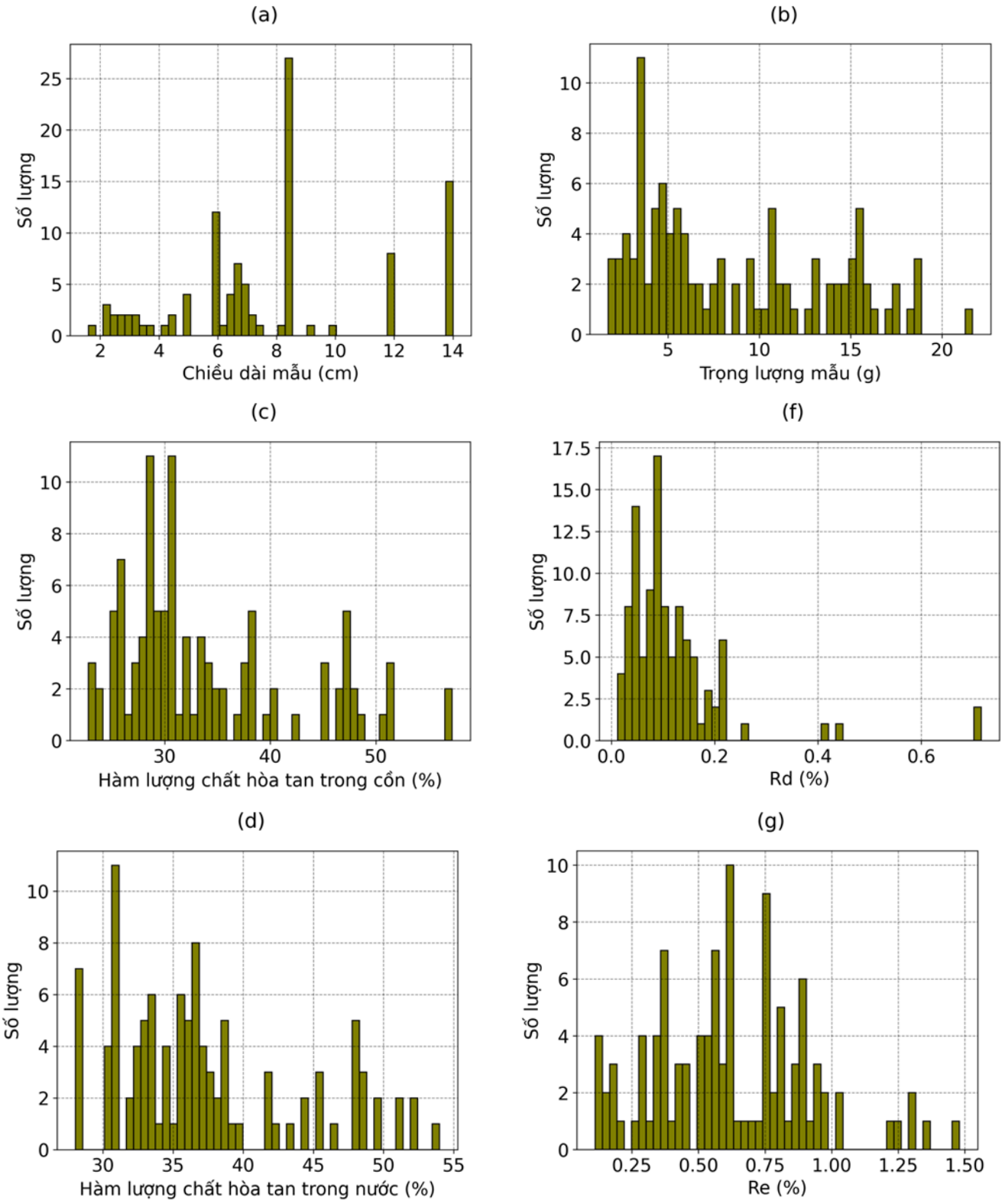
Bảng 1. Phân tích thống kê cơ sở dữ liệu

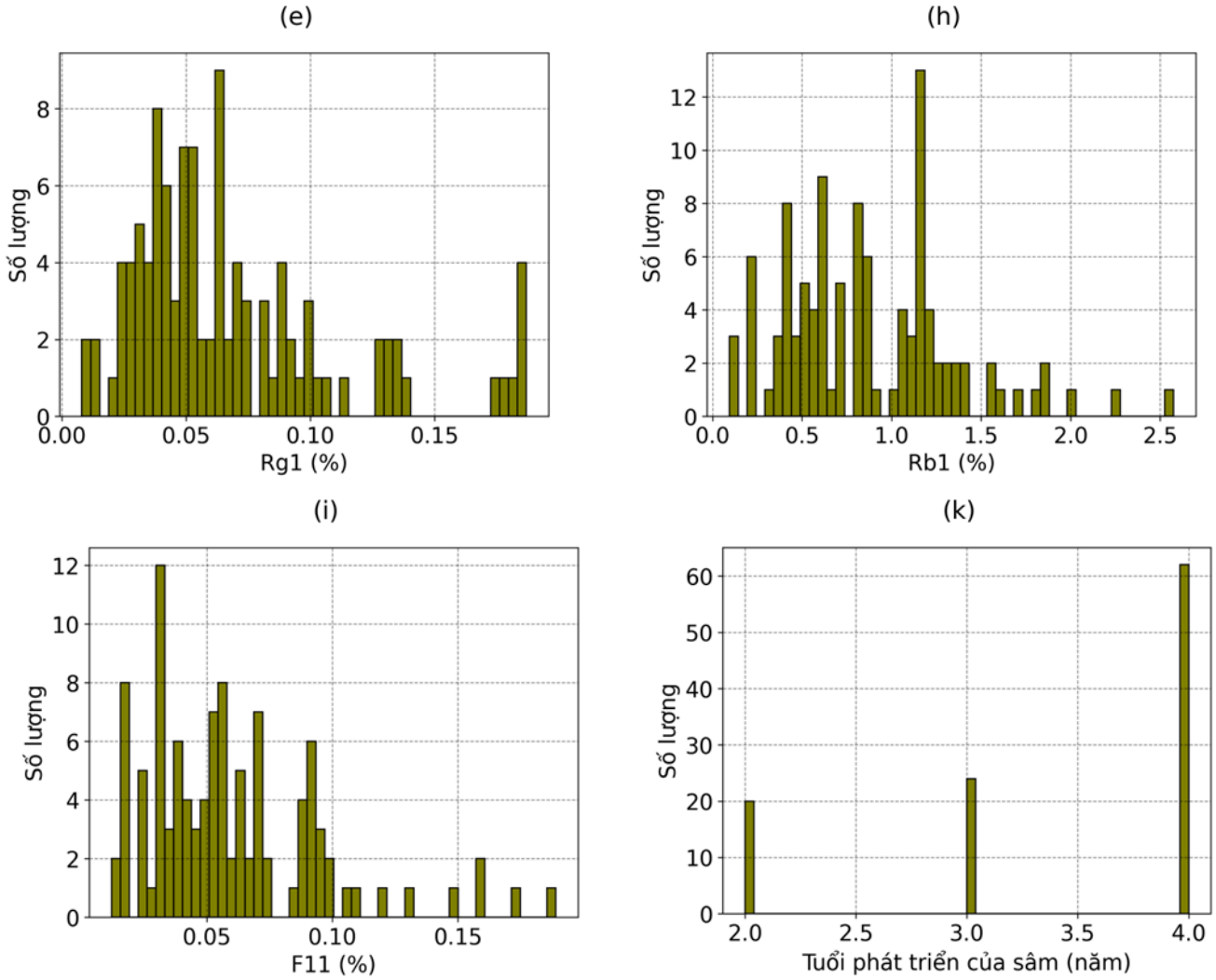
Tên	Chiều dài mẫu	Trọng lượng mẫu	Chất hòa tan trong cồn	Chất hòa tan trong nước	Rg1	Rd	Re	Rb1	F11	Tuổi phát triển của sâm
Đơn vị	(cm)	(g)	(%)	(%)	(%)	(%)	(%)	(%)	(%)	(năm)
Ký hiệu	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	Y
Vai trò	Đầu vào	Đầu vào	Đầu vào	Đầu vào	Đầu vào	Đầu vào	Đầu vào	Đầu vào	Đầu vào	Đầu ra
Số lượng	106	106	106	106	106	106	106	106	106	106
Trung bình	7,938	8,543	33,899	37,251	0,068	0,120	0,623	0,880	0,059	3,396
Độ lệch chuẩn	3,419	5,162	8,314	6,590	0,043	0,109	0,294	0,477	0,035	0,789
Nhỏ nhất	1,6	1,72	22,772	28,031	0,008	0,012	0,112	0,093	0,012	2
10%	3,2	3,145	25,563	30,62	0,028	0,039	0,2695	0,366	0,023	2
20%	6	3,6	27,664	31,09	0,036	0,049	0,377	0,444	0,03	3
30%	6	4,625	28,613	32,943	0,042	0,0735	0,454	0,572	0,037	3
40%	6,75	5,4	30,05	34,749	0,049	0,085	0,561	0,651	0,045	3
50%	7,8	6,69	30,908	35,9295	0,0545	0,093	0,609	0,824	0,053	4
60%	8,5	9,62	33,418	36,76	0,064	0,103	0,672	1,035	0,059	4
70%	8,5	11,335	36,3285	38,5595	0,0745	0,131	0,764	1,17	0,069	4
80%	12	14,19	40,641	43,246	0,092	0,157	0,876	1,186	0,088	4
90%	14	15,74	47,43	48,252	0,1315	0,211	0,9445	1,407	0,096	4
Lớn nhất	14	21,68	57,226	53,993	0,187	0,716	1,48	2,576	0,189	4

Cơ sở dữ liệu thu thập được được tách ngẫu nhiên thành hai phần để phát triển các thuật toán ML (phần dữ liệu đào tạo và phần dữ liệu kiểm tra). Phần dữ liệu đào tạo chiếm 70% tổng dữ liệu và phần dữ liệu kiểm tra chiếm 30% dữ liệu còn lại. Tỷ lệ 70/30 để tạo ra 2 phần dữ liệu được chọn dựa trên kinh nghiệm của một số nhà nghiên cứu [15], [16]. Phần dữ liệu đào tạo sử dụng để huấn luyện và xác nhận chéo mô hình, nhằm lựa chọn siêu tham số của các mô hình ML. Phần dữ liệu thử nghiệm sử dụng để đánh giá độ chính xác của các mô hình được đề xuất.

Sự phân bố dữ liệu đầu vào được sử dụng trong nghiên cứu này thể hiện trên Hình 1. Có thể thấy rằng hầu hết các biến đầu vào của sở dữ liệu đều có giá trị biến thiên trong một khoảng rộng. Chiều dài của mẫu là 1,6÷14 (cm). Trọng lượng mẫu chủ yếu trong khoảng 1,7÷21 (g). Hàm lượng các chất hòa tan trong cồn nằm trong khoảng 22÷52 (%), với một vài giá trị cao hơn 57%. Ngược lại, hàm lượng các chất hòa tan trong nước chủ yếu nằm trong khoảng 30÷40 (%), một số giá trị

nằm rải rác giữa 42 và 53%. Rg1 thay đổi từ 0,008 đến 0,187 (%) nhưng các giá trị chủ yếu nằm trong khoảng 0,008÷0,1 (%). Rd dao động trong phạm vi từ 0,012 đến 0,716 (%) nhưng hầu hết các giá trị đều nằm trong khoảng 0,012÷0,22 (%). Biến Re dao động từ 0,112 đến 1,48 (%) và Rb1 biến thiên từ 0,093 đến 2,576%. F11 thay đổi từ 0,012 đến 0,189 (%) nhưng các giá trị chủ yếu nằm trong khoảng 0,012÷0,1 (%). Tương ứng với các giá trị này, CAG có các giá trị là 2, 3 và 4 năm tuổi.





Hình 1. Biểu đồ tần suất phân bố của các tham số đầu vào và đầu ra

Hình 2 là biểu đồ ma trận tương quan giữa các biến đầu vào và biến đầu ra. Đây là một ma trận 10x10 được tạo ra được tạo ra để khám phá mối tương quan thống kê tuyến tính giữa các biến trong cơ sở dữ liệu. Trục tung và trục hoành chỉ ra 9 biến đầu vào (X_i) và biến đầu ra (tuổi phát triển của sâm, Y). Biểu đồ này được xây dựng dựa trên hệ số tương quan thứ hạng của Spearman (r_s) giữa mỗi biến theo từng cặp. Trong đó mối tương quan giữa tất cả các thông số được vẽ rõ ràng và chính xác, các màu sắc khác nhau thể hiện các giá trị tương quan khác nhau. Căn cứ vào giá trị của r_s có thể chia mức độ tương quan thành các cấp độ như sau:

r_s = 0÷0,19: tương quan rất yếu

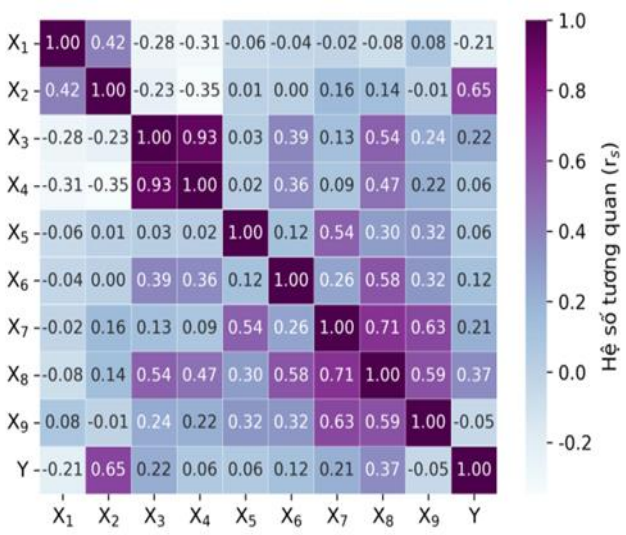
r_s = 0,2÷0,39: tương quan yếu

r_s = 0,4÷0,59: tương quan vừa phải

r_s = 0,6÷0,79: tương quan mạnh

r_s = 0,8÷1: tương quan rất mạnh

Quan sát Hình 2 cho thấy, đường chéo của ma trận đều có giá trị bằng 1 (vì hệ số tương quan của biến một biến với chính nó luôn bằng 1). Hầu hết các mối tương quan giữa các biến ở mức độ tương quan rất yếu, yếu và vừa phải (ứng với chỉ số r_s nằm trong khoảng từ 0 đến 0,65). Chỉ có một vài tương quan ở mức độ mạnh, rất mạnh như X_4 với X_3 (r_s=0,93) và X_8 với X_7 (r_s =0,71). Phân tích cho thấy, có thể coi 9 tham số đầu vào và 1 tham số đầu ra của tập dữ liệu là các biến độc lập. Vì vậy trong nghiên cứu này tất cả các biến đều được sử dụng để phát triển mô hình máy học và ước tính tầm quan trọng của tính năng.



Hình 2. Ma trận tương quan giữa các biến đầu vào và biến đầu ra

3. Các phương pháp máy học sử dụng trong nghiên cứu

3.1. Mô hình tăng độ dốc - Gradient Boosting (GB)

Thuật toán tăng độ dốc (GB) sử dụng một số cây phân loại hoặc hồi quy để cung cấp một kết quả đáng tin cậy và mong muốn. Kỹ thuật này xây dựng cây cơ bản (đôi khi được gọi là "người học cơ sở"), hết cây này đến cây khác để cải thiện hiệu suất của thuật toán. GB được Friedman thiết kế để sử dụng cho các bài toán hồi quy cũng như các bài toán về phân loại, mặc dù mục đích ban đầu của nó là chỉ được sử dụng cho các bài toán phân loại [17]. Việc kết hợp các yếu tố dự báo khác nhau từ mỗi lần lặp có thể nâng cao hiệu suất của mô hình, giảm thiểu sai số mô hình tổng thể. Từ đó, hiện tượng “quá khớp” có thể được giảm bớt. Trong kỹ thuật GB, cây hồi quy được sử dụng như những người học yếu, và đường xuống dốc ngẫu nhiên được sử dụng để huấn luyện mô hình trong mỗi lần lặp để giảm thiểu sai số [18]. Về bản chất, phương pháp này chia dữ liệu thành nhiều phần khác nhau. Một thuật toán được triển khai để xác định sự khác biệt giữa các giá trị mục tiêu và dự đoán tại mỗi điểm phân tách. Các sai số cũng được tính toán và điểm phân tách được chọn bằng cách sử dụng biến có giá trị thấp nhất cho chức năng thích hợp trước khi hoạt động được lặp lại [19].

3.2. Tăng cường độ dốc nhẹ -Light Gradient Boosting (LGB)

LGB là một khung tăng cường độ dốc sử dụng các thuật toán học cây. Nó được thiết kế để phân phối và hiệu quả bằng cách sử dụng hai kỹ thuật mới: Lấy mẫu một phía dựa trên Gradient (GOSS) và Gói tính năng độc quyền (EFB). So với các phương pháp CB có sẵn khác, LGB có một số ưu điểm như tốc độ đào tạo nhanh hơn, hiệu quả cao hơn, sử dụng bộ nhớ thấp hơn, độ chính xác tốt hơn, khả năng xử lý dữ liệu quy mô lớn và hỗ trợ học song song. Phương pháp này là một khung tăng cường độ dốc nhanh, phân tán, hiệu suất cao dựa trên thuật toán cây quyết định. Nó được sử dụng để xếp hạng, phân loại và nhiều nhiệm vụ khác trong lĩnh vực máy học [20].

3.3. Tăng cường độ dốc cực cao- Extreme Gradient Boosting (XGB)

XGB là một thuật toán được nâng cấp từ thuật toán Gradient Tree Boosting được phát triển bởi Friedman et al. vào năm 2000 [17]. Ý tưởng của thuật toán này là "Đẩy", có nghĩa là nó kết hợp tất cả các dự đoán của một nhóm người học "yếu" để xây dựng một người học "mạnh" thông qua các chiến lược đào tạo bổ sung. Một thành phần (được gọi là chính quy hóa) được đưa vào hàm mục tiêu và hàm mất mát của XGB để nâng cao hiệu suất của mô hình bằng cách làm mịn các trọng số cuối cùng và giảm thiểu các lần lặp không cần thiết. Thống kê độ dốc bậc nhất và bậc hai được sử dụng để tối ưu hóa hàm mất mát. Ngoài ra, trong thời gian đào tạo, các tính toán song song cho các chức năng trong XGB được thực hiện tự động. Do đó, các lợi ích của thuật toán XGB là tính linh hoạt và hiệu quả cao.

3.4. Kỹ thuật xác thực chéo -Cross validation (CV)

Trong lĩnh vực máy học, hiện tượng mô hình quá khớp (overfitting) là hiện tượng mô hình được tìm thấy phù hợp quá mức với dữ liệu đào tạo. Hiện tượng này có thể dẫn đến dự đoán không chính xác, nhiễu và mô hình có hiệu suất dự đoán thấp

trên dữ liệu xác thực. Kỹ thuật xác thực chéo thường được sử dụng để giải quyết vấn đề này.

Đối với các mô hình máy học mà quá trình huấn luyện sử dụng xác thực chéo với K nếp gấp thì toàn bộ cơ sở dữ liệu được phân chia ngẫu nhiên thành hai phần: tập huấn luyện (70% tổng số dữ liệu) và tập kiểm tra (với 30% dữ liệu còn lại). Tập dữ liệu kiểm tra sẽ được giữ riêng cho giai đoạn kiểm chứng mô hình. Trong quá trình đào tạo mô hình, tập dữ liệu kiểm tra hề được mô hình biết đến. Tập dữ liệu huấn luyện bao gồm các quá trình đào tạo và xác thực của mô hình. Điều này sẽ được thực hiện bằng cách chia ngẫu nhiên tập huấn luyện thành K phần bằng nhau. Mô hình sẽ được đào tạo K lần, với mỗi lần đào tạo chọn một phần làm dữ liệu xác thực và (K-1) phần còn lại làm dữ liệu đào tạo. Kết quả đánh giá cuối cùng sẽ là giá trị trung bình của K lần đào tạo. Nói chung, không nên chọn K quá lớn vì K cao dẫn đến tập dữ liệu đào tạo lớn hơn nhiều so với tập dữ liệu xác thực. Khi đó, các kết quả đánh giá sẽ không còn thể hiện chính xác bản chất máy học, đặc biệt là với các tập dữ liệu lớn. Trong nghiên cứu này kỹ thuật xác thực chéo với số nếp gấp K = 5 được lựa chọn.

3.5. Mức độ quan trọng của tính năng (Feature importance)

Mức độ quan trọng của tính năng phản ánh các chiến lược đánh giá mức độ hữu ích của thông tin đầu vào trong việc dự đoán một biến mục tiêu. Ý nghĩa của các tính năng là điều cần thiết trong các dự án mô hình dự báo vì chúng cung cấp cái nhìn sâu sắc về dữ liệu và thông tin về mô hình và là nền tảng để giảm kích thước và sự lựa chọn của các tính năng, tăng hiệu suất và hiệu quả của mô hình dự báo [21]. Kỹ thuật tầm quan trọng của tính năng được sử dụng ở đây đề cập đến tầm quan trọng của hoán vị, là một hàm tích hợp sẵn của tất cả các mô hình máy học dạng cây đề xuất được xem xét.

3.6. Các tiêu chí đánh giá năng lực dự báo của mô hình (RMSE, MAE, R^2)

Đối với các bài toán dự báo nói chung, năng lực dự báo của mô hình là quan trọng nhất. Nó được thể hiện thông qua các chỉ tiêu đánh giá sai số. Trong nghiên cứu này, ba chỉ số được sử dụng để đánh giá hiệu suất của các mô hình máy học, đó là hệ số xác định (R^2), sai số tuyệt đối trung bình (MAE) và sai số toàn phương trung bình (RMSE). Trong đó, RMSE đo lường sự khác biệt giữa giá trị thực tế và giá trị dự đoán, MAE đại diện cho sai số trung bình giữa giá trị thực và giá trị dự đoán. Giá trị RMSE và MAE càng thấp, độ chính xác của các mô hình càng cao hay hiệu suất dự báo của mô hình càng tốt. Ngược lại, giá trị R^2 cao hơn cho thấy hiệu suất mô hình tốt hơn. Các tiêu chí này được xác định như sau:

$$MAE=\frac{1}{n}\sum_{i=1}^{n}|g_i - g'_i| \quad (1)$$

$$RMSE=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(g_i - g'_i)^2} \quad (2)$$

$$R^2 = 1-\frac{\sum_{i=1}^{n}(g_i - g'_i)^2}{\sum_{i=1}^{n}(g_i - \bar{g})^2} \quad (3)$$

trong đó g_i và g'_i lần lượt biểu thị các giá trị thực tế và dự đoán của mẫu thứ i; $\bar{g}$ là giá trị trung bình của tất cả các giá trị thực tế (g_i), n là tổng số mẫu.

4. Kết quả và thảo luận

Trong nghiên cứu này, ba mô hình máy học bao gồm XGB, LGB và GB đã được phát triển để dự đoán CAG. Mục tiêu chính của quy trình này là tìm kiếm các giá trị của các siêu tham số để cung cấp cho các mô hình máy học hoạt động tốt nhất. Một tập hợp các tham số điển hình được chọn để tối ưu hóa. Để tối ưu hóa các tham số quan trọng của mô hình, giá trị của chúng thay đổi trong một phạm vi nhất định, trong khi các tham số ít quan trọng hơn được chọn theo giá trị mặc định. Để xác định cấu trúc tối ưu của mô hình máy học, năng lực dự báo và sự ổn định của mô hình được đánh giá dựa trên tiêu chí R^2 và độ lệch chuẩn (Std) tương

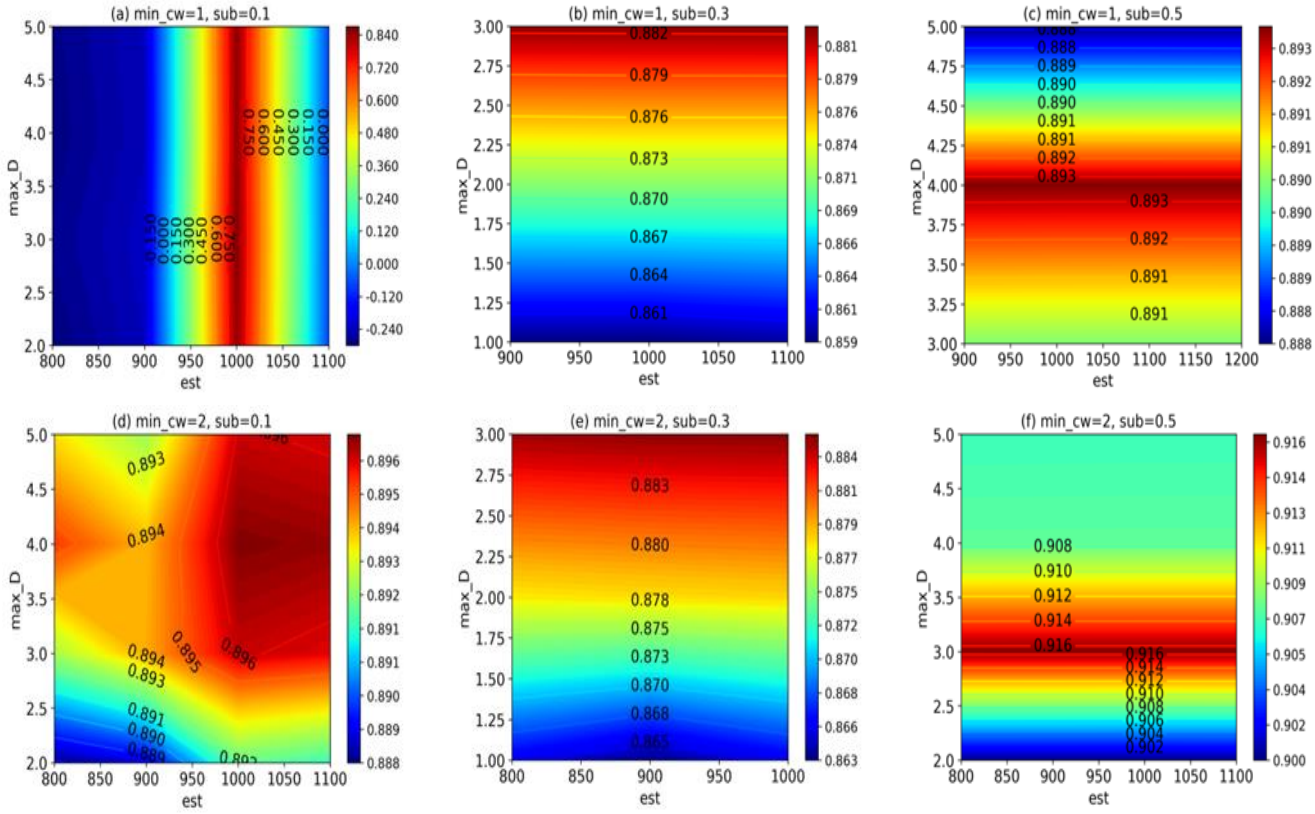
ứng. Giá trị R^2 được xác định bằng cách lấy trung bình của 5 lần xác thực. Điều quan trọng cần lưu ý là quá trình xác thực 5 lần chỉ được sử dụng trên tập dữ liệu huấn luyện (chiếm 70% tổng số toàn bộ dữ liệu) chứ không sử dụng trên tập dữ liệu kiểm tra (30% tổng số dữ liệu còn lại). Các mô hình không hề biết đến tập dữ liệu kiểm tra trong suốt quá trình đào tạo và xác nhận mô hình.

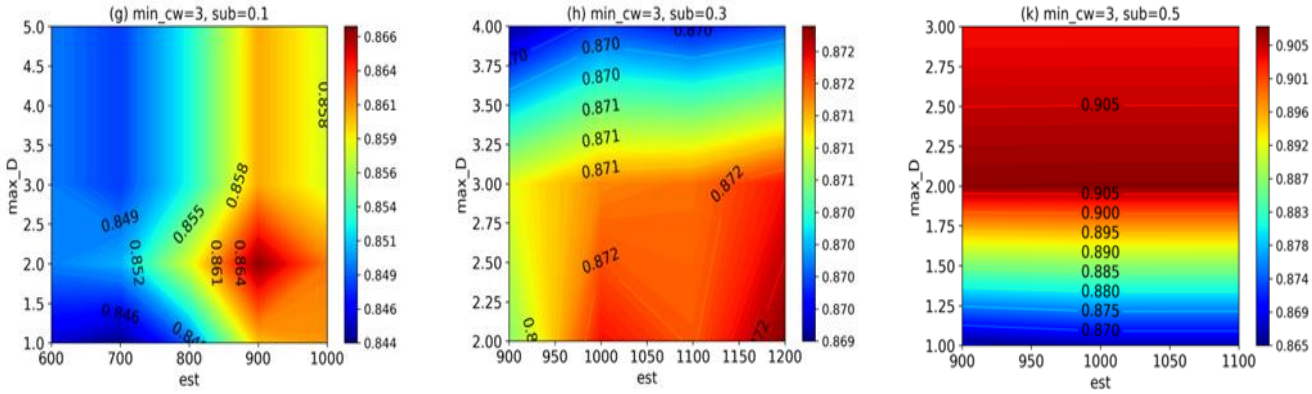
Tuy nhiên, do hạn chế về thời gian, trong nghiên cứu này chỉ trình bày chi tiết quá trình tối ưu hóa các tham số của mô hình XGB. Sau khi đã tối ưu hóa các tham số, hiệu suất dự báo của 3 mô hình XGB, LGB và GB tối ưu nhất được so sánh để chọn ra mô hình có hiệu suất tốt nhất nhằm dự đoán CAG.

4.1. Tối ưu hóa các tham số của mô hình XGB

Quá trình tối ưu hóa các tham số của mô hình XGB được trình bày trong phần này. Hiệu suất dự báo của mô hình XGB phụ thuộc vào nhiều tham số, trong đó 4 tham số chính là: Độ sâu tối đa của cây (max_D), Số lượng cây được tăng cường độ dốc (est), tổng trọng lượng tối thiểu (min_cw), Tỷ lệ mẫu con (sub). Vì vậy 4 tham số này sẽ được chọn để tối ưu hóa. Kết quả của việc tối ưu hóa các tham số của mô hình XGB được thể hiện trên hình 3.

Quan sát biểu đồ nhiệt Hình 3, trục tung thể hiện giá trị tham số max_D, trục hoành thể hiện giá trị tham số est, các giá trị min_cw và sub thay đổi trên các Hình 3 a, b, c , d, e, f, g, h, k. Miền màu đỏ thể hiện R^2 cao tức mô hình có hiệu suất dự báo tốt, trái lại, miền màu xanh thể hiện R^2 thấp tức mô hình có hiệu suất dự báo thấp. Kết quả là, Hình 3f mô tả trường hợp mô hình XGB có hiệu suất dự báo tốt nhất (R^2 =0,916) ứng với với các siêu tham số là max_D=3, est=1000 , min_cw=2, sub=0,5 và các tham số còn lại lấy giá trị mặc định. Đây là mô hình XGB tối ưu khi dự báo tuổi của sâm. Tiến hành tương tự cho 2 mô hình còn lại là LGB và GB. Sau khi tối ưu hóa, 3 mô hình XGB, LGB và GB với các siêu tham số được xác định. Hiệu suất dự báo CAG của 3 mô hình này sẽ được trình bày ở phần tiếp theo.





Hình 3. Kết quả tối ưu hóa các tham số của mô hình XGB

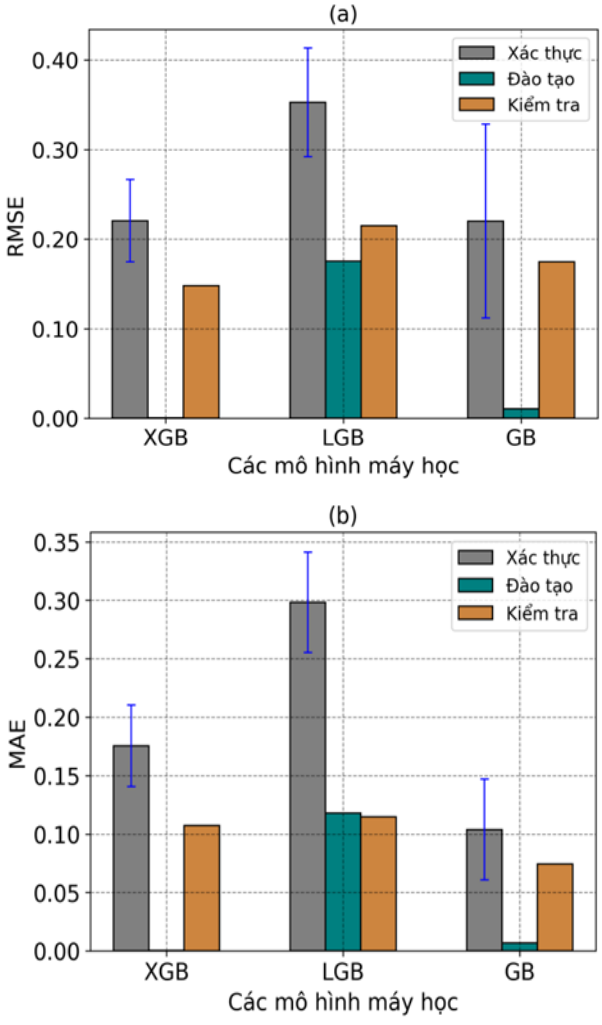
4.2. So sánh hiệu suất dự báo của ba mô hình đã tối ưu các tham số XGB, LGB và GB

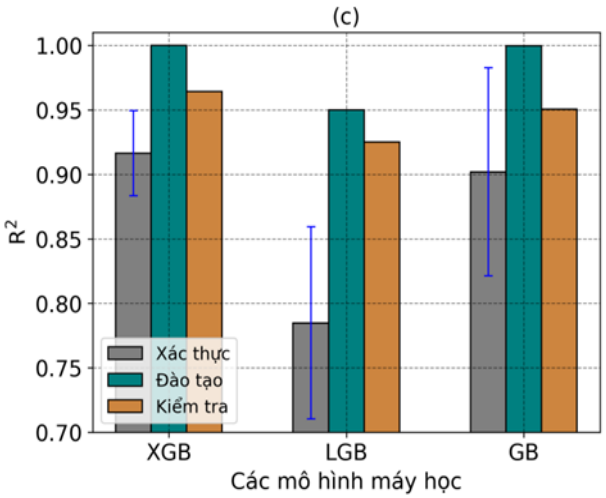
Nội dung phần này sẽ so sánh hiệu suất của 3 mô hình XGB, LGB và GB khi dự báo tuổi phát triển của sâm. Hiệu suất dự báo của 3 mô hình sẽ được đánh giá dựa trên ba chỉ số RMSE (Hình 4a), MAE (Hình 4b) và R^2 (Hình 4c).

Như đã trình bày ở mục 3.6, mô hình có khả năng dự báo càng tốt khi chỉ tiêu đánh giá R^2 càng cao và các sai số MAE, RMSE càng thấp. Quan sát Hình 4a cho thấy, ở cả 3 giai đoạn đào tạo, xác thực và kiểm tra mô hình XGB đều có giá trị RMSE nhỏ nhất trong tổng số 3 mô hình nghiên cứu và các giá trị này rất nhỏ thể hiện ở $RMSE_{đàotạo}$=0,0005, $RMSE_{xácthực}$=0,2206, $RMSE_{kiểmtra}$ = 0,1481. Kết quả này chứng tỏ nếu dựa trên tiêu chí RMSE thì mô hình XGB có khả năng dự báo tốt nhất trong 3 mô hình và hiệu suất dự báo tuổi phát triển sâm có sai số rất nhỏ.

Dựa vào tiêu chí MAE (Hình 4b), nếu ở giai đoạn đào tạo mô hình XGB có sai số MAE thấp nhất (0,00037) thể hiện khả năng dự báo tốt nhất. Tuy nhiên, nếu xét ở giai đoạn xác thực và kiểm tra thì thứ tự này có sự thay đổi, mô hình GB khả năng dự báo tốt nhất với $MAE_{kiểmtra}$= 0,0744, tiếp theo đến mô hình XGB với $MAE_{kiểmtra}$= 0,1071 và cuối cùng là mô hình LGB với $MAE_{kiểmtra}$ = 0,1148.

Tiếp theo, Hình 4c thể hiện giá trị R^2 của 3 mô hình nghiên cứu, kết quả tương tự như ở hình 4a, khả năng dự báo của các mô hình theo thứ tự giảm dần là XGB, GB và thấp nhất là LGB cho cả ba giai đoạn đào tạo, xác thực và kiểm tra. Đặc biệt ở giai đoạn kiểm tra, R^2 của mô hình XGB khá cao ($R^2_{kiểmtra}$=0,964) thể hiện năng lực dự báo của mô hìn XGB rất tốt. Ngoài ra, ở giai đoạn xác thực mô hình XGB có giá trị Std nhỏ nhất so với 2 mô hình LGB, GB chứng tỏ hiệu suất dự báo của mô hình XGB là ổn định nhất.





Hình 4. Kết quả so sánh hiệu suất dự báo CAG của 3 mô hình XGB, LGB, GB theo các tiêu chí đánh giá (a) RMSE, (b) MAE, (c) R^2

Như vậy, sau khi so sánh khả năng dự báo của 3 mô hình XGB, GB, LGB, có thể kết luận rằng XGB có hiệu suất dự báo cao nhất và ổn định nhất. Điều này hoàn toàn phù hợp vì trong 3 mô hình trên, LGB là mô hình đã được đơn giản hoá nên năng lực dự báo sẽ hạn chế, mô hình XGB (có sự tăng cường) nên khả năng dự báo sẽ rất tốt. Hiệu suất dự báo của 3 mô hình XGB, LGB và GB được thể hiện chi tiết trong Bảng 2.

Bảng 2. Kết quả dự báo của 3 mô hình XGB, LGB, GB cho các giai đoạn đào tạo, xác thực và kiểm tra

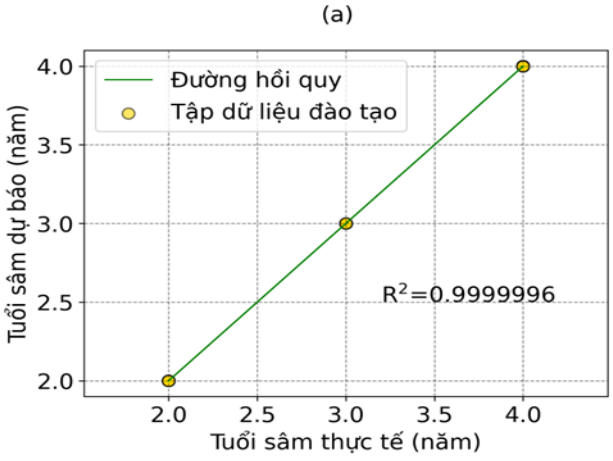
	Tiêu chí	XGB	LGB	GB
Xác thực	RMSE	0.2205960	0.3526249	0.2202240
	MAE	0.1754708	0.2981771	0.1038067
	R^2	0.9164411	0.7848670	0.9020409
Đào tạo	RMSE	0.0004915	0.1754760	0.0105320
	MAE	0.0003700	0.1179378	0.0069161
	R^2	0.9999996	0.9500693	0.9998201
Kiểm tra	RMSE	0.1481402	0.2147620	0.1746620
	MAE	0.1071492	0.1147797	0.0743862
	R^2	0.9643864	0.9251511	0.9504929

4.3. Kết quả điển hình khi dự báo tuổi của sâm bằng mô hình tốt nhất (XGB)

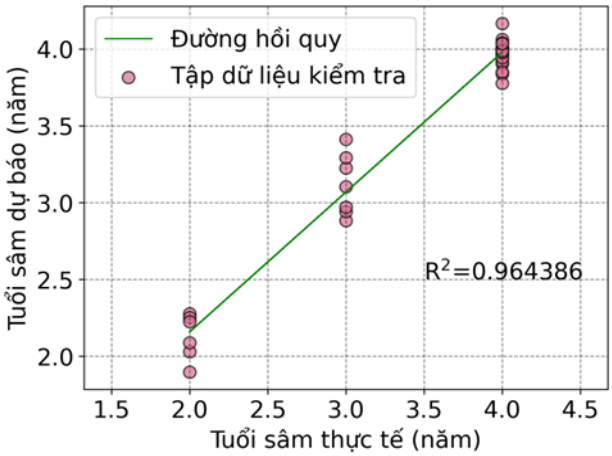
Ở phần trên, khả năng dự báo của 3 mô hình XGB, LGB, GB đã được so sánh, kết quả là mô hình XGB có hiệu suất dự báo cao nhất. Vì vậy ở mục này sẽ trình bày kết quả dự báo CAG của mô hình tiêu biểu (mô hình XGB).

Hình 5 là biểu đồ hồi quy của mô hình XGB khi dự báo CGA. Biểu đồ hồi quy thể hiện tương quan kết quả giữa giá trị tuổi sâm do mô hình XGB dự báo và giá trị tuổi sâm thực tế cho tập dữ liệu đào tạo (Hình 5a) và tập dữ liệu kiểm tra (Hình 5b). Trong đó, trục hoành đại diện cho kết quả của thực tế được thu thập, trục tung đại diện cho kết quả được dự báo theo mô hình XGB. Quan sát cho thấy, khả năng đào tạo của mô hình gần như lý tưởng (Hình 5a) với $R^2_{đàotạo}$= 0,999. Ở giai đoạn kiểm tra (Hình 5b), hầu hết các mẫu có kết quả dự báo rất gần với kết quả thực tế, thể hiện giá trị $R^2_{kiểmtra}$ =0,964 cho thấy năng lực dự báo CAG rất tốt của mô hình XGB.

Tiếp theo, Hình 6 minh họa sự trùng khớp của các các giá trị CAG được dự báo bởi mô hình XGB và các giá trị CAG thực tế. Quan sát Hình 6a cho thấy, 74 mẫu trong tập dữ liệu đào tạo có giá trị CAG dự báo gần như trùng khít với đường giá trị CAG thực tế. Hình 6b thể hiện hầu hết các mẫu trong số 32 mẫu của tập dữ liệu kiểm tra có giá trị CAG dự báo trùng với CAG thực tế, chỉ một vài mẫu có sự sai lệch nhỏ như mẫu số 25, mẫu số 27. Tuy nhiên, số lượng mẫu có sự sai lệch này là không đáng kể so với tổng số 32 mẫu trong tập dữ liệu kiểm tra nên kết quả dự báo hoàn toàn đáng tin cậy.

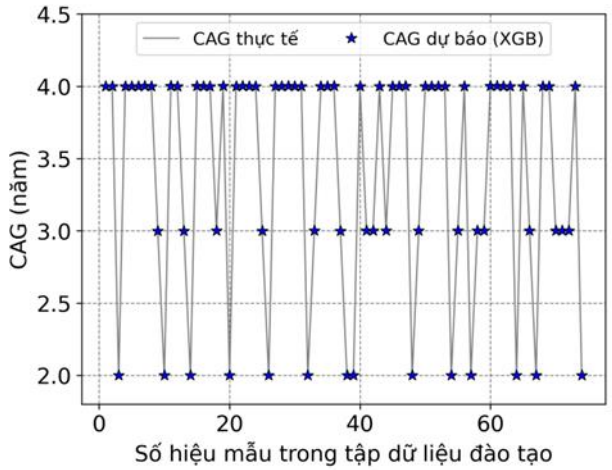


(b)

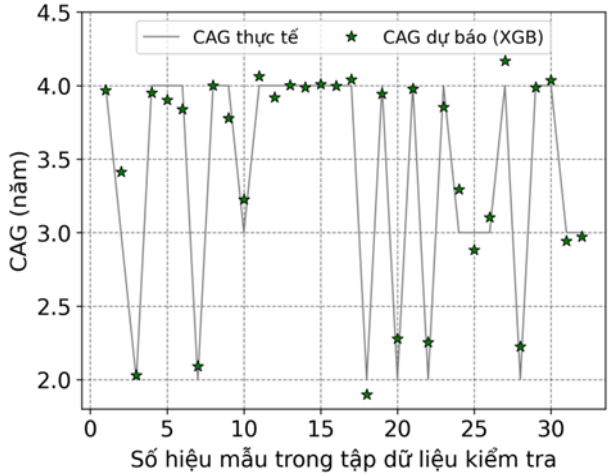


Hình 5. Biểu đồ hồi quy của mô hình XGB khi dự báo CGA: (a) giai đoạn đào tạo, (b) giai đoạn kiểm tra

(a)



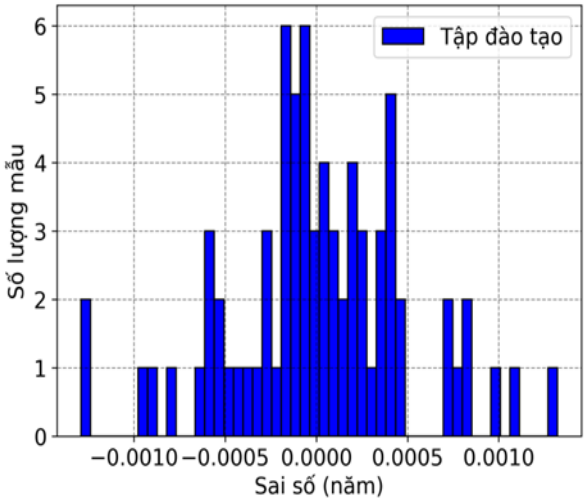
(b)



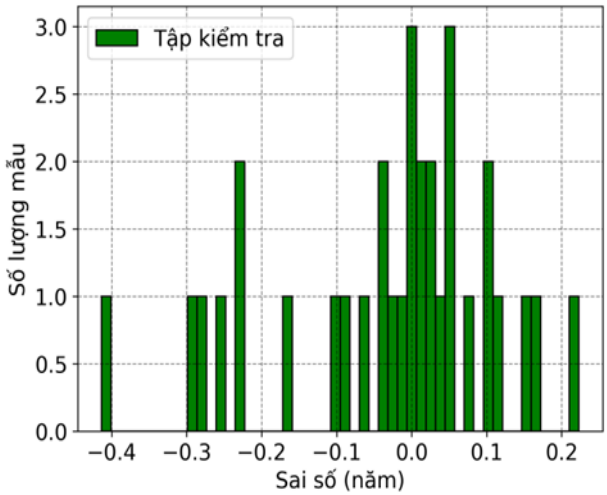
Hình 6. So sánh giữa giá trị CAG thực tế và CAG dự báo bởi mô hình XGB cho (a) tập dữ liệu đào tạo, (b) tập dữ liệu kiểm tra

Bên cạnh đó, sai số giữa CAG thực tế với CAG dự báo cho từng mẫu được mô tả trên Hình 7 ở giai đoạn đào tạo (Hình 7a) và giai đoạn kiểm tra (Hình 7b). Sai số càng nhỏ, thể hiện giá trị dự đoán càng gần giá trị thực tế hay khả năng dự báo của mô hình XGB càng chính xác. Quan sát hình cho thấy, phần lớn các mẫu có sai số nhỏ, cụ thể là [-0,001÷ 0,001] (năm) với giai đoạn đào tạo và [-0,3÷ 0,2] (năm) đối với giai đoạn kiểm tra. Đa số các mẫu có sai số nhỏ chứng tỏ kết quả dự báo của mô hình XGB hoàn toàn đáng tin cậy.

(a)



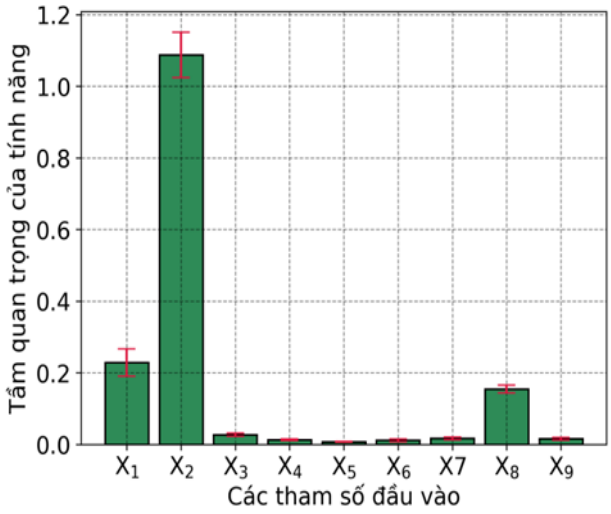
(b)



Hình 7. Tần suất sai số giữa CAG dự báo bởi mô hình XGB và CAG thực tế cho: (a) tập dữ liệu đào tạo và (b) tập dữ liệu kiểm tra

4.4. Phân tích mức độ ảnh hưởng của các tham số đầu vào đến CAG

Trong phần này, mức độ ảnh hưởng của các tham số đầu vào đến CAG được phân tích dựa vào đánh giá tầm quan trọng của tính năng (Feature importance). Kết quả phân tích được mô tả trên Hình 8. Trục hoành biểu thị 9 yếu tố đầu vào là (từ X_1 đến X_9), trục tung thể hiện tầm quan trọng của tính năng. Yếu tố đầu vào nào có giá trị tầm quan trọng tính năng càng cao thể hiện yếu tố đó có ảnh hưởng càng nhiều tới CAG. Quan sát Hình 8 cho thấy, 3 yếu tố có ảnh hưởng nhiều đến CAG là trọng lượng mẫu (X_2), chiều dài mẫu (X_1) và Rb1 (X_8). Trong đó X_2 có ảnh hưởng vượt trội đến CAG. Đây là yếu tố ảnh hưởng lớn nhất đến CAG trong tổng số 9 yếu tố đầu vào nghiên cứu. Sáu yếu tố còn lại có ảnh hưởng ít hơn đến CAG, xếp theo thứ tự giảm dần là Hàm lượng chất hòa tan trong cồn (X_3) > Re (X_7) > Hàm lượng chất hòa tan trong nước (X_4) > F11 (X_9) > Rd (X_6) > Rg1(X_5).



Hình 8. Các yếu tố ảnh hưởng tới CAG phân tích bằng mô hình XGB

5. Kết luận

Mục tiêu của nghiên cứu này là đề xuất một mô hình học máy tối ưu để dự báo tuổi phát triển của sâm một cách nhanh chóng, hiệu quả dựa trên cơ sở dữ liệu thực nghiệm thu thập được. Để đạt được mục tiêu này, 3 mô hình học máy XGB, LGB, GB được lựa chọn nghiên cứu. Để nâng cáo hiệu suất dự báo, các tham số quan của 3 mô hình được tối ưu hóa để lựa chọn ra các siêu tham số. Dựa trên 3 tiêu chí đánh giá là R^2, MAE và RMSE, hiệu suất dự báo của 3 mô hình XGB tối ưu, LGB, GB được so sánh. Mô hình XGB được đề xuất là mô hình tốt nhất và ổn định nhất trong dự báo CAG, với kết quả dự rất tốt thể hiện ở $R^2_{\text{đào tạo}}$=0,999 và $R^2_{\text{kiểm tra}}$=0,964. Ngoài ra, tầm quan trọng của các tính năng được sử dụng để phân tích tác động của chín tham số đầu vào đối với tuổi sâm dự báo. Kết quả phân tích chỉ ra rằng trọng lượng mẫu (X_2) có ảnh hưởng lớn nhất tới CAG. Ngoài ra, 8 tham số còn lại có mức độ ảnh hưởng xếp theo thứ tự giảm dần là: chiều dài mẫu, Rb1, hàm lượng chất hòa tan trong cồn, Re, hàm lượng chất hòa tan trong nước, F11, Rd, Rg1. Kết quả nghiên cứu này là cơ sở để xây dựng một công cụ phần mềm đáng tin cậy trong dự báo CAG một cách nhanh chóng, chính xác và tiết kiệm chi phí dựa trên cơ sở dữ liệu thực nghiệm sẵn có.

Tài liệu tham khảo

[1] H. Zhao, J. Xu, H. Ghebrezadik, and P. J. Hylands, ‘Metabolomic quality control of commercial Asian ginseng, and cultivated and wild American ginseng using (1)H NMR and multi-step PCA’, J Pharm Biomed Anal, vol. 114, pp. 113–120, Oct. 2015, doi: 10.1016/j.jpba.2015.05.010.

[2] R. B. Duda, Y. Zhong, V. Navas, M. Z. Li, B. R. Toy, and J. G. Alavarez, ‘American ginseng and breast cancer therapeutic agents synergistically inhibit MCF-7 breast cancer cell growth’, J Surg Oncol, vol. 72, no. 4, pp. 230–239, Dec. 1999, doi: 10.1002/(sici)1096-9098(199912)72:4<230::aid-jso9>3.0.co;2-2.

[3] Z.-H. Shao et al., ‘Antioxidant effects of American ginseng berry extract in cardiomyocytes exposed to acute oxidant stress’, Biochim Biophys Acta, vol. 1670, no. 3, pp. 165–171, Feb. 2004, doi: 10.1016/j.bbagen.2003.12.001.

[4] A. Scholey et al., ‘Effects of American ginseng (Panax quinquefolius) on neurocognitive function: an acute, randomised, double-blind,

placebo-controlled, crossover study', Psychopharmacology (Berl), vol. 212, no. 3, pp. 345–356, Oct. 2010, doi: 10.1007/s00213-010-1964-y.

[5] S. I. Chung, S. J. Nam, M. Xu, M. Y. Kang, and S. C. Lee, 'Aged ginseng (Panax ginseng Meyer) reduces blood glucose levels and improves lipid metabolism in high fat diet-fed mice', Food Sci Biotechnol, vol. 25, no. 1, pp. 267–273, 2016, doi: 10.1007/s10068-016-0039-1.

[6] I.-M. Chung, J.-W. Kim, P. Seguin, Y.-M. Jun, and S.-H. Kim, 'Ginsenosides and phenolics in fresh and processed Korean ginseng (Panax ginseng C.A. Meyer): Effects of cultivation location, year, and storage period', Food Chemistry, vol. 130, no. 1, pp. 73–83, Jan. 2012, doi: 10.1016/j.foodchem.2011.06.056.

[7] X. Chang et al., 'Nontargeted metabolomics approach for the differentiation of cultivation ages of mountain cultivated ginseng leaves using UHPLC/QTOF-MS', J Pharm Biomed Anal, vol. 141, pp. 108–122, Jul. 2017, doi: 10.1016/j.jpba.2017.04.009.

[8] M. C. Ichim and H. J. de Boer, 'A Review of Authenticity and Authentication of Commercial Ginseng Herbal Medicines and Food Supplements', Front Pharmacol, vol. 11, p. 612071, 2020, doi: 10.3389/fphar.2020.612071.

[9] E.-J. Lee et al., 'Quality Assessment of Ginseng by 1H NMR Metabolite Fingerprinting and Profiling Analysis', ACS Publications, Aug. 05, 2009. https://pubs.acs.org/doi/pdf/10.1021/jf901675y (accessed Jul. 04, 2022).

[10] S.-O. Yang et al., 'NMR-based metabolic profiling and differentiation of ginseng roots according to cultivation ages', J Pharm Biomed Anal, vol. 58, pp. 19–26, Jan. 2012, doi: 10.1016/j.jpba.2011.09.016.

[11] X. Hu et al., 'Machine learning methods to predict the cultivation age of Panacis Quinquefolii Radix', Chin Med, vol. 16, no. 1, p. 100, Oct. 2021, doi: 10.1186/s13020-021-00511-5.

[12] H.-V. T. Mai, T.-A. Nguyen, H.-B. Ly, and V. Q. Tran, 'Investigation of ANN Model Containing One Hidden Layer for Predicting Compressive Strength of Concrete with Blast-Furnace Slag and Fly Ash', Advances in Materials Science and Engineering, vol. 2021, p. e5540853, Jun. 2021, doi: 10.1155/2021/5540853.

[13] S.-E. Park et al., 'Metabolomic Approach for Discrimination of Cultivation Age and Ripening Stage in Ginseng Berry Using Gas Chromatography-Mass Spectrometry', Molecules, vol. 24, no. 21, p. E3837, Oct. 2019, doi: 10.3390/molecules24213837.

[14] S. Pan, H. Zhang, Z. Li, and T. Chen, 'Classification of Ginseng with different growth ages based on terahertz spectroscopy and machine learning algorithm', Optik, vol. 236, p. 166322, Jun. 2021, doi: 10.1016/j.ijleo.2021.166322.

[15] M. S. Khorsheed and A. O. Al-Thubaity, 'Comparative evaluation of text classification techniques using a large diverse Arabic dataset', Lang Resources & Evaluation, vol. 47, no. 2, pp. 513–538, Jun. 2013, doi: 10.1007/s10579-013-9221-8.

[16] 'Neural network classifier optimization using Differential Evolution with Global Information and Back Propagation algorithm for clinical datasets - ScienceDirect'. https://www.sciencedirect.com/science/article/abs/pii/S1568494616303866 (accessed Jan. 12, 2022).

[17] J. H. Friedman, 'Greedy Function Approximation: A Gradient Boosting Machine', The Annals of Statistics, vol. 29, no. 5, pp. 1189–1232, 2001.

[18] Z. Wei, Y. Meng, W. Zhang, J. Peng, and L. Meng, 'Downscaling SMAP soil moisture estimation with gradient boosting decision tree regression over the Tibetan Plateau', Remote Sensing of Environment, vol. 225, pp. 30–44, May 2019, doi: 10.1016/j.rse.2019.02.022.

[19] W. Ben Chaabene, M. Flah, and M. L. Nehdi, 'Machine learning prediction of mechanical properties of concrete: Critical review', Construction and Building Materials, vol. 260, p. 119889, Nov. 2020, doi: 10.1016/j.conbuildmat.2020.119889.

[20] G. Ke et al., 'LightGBM: A Highly Efficient Gradient Boosting Decision Tree', in Advances in Neural Information Processing Systems, 2017, vol. 30. Accessed: Jan. 18, 2022. [Online]. Available: https://proceedings.neurips.cc/paper/2017/hash/6449f44a102fde848669bdd9eb6b76fa-Abstract.html

[21] K.-Q. Shen, C.-J. Ong, X.-P. Li, and E. P. V. Wilder-Smith, 'Feature selection via sensitivity analysis of SVM probabilistic outputs', Mach Learn, vol. 70, no. 1, pp. 1–20, Jan. 2008, doi: 10.1007/s10994-007-5025-7.